Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 14, Número 2 (2022)

ISSN: 1647-0818

Volume 14, Número 2 – 2022

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Artigos de Investigação

Artigos Técnicos

Revisão

A comissão científica da **Linguamática** pode ser consultada na página Web da revista, em https://linguamatica.com/index.php/linguamatica/about/editorialTeam.

Para esta edição, colaboraram os seguintes investigadores:

- **Alberto Álvarez Lugrís**, Universidade de Vigo, Spain
- **Álvaro Iriarte Sanroman**, Universidade do Minho, Portugal
- **Antoni Oliver González**, Universitat Oberta de Catalunya, Spain
- **Arkaitz Zubiaga**, University of Warwick, United Kingdom
- **Hugo Gonçalo Oliveira**, Universidade de Coimbra, Portugal
- **Iria da Cunha**, Universidad Nacional de Educación a Distancia, Spain
- **Itziar Gonzalez-Dios**, Euskal Herriko Unibertsitatea, Spain
- **Jorge Baptista**, Universidade do Algarve, Portugal
- **Kepa Sarasola**, Euskal Herriko Unibertsitatea, Spain
- **Manex Agirrezabal**, University of Copenhagen, Denmark
- **Mercè Lorente Casafont**, Universitat Pompeu Fabra, Spain
- **Mikel L Forcada**, Universitat d'Alacant, Spain
- **Pablo Gamallo**, Universidade de Santiago de Compostela

Artigos de Investigação

Proposta recebida em Abril 2022 e aceite para publicação em Novembro 2022.

PROPOE (prose to poetry): Geração Computacional de Poemas Metrificados a partir da Prosa Literária em Língua Portuguesa

PROPOE (prose to poetry): Computational Generation of Metrified Poems from Literary Prose in Portuguese Language

Ana Cleyge de Azevedo
Universidade Estadual de Feira de Santana

João Queiroz
Universidade Federal de Juiz de Fora

Angelo C. Loula
Universidade Estadual de Feira de Santana

Resumo

A geração computacional do que pode ser concebido e reconhecido como 'poesia verbal' é explorada, há muitas décadas, em muitas línguas naturais. Mas apenas projetos recentemente desenvolvidos possuem aplicação intensiva da computação, considerando muitos níveis de organização, linguísticos e paralinguísticos, fonológicos, rítmicos, sintáticos, semânticos, e até pragmáticos. O que apresentamos aqui é um sistema de geração computacional de poemas, o PROPOE (*Prose to Poem*). Ele trabalha em conjunto com uma ferramenta de 'mineração' de estruturas de versificação na prosa de língua portuguesa, o MIVES (*Mining Verse Structure*). O PROPOE gera poemas em língua portuguesa a partir de sentenças versificadas (estruturas heterométricas de versificação) identificadas e classificadas pelo MIVES, e extraídas da prosa literária. O PROPOE combina sentenças, gerando poemas baseados na otimização de critérios rítmicos e fonológicos. É aplicado um 'algoritmo guloso' (*greedy algorithm*) cujo propósito é identificar a melhor combinação das sentenças, considerando normas rítmicas estabelecidas para o português. Em uma etapa final, realiza-se uma avaliação automatizada do resultado, atribuindo uma pontuação de acordo com a identificação de algum padrão considerado ótimo em poemas com métricas regulares, tendo como base esquema rítmico e adequação a estruturas rímicas.

Palavras chave

geração de linguagem natural; poema; prosa literária; padrões rítmicos; rima

Abstract

The computational generation of what can be recognized as 'verbal poetry' has been explored for many decades in many natural languages. But only recently developed projects have intensive application of computation, considering many levels of organization, linguistic and paralinguistic, phonological, rhythmic, syntactic, semantic, and even pragmatic. Here we present a computational poem generation system, PROPOE (Prose to Poem). It works in conjunction with a tool for 'mining' versification structures in Portuguese prose, MIVES (Mining Verse Structure). PROPOE generates poems in Portuguese from versified sentences (heterometric versification structures) identified and classified by MIVES, and extracted from literary prose. PROPOE combines sentences, generating poems based on the optimization of rhythmic and phonological criteria. A greedy algorithm is applied to identify the best combination of sentences, considering rhythmic norms established for Portuguese. In a final step, an automated evaluation of the result is carried out, assigning a score according to the identification of patterns considered optimal in poems with regular metrics, based on rhythmic scheme and adequacy to rhyme structures.

Keywords

natural language generation; poem; literary prose; rhythmic patterns; rhyme

1. Introdução

A geração computacional do que pode ser considerado 'poesia verbal versificada' é conhecida, ao menos, desde 1959, quando é publicado o 'Stochastische Texte' de Lutz (1959). No escopo mais amplo e histórico da 'arte algorítmica', ou 'arte computacional' (Antonio, 2009; D'Ambrosio, 2018), algoritmos 'associam' palavras para gerar versos (Lutz, 1959), caligramas ou estruturas verbo-visuais (Souza, 1991; Antonio, 2009), poemas baseados em sentenças extraídas de notícias de jornais, crônicas e de outros poemas (Balestrini, 1962; Antonio, 2009).

DOI: 10.21814/lm.14.2.369

Em ciência da computação, o tema tornou-se um tópico de interesse no domínio da Geração de Linguagem Natural (Gatt & Krahmer, 2018) e Criatividade Computacional (Boden, 2009; Colton & Wiggins, 2012). Trabalhos pioneiros, baseados no uso intensivo de técnicas computacionais, foram propostos por Gervás (2000), Manurung et al. (2000) e Gonçalo Oliveira et al. (2007). Na última década, observamos um notável crescimento no número de trabalhos (ver revisão de Gonçalo Oliveira 2017b), uso de diversas técnicas, e o desenvolvimento de sistemas capazes de gerar poemas de variadas morfologias, atentos a diversos níveis de descrição – sintáticos, fonéticos, rítmicos, semânticos e até pragmáticos.

O desenvolvimento de um gerador de poemas consiste tanto na escolha geral de certas abordagens, técnicas associadas à eficiência e desempenho de processamento, quanto na construção de algoritmos capazes de seguir regras de combinação frásica, padrões de organização e versificação, e verificação do significado das palavras. Muitas vezes são eleitos apenas alguns níveis de descrição – sintático, gramatical, fonológico ou morfológico (Gonçalo Oliveira, 2017b). A reutilização, e combinação, de materiais previamente disponíveis (parágrafos, sentenças, versos), para geração de poemas, é uma estratégia usada por muitos pesquisadores (Gonçalo Oliveira, 2017b). Há muitos antecedentes na história recente da poesia do último século, em muitos idiomas (Perloff, 2013), e ela pode ser descrita como uma abordagem de 'criatividade combinatorial', como define Boden (2009), uma 'produção de combinações não familiares de ideias familiares'.

Apresentamos um sistema de geração computacional de poemas versificados (PROPOE, *Prose to Poem*). Ele trabalha em conjunto com uma ferramenta computacional de mineração de estruturas de versificação, ou sentenças metrificadas, na prosa de língua portuguesa, MIVES (*Mining Verse Structure*)[1](Carvalho et al., 2020). O PROPOE então seleciona e combina sentenças metrificadas, ou excertos de sentenças metrificadas, extraídas da prosa literária. Este fenômeno — ocorrência de estruturas de versificação na prosa literária, em língua portuguesa — foi identificado por Guilherme de Almeida (1946) em *Os Sertões*, de Euclides da Cunha. Almeida (1946) escandiu diversas sentenças metrificadas na obra de Euclides, e identificou uma grande variedade de estruturas, com ênfase em decassílabos heroicos e sáficos[2], e dodecassílabos[3]. Anos depois, Augusto de Campos (1996) produziu uma lista mais ampla, relacionou a operação com um método que chamou de 'leitura verso-espectral', e identificou 'mais de 500 decassílabos significativos, com predominância de sáficos (acentuados na quarta e oitava sílabas) e heróicos (acentuados na sexta sílaba)', e pouco mais de duas centenas de dodecassílabos, entre os quais muitos alexandrinos perfeitos. Para Almeida (1946), 'se certas passagens d'Os Sertões, em vez de compostas tipograficamente em forma de prosa, o fossem em forma de versos livres' seria possível encontrar poemas 'autênticos' e 'legítimos'. Ele próprio, Almeida (1946), 'gerou' um poema versificado, 'A Vaquejada', através da redistribuição gráfica de um trecho:

> A vaquejada
>
> De repente estruge ao lado
> um estrídulo tropel de cascos sobre pedras,
> um estrépito de galhos estalando,
> um estalar de chifres embatendo;
> tufa nos ares, em novelos,
> uma nuvem de pó;
> rompe, a súbitas, na clareira,
> embolada, uma ponta de gado;
> e, logo após,
> sobre o cavalo que estaca esbarrado,
> o vaqueiro, teso nos estribos...
>
> (Almeida, 1946, p.212-213)

As estruturas métricas definem a medida do verso (Spina, 2003, p.29), que pode ter diversas extensões. Em português, o sistema de estruturas de versificação é silábico-acentual — conta-se o número de sílabas poéticas de cada verso e verifica-se a alternância entre sílabas fortes (tônicas) e fracas, até a última tônica. Este processo é chamado de escansão. Tal alternância estabelece um certo número de padrões que, combinados às repetições posicionais das sílabas poéticas, cria segmentos internos, estabelecendo as regras de versificação ou metrificação (Ali, 1999). O MIVES realiza a escansão de segmentos frásicos na prosa, separa as sílabas poéticas e classifica as estruturas encontradas de acordo com padrões rítmicos normatizados. Estas sentenças (metrificadas) fornecem material para o sistema PROPOE gerar os poemas, considerando aspectos rítmicos e fonológicos. O PROPOE usa um 'algoritmo guloso' para identificação de semelhanças entre padrões rítmicos. Explora di-

[1] https://mivestool.wordpress.com/

[2] Decassílabos heroicos e sáficos, correspondem a versos com dez sílabas poéticas cujas tônicas são identificadas nas posições 3, 6, 10, e 4, 8, 10 respectivamente (Spina, 2003)

[3] Dodecassílabos são versos compostos por doze sílabas poéticas (Spina, 2003)

versos tipos de rima, considera várias estruturas métricas, combinando diferentes formas de busca através de similaridade entre versos, padrões rítmicos e posicionamento de tônicas. Baseado em critérios estabelecidos para avaliação dos poemas, o sistema avalia padrões fonéticos, rítmicos e similaridade entre as sílabas tônicas dos versos.

Nas próximas seções, apresentamos diversos projetos de geração automática de poemas, com ênfase em trabalhos considerados pioneiros, aplicações em vários idiomas, e destacamos alguns sistemas baseados na reutilização de textos. Em seguida, detalhamos o PROPOE, suas etapas de processamento, incluindo filtragem, montagem e avaliação dos poemas gerados. Finalmente, exibimos resultados de alguns poemas gerados pelo sistema, a partir de sentenças extraídas da prosa literária brasileira. Nossos resultados sugerem que há variações nas pontuações de avaliação obtidas pelos poemas, com possível relação com o volume de sentenças disponíveis, e que os diferentes critérios individuais de avaliação impõem níveis distintos de dificuldade para pontuação na avaliação.

2. Trabalhos Relacionados

Como mencionamos, são pioneiros os trabalhos de geração computacional de poemas desenvolvidos por Lutz (1959). Ele concebeu um sistema incapaz de combinar trechos de sentenças por meio de um algoritmo gerador de números aleatórios (Ryan et al., 2014). Mas este tópico tornou-se realmente científico, para a comunidade de Ciência da Computação, apenas a partir dos anos 2000, com tratamentos mais rigorosamente sistemáticos de muitos níveis, linguísticos e paralinguísticos (Gonçalo Oliveira et al., 2007). Os trabalhos de Manurung et al. (2000) e de Gervás (2000) deram início a novos projetos computacionais, muitos dos quais ganharam relevância através da aplicação de diversas técnicas de inteligência artificial.

Entre os projetos considerados mais promissores, e pioneiros, Manurung et al. (2000) aplicaram o modelo estocástico de *hill climbing*[4] para implementar um sistema capaz de gerar poemas em inglês, a partir de buscas estocásticas por sentenças derivadas de regras gramaticais, lexicais e proposições semânticas. Em outra implementação, Manurung (2004) e Rahman & Manurung (2011) desenvolveram um sistema gerador de poemas em inglês, o Mcgonagall (exemplo (a) na Tabela 1), que aplica algoritmos evolutivos[5] para avaliação de padrões de tônicas e similaridade (métrica e semântica), utilizando operadores genéticos para variar características gramaticais, semânticas e lexicais de sentenças candidatas, representadas por árvores de derivação inicializadas aleatoriamente. Rashel & Manurung (2014) desenvolveram um sistema, o Pemuisi, que gera poemas em indonésio utilizando uma abordagem baseada na 'satisfação de restrições' (*constraint satisfaction approach*)[6], tais como número de versos, contagem de sílabas e rima, a partir de notícias publicadas em jornais, e utiliza sentenças construídas a partir de modelos e palavras chaves obtidas nas notícias. Tobing & Manurung (2015) propuseram um sistema que não utiliza modelos de sentença, mas extrai sentenças de notícias online e aplica um 'gerador tabular' (*chart generation*)[7] para produzir variações que atendam a um padrão determinado (exemplo (b) na Tabela 1).

O sistema WASP (Gervás, 2000) usa uma abordagem baseada em regras de 'raciocínio avançado' (*forward reasoning*)[8] para produzir versos em espanhol a partir de um conjunto de palavras e padrões de versificação. Aplica-se um método para gerar e testar rascunhos de versos através de diversas restrições, produzindo versos independentes, ou poemas na forma de *romances*, *cuartetos*, e *tercetos encadenados*. Gervás (2001) incrementou sua implementação no sistema ASPERA (exemplo (c) na Tabela 1). O sistema solicita do usuário um fragmento de texto para referência, alguns parâmetros, seleciona métrica e estrofe em uma base de conhecimento, e aplica um método de 'raciocínio baseado em casos' (*Case-Base Reasoning, CBR*)[9]. Em outro momento, Gervás (2016) retomou seu projeto WASP para adicionar restrições semânticas (exemplo (d) na Tabela 1. Um *corpus* de poemas mexicanos, e notícias retiradas de jornais, foram utilizados para extrair sequências de palavras,

[4] Subida da Encosta (Hill Climbing) é um método de busca local e iterativa de soluções para problemas de maximização, com avaliação e escolha pelos melhores estados com base no estado atual (Castro, 2006).

[5] Algoritmos Evolutivos são algoritmos de busca e otimização de uma população de soluções que passa por um processo iterativo de seleção competitiva seguida de variação (Eiben & Smith, 2003).

[6] Abordagens de 'satisfação de restrições' buscam soluções para atribuir valores para variáveis satisfazendo restrições que limitam seus possíveis valores (Brailsford et al., 1999).

[7] 'Geração tabular' (chart generation) é um método para construir sentenças sintaticamente bem formadas de um enunciado de linguagem natural (Kay, 1996).

[8] 'Encadeamento para frente' é uma forma de acionamento de regras em mecanismos de inferência (Durkin, 1994).

[9] 'Raciocínio baseado em casos' (case-based reasoning) é uma técnica que resolve problemas buscando casos similares anteriores (Aamodt & Plaza, 1994).

A very african lion,
who is african, dwells in a waste.
Its head, that is big,
is very big.
A waist, that is its waist, it is small.
(a) (Manurung, 2004)

Ask in french surface
Call her years, check her
Think were toy tennis
Skid in chase, land her
(b) (Tobing & Manurung, 2015)

Ladrará la verdad el viento airado
en tal corazón por una planta dulce
al arbusto que volais mudo o helado.
(c) (Gervás, 2001)

Toda era una ave larga
que cuando se conforman.
Admitió que se tienen
registradas personas
(d) (Gervás, 2016)

delatar sempre causa delação
delação negra sem acusação
acusação em meia delação
sem achar cita, nem citação
(e) (Gonçalo Oliveira, 2017a)

Tabela 1: Exemplos de poemas gerados por diferentes sistemas.

que são modificadas segundo critérios específicos, produzindo fragmentos de poemas, combinados em estrofes.

PoeTryMe é uma plataforma com arquitetura modular para geração de poemas em português desenvolvido por Gonçalo Oliveira (2012) (ver exemplos na Tabela 2). A plataforma disponibiliza várias estratégias de geração, incluindo sentenças aleatórias, sentenças selecionadas por método 'gerar-e-testar' (*generate-and-test*)[10], ou sentenças obtidas por algoritmos evolutivos. Baseado em 'palavras semente' (*seed words*)[11], e utilizando redes semânticas construídas por dicionários, são gerados poemas conforme modelos (quantidade de estrofes, quantidade de linhas por estrofe e métrica) selecionados pelo usuário. Esta mesma plataforma foi utilizada por Gonçalo Oliveira (2017a) para gerar poemas baseados em postagens do Twitter, resultando no sistema *Poeta Artificial 2.0* (ver exemplo na Tabela 1e). O *Poeta 2.0* reutiliza fragmentos de sentenças de postagens, com possibilidade de troca de palavras por sinônimos para satisfazer certos padrões métricos.

Haiku
ah paixão afecto
não tem paixão nem objecto
sem na arte modos

(palavras semente: paixão, arte)

Quadra
ah escultura representação
e os que dão ao diabo o movimento da convulsão
sua composição de preparação
é destino estar preso por orientação

(palavras semente: paixão, arte)

Soneto
e não há deus nem preceito nem arte
um palácio de arte e plástica
as máquinas pasmadas de aparelhos
num mundo de poesias e versos

o seu macaco era duas máquinas
horaciano antes dos poetas
para as consolas dos computadores
num mundo de poesias e carmes

longas artes apografias cheias
tenho poesias como a harpa
poema em arte modelação

somos artificiais teatrais
máquinas engenhocas repetido
um poeta de líricas doiradas

(palavras semente: computador,
máquina, poeta, poesia, arte,
criatividade, inteligência,
artificial)

Tabela 2: Exemplos de poemas gerados pelo *PoeTryMe*(Gonçalo Oliveira, 2012).

Outros pesquisadores desenvolveram geradores em diferentes línguas, baseados em diversas escolhas, relacionadas ao tratamento de diferentes fontes de dados, processamento e dados de saída. O sistema de Agirrezabal et al. (2013) gera poemas em basco (*euskera*), de tradição *bertsolaritza*, utilizando modelos de etiquetagem morfossintática (*POS tagging*)[12] obtidos de corpora de poemas. Das & Gambáck (2014) desenvolveram

[10] Em uma heurística de 'gerar-e-testar', soluções são produzidas de forma estocástica e então avaliadas por critérios específicos do problema.

[11] Palavras de entrada inicial no sistema *PoeTryMe*, que servem como tema para geração do poema (Gonçalo Oliveira, 2012).

[12] 'Etiquetagem morfossintática' corresponde à atribuição de classes morfossintáticas às palavras de uma sentença, conforme sua contextualização. Para isso, utiliza-se tanto o léxico de palavras, e processos, categorizados pelo conjunto de etiquetas de um sistema, quanto o conjunto de possibilidades semânticas para categorias presentes no

um sistema de geração de poemas em bengali com sequências rítmicas determinadas a partir de uma linha de entrada fornecida pelo usuário, modelo da linha inicial e geração de linhas compatíveis a partir de um conjunto de poemas. A utilização de materiais previamente produzidos também é uma abordagem bastante conhecida. Tobing & Manurung (2015) e Gonçalo Oliveira (2017a) usam fragmentos de sentenças, modificados através de substituições de palavras. Outros projetos não modificam as sentenças, ou trechos de sentenças originais, e selecionam versos baseados em repositórios de sentenças. A partir de textos de blogs, o sistema de Wong et al. (2008) cria um acervo de fragmentos de sentenças com quatro palavras. Buscas por palavras-chave são realizadas neste acervo para obter haicais[13], combinando sentenças através da similaridade de palavras. Netzer et al. (2009) cria um repositório de sequências de palavras obtidas de uma base do Google e do Projeto Gutenberg para construir haicais. Deste repositório são extraídas sentenças candidatas que satisfazem estruturas sintáticas, gerando haicais. Eles são, em seguida, submetidos a um ranqueamento pela presença de associação entre palavras para seleção de um poema final.

Além das abordagens citadas, pesquisas sobre a utilização de redes neurais artificiais, particularmente redes profundas, para geração de linguagem natural, têm aumentado nos últimos anos, embora a geração de poemas através dessa abordagem tenha sido indicada como a tarefa mais difícil por exigir a consideração de requisitos estruturais (Otter et al., 2020). Uma rede neural treinada com sequências de sílabas da *Divina Comédia*, de Dante Alighieri, foi utilizada por Zugarini et al. (2019) para gerar poemas em italiano em tercetos a partir de uma sequência inicial de sílabas aleatórias. O modelo pré-treinado da GPT-2 (Radford et al., 2019), treinado em 40GB de textos da Internet, também já foi aplicado para geração de poemas. Bena & Kalita (2019) utilizam este modelo com 774 milhões de parâmetros para gerar poemas em inglês, embora sem considerar requisitos estruturais. Hämäläinen et al. (2022) utilizou uma rede GPT-2, treinada com 60GB de textos em belga, associada a uma segunda rede neural codificadora, para produzir poemas em francês, mas para obter rimas foi necessário utilizar um algoritmo que escolhesse uma palavra entre um conjunto de possibilidades indicadas pela rede GPT-2.

Há, portanto, uma grande variedade de sistemas de geração de poemas. Quanto às técnicas aplicadas, encontramos heurísticas como *hill climbing* e algoritmos evolutivos, sistemas baseados em regras, raciocínio baseado em casos, geradores tabulares, redes neurais, entre outras. Em relação aos níveis linguísticos, e paralinguísticos, de organização, são considerados aspectos gramaticais, morfológicos, fonéticos, rítmicos, semânticos, e até pragmáticos. Em geral, eles não são tratados conjuntamente, variando conforme a ênfase da pesquisa. O aproveitamento de materiais para início do processo de geração é uma estratégia frequente, como 'palavras sementes' (*seed words*) (Gonçalo Oliveira, 2017a), sentenças que sofrem ajustes, ou são preservadas (fragmentos) como versos.

O PROPOE utiliza sentenças extraídas da prosa literária. Isso é particularmente importante, no contexto de geradores automáticos, se considerarmos que a prosa literária 'ocupa um lugar intermediário entre a poesia enquanto tal e a língua de comunicação comum, prática' (Jakobson & Pomorska, 1985, p.106). Trata-se, também, de uma estratégia baseada na ideia de reutilização de paralelismos estruturais (fonológicos, rítmicos) já realizados, muitos dos quais já historicamente canonizados no acervo da fortuna crítica-literária em língua portuguesa. Nossas experimentações iniciais incluem a prosa de Euclides da Cunha, e dos modernistas Mário e Oswald de Andrade.

A reutilização de estruturas de uma fonte, sem modificação ou reposicionamento de palavras, aproxima o PROPOE de sistemas propostos por Netzer et al. (2009) ou Wong et al. (2008). Entretanto, estes dois sistemas anteriores reutilizam sequências curtas, de poucas palavras. O PROPOE é capaz de reutilizar sentenças inteiras, trechos iniciais ou finais de sentenças. Por outro lado, a construção de um poema é tratado pelo PROPOE como um problema de otimização, com foco em restrições e critérios rítmicos, a ser resolvido pela maximização de critérios definidos considerando restrições iniciais, como Manurung (2004) e Rahman & Manurung (2011).

léxico (Carvalho et al., 2012)

[13]Poema de origem japonesa, o haicai descende de um processo de adaptação cultural milenar que resultou numa vasta produção criativa (Sousa, 2007). Inicialmente chamado de waka (wa = expressão designativa de Japão e ka = poema, canto), aquilo que vem a se tornar o haicai é registrado pela primeira vez numa compilação de cerca de quatro mil e quinhentos poemas, distribuídos em vinte volumes, no primeiro e mais importante documento-livro poético da história do país, o Man´yôshu, ou *Antologia de dez mil folhas*, elaborado entre os séculos VII e VIII.

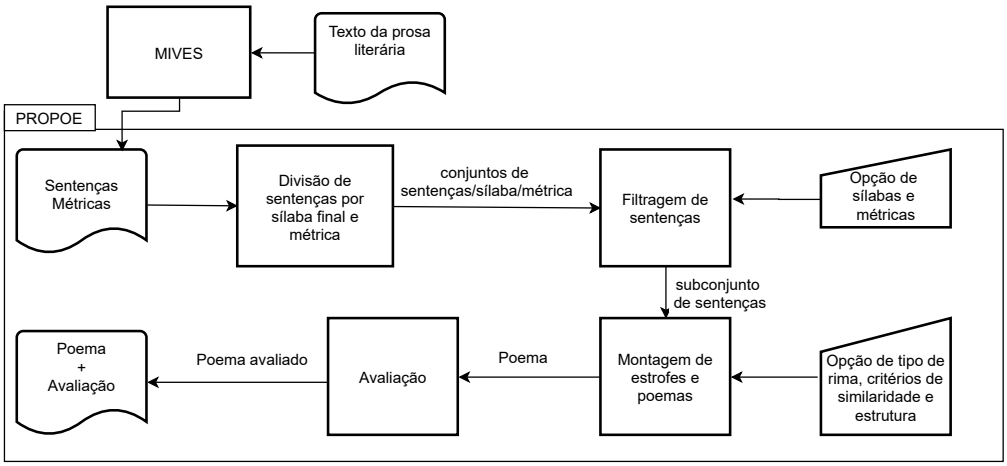

Figura 1: Etapas de processamento do sistema PROPOE.

3. PROPOE

O PROPOE[14] (*Prose to Poem*) gera poemas versificados a partir de sentenças heterométricas (estruturas versificadas) identificadas e classificadas pelo MIVES (ver descrição abaixo). Estas estruturas são identificadas, classificadas e extraídas da prosa literária de língua portuguesa. O sistema combina tais estruturas e gera poemas baseados na otimização de critérios rítmicos e fonológicos. É aplicado um 'algoritmo guloso' (*greedy algorithm*) em que são consideradas normas rítmicas e rímicas estabelecidas para o português (Silva, 2014). Em uma etapa final, é realizada uma avaliação automatizada, com atribuição de uma pontuação baseada em critérios relacionados a padrões rítmicos.

A arquitetura do sistema, e suas etapas de processamento, são apresentadas abaixo (Figura 1). Após o processamento pelo MIVES de um texto em prosa, as sentenças versificadas são enviadas ao PROPOE, que realiza a divisão e organização destas estruturas, indexadas de acordo com a última sílaba da última palavra da sentença e sua estrutura métrica. A partir das escolhas de sílabas finais e da estrutura métrica, é obtido um subconjunto de sentenças filtradas. Em seguida, são escolhidos os parâmetros estruturais que regulam a montagem das estrofes. De acordo com os critérios estabelecidos para avaliação, são exibidos os resultados. Detalhamos, em seguida, cada uma das etapas.

[14]O código fonte do PROPOE está disponível em https://github.com/lasicuefs/propoe

3.1. Prosa baseada em estruturas de versificação

Um texto em prosa de língua portuguesa é fornecido ao MIVES (Carvalho et al., 2020) que realiza a segmentação e o processamento de sentenças, ou fragmentos de sentenças, delimitadas por ponto final, ou outros sinais de pontuação. O MIVES identifica e classifica sentenças versificadas completas, trechos iniciais ou finais. Uma sentença completa começa após um ponto final, ou início de parágrafo, e termina com um ponto final ou final de parágrafo. Um trecho inicial também começa em um ponto final e pode finalizar em qualquer sinal de pontuação, ponto final ou não. Um trecho final começa após qualquer sinal de pontuação e termina em um ponto final. A busca por trechos iniciais, ou finais, pode fornecer partes de uma sentença mas também sentenças completas delimitadas por pontos finais. Na etapa seguinte, ocorre a separação silábica ortográfica de cada palavra da sentença, e a escansão das sílabas poéticas considerando diversos fenômenos fonológicos, como a identificação de posicionamento das sílabas tônicas, fusões e separações inter-silábicas. Um exemplo de trecho inicial é 'Sentou no fundo da igarité virada', obtido da sentença original 'Sentou no fundo da igarité virada, esperando.' de *Macunaíma, o herói sem nenhum caráter*, de Mário de Andrade, 1928. O exemplo de uma sentença completa da mesma obra é 'Acabou-se a história e morreu a vitória.'

Na etapa seguinte, ocorre a separação silábica ortográfica de cada palavra da sentença, e a escansão das sílabas poéticas, considerando vários fenômenos fonológicos, como a identificação de posicionamento das sílabas tônicas, fusões e separações inter-silábicas. 'Escansão' é o processo de contagem, identificação e classificação de sílabas poéticas. Em língua portuguesa, a

Sentença	Escansão	Métrica	Tônicas
Sentou no fundo da igarité virada,	Sen/t#o/u/ no/ f#un/do/ da i/ga/ri/t#é/ vi/r#a/da,	12	2 5 10 12
Acabou-se a história e morreu a vitória.	A/ca/b#ou-/se a his/t#ó/ri/a e/ mo/rr#e/u a/ vi/t#ó/ria.	12	3 5 9 12
Acabou-se a história e morreu a vitória.	A/ca/b#ou-/se a his/t#ó/ria e/ mo/rr#eu a/ vi/t#ó/ria.	10	3 5 8 10
Acabou-se a história e morreu a vitória.	A/ca/b#ou-/se a his/t#ó/ria e/ mo/rr#e/u a/ vi/t#ó/ria.	11	3 5 8 11
Ia pro céu viver com a marvada.	#I/a/ pr#o/ c#éu/ vi/v#er/ com/ a/ mar/v#a/da.	10	1 3 4 6 10
O herói não maliciou nada.	O he/r#ó/i/ n#ão/ ma/li/ci/#o/u/ n#a/da.	10	2 4 8 10
De certo era dinheiro enterrado.	De/ c#er/to/ #e/ra/ di/nh#e/i/ro/ en/te/rr#a/do	12	2 4 7 12
Passou um tempo de silêncio sagrado.	Pa/ss#o/u um/ t#em/po/ de/ si/l#ên/ci/o/ sa/gr#a/do	12	2 4 8 12

Tabela 3: Exemplos de sentenças extraídas pelo MIVES da obra *Macunaíma*

estrutura do verso é determinada pela relação entre as sílabas poéticas, e alternância de acentos até a última sílaba acentuada. Fenômenos fonológicos diversos podem unir ou separar as sílabas poéticas. No processo de identificação de sílabas, o MIVES considera fenômenos de sinérese, diérese, sinalefa, elisão e crase.[15]

Como o resultado da escansão não é unívoca, uma mesma sentença pode ser escandida pelo MIVES com diferentes medidas. O resultado é um conjunto de sentenças que atende aos limites mínimos e máximos de estrutura métrica estabelecidos pelo usuário, permitindo a exportação de sentenças identificadas na obra literária; cada sentença é associada à sua forma original e às suas possíveis versões escandidas com separação de sílabas poéticas e marcação de tônicas. A Tabela 3 exemplifica algumas sentenças e suas escansões obtidas pelo MIVES na obra *Macunaíma*.

3.2. Divisão e agrupamento das sentenças

O PROPOE recebe o resultado do MIVES e inicia o processamento separando as sentenças pela última sílaba da última palavra de cada sentença. Assim, todas as sentenças que terminam com a mesma sílaba final são agrupadas, facilitando a recuperação de sentenças por sílaba final para etapa de filtragem em seguida. Cada sentença tem associada a ela uma ou mais medidas de metro, resultado de sua escansão, marcação de tônicas e padrão rítmico. Assim, por exemplo, a partir de uma sílaba final 'da' e a lista de sentenças da Tabela 3 obtém-se a sub-lista de sentenças 'Sentou no fundo da igarité virada', 'Ia pro céu viver com a marvada' e 'O herói não maliciou nada' com suas escansões, métricas e esquema rítmico (posicionamento de tônicas). Como a mesma sentença pode ser escandida diferentemente, ela pode aparecer muitas vezes na lista de sentenças para uma determinada sílaba. A sentença 'Acabou-se a história e morreu a vitória' pode, por exemplo, ser escandida com 10 sílabas (decassílabo), 11 sílabas (hendecassílabo), ou 12 (dodecassílabo), conforme Tabela 3, e cada versão da escansão é associada à sílaba final 'te'.

3.3. Filtro das sentenças

Devido à grande quantidade de sentenças metrificadas que podem ser encontradas, o PROPOE realiza um processo de filtragem para reduzir o número de sentenças a serem manipuladas na montagem do poema. Este processo produz uma sub-lista de sentenças, com sentenças correspondentes à parametrização escolhida, minimizando o custo computacional. Para exemplificar, o número total de sentenças, ou trechos de sentenças, decassilábicas, identificadas em *Os Sertões* (Cunha, 1902) é 1.952. Existem, portanto, 14.518.416.572.416 possíveis arranjos de 4 sentenças. Deste modo, a montagem baseia-se em combinações de sentenças somente entre um número reduzido delas.

A filtragem de sentenças é feita por sílaba, por estrutura do poema e por métrica. Através do filtro silábico, duas ou quatro sílabas são escolhidas, produzindo um subconjunto de sentenças que possuem sílabas selecionadas como última sílaba. Este filtro reduz o número de sentenças a serem consideradas na produção de estrofes com rimas externas[16] na fase de montagem. Um recurso utilizado para montagem do verso, que também afeta a filtragem de sentenças, é a escolha do número de estrofes do poema e do número de versos por estrofes. A partir desses parâmetros, são determinadas sentenças candidatas em quantidade que podem atender ao número total de versos na montagem do poema,

[15] Sinérese: união de duas vogais, formando uma unidade sonora. Diérese: separação de duas vogais em uma mesma sílaba. Sinalefa: transforma duas sílabas formadas por vogais em uma só sílaba. Elisão: união em uma sílaba métrica de diferentes vogais, em uma só emissão de voz. Crase: união em uma sílaba métrica de vogais idênticas. (Silva, 2014)

[16] A rima externa é a semelhança fonética entre duas palavras no final de dois versos. Já a rima interna, é a semelhança fonética entre duas palavras no meio dos versos (Goldstein, 2008).

evitando a repetição de uma mesma sentença no mesmo poema, por falta de sentenças distintas. Por exemplo, se a estrutura do poema tem duas estrofes de quatro versos considerando somente duas sílabas finais, então serão necessárias pelo menos quatro sentenças para cada sílaba.

O filtro de métrica permite selecionar um conjunto de sentenças a partir dos metros obtidos na escansão de sílabas poéticas. Os metros disponíveis no sistema PROPOE variam de uma a treze sílabas. Ao selecionar um metro específico, são filtradas somente as sentenças que tenham o metro associado, produzindo um poema com versos isométricos. Este filtro é opcional. Se não for escolhida um metro, todos os metros são considerados e é gerado um poema heterométrico.

3.4. Montagem dos versos

A montagem final do poema corresponde a organização de versos em estrofes a partir de uma lista de sentenças filtradas na etapa anterior. O problema computacional nesta fase está relacionado a escolha de versos entre as sentenças candidatas, já filtradas por sílaba, estrutura do poema e métrica conforme etapa anterior, e posicionamento dos versos em estrofes. O PROPOE trata a seleção e a ordenação de sentenças como um problema de otimização combinatória de produção de um arranjo de k sentenças entre n possíveis, de modo a maximizar critérios rítmicos do poema. Uma estratégia possível para esta otimização combinatorial poderia testar todas combinações para obter o melhor padrão rítmico no poema, ou seja, aplicando uma estratégia de otimização global ao poema. Porém, teria um alto custo computacional avaliar todas as combinações, mesmo após a filtragem, dado o grande número de sentenças possíveis.

A abordagem do PROPOE aplica uma técnica 'dividir-e-conquistar' (*divide-and-conquer*) e 'gulosa' (*greedy*), para reduzir o custo computacional (Cormen et al., 2009). Através de divisão e conquista, é feita uma fragmentação do problema de geração de um poema em partes. Em seguida, as soluções obtidas para as estrofes são combinadas para obter uma solução geral. Já as estrofes são geradas através de uma abordagem gulosa, que busca otimizar localmente os critérios rítmicos, comparando somente pares de versos, para efetuar a seleção de um verso por vez em cada posição.

A montagem do poema é apresentada na Figura 2. Ela tem início através de um conjunto de sentenças já filtradas em uma etapa anterior, e baseia-se nos parâmetros: escolha de sílabas finais, número de estrofes e de versos por estrofe, tipos de rima externa, critério de comparação e similaridade. O poema resulta da justaposição sequencial das estrofes montadas.

3.4.1. Montagem da Estrofe

Cada estrofe é montada separadamente e de acordo com os parâmetros fornecidos. A quantidade de sílabas finais e de estrofes determinam quais sentenças candidatas são exploradas em cada estrofe. Se são escolhidas duas sílabas finais, os dois conjuntos de sentenças associadas a essas sílabas são utilizadas em todas as estrofes. Se são escolhidas quatro sílabas, as estrofes ímpares utilizam as sentenças das 1ª e 2ª sílabas e as estrofes pares utilizam as sentenças das 3ª e 4ª sílabas.

O primeiro verso da estrofe é escolhido aleatoriamente entre as sentenças candidatas terminadas com uma das sílabas finais escolhidas pelo usuário. Este primeiro verso torna-se uma referência que regula a seleção dos demais versos da estrofe, através de sucessivas comparações rítmicas. O posicionamento dos versos seguintes depende da escolha do tipo de rima externa (homofonia externa) para as estrofes. No português, a rima externa ocorre através da repetição de sons no final dos versos, que é o parâmetro estrutural mais facilmente identificado. Nesta etapa, consideramos que a rima entre versos ocorre quando as sentenças candidatas possuem a mesma última sílaba poética.[17]

Para estrofes em quadras (quatro versos), o PROPOE pode gerar poemas de rima cruzada, com versos alternados ABAB, onde A e B correspondem a uma sentença com uma sílaba final A ou B; rima emparelhada, mantendo uma sequência de cada dois versos, AABB; rima interpolada, ao rimar o primeiro verso com o quarto, e o segundo com o terceiro, ABBA; rima mista, formado sem imposição de ordenação pré-definida dos versos na estrofe. Na rima mista, é sorteado, a cada verso, se ele é do tipo A ou B, expandindo as possibilidades de combinações entre as rimas, podendo gerar AAAA, AAAB, AABA, ABAA, BAAA, BBBB, BBBA, BBAB, ou qualquer combinação de rimas anteriormente descritas. Em poemas de rima cruzada, emparelhada ou interpolada, um fator que pode alterar o resultado final é o número de versos por estrofe. Embora o PROPOE gere poemas com variado número de versos por estrofe, a rima baseia-se em uma es-

[17] Um caso específico de sentenças com a mesma sílaba final seriam sentenças com a mesma palavra final, e esta possibilidade não é excluída durante a montagem.

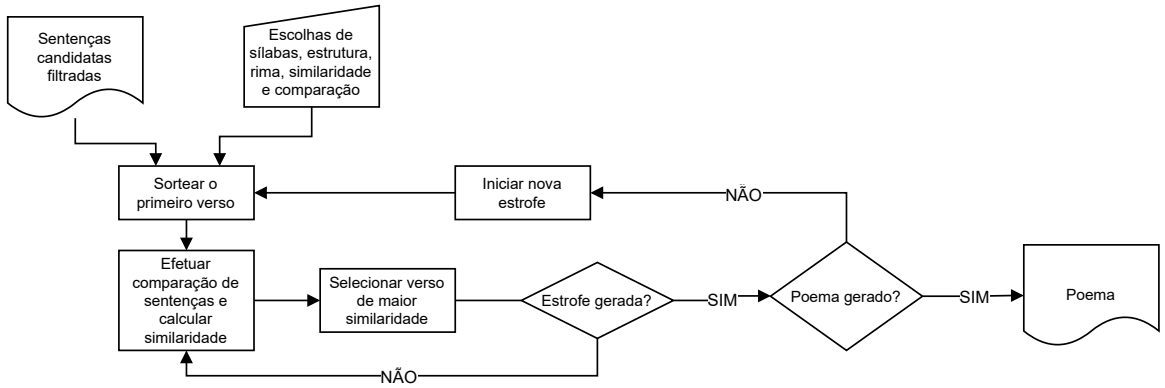

Figura 2: Etapas da montagem do poema.

trutura de quartetos. Se a estrutura é menor, o padrão é interrompido antes de ser completado; por exemplo, um terceto de rima cruzada ABAB é interrompido em ABA. Da mesma forma, se a estrutura é maior, o padrão é repetido total ou parcialmente; assim, um sexteto de rima emparelhada AABB é AABBAA.

Se o primeiro verso de cada estrofe é selecionado aleatoriamente, os demais são escolhidos por similaridade rítmica com o verso de referência. O verso de referência pode ser o primeiro verso da estrofe ou o verso imediatamente anterior, conforme escolha do usuário para este parâmetro da montagem. Para evitar a repetição de versos, cada sentença posicionada no poema é marcada como já utilizada na lista de sentenças candidatas para não ser utilizada em novas seleções. Após comparação das sentenças candidatas com o verso de referência, o verso selecionado é aquele que possui maior similaridade rítmica com ele. Como parâmetro do sistema, esse critério pode ser de similaridade por padrão rítmico ou de acentuação da última tônica. O padrão rítmico corresponde à distribuição posicional das sílabas fortes (tônicas) na sentença; a regra de similaridade da acentuação corresponde à posição da última sílaba tônica nas palavras que rimam. É efetuado um cálculo de similaridade de acordo com o critério escolhido para determinar a maior similaridade entre as sentenças candidatas. Em caso de empate, o outro critério é utilizado.

Para obter uma similaridade de padrão rítmico, o PROPOE utiliza o 'coeficiente de Jaccard' para comparar os conjuntos U e V de posições de tônicas de cada verso:

$$Jaccard(U,V) = \frac{|U \cap V|}{|U \cup V|} \qquad (1)$$

O coeficiente de Jaccard mede a similaridade entre dois conjuntos finitos e é definido como a razão entre a cardinalidade da interseção dos conjuntos e a cardinalidade da união dos conjuntos. Se considerarmos o esquema rítmico de um verso como um conjunto de números correspondentes, a posição das sílabas tônicas de cada verso, é possível calcular o coeficiente de Jaccard entre esquemas rítmicos e assim obter um valor de similaridade entre eles. Se o coeficiente tiver o valor 1.0, os esquemas são iguais; se o coeficiente é 0.0, não há qualquer posição tônica similar.

Como exemplo, as sentenças 'De certo era dinheiro enterrado' e 'Passou um tempo de silêncio sagrado', da Tabela 3, tem tônicas nas posições {2, 4, 7, 12} e {2, 4, 8, 12}, respectivamente. A interseção das posições das tônicas então seria {2, 4, 12} enquanto a união seria {2, 4, 7, 8, 12}, e assim a similaridade de padrão rítmico seria $3/5 = 0.6$.

Quanto à similaridade de acentuação da última tônica, o PROPOE determina para cada verso qual a posição da sílaba tônica na última palavra. A acentuação pode ser aguda (oxítona), grave (paroxítona) ou esdrúxula (proparoxítona). Se duas sentenças têm a mesma acentuação, sua casa similaridade é 1.0, caso contrário a similaridade é 0.0. Considerando como exemplo as mesmas sentenças anteriores, ambas têm acentuação grave (paroxítona), logo a similaridade é de 1.0.

O processo de montagem das estrofes é regulado por um 'procedimento guloso', uma vez que as comparações de similaridade são feitas em pares, de maneira iterativa, e não entre todas as sentenças da estrofe, ou do poema, que produziria um custo computacional muito maior. O poema final é resultante portanto da associação de estrofes construídas por esta abordagem de otimização local. Uma avaliação global das características rítmicas do poema é o último passo do processo de sua geração.

3.5. Avaliação do poema

A avaliação automatizada de poemas é um problema abordado em muitos trabalhos, como indica Gonçalo Oliveira (2017b), que concentra-se em dimensões objetivas de avaliação. A avaliação produzida pelo PROPOE é realizada pelo próprio sistema segundo padrões rítmicos. Para isso, é feita uma avaliação com base em critérios que calculam as pontuações, normalizadas no intervalo [0; 1], somando-as para obter uma avaliação final. Como são avaliados quatro critérios no poema, a pontuação total pode variar de 0.0 até 4.0 pontos. Os critérios estabelecidos para avaliação do poema estão relacionados à similaridade do esquema rítmico, correspondência das sílabas nas posições de tônicas, acentuação de última tônica, e similaridade entre as palavras que rimam. Cada um destes critérios contribuem para a avaliação do poema.

Sobre o critério de similaridade do esquema rítmico, é analisado o padrão rítmico geral do poema que contribui para sua avaliação porque produz um padrão de variação de intensidade sonora, de acordo com a posição das sílabas átonas e tônicas nos versos. É feita a comparação e cálculo de similaridade de Jaccard do esquema rítmico de cada verso com o verso de referência, de maneira análoga ao cálculo de similaridade de padrão rítmico efetuado na montagem. Em seguida, é feita a soma dos coeficientes encontrados para todos os versos de todas as estrofes do poema. Para normalizar este valor, esta soma é dividida pelo número de comparações realizadas. Por exemplo, se um poema tiver duas estrofes com quatro versos cada, serão realizadas seis comparações para avaliação de similaridade, uma vez que cada verso é avaliado somente uma vez, comparando com o verso de referência e o primeiro verso de cada estrofe não tem verso de referência.

O critério relacionado às sílabas tônicas idênticas avalia outro efeito sonoro decorrente de sílabas idênticas que ocupam a mesma posição tônica nos versos. Duas sílabas tônicas são consideradas idênticas se possuem as mesmas letras e estão na mesma posição. Feita a identificação das sílabas tônicas idênticas, o poema recebe um ponto para cada sílaba igual identificada. O resultado da soma destas pontuações é dividida pelo número de comparações realizadas para normalizar a pontuação. Por exemplo, ao comparar os versos 'A palha o fosfre e o goiano' (A/ p#a/lha/ o/ f#os/fre/ e o/ go/i/#a/no) e 'Faz com que os ventos do oceano' (F#az/ com/ que/ os/ v#en/tos/ do/ o/ce/#a/no), há duas tônicas em mesma posição, {5, 10}, e a sílaba tônica na posição 10, /#a/, é coincidente então a pontuação desta comparação será 1/2=0.5.

Outro critério de avaliação está relacionado à acentuação das últimas palavras nos versos. Há uma pequena diferença de referência, neste caso, em relação ao critério de acentuação para montagem. Para fins de avaliação, a comparação da acentuação é feita com base nas palavras que rimam, ou seja, cada verso é comparado com o verso anterior de mesma rima. Se duas sentenças têm a mesma acentuação, sua casa similaridade é 1.0, caso contrário a similaridade é 0.0. O número de coincidências de acentuação existentes em um poema é dividido pelo total de comparações realizadas para obter a pontuação.

O último critério de avaliação corresponde à similaridade de letras das palavras que rimam ao final de cada verso. Assim, duas palavras finais de versos que rimam na estrofe são comparadas, letra a letra, em sentido inverso, e cada correspondência de letra é avaliada. Esta contagem é dividida pelo tamanho da menor das palavras comparadas. Por exemplo, ao comparar a palavra final 'sagrado' de um verso com a palavra final 'enterrado' de outro verso, há três letras finais iguais e a menor palavra tem sete letras, então a pontuação desta comparação será 3/7=0.43. Tal procedimento é aplicado para cada par de versos que rimam, somando-se os valores obtidos a cada comparação. Em seguida, o valor total da soma é dividido pelo número de versos comparados.

As pontuações obtidas por cada um dos quatro critérios de avaliação, ao fim, são somadas e este valor define a pontuação do poema.

4. Resultados

O PROPOE, como descrito, gera poemas isométricos e heterométricos (deca e dodecassilábicos), baseados em estruturas rítmicas regulares, a partir de sentenças metrificadas extraídas de obras da prosa em língua portuguesa, pelo MIVES. A geração baseia-se na aplicação de filtros e na montagem de sentenças a partir de padrões rítmicos estabelecidos. O poema é avaliado pelo próprio sistema, através de pontuações relacionadas ao padrão rítmico e fonético (rima), e acentuação.

Selecionamos quatro obras da literatura brasileira: *Os Sertões*, de Euclides da Cunha, 1902; *Macunaíma, o herói sem nenhum caráter*, de Mário de Andrade, 1988; *Serafim Ponte Grande*, de Oswald de Andrade, 1933; *Grande Sertão: Veredas*, de Guimarães Rosa, 1956. Elas foram processadas pelo MIVES, e os resultados são apresentados na Tabela 4, com um resumo descritivo

Obra	Total de Sentenças	Tipo de Sentença	Sentenças Métricas (6 a 13)	Decassílabos	Dodecassílabos
Os Sertões	8.361	Completas	3.182	409	414
		Trechos Iniciais	5.391	912	1023
		Trechos Finais	5.838	730	1284
Serafim Ponte Grande	1803	Completas	828	115	111
		Trechos Iniciais	926	198	76
		Trechos Finais	1172	210	205
Macunaíma	2897	Completas	1047	154	149
		Trechos Iniciais	1596	269	250
		Trechos Finais	1714	225	355
Grande Sertão: Veredas	14725	Completas	6401	954	912
		Trechos Iniciais	12481	2006	2264
		Trechos Finais	12122	1433	2544

Tabela 4: Quantidade de sentenças métricas nas obras selecionadas.

Variações	Quantidade de Poemas	Pontuação				
		Esquema Rítmico	Sílabas Tônicas Iguais	Acentuação	Letras Finais Iguais	Geral
Obra Literária						
Os Sertões	25	0.65±0.09	0.03±0.05	1.00±0.02	0.57±0.09	2.24±0.19
Serafim Ponte Grande	25	0.50±0.11	0.01±0.02	0.89±0.20	0.45±0.16	1.86±0.30
Macunaíma	25	0.58±0.11	0.03±0.04	1.00±0.00	0.48±0.14	2.10±0.22
Grande Sertão Veredas	25	0.69±0.12	0.06±0.07	1.00±0.00	0.67±0.14	2.41±0.18
Número de Estrofes						
1	32	0.57±0.14	0.02±0.04	0.92±0.18	0.52±0.20	2.03+0.39
2	57	0.63±0.12	0.03±0.04	0.99±0.04	0.56±0.13	2.21±0.23
3	11	0.59±0.12	0.05±0.09	1.00±0.00	0.55±0.15	2.19±0.22

Tabela 5: Avaliação de poemas gerados pelo PROPOE com pontuação média e desvio padrão.

das sentenças metrificadas entre hexassílabos e bárbaros (métricas de seis a treze)). É importante salientar que uma mesma sentença pode ser escandida de diferentes formas, podendo aparecer na lista com diferentes medidas.

As sentenças metrificadas identificadas pelo MIVES, nestas obras, foram enviadas ao PROPOE para geração de poemas, através de variações métricas de versificação, número de estrofes, número de versos por estrofe, tipos de rima, uso de sentenças completas ou trechos iniciais e finais. Para avaliar os resultados da geração de poemas pelo PROPOE, foram produzidos 100 poemas, 25 poemas para cada obra com variações estruturais. Os resultados de pontuação de avaliação estão resumidos na Tabela 5.

Uma análise geral das pontuações dos 100 poemas gerados indica que a pontuação geral dos poemas está, em maior parte, entre 1.8 e 2.6. Há, no entanto, uma diferenciação da pontuação geral a depender da obra ou do tamanho do poema. Obras com maior número de sentenças candidatas, como *Os Sertões* e *Grande Sertão: Veredas*, possuem pontuação geral média maior do que obras mais curtas, como *Macunaíma* e *Serafim Ponte Grande*. Pode ser esperada uma relação entre quantidade de sentenças candidatas e a pontuação geral, uma vez que quanto maior a oferta de possibilidades para combinações, maior a chance de obter uma sentença que maximize critérios rítmicos. Quanto ao tamanho do poema gerado, em número de estrofes, a pontuação geral média de poemas curtos (1 estrofe) tende a ser inferior aos poemas mais longos (2 ou 3 estrofes). Como as obras mais curtas oferecem menor número de sentenças candidatas, é mais difícil obter poemas longos pela falta de versos suficientes. Então, grande parte dos poemas mais longos resultam de obras com mais sentenças candidatas porque permitem mais possibilidades de otimização.

A análise dos critérios individuais de avaliação evidenciam que há uma diferenciação da tendência de pontuação conforme o critério. Há um critério mais fácil de ser atendido, que é a pontuação por semelhança de acentuação, que tende a pontuações mais altas, possivelmente devido a pouca variação (três possibilidades) de posicionamento da tônica na última palavra. Já a pontuação por sílabas tônicas iguais tem valores consistentemente baixos, indicando que este é um critério difícil de ser satisfeito, por não ser tratado explicitamente durante a montagem. Por outro lado, as pontuações por similaridade de esquema rítmico e por letras iguais na última palavra tendem a pontuações intermediárias, uma vez que são considerados, ao menos parcialmente, na montagem. Para exemplificar os resultados obtidos na geração e pontuação de poemas individuais pelo sistema, com diferentes parametrizações, alguns foram selecionados para análise do processo de montagem e poemas.

Os três primeiros poemas são mais curtos, de uma única estrofe de 4 versos (quadra), com diversos tipos de sentenças (completas ou extrações parciais), métricas, parâmetros de filtragem e montagem. O Poema 1 é uma quadra gerada a partir de sentenças completas de *Macunaíma*, filtradas para obter somente versos dodecassilábicos, com sílabas finais 'do' e 'da' (Tabela 6). O PROPOE recebeu 151 sentenças completas, processadas do *Macunaíma*, sendo 88 terminadas com 'do' e 44 com 'da'. A filtragem dodecassilábica reduziu este número para 14 sentenças terminadas com 'do' e 8 com 'da'. A montagem foi feita para uma estrutura de uma estrofe de quatro versos, com rima emparelhada (AABB), seguindo um critério de similaridade por padrão rítmico (posição das tônicas), e comparação com o primeiro verso da estrofe para seleção de cada verso subsequente.

Para exemplificar o funcionamento das etapas de montagem, descrevemos os passos seguidos até obter este poema. O primeiro verso da estrofe do Poema 1 foi sorteado entre as sentenças candidatas. Uma vez que a primeira sílaba sorteada foi 'do', foi sorteado o verso 'De certo era dinheiro enterrado' para esta posição. Para selecionar o verso seguinte, outras 13 sentenças disponíveis terminadas com 'do' foram comparadas com o primeiro verso, através de cálculo de similaridade do padrão rítmico. O primeiro verso possui um padrão de posição de tônica {2, 4, 7, 12}. Entre as opções candidatas de maior similaridade, encontra-se o 'Passou um tempo de silêncio sagrado.' (Pa/ss#o/u um/ t#em/po/ de/ si/l#ên/ci/o/ sa/gr#a/do) com tônicas nas sílabas {2, 4, 8, 12} e o 'Porém Oibê não fez caso e veio vindo.' (Po/r#ém/ Oi/b#ê/ n#ão/ f#ez/ c#a/so e/ v#e/i/o/ v#in/do.) com tônicas em {2, 4, 5, 6, 7, 9, 12}. Como o primeiro verso candidato tem similaridade de 0.6, com relação ao verso de referência, e o segundo verso candidato tem similaridade de 0.57, a primeira opção foi selecionada. A terceira posição da estrofe é preenchida pela seleção entre as oito sentenças terminadas com 'da', considerando a rima emparelhada, e avaliando similaridade de esquema rítmico com o primeiro verso da estrofe. O verso candidato 'Pulou nesse e abriu na galopada' (Pu/l#o/u/ n#e/sse/ e a/br#i/u/ na/ ga/lo/p#a/da) tem padrão {2, 4, 7, 12} e similaridade máxima de 1.0, com todas as tônicas em posições idênticas ao do primeiro verso, e selecionado para a terceira posição. Em seguida, outro verso terminado em 'da', 'A companheira não trabalhara nada' (A/ com/pa/nh#e/i/ra/ n#ão/ tra/ba/lh#a/ra/ n#a/da) com tônicas nas posições {4, 7, 10, 12} obtendo similaridade de 0.6, maior valor entre as sentenças restantes.

Após a montagem, o Poema 1 foi avaliado pelo sistema obtendo escore de 2.39, de um máximo de 4.0, na soma dos quatro critérios. A pontuação de similaridade do esquema rítmico foi 0.73, obtida pela média das similaridades das posições das tônicas entre os versos comparados com o verso de referência (neste caso, o pri-

De certo era dinheiro enterrado
Passou um tempo de silêncio sagrado
Pulou nesse e abriu na galopada
A companheira não trabalhara nada

escansão	**métrica**	**tônicas**
De/ c#er/to/ #e/ ra/ di/nh#e/i/ro/ en/te/rr#a/do	Dodecassílabo	2 4 7 12
Pa/ss#o/u um/ t#em/po/ de/ si/l#ên/ci/o/ sa/gr#a/do	Dodecassílabo	2 4 8 12
Pu/l#o/u/ n#e/sse/ e a/br#i/u/ na/ ga/lo/p#a/da	Dodecassílabo	2 4 7 12
A/com/pa/nh#e/i/ra/ n#ão/ tra/ba /lh#a/ra/ n#a/da	Dodecassílabo	4 7 10 12

Tabela 6: Poema 1 gerado pelo PROPOE a partir de *Macunaíma*.

De noite continuou chorando
Agora o herói está fatigado
A palha o fosfre e o goiano
Faz com que os ventos do oceano

escansão	**métrica**	**tônicas**
De/ n#oi/te/ con/ti/nu/#o/u/ cho/r#an/do	Decassílabo	2 7 10
A/g#o/ra/ o he/r#ó/i es/t#á/ fa/ti/g#a/do	Decassílabo	2 5 7 10
A/ p#a/lha/ o/ f#os/fre/ e o/ go/i/#a/no	Decassílabo	2 5 10
F#az/ com/ que/ os/ v#en/tos/ do/ o/ce/#a/no	Decassílabo	1 5 10

Tabela 7: Poema 2 gerado pelo PROPOE a partir de *Macunaíma*.

meiro) – 0.6, 1.0 e 0.6. A pontuação por sílabas tônicas idênticas foi 0.0, o que reforça a dificuldade de pontuar neste critério, uma vez que não há sílabas idênticas na mesma posição (tônica) em comparação com o verso de referência. A pontuação por acentuação das últimas palavras obteve pontuação máxima de 1.0, resultado da comparação da acentuação das palavras que rimam, uma vez que todos os versos têm acentuação grave (tônica na penúltima sílaba). A pontuação por similaridade de letras das últimas palavras foi 0.66, acima da média dos poemas advindos da obra *Macunaíma*, pois 'sagrado' e 'enterrado' têm quatro letras idênticas (em ordem inversa); 'nada' e 'galopada' possuem três letras idênticas.

O Poema 2 (Tabela 7) também obteve sentenças completas de *Macunaíma* para gerar uma quadra. Mas, diferente do Poema 1, foram filtradas sentenças decassilábicas, sílabas finais 'do' e 'no', e alterado o verso de referência de similaridade para o verso imediatamente anterior. A filtragem decassilábica forneceu 15 sentenças que terminaram com 'do', e 4 sentenças com 'no'. A montagem foi configurada para uma estrutura de 1 estrofe de quatro versos, rima emparelhada (AABB), e critério de similaridade por padrão rítmico (posição das tônicas) para comparação com o verso imediatamente anterior, regulando a seleção de cada verso subsequente.

A avaliação do Poema 2, após a montagem, indicou um escore total de 2.15. A pontuação de similaridade do esquema rítmico foi 0.67, considerando a média das pontuações 0.75, 0.75 e 0.5, nas comparações de cada sentença com a anterior. Havia muitas sentenças para a primeira sílaba final, levando a pontuações mais altas, porém poucas para a segunda sílaba final, limitando a otimização deste critério. A pontuação por sílabas tônicas idênticas foi 0.11, um valor maior que o obtido no poema anterior, já que, entre as nove comparações de letras de sílabas tônicas, foi identificada uma sílaba tônica idêntica, 'a' entre o terceiro e o quarto versos na décima posição. A pontuação por acentuação das últimas palavras obteve pontuação máxima de 1.0 já que todos os versos têm acentuação grave. E a pontuação por similaridade de letras das últimas palavras obteve pontuação de 0.38, pois 'chorando' e 'fatigado' possuem três letras idênticas, como 'goiano' e 'oceano'.

Serafim Ponte Grande, de Oswald de Andrade, forneceu sentenças completas para geração do Poema 3 (Tabela 8). Foram filtradas sentenças dodecassilábicas, terminadas com 'do' e 'te', obtendo sete candidatas para cada sílaba. A montagem novamente foi configurada para uma quadra de rima emparelhada, seguindo um critério de similaridade por padrão rítmico (posição das tônicas) em comparação com o verso imediatamente anterior. A montagem selecionou três versos alexandrinos (tônicas em 6 e 12). Na avaliação final, o poema obteve escore de 2.11. A pontuação de similaridade do esquema rítmico foi 0.59, considerando a média das pontuações 0.67, 0.67 e 0.43, nas comparações de similaridade de posicionamento das tônicas de cada sentença com a anterior. Não foi obtida qualquer sílaba tônica idêntica na comparação de sentenças, com pontuação zero nesse critério. Obteve-se pontuação máxima de acentuação da última palavra, já que todos os versos terminam com acentuação grave. Para a similaridade de letras das palavras que rimam, a pontuação foi 0.52, já que 'frente' e 'silente' possuem quatro letras finais idênticas (e a menor palavra tem seis letras), e 'enrugado' e 'atentado' possuem três letras finais iguais (ambas possuem oito letras).

O Poema 4 (Tabela 9) baseou-se em *Os Sertões*, de Euclides da Cunha. Foram consideradas sentenças finais. A filtragem selecionou sentenças decassilábicas, terminadas com as sílabas finais 'dos' e 'tes', e obteve 24 e 22 candidatas para cada sílaba, respectivamente, uma quantidade maior de possibilidades do que aquelas exibidas nos cenários anteriores. Foi montada uma quadra de rima emparelhada aplicando o

Nosso herói tende ao anarquismo enrugado
Contra ele preparo um imenso atentado
Tudo conspira nesta cidade silente
O meu relógio anda sempre para a frente

escansão	**métrica**	**tônicas**
N#o/sso/ he/r#ó/i/ t#en/de ao a/nar/qu#is/mo en/ru/g#a/do	Dodecassílabo	1 4 6 9 12
C#on/tra/ #e/le/ pre/p#a/ro um/ i/m#en/so a/ten/t#a/d	Dodecassílabo	1 3 6 9 12
T#u/do/ cons/p#i/ra/ n#es/ta/ ci/d#a/de/ si/l#en/te	Dodecassílabo	1 4 6 9 12
O/ m#eu/ re/l#ó/gi/o/ #an/da/ s#em/pre/ para a/ fr#en/te	Dodecassílabo	2 4 7 9 12

Tabela 8: Poema 3 gerado pelo PROPOE a partir de *Serafim Ponte Grande*.

critério de similaridade por padrão rítmico, em comparação com o verso imediatamente anterior. Em razão da grande quantidade de sentenças candidatas, foi obtida uma maior variedade de posicionamento de tônicas. A busca por similaridade obteve todos os versos exatamente com o mesmo esquema rítmico {2, 7, 10}. Assim, na avaliação final, de 2.52, a pontuação de similaridade do esquema rítmico contribuiu com o valor 1.0, evidenciando que a maior quantidade de sentenças candidatas pode permitir melhor otimização deste critério. Entretanto, a ausência de sílabas tônicas idênticas na comparação de sentenças recebeu uma pontuação zero nesse critério; ou seja, mesmo com muitas sentenças, este critério permanece difícil de pontuar. A pontuação foi máxima para acentuação das últimas palavras, todas graves. A similaridade de letras das palavras que rimam produziu uma pontuação de 0.52, já que 'verdejantes' e 'retumbantes' possuem cinco letras finais idênticas, e 'unidos' e 'isolados' possuem três.

O Poema 5 (Tabela 10) foi obtido de *Os Sertões* utilizando sentenças completas para construir dois tercetos de dodecassílabos de estrutura rímica ABB. As sílabas finais escolhidas foram 'va' e 'so', em 5 e 7 sentenças candidatas dodecassilábicas, respectivamente. Mesmo em uma obra com grande quantidade total de sentenças candidatas, há sílabas com quantidade limitada de possibilidades. A montagem foi feita por comparação de similaridade de esquema rítmico, considerando a última palavra de cada verso, utilizando como referência o verso anterior. Como não havia sentenças métricas com as mesmas posições de tônicas, foram selecionadas as opções mais similares para geração do poema. A avaliação final do poema foi 2.15, com 1.0 ponto pela acentuação, já que são graves todas as palavras finais. Como não foram encontradas sílabas tônicas idênticas, não foi pontuado este critério. A avaliação fonética examinou apenas as sentenças terminadas com 'so', já que nesta estrutura de três versos por estrofe não há outra sentença com 'va'. Assim, a pontuação foi 0.56 considerando três letras finais idênticas entre 'nervoso' e 'desvalioso', e duas letras finais idênticas entre 'acaso' e 'lastimoso'. A avaliação de esquema rítmico obteve pontuação intermediária, de 0.58, valor esperado em consequência da limitação de sentenças candidatas.

A partir de sentenças completas de *Grande Sertão: Veredas*, foi gerado o Poema 6 (Tabela 11) com dois quartetos de decassílabos, rimas cruzadas, e sílabas finais 'dos' (17 sentenças candidatas) e 'go' (16 sentenças candidatas). A montagem foi feita por comparação de similaridade de esquema rítmico, utilizando como referência o verso anterior. O poema possui versos de padrões rítmicos clássicos — dois

ao cabo, completamente isolados
perdiam-se nos casebres unidos
as linhas de ordens do dia retumbantes
ilhados, os igapós verdejantes

escansão	**métrica**	**tônicas**
ao/ c#a/bo,/ com/ple/ta/m#en/te i/so/l#a/dos	Decassílabo	2 7 10
per/d#i/am-/se/ nos/ ca/s#e/bres/ u/n#i/dos	Decassílabo	2 7 10
as/ l#i/nhas/ de or/dens/ do/ d#ia/ re/tum/b#an/tes	Decassílabo	2 7 10
i/lh#a/dos,/ os/ i/ga/p#ós/ ver/de/j#an/tes	Decassílabo	2 7 10

Tabela 9: Poema 4 gerado pelo PROPOE a partir de *Os Sertões*.

Fez-se, porém, uma última tentativa
Determinou-a incidente desvalioso
Logo depois vibrava um abalo nervoso
Servidão inconsciente; vida primitiva
Uauá patenteava quadro lastimoso
Tomemos um fato, entre muitos, ao acaso

escansão	**métrica**	**tônicas**
F#ez/se,/ po/r#ém,/ #u/ma/ #úl/ti/ma/ ten/ta/t#i/va	Dodecassílabo	1 4 5 7 12
De/ter/mi/n#o/ua in/ci/d#en/te/ des/va/li/#o/so	Dodecassílabo	4 7 12
L#o/go/ de/p#ois/ vi/br#a/va um/ a/b#a/lo/ ner/v#o/so	Dodecassílabo	1 4 6 9 12
Ser/vi/d#ão in/cons/ci/#en/te;/ v#i/da/ pri/mi/t#i/va	Dodecassílabo	3 6 8 12
U/au#á/ pa/ten/te/#a/va/ qu#a/dro/ las/ti/m#o/so	Dodecassílabo	2 6 8 12
To/m#e/mos/ um/ f#a/to, #en/tre/ m#u/i/tos,/ ao a/c#a/so	Dodecassílabo	2 5 6 8 12

Tabela 10: Poema 5 gerado pelo PROPOE a partir de *Os Sertões*.

decassílabos heróicos (tônicas em 6 e 10), um martelo (tônicas em 3, 6 e 10), e versos de gaita galega (decassílabos com tônicas em 4, 7 e 10). Ele obteve pontuação final de 2.55. A acentuação teve 1.0, já que todas palavras finais são graves, confirmando a facilidade de pontuação máxima neste critério. A avaliação fonética teve uma pontuação 0.73, já que 'digo' e 'contigo' possuem 3 letras finais idênticas, enquanto os demais pares de palavras têm somente a sílaba final com rima idêntica. A avaliação de sílabas tônicas idênticas teve pontuação de 0.08, em 'Eu' no segundo e terceiro versos e 're' no primeiro e segundo versos da segunda estrofe, indicando que é possível alguma melhora nesse difícil critério, mas a pontuação tende a ser baixa devido a quantidade de tônicas diferentes. A pontuação de esquema rítmico obteve valor final de 0.73 com muitas intersecções de posições de tônicas em comparação com o verso anterior, em uma situação de maior quantidade de sentenças disponíveis.

Grande Sertão: Veredas também forneceu sentenças para o Poema 7 (Tabela 12), completas e parciais finais, em uma estrutura de dois quartetos, em versos dodecassílabos e 'rima mista', resultando numa escolha aleatória para sílaba final de cada verso, em um padrão ABAA e AAAA. As sílabas finais escolhidas foram 'te' e 'to', com

E uns outros, muito estampidos
Eu olhava para a beira do rego
Eu sabia do respirar de todos
Minha cara estava pegando fogo
Depois, tirei a dureza dos dedos
Calado, considerei comigo
Passei quase para a frente de todos
Ao dôido, doideiras digo

escansão	**métrica**	**tônicas**
E/ uns/ #o/u/tros,/ m#u/i/to es/tam/p#i/dos	Decassílabo	3 6 10
Eu/ o/lh#a/va/ para a/ b#e/i/ra/ do/ r#e/go	Decassílabo	1 3 6 10
Eu/ sa/b#i/a/ do/ res/pi/r#ar/ de/ t#o/dos	Decassílabo	1 3 8 10
M#i/nha/ c#a/ra es/t#a/va/ pe/g#an/do/ f#o/go	Decassílabo	1 3 5 8 10
De/p#ois,/ ti/r#e/i a/ du/r#e/za/ dos/ d#e/dos	Decassílabo	2 4 7 10
Ca/l#a/do,/ con/si/de/r#e/i/ co/m#i/go	Decassílabo	2 7 10
Pa/ss#e/i/ qu#a/se/ para a/ fr#en/te/ de/ t#o/dos	Decassílabo	2 4 7 10
Ao/ d#ô/i/do,/ do/i/d#e/i/ras/ d#i/go	Decassílabo	2 7 10

Tabela 11: Poema 6 gerado pelo PROPOE a partir de *Grande Sertão: Veredas*.

e o risco extenso dágua, de parte a parte		
o senhor se entende? Deixe avante; conto		
A razão dele era do estilo acinte		
reconheço, ele tem tido uma sorte		
sem prazo, isto é, o quase calados, somente		
naquela sanha, é ver cachorrada caçante		
por medo ter? Eu não campeava a morte		
que carreguei o rifle, escorei, repetente		
escansão	**métrica**	**tônicas**
e/ o/ r#is/co ex/t#en/so/ d#á/gua,/ de/ p#ar/te a/ p#ar/te	Dodecassílabo	3 5 7 10 12
o/ se/nh#or/ se en/t#en/de?/ D#e/i/xe a/v#an/te;/ c#on/to	Dodecassílabo	3 5 7 10 12
A/ ra/z#ã/o/ d#e/le/ #e/ra/ do es/t#i/lo a/c#in/te	Dodecassílabo	3 5 7 10 12
re/co/nh#e/ço,/ #e/le/ t#em/ t#i/do/ #u/ma/ s#or/te	Dodecassílabo	3 5 7 8 10 12
sem/ pr#a/zo,/ #is/to é, o/ qu#a/se/ ca/l#a/dos,/ so/m#en/te	Dodecassílabo	2 4 6 9 12
na/qu#e/la/ s#a/nha, é/ v#er/ ca/cho/rr#a/da/ ca/ç#an/te	Dodecassílabo	2 4 6 9 12
por/ m#e/do/ t#er?/ #Eu/ n#ão/ cam/pe/#a/va/ a/ m#or/te	Dodecassílabo	2 4 5 6 9 12
que/ ca/rre/gu#e/i o/ r#i/fle, es/co/r#ei,/ re/pe/t#en/te	Dodecassílabo	4 6 9 12

Tabela 12: Poema 7 gerado pelo PROPOE a partir de *Grande Sertão: Veredas*.

um grande número de sentenças candidatas, respectivamente, 75 e 68. Isso permitiu gerar uma grande variedade de sentenças candidatas distintas para o poema. A sílaba 'te' forneceu seis sentenças, um quantitativo maior do que em outros exemplos, devido ao padrão de rima, uma exigência muito maior que a observada nos poemas anteriores. Nota-se que ainda assim foi possível obter uma segunda estrofe de versos alexandrinos. Sua pontuação final foi 2.52. O esquema rítmico contribuiu com 0.91 pontos, devido a repetição de posicionamento de tônicas, mas não obteve pontuação máxima como no Poema 4, apesar do número muito maior de sentenças disponíveis, o que indica que mesmo com grande quantidade de sentenças há desafios na otimização de critérios rítmicos. A avaliação fonética teve pontuação de 0.58 pontos, já que as palavras 'somente' e 'caçante' tiveram apenas uma letra idêntica, além da sílaba final. A acentuação teve 1.0 uma vez que todas palavras finais são graves; a avaliação de sílabas tônicas iguais teve pontuação de 0.03 pela presença da sílaba 'ten' nas palavras 'extenso' e 'entende'.

O Poema 8 (Tabela 13) foi gerado a partir de sentenças completas de *Serafim Ponte Grande*. A estrutura foi de 2 quartetos com rima mista, com escolhas aleatórias de sílaba final, produzindo um padrão AABA ABBA. Para filtragem de sentenças candidatas, não foi estabelecida qualquer restrição métrica, gerando um poema heterométrico: 3 dodecassílabos, 1 decassílabo, 2 eneassílabos e 2 heptassílabos. Já que a obra possui o menor número total de sentenças, todas as métricas poderiam ser consideradas. Encontravam-se, disponíveis, uma quantidade razoável de sentenças candidatas, com 27 sentenças completas terminadas em 'te' e 28 em 'a'. A pontuação final foi 2.19 pontos, uma pontuação razoável principalmente porque se trata de um poema heterométrico. A acentuação grave de todas as últimas palavras da primeira estrofe, e todas as esdrúxulas na segunda, resultou numa pontuação máxima de acentuação. Isto evidencia que este é um critério de fácil pontuação, mesmo em um poema com várias métricas. Por outro lado, a pontuação pelo esquema rítmico, intermediária, foi de valor 0.61, já que a escolha pelo padrão heterométrico obteve sentenças de comprimentos diversos diminuindo as possibilidades de coincidência de posições entre as tônicas. Houve apenas uma coincidência de sílaba idêntica na mesma posição tônica, 'men' ('inclemente' e 'comovidamente'), com pontuação limitada a 0.03. Já na avaliação fonética, foram encontrados letras iguais para além das sílabas finais em pares de versos, gerando uma pontuação 0.55, com quatro letras finais idênticas ('ente') entre 'inclemente' e 'frente' e entre 'frente' e 'comovidamente', além das duas 'ia' entre 'platéia' e 'obrigatória' e três letras 'ite' entre 'leite' e 'noite'.

Os resultados apresentados exemplificam como poemas gerados a partir da prosa literária recebem avaliações finais variadas. O tamanho da obra fonte, com maior número de sentenças candidatas, pode favorecer a otimização

A noite desce qual um pássaro inclemente		
O meu relógio anda sempre para a frente		
Mostrou-nos uma carta e uma fotografia		
A nova se espalha comovidamente		
Amélia - Minha ama de leite		
Rosáceas sobre aspargos da platéia		
Vacina Obrigatória		
Lalá passou mal a noite		
escansão	**métrica**	**tônicas**
A/ n#oi/te/ d#es/ce/ qu#al/ um/ p#á/ssa/ro in/cle/m#en/te	Dodecassílabo	2 4 6 8 12
O/ m#eu/ re/l#ó/gi/o #an/da/ s#em/pre/ para a/ fr#en/te	Hendecassílabo	2 4 6 8 11
Mos/tr#ou/nos/ #u/ma/ c#ar/ta e #u/ma/ fo/to/gra/f#i/a	Dodecassílabo	2 4 6 7 12
A/ n#o/va/ se/ es/p#a/lha/ co/mo/vi/da/m#en/te	Dodecassílabo	2 6 12
A/m#é/li/a -/ M#i/nha/ #a/ma/ de/ l#e/i/te	Decassílabo	2 5 7 10
Ro/s#á/ce/as/ s#o/bre as/p#ar/gos/ da/ pla/t#é/i/a	Hendecassílabo	2 5 7 11
Va/c#i/na/ O/bri/ga/t#ó/ri/a	Heptassílabo	2 7
La/l#á/ pa/ss#ou/ m#al/ a/ n#o/i/te	Heptassílabo	2 4 5 7

Tabela 13: Poema 8 gerado pelo PROPOE a partir de *Grande Sertão: Veredas*

rítmica, embora não em todos casos e nem em todos critérios individuais considerados na avaliação. Os poemas exemplificados seguiram diferentes parâmetros de filtragem e de montagem, com quantidades maiores ou menores de estrofes. Mesmo com estas variações, baseadas na existência de uma quantidade suficiente de sentenças candidatas, as pontuações obtidas na avaliação rítmica exibiram mudanças pequenas de valores. Por outro lado, quando há uma quantidade limitada de sentenças, as pontuações tendem a exibir mudanças maiores de valores.

5. Considerações Finais

PROPOE é o primeiro sistema computacional a gerar poemas a partir de sentenças versificadas extraídas da prosa literária em língua portuguesa. A montagem dos poemas baseia-se na otimização combinatória de versos similares, por meio de um algoritmo guloso, que realiza a seleção de cada verso de acordo com critérios rítmicos e fonológicos. Tal estratégia permite reduzir o custo computacional, através da comparação de sentenças candidatas, gerando poemas com vários tipos de rima. Os poemas, de versos isométricos ou heterométricos, são avaliados pelos parâmetros que os caracterizam como poemas versificados. Os resultados mostram grande variação morfológica na construção dos poemas, provenientes de uma mesma obra, processada pelo MIVES. É importante enfatizar que a associação sequencial das sentenças (versos) não sofre qualquer restrição semântica; isso é, a sucessão dos versos não se submete a critérios semânticos de organização. Entretanto, a geração de poemas, a partir da mesma obra em prosa, fornece um contexto e uma certa coesão temática.

Outro aspecto importante é que, diferente do que observamos em muitos sistemas, as sentenças são integralmente preservadas e tratadas como unidades candidatas. As sentenças são mantidas, em suas formas originais, preservando os paralelismos estruturais, especialmente fonológicos e rítmicos, encontrados na obra em prosa de que foram extraídos. Ainda sobre a organização, são desconsideradas quaisquer descontinuidades sintáticas entre os versos (enjambment). Tanto os versos resultantes, quanto as estrofes, exibem uma estrutura sem ordenamento sintático. Os poemas gerados são submetidos a avaliação do próprio sistema, que baseia-se (i) na coincidência entre letras das rimas externas, (ii) sílabas tônicas e seus posicionamentos, e (iii) esquema rítmico de pares de versos. Em muitos casos os poemas recebem pontuação nula em certos critérios, resultado, por exemplo, da ausência de coincidência entre sílabas tônicas dos esquemas rítmicos, pontuando em outros. Embora, como já afirmamos, a avaliação desconsidere aspectos semânticos, concentrando-se em propriedades estruturais (especialmente padrões métricos e rítmico-fonológicos), encontramos alguma coesão temática nas construções, resultado, como afirmamos acima, de extrações da mesma obra de origem, em todos os casos.

Novos desenvolvimentos estão planejados para o PROPOE, principalmente ampliação e flexibilização de critérios para montagem dos poemas. Ao pontuar sentenças candidatas, além do esquema rítmico, consideramos a inclusão de critérios para possíveis rimas toantes e consoantes, rima interna, assim como similaridade semântica, em diversas escalas, parágrafos, sentenças ou palavras. Baseados em múltiplos critérios, tais padrões podem ser ponderados de formas diversas pelo usuário, por meio de pesos numéricos, permitindo obter diferentes resultados na otimização da geração dos poemas.

Também planejamos ampliar ou diversificar as fontes de sentenças, com possibilidade de compor um corpora de sentenças versificadas oriundas de mais de uma obra, por exemplo, do mesmo autor, do mesmo movimento literário, e mesmo de diferentes períodos e movimentos, produzindo maiores variações e novas possibilidades para geração de poemas.

Referências

Aamodt, Agnar & Enric Plaza. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications* 7(1). 39–59. 10.3233/AIC-1994-7104.

Agirrezabal, Manex, Bertol Arrieta, Aitzol Astigarraga & Mans Hulden. 2013. POS-tag based poetry generation with wordnet. Em *14th European Workshop on Natural Language Generation*, 162–166.

Ali, Manuel Said. 1999. *Versificação Portuguesa*. EDUSP.

Almeida, Guilherme de. 1946. A poesia d'Os Sertões. *Diário de São Paulo* 18.

Andrade, Mário de. 1988. *Macunaíma: o herói sem nenhum caráter*. Oficinas Gráficas de Eugenio Cupolo.

Andrade, Oswald de. 1933. *Serafim ponte grande*. Ariel, Editora Ltda.

Antonio, Jorge Luiz. 2009. Procedimentos poéticos com o(s) computador(es). *SIBILA: Revista de poesia e crítica literária* 21.

Balestrini, N. 1962. Tape Mark I. Em *Almanacco Letterario Bompiani*, V. Bompiani e C.

Bena, Brendan & Jugal Kalita. 2019. Introducing aspects of creativity in automatic poetry generation. Em *16th International Conference on Natural Language Processing*, 26–35.

Boden, Margaret A. 2009. Computer models of creativity. *AI Magazine* 30(3). 23–23. 10.1609/aimag.v30i3.2254.

Brailsford, Sally C, Chris N Potts & Barbara M Smith. 1999. Constraint satisfaction problems: Algorithms and applications. *European Journal of Operational Research* 119(3). 557–581. 10.1016/S0377-2217(98)00364-6.

de Campos, Augusto. 1996. Transertões. *Folha de São Paulo* 3 de novembro de 1996.

Carvalho, Cid Ivan da Costa, Davis Macedo Vasconcelos & Leonel Figueiredo de Alencar Araripe. 2012. Superando o estado da arte na etiquetagem morfossintática por meio de regras de pós-etiquetagem. Em *X Encontro de Linguística de Corpus: Aspectos metodológicos dos estudos de corpora*, 122–134.

Carvalho, Ricardo, Angelo Loula & João Queiroz. 2020. Identificação computacional de estruturas métricas de versificação na prosa de euclides da cunha / computational identification of versification metric structures in euclides da cunha's prose. *Revista de Estudos da Linguagem* 28(1). 41–68.

Castro, Leandro N. 2006. *Fundamentals of natural computing: basic concepts, algorithms, and applications*. Chapman and Hall/CRC.

Colton, Simon & Geraint A. Wiggins. 2012. Computational creativity: The final frontier? Em *20th European Conference on Artificial Intelligence*, 21–26.

Cormen, Thomas H, Charles E Leiserson, Ronald L Rivest & Clifford Stein. 2009. *Introduction to algorithms*. MIT press.

Cunha, Euclides da. 1902. *Os Sertões*. Fundação Biblioteca Nacional.

D'Ambrosio, Matteo. 2018. The early computer poetry and concrete poetry. *MATLIT: Materialidades da Literatura* 6. 51–72.

Das, Amitava & Björn Gambäck. 2014. Poetic machine: Computational creativity for automatic poetry generation in bengali. Em *5th International Conference on Computational Creativity (ICCC)*, 230–238.

Durkin, John. 1994. *Expert Systems: Design and development*. Prentice Hall PTR.

Eiben, Agoston E. & James E. Smith. 2003. *Introduction to evolutionary computing*. Springer. 10.1007/978-3-662-05094-1.

Gatt, A. & E. Krahmer. 2018. Survey of the state of the art in natural language generation: Core

tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61. 65–170.

Gervás, Pablo. 2000. WASP: Evaluation of different strategies for the automatic generation of spanish verse. Em *AISB-00 Symposium on Creative & Cultural Aspects of AI*, 93–100.

Gervás, Pablo. 2001. An expert system for the composition of formal spanish poetry. Em *Applications and Innovations in Intelligent Systems VIII*, 19–32. Springer.

Gervás, Pablo. 2016. Constrained creation of poetic forms during theme-driven exploration of a domain defined by an n-gram model. *Connection Science* 28(2). 111–130. 10.1080/09540091.2015.1130024.

Goldstein, Norma Seltzer. 2008. *Versos, sons, ritmos*. Editora Ática.

Gonçalo Oliveira, Hugo. 2012. PoeTryMe: a versatile platform for poetry generation. *Computational Creativity, Concept Invention, and General Intelligence* 1. 21.

Gonçalo Oliveira, Hugo. 2017a. O poeta artificial 2.0: Increasing meaningfulness in a poetry generation twitter bot. Em *Workshop on Computational Creativity in Natural Language Generation (CC-NLG)*, 11–20.

Gonçalo Oliveira, Hugo. 2017b. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. Em *10th International Conference on Natural Language Generation*, 11–20. 10.18653/v1/W17-3502.

Gonçalo Oliveira, Hugo, F. Amilcar Cardoso & Francisco C. Pereira. 2007. Tra-la-Lyrics: An approach to generate text based on rhythm. Em *4th International Joint Workshop on Computational Creativity*, s/pp.

Hämäläinen, Mika, Khalid Alnajjar & Thierry Poibeau. 2022. Modern french poetry generation with roberta and gpt-2. Em *International Conference on Computational Creativity (ICCC)*, s/pp.

Jakobson, Roman & Krystyna Pomorska. 1985. *Diálogos*. Editora Cultrix.

Kay, Martin. 1996. Chart generation. Em *34th Annual Meeting on Association for Computational Linguistics*, 200–204. 10.3115/981863.981890.

Lutz, Theo. 1959. Stochastische texte. *augenblick* 4(1). 3–9.

Manurung, Hisar. 2004. *An evolutionary algorithm approach to poetry generation*: University of Edinburgh. College of Science and Engineering. Tese de Doutoramento.

Manurung, Hisar, Graeme Ritchie & Henry Thompson. 2000. Towards a computational model of poetry generation. Relatório técnico. The University of Edinburgh.

Netzer, Yael, David Gabay, Yoav Goldberg & Michael Elhadad. 2009. Gaiku: Generating Haiku with word associations norms. Em *Workshop on Computational Approaches to Linguistic Creativity*, 32–39.

Otter, Daniel W., Julian R Medina & Jugal K. Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems* 32(2). 604–624. 10.1109/TNNLS.2020.2979670.

Perloff, Marjorie. 2013. *O gênio não original: poesia por outros meios no novo século*. UFMG.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever. 2019. Language models are unsupervised multitask learners. openai blog. 2019; 1 (8). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rahman, Fahrurrozi & Ruli Manurung. 2011. Multiobjective optimization for meaningful metrical poetry. Em *International Conference on Computational Creativity (ICCC)*, 4–9.

Rashel, Fam & Ruli Manurung. 2014. Pemuisi: a constraint satisfaction-based generator of topical indonesian poetry. Em *International Conference on Computational Creativity (ICCC)*, 82–90.

Rosa, João Guimarães. 1956. *Grande sertão: veredas*. Livraria José Olympio Editora.

Ryan, Marie-Laure, Lori Emerson & Benjamin J Robertson. 2014. *The Johns Hopkins guide to digital media*. JHU Press.

Silva, Pedro Paulo da (ed.). 2014. *Teoria da Literatura I*. Editora Pearson.

Sousa, Tatiane. 2007. *Haicais de Bashô: O oriente traduzido no ocidente*: Universidade Estadual do Ceará, Fortaleza. Tese de Mestrado.

Souza, Erthos Albino de. 1991. Le Tombeau de Mallarmé. Em *Mallarmé*, Perspectiva.

Spina, Segismundo. 2003. *Manual de versificação românica medieval*. Ateliê Editorial.

Tobing, Berty Chrismartin Lumban & Ruli Manurung. 2015. A chart generation system for topical metrical poetry. Em *International Conference on Computational Creativity (ICCC)*, 308–314.

Wong, Martin Tsan, Andy Hon, Wai Chun & Tat Chee. 2008. Automatic haiku generation using VSM. Em *7th WSEAS International Conference on Applied Computer and Applied Computational Science*, 318–323.

Zugarini, Andrea, Stefano Melacci & Marco Maggini. 2019. Neural poetry: Learning to generate poems using syllables. Em *International Conference on Artificial Neural Networks*, 313–325. doi 10.1007/978-3-030-30490-4_26.

Propuesta recibida en junio 2022 y aceptada para publicación en diciembre 2022.

ARAPP: Análisis y Resumen Automático de Políticas de Privacidad

Analysis and Automatic Summary of Privacy Policies

Rodrigo Alfaro
Pontificia Universidad Católica de Valparaíso

René Venegas
Pontificia Universidad Católica de Valparaíso

Alan Bronfman
Pontificia Universidad Católica de Valparaíso

Miguel Valenzuela
Pontificia Universidad Católica de Valparaíso

Stephanie Riff
Pontificia Universidad Católica de Valparaíso

Enrique Sologuren
Universidad del Desarrollo

Resumen

Un derecho fundamental de los usuarios de aplicaciones informáticas es que puedan conocer las políticas de privacidad (PP) que tales aplicaciones establecen, en particular es relevante que conozcan acerca del tratamiento que aceptan sobre el uso de sus datos. No obstante, estas PP son muy extensas y escritas en un lenguaje administrativo-jurídico y comercial, lo que dificulta su lectura y comprensión. El objetivo de este artículo es resumir automatizadamente las PP de cinco aplicaciones de redes sociales (Facebook, Twitter, TikTok, Snapchat e Instagram) en español, a través de técnicas extractivas y abstractivas. Para ello se utilizan tres aproximaciones de representación desde el Procesamiento de Lenguaje Natural, estas son: Teoría de Grafos, TF-IDF y Gensim. A partir de ellas, se generan automáticamente 15 resúmenes, los que son evaluados por un experto en derecho, para medir la legibilidad y relevancia en base a 20 preguntas confeccionadas por un estudio de la Universidad de Austin, Texas (Zaeem et al., 2018). Por último, a partir de una clasificación de cada política de privacidad, según distintos factores de riesgos, se comprueba que el método Gensim es el más adecuado para la representación y resumen. Además se identifica a Snapchat como la aplicación que mejor cumple dichos factores.

Palabras clave

resumen automático; políticas de privacidad; factores de riesgo; textos jurídicos; Gensim; redes sociales

Abstract

A fundamental right of the users of computer applications is that they can know the privacy policies (PP) that such applications establish. It is particularly relevant that they know about the treatment that they accept regarding the use of their data. However, these PP are very extensive and written in administrative-legal and commercial language, which makes them difficult to read and understand. The aim of this paper is to automatically summarize the PPs of five social network applications (Facebook, Twitter, TikTok, Snapchat and Instagram) in spanish, through extractive and abstractive techniques. For this purpose, three representation approaches from Natural Language Processing are used, these are: Graph Analysis, TF-IDF and Gensim. Fifteen summaries were automatically generated and evaluated in order to measure the readability and relevance, by an expert in law, based on 20 questions prepared by a study of the University of Austin, Texas (Zaeem et al., 2018). Finally, based on a classification of each privacy policy according to different risk factors, the Gensim method is found to be the most suitable for the representation and summarization of the PP's. The PP of Snapchat is also identified as the application that best meets these risk factors.

Keywords

automatic summarization; private policies; risk factors; legal texts; Gensim; social networks

1. Introducción

La privacidad es entendida como todo ámbito de la vida privada que se tiene derecho a proteger de cualquier intromisión (RAE, 2021). Hacia fines del siglo XIX el derecho a la privacidad fue entendido como *"the right to be let alone"*, el derecho a ser dejado solo (Warren & Brandeis, 1890). Este derecho protege la vida privada como un ámbito o espacio libre de intrusión o invasión por parte de un tercero, sin perjuicio de intromisiones justificadas por una necesidad manifiesta de una comunidad que vive en un estado de derecho (Corral, 2000). En este sentido, el reconocimiento y garantía de la esfera privada no solo atiende a una dimensión individual o subjetiva, sino que tam-

DOI: 10.21814/lm.14.2.375

bién considera la defensa de la privacidad desde una dimensión colectiva y social, que coadyuva al mantenimiento y avance del sistema democrático (Saldaña, 2012).

Un elemento distintivo de aquello que forma parte de la vida privada es el control sobre el acceso, el cual pertenece a su titular. La pérdida o menoscabo de dicho control constituye una pérdida o amenaza sobre el derecho a la vida privada.

El desarrollo de la informática y el tratamiento de datos en la segunda mitad del siglo pasado puso de manifiesto su incidencia en ámbitos que forman parte de la vida privada, lo que impulsó la construcción del concepto de datos personales y el establecimiento de medios para su protección. Hace más de cuarenta años, leyes tales como la *Datenshutz* (Alemania, 1970), *Data Lag* (Suecia, 1973) o la *Privacy Act* (Estados Unidos, 1974), entre otras, reconocieron la necesidad de proteger los datos personales mediante la regulación de su uso, almacenamiento y transmisión (de la Cueva & Fernández-Miranda, 1990; Pérez Luño, 1984). Así, por ejemplo, en 1983 un fallo del Tribunal Constitucional alemán precisó esta tutela con la idea de *autodeterminación informativa*, la que constituye un derecho que permite al individuo decidir por sí mismo cuándo y dentro de qué límites procede revelar situaciones referidas a su propia vida (Schwabe, 2009; TCCh, 1989).

El derecho a la vida privada, entonces, reconoce a cada persona el poder de decidir acerca del uso, almacenamiento y transmisión de sus datos personales. Esta facultad de decidir requiere conocer qué datos personales serán registrados en las bases de datos de cada aplicación y cómo serán tratados y utilizados. Las políticas de privacidad (PP) contienen una declaración pública, de base jurídica y estándar, de los propietarios o controladores de una aplicación dirigida a sus potenciales usuarios respecto a los datos que ella utilizará y su tratamiento. Dichas PP han sido redactadas de modo unilateral por los propietarios o controladores de la aplicación de acuerdo con su conveniencia e interés y, por lo mismo, no contienen necesariamente la información pertinente y relevante acerca de la afectación de la privacidad derivada del uso de la aplicación. Es claro que el registro y tratamiento de datos por parte de estas aplicaciones o plataformas genera instrumentos precisos de focalización publicitaria que son difíciles de compatibilizar con la tutela jurídica de la vida privada. En las dimensiones globales y potencial económico del mercado creado por la red internet, la opacidad de las PP ofrece un claro beneficio para los propietarios o controladores de las aplicaciones.

Ahora bien, las PP son una herramienta esencial para la toma de decisiones del potencial usuario en lo que a su derecho a la vida privada se refiere (Anastasopoulou et al., 2017). Muchas veces son extensas, confusas y utilizan de modo reiterado términos técnicos, lo que dificulta su lectura. Con frecuencia son aceptadas sin ser comprendidas por los usuarios que desean o necesitan acceder a los servicios de la aplicación ofrecida. De acuerdo con los estudios que informan acerca de los niveles de lecturabilidad, la comprensión de este tipo de documento requiere de personas que al menos hayan completado dos años de estudios universitarios (Zaeem et al., 2018). La complejidad de estos textos genera costos de oportunidad del orden de U$D 781 mil millones de dólares (McDonald & Cranor, 2009), si se considera el tiempo de lectura que requieren, los cuales ascienden de forma individual a 201 horas por año. Lo que excede el porcentaje de tiempo dedicado en promedio a la navegación en la web (Meier et al., 2020). Este costo podría reducirse si es posible comunicar al usuario de modo más breve y rápido aquello que es esencial para decidir sobre la tutela de su vida privada.

En este marco, adquieren especial importancia las técnicas para el análisis de los textos que utilizan lenguaje jurídico o legal y que exhiben una alta dificultad para su comprensión en las personas debido a su opacidad discursiva (Meza et al., 2021; Montolío & López Samaniego, 2008). Dentro de las técnicas disponibles para el estudio de estas formas textuales, en este ejercicio, utilizamos aquellas dirigidas a evaluar la posibilidad de generar información sintética y comprensible para el usuario de aplicaciones tecnológicas como insumo para el ejercicio de su derecho a la privacidad (Montolío & Tascón, 2020). Ello porque solo la comprensión cabal del uso, almacenamiento y transmisión de los datos personales asociado a la utilización de cada aplicación permite, desde el punto de vista jurídico, otorgar un consentimiento válido para el acceso al ámbito protegido por el derecho a la vida privada.

El enfoque de la comunicación clara agrega otro paradigma aplicable al análisis de las PP. La comunicación clara precave litigios y conflictos entre empresas y los usuarios de sus servicios, en este caso, gracias a la oportuna entrega de información completa e inteligible sobre la afectación de la privacidad ocasionada por el uso de una aplicación. La comunicación clara enfatiza un cambio de cultura comunicativa en las empresas y organizaciones, promoviendo que las cláusulas de los contratos sean claras, precisas y con una estructura lo más transparente y visual

posible. Todo ello con el fin de mejorar la experiencia del usuario así como de que cualquier cliente pueda entenderlas y saber a qué está adhiriendo (da Cunha & Escobar, 2021; Montolío & López Samaniego, 2008).

Los estudios en el ámbito del lenguaje jurídico claro son de reciente data en el idioma español. Si bien existen propuestas de modernización, clarificación y normalización del discurso jurídico en España (García Calderón, 2019; Montolío & López Samaniego, 2008) y Latinoamérica (Cucatto, 2011; Meza et al., 2021; Poblete & Fuenzalida González, 2018), los intentos por solucionar estos problemas con ayuda de herramientas automatizadas son aún escasos Ruidías et al. (2018). En este sentido, el trabajo presentado por Valenzuela et al. (2020) y el realizado en este artículo busca llenar este vacío, en especial considerando la potencial utilidad de las PP en la protección efectiva de la privacidad de millones de usuarios de las aplicaciones más populares. Según el informe *Global Digital Overview* 2022 (DataReportal, 2022), elaborado por las empresas *We are Social* y *Hootsuite*, en el mes de abril, el 63 % de las personas del planeta usaron internet y, de ellas, más de 326 millones utilizaron las redes sociales. En el caso de facebook, por ejemplo, el número mensual de usuarios activos de en el año 2021 fue de 2,9 billones (DataReportal, 2022).

En este estudio se somete a las PP en español de Twitter, Instagram, Facebook, Snapchat y TikTok a métodos automatizados de resumen con el propósito de mejorar el conocimiento y comprensión de los usuarios sobre el tratamiento que recibirán sus datos personales. Para esta labor se intenta identificar los factores de riesgo más importantes de las políticas de privacidad declaradas y, sobre esta base, seleccionar el resumen automático de mayor calidad.

En la identificación de los factores de riesgo se utiliza el estudio de Zaeem et al. (2020), que fundamenta la herramienta que se ofrece con el nombre de *PrivacyCheck*. *PrivacyCheck* es una aplicación innovadora que analiza las PP mediante la extracción automática de textos en línea, utilizando modelos de minería de textos. Si bien esta herramienta presenta una propuesta para evaluar PP en idioma inglés, no se presenta un resumen para que usuarios comprendan qué están aprobando cuando aceptan una PP.

Como se indicó arriba, el *input* del proceso son las PP de Twitter, Instagram, Facebook, Snapchat y TikTok que se utilizaron en Chile durante el período octubre 2020–enero 2021. En el proceso se aplican dos métodos de técnica extractiva (Análisis de Grafos y TF-IDF) y uno de técnica abstractiva (Gensim), para obtener como resultado quince resúmenes automáticos distintos. Las dos primeras técnicas tienen en común la extracción y priorización de oraciones de los textos originales, en tanto que en la tercera se hace uso de técnicas de similitud semántica, para generar oraciones parafraseadas, y luego se priorizan las más relevantes (ver 3.3). Las variables analizadas en los textos son la cantidad de palabras y el tiempo de lectura; así como la legibilidad y la relevancia (van de Luijtgaarden, 2019). La primera se define como la fluidez, gramaticalidad y coherencia, en tanto la segunda como presencia de información importante en el resumen. Además se complementa el análisis con las veinte preguntas proporcionadas por el estudio de *PrivacyCheck* para medir el riesgo en los resúmenes de las PP.

Cabe señalar que el desarrollo de esta investigación enfrentó tres dificultades. Por una parte, existen pocos estudios relacionados al análisis y resumen automático de las políticas de privacidad (ARAPP) en español, por lo que no se cuenta con un *corpus* público de políticas de privacidad, especialmente de las redes sociales y, menos aún, con un algoritmo eficiente para su desarrollo dentro de América Latina. Por otra parte, no se cuenta con resúmenes de referencia elaborados por humanos que sirvan para construir una evaluación computacional de los resúmenes automáticos. Finalmente, tampoco existe una regulación de protección de datos que sea imperativo cumplir por las aplicaciones analizadas (sin perjuicio de la existencia de normas y estándares nacionales e internacionales que no son vinculantes para ellas) y que nos permita determinar el correcto contenido de una PP.

La exploración interdisciplinar que proponemos es el inicio de futuras investigaciones y podría generar beneficios para los usuarios de aplicaciones informáticas como también para las empresas que buscan avanzar hacia una relación comunicativa más transparente, leal y efectiva con sus usuarios.

2. Marco Teórico

El Procesamiento de Lenguaje Natural (PLN) es una subdisciplina aplicada que se centra en investigar y formular soluciones computacionales que faciliten la interrelación hombre-máquina mediante la automatización de procesos relacionados con la comunicación humana (Zhang & Lu, 2021). Por lo tanto, el PLN es la manipulación informática del lenguaje natural que integra la lingüística y la matemática (Rodrigo & Allende, 2020) y que puede aplicarse con diversos matices

(Bird et al., 2009). Entre sus potenciales aplicaciones se encuentran la recuperación de información, la clasificación automatizada, la generación automática de resúmenes, entre otras.

La generación automática de resúmenes nace con Luhn (1958) y es un área del PLN que según Maybury (1995), puede entenderse como un conjunto de procesos que "destila la información más importante de una fuente (o fuentes) para producir una versión abreviada de la información original para un usuario o tareas en particular" (p. 735). Existen dos tipos de técnicas de resumen automático: extractivas y abstractivas. Rane & Govilkar (2019) explican que las técnicas extractivas generan resúmenes mediante la frecuencia de las frases del texto original para ser clasificadas por prioridad, la cual está basada en características lingüísticas y estadísticas, para luego combinarlas en un solo texto de salida. Según Hernández (2017), su implementación es más fácil que su contraparte abstractiva y garantiza la coherencia mínima de la información generada. En cuanto a la técnica abstractiva, las oraciones generadas son parafraseadas del texto de entrada, esto es, se incluyen palabras que no aparecen en el texto original (Widyassari et al., 2022), obteniéndose un resultado similar a un resumen creado por una persona. Esta técnica es más compleja, ya que necesita procesamiento de lenguaje natural y recursos de lenguaje para obtener palabras similares provenientes de otros textos (Gambhir & Gupta, 2016).

Entre las técnicas extractivas se cuenta *Term frequency - Inverse document frecuency* (TF-IDF) Christian et al. (2016), que calcula la estadística de una palabra dentro de un documento y un *corpus* de documentos (Salton & Buckley, 1988), y representa aquellas palabras más relevantes del texto y subrepresenta aquellas que se repiten en todos los textos del corpus. El término de frecuencia (TF) significa la frecuencia bruta de un término en un documento y el término de la frecuencia inversa en el corpus (IDF) es una medida que indica si la palabra es común o rara en todos los documentos analizados.

Otra técnica es el método basado en grafos, que fue planteado para resúmenes automáticos por Erkan & Radev (2004). La técnica se denomina *LexRank* y define las aristas como relaciones de co-ocurrencia de términos entre oraciones. Ella ha sido perfeccionada en el tiempo, evolucionando desde los modelos bipartitos (Wan et al., 2021) hasta los modelos de hipergrafo. El método grafo-analítico se concibe como un modelo más preciso, en contraposición al método por frecuencias. Como es posible notar, el método grafo-analítico requiere más tiempo de procesamiento a medida que los textos de entrada se incrementan. Aun así, es capaz de mostrar información relevante (Karmaker & Hossen, 2019) y, por lo mismo, más preciso que el método por frecuencias. Lierde & Chow (2019) presentan el método grafo en el que los nodos son oraciones y las aristas el número de palabras en común entre las oraciones. Ellos proponen un nuevo modelo de hipergrafo para capturar la relación semántica entre oraciones. En él se utilza un enfoque de selección de oraciones, basado en la maximización de la relevancia de oraciones individuales y la cobertura temática, y un algoritmo polinomial basado en la teoría de funciones submodulares para resolver problemas de optimización.

Una técnica abstractiva actual es la proporcionada por Řehůřek (2021), denominada Gensim, la cual está documentada en lenguaje de programación para su utilización mediante una librería informática. Esta es una librería abierta (*open-source*), desarrollada en Python para representar documentos en vectores semánticos de manera eficiente computacionalmente. Está diseñada para ser de fácil utilización. Gensim procesa textos planos con algoritmos de aprendizaje de máquina no supervisados, entre ellos, Word2Vec, FastText, Latent Semantic Indexing (LSI, LSA, LsiModel), Latent Dirichlet Allocation (LDA, LdaModel), etc. Estos permiten descubrir automáticamente la estructura semántica de los textos utilizados para el entrenamiento de los algoritmos, a través del análisis de los patrones estadísticos de co-ocurrencia de la información lingüística del texto. Estos algoritmos no requieren información aportada por el analista humano y solo necesitan el texto plano como entrada del modelo. Así se evita el sesgo humano en el análisis. Una vez identificados los patrones estadísticos de palabras, frases u oraciones, estos pueden ser representados semánticamente y ser utilizados para la comparación de documentos, a través de las técnicas de similitud de tópicos. Para el caso del resumen automatizado, se ofrece un módulo en el que dado un texto se extraen de este una o más oraciones relevantes, así como las palabras clave y se entrega un resumen. Se utiliza para ello una variación del algoritmo TextRank (Barrios et al., 2016).

En relación con la calidad de los resúmenes que puedan ser producidos automáticamente, van de Luijtgaarden (2019) propone una metodología que identifica la legibilidad y la relevancia en una escala del 1 al 10, de acuerdo con evaluadores humanos, en nuestro caso un abogado.

Un segundo estándar de calidad podría construirse desde la regulación. En Europa en ma-

yo de 2018 entró en vigor el *Reglamento General de Protección de Datos* (RGPD) (2016/679), que contiene un conjunto de principios que las empresas deben observar, tales como el de responsabilidad, la transparencia y la protección de datos personales. Este reglamento europeo dispone de medidas técnicas y organizativas para garantizar que los datos sean únicamente objeto de tratamiento para los fines específicos que sean entregados, reduciendo la extensión del tratamiento, limitando el plazo de conservación y su accesibilidad. En el estado de California en Estados Unidos rige la Ley de privacidad del consumidor (CCPA) (Becerra, 2020), la que estipula que las empresas reguladas deberán proteger la información personal de los consumidores en California y en Chile existe la (Ley Chile, 2014) 19.628 para la protección de datos de carácter personal. También existen directrices de la OCDE sobre protección de la privacidad y los flujos transfronterizos (OCDE, 2002). El estándar de calidad que podría nacer de la identificación en los resúmenes de referencias relevantes a las materias objeto de reglas y principios podría adolecer de problemas de generalidad (proveniente de las propias normas y la técnica legislativa) y territorialidad, toda vez que dicho estándar solo tendría valor para el territorio estatal donde la correspondiente regulación tiene vigencia y aplicación.

Por último, Zaeem et al. (2018) proponen *Privacy Check*, una herramienta que aborda el análisis de las políticas de privacidad mediante la extracción automática de resúmenes de las políticas de privacidad en línea, utilizando modelos de minería de textos que responden a diez preguntas dirigidas a evaluar la privacidad y seguridad de los datos del usuario. Las preguntas (traducidas al español) son:

1. ¿Qué tan bien protege este sitio web su dirección de correo electrónico?
2. ¿Qué tan bien protege este sitio web la información y la dirección de su tarjeta de crédito?
3. ¿Qué tan bien maneja este sitio web su número de seguro social?
4. ¿Este sitio web utiliza o comparte su información de identificación personal con fines de marketing?
5. ¿Este sitio web rastrea o comparte su ubicación?
6. ¿Este sitio web recopila información de identificación personal de niños menores de 13 años?
7. ¿Este sitio web comparte su información con las fuerzas del orden?
8. ¿Este sitio web notifica o le permite darse de baja después de cambiar su política de privacidad?
9. ¿Este sitio web le permite editar o eliminar su información de sus registros?
10. ¿Este sitio web recopila o comparte datos agregados relacionados con su identidad o comportamiento?

Este modelo se entrenó con cuatrocientas empresas de diversas industrias. Su implementación utiliza una extensión en el navegador Chrome, gratuita y disponible para todo público, aplicable a cualquier política de privacidad en línea. Los resultados demostraron que los resúmenes de *PrivacyCheck* son precisos entre el 40 % y el 73 %. Finalmente, *PrivacyCheck* clasifica el riesgo de las Políticas de Privacidad (PP) en tres niveles (verde, amarillo o rojo) para cada pregunta (factor de riesgo) en más de cuatrocientas compañías independientes.

En 2020 se presentó *PrivacyCheck v2*, que agregó diez nuevos factores de riesgo a partir del Reglamento General de Protección de Datos (RGPD), arriba mencionado, e incorporó una herramienta de análisis de las PP. Las 10 preguntas que se incorporaron son:

1. ¿Comparte la información del usuario con otros sitios web solo con el consentimiento del usuario?
2. ¿Revela dónde se encuentra la empresa / dónde se procesará y transferirá la información de identificación personal del usuario?
3. ¿Apoya el derecho al olvido?
4. Si retienen información de identificación personal con fines legales después de la solicitud del usuario de ser olvidado, ¿informarán al usuario?
5. ¿Permite al usuario rechazar el uso de la información de identificación personal del usuario?
6. ¿Restringe el uso de información de identificación personal de niños menores de 16 años?
7. ¿Advierte al usuario que sus datos están encriptados incluso en reposo?
8. ¿Solicita el consentimiento informado del usuario para realizar el procesamiento de datos?
9. ¿Implementa todos los principios de protección de datos por diseño y por defecto?
10. ¿Notifica al usuario las violaciones de seguridad sin demoras indebidas?

Si bien, la herramienta *PrivacyCheck* permite evaluar riesgos de las PP en idioma inglés, no se aplica al idioma español, ni tampoco permite a los usuarios comprender, mediante un texto resumido, qué están aprobando cuando aceptan una PP.

3. Metodología

En primer lugar, se recopilaron los documentos a ser resumidos por medio de la identificación de las principales aplicaciones informáticas, con base en un criterio intencionado por los investigadores. La recopilación, en este sentido, fue de carácter manual. Luego, se aplicaron las técnicas de elaboración automatizada de resumen consideradas en esta investigación. Finalmente, se comparó la calidad de los resúmenes, utilizando criterio de experto.

3.1. TF-IDF

Se utilizan programas computacionales implementados en las siguientes librerías: numpy, pandas, nltk, heapq y re. Se inició con la tokenización del texto original por palabras. Se procede con la limpieza de los datos, se remueven los caracteres no alfabéticos y se reemplazan las mayúsculas por minúsculas, para ser almacenados en variables tipo texto. Luego de generar las condiciones de trabajo, se obtiene el puntaje TF-IDF, que consiste en encontrar la frecuencia de las palabras dentro del texto, para calcular la frecuencia ponderada de cada término y en consecuencia calcular puntuaciones de las oraciones a partir de la frecuencia de aparición de palabras. Finalmente, se genera el resumen a partir de las diecisiete oraciones con mejor puntaje (Christian et al., 2016).

3.2. Análisis de Grafos

Se utiliza las librerías numpy, pandas, nltk y re para resumir el texto a través de *text rank*. Se inicia con la limpieza de datos, removiendo las *stopwords*, corchetes y espacios extras, caracteres que no son letras y cambios de mayúsculas a minúsculas. Luego, se procede con la representación del vector de oración, el cual se crea utilizando el modelo *word2vec* previamente entrenado de Gensim, es decir, se componen las *word embeddings* (representación de palabras como vectores de números reales). Se conforman los grafos a partir de las oraciones, para ello se consideran las oraciones como los nodos y las aristas se ponderan a partir de la similitud del coseno entre las oraciones. Finalmente, se producen las oraciones utilizando el algoritmo *PageRank*, imprimiendo las cinco oraciones mejor clasificadas.

3.3. Gensim

Como ya se mencionó, este es un método de representación de textos, el que, a partir de la co-ocurrencia de oraciones, frases y palabras permite realizar resúmenes automáticos de manera no supervisada. Para su implementación se utiliza las librerías gensim, logging, numpy, pandas y re. Su implementación es más simple que los dos anteriores, los datos se leen en la fase inicial, se ingresan al método de resumen y se obtienen resultados variando los valores del parámetro *split*. Los parámetros de este método son:

- Text (str): Texto de entrada.
- Ratio (flotante, opcional): Número entre 0 y 1 que determina la proporción del número de frases del texto original que se elegirán para el resumen.
- Word_count (int o None, opcional): determina cuántas palabras contendrá la salida. Si se proporcionan ambos parámetros, la relación se ignorará.
- Split (*bool*, opcional): si es *True*, se devolverá la lista de oraciones. De lo contrario, las cadenas unidas se devolverán.

La salida de este algoritmo es un resumen que tiene como parámetro el porcentaje de palabras a incorporar en él, en nuestro caso utilizamos el 10 %.

3.4. Evaluación de los Resúmenes

Se utiliza una metodología de evaluación y recolección de datos sobre la base de dos formularios, utilizando la plataforma *Google Forms*. El primero se compone de tres secciones. En la primera se identifica el documento evaluado según el Cuatro 1 y se señala el tiempo de lectura.

TF-IDF	Gensim	Grafos
001-Twitter	011-Twitter	101-Twitter
002-Facebook	012-Facebook	102-Facebook
003-Instagram	013-Instagram	103-Instagram
004-Snapchat	014-Snapchat	104-Snapchat
005-TikTok	015-TikTok	105-TikTok

Cuadro 1: Codificación de los resúmenes automáticos.

En la siguiente sección se evalúan los resúmenes en una escala del 0 al 10 de legibilidad, la que se define como la fluidez, gramaticalidad y coherencia del resumen. Luego se evalúa la relevancia, es decir, si contiene información importante y destacada de la política de privacidad, siendo correcto al evitar información contradictoria, no relacionada y evita información repetida o redundante. Por último, se finaliza con una descripción de las fortalezas, debilidades, oportunidades y amenazas, incluyendo apreciaciones generales respecto de los resúmenes. La segunda encuesta, comienza de la misma manera y considera una sección en la que se identifica el contenido de las PP ideales, a partir de los factores de la Tabla 4, clasificándolos del 1 al 10 (no excluyentes), además de tener la posibilidad de incluir otros. Es importante destacar que el evaluador, experto en derecho, se enfrenta a los resúmenes sin saber qué método fue el utilizado, con el fin de disminuir el sesgo de la persona.

Los datos se analizan utilizando el Principio de Pareto, también conocido como la regla del 80–20, que plantea que, en diferentes fenómenos, una pequeña proporción de los elementos es la que explica la mayor parte del efecto. Lo anterior es complementado con el análisis FODA por método y por política. Bajo este análisis se busca identificar las fortalezas y debilidades de los métodos, así como las dificultades y oportunidades para el trabajo con textos de PP. A partir de lo anterior, se logra la sinergia entre los análisis cuantitativo y cualitativo. Finalmente, con el fin de encontrar relaciones entre las variable de relevancia, cantidad de palabras y tiempo de lectura, se utiliza la correlación de Pearson.

Se utilizó la herramienta *Google Colaboratory* y el lenguaje de programación *Python*, dada su fácil implementación desde cualquier ordenador con conexión a internet, mediante la plataforma de *Google Drive*. Los códigos computacionales de las librerías para la implementación de las técnicas se obtuvieron desde el repositorio *GitHub* [1].

4. Resultados

Se generó un corpus de cinco políticas de privacidad de las redes sociales en cuestión, que en promedio tienen 5.042,8 de palabras, tal como se presenta en la Figura 1, los que para ser procesados deben tener un formato `*.txt`.

Además, se implementan tres métodos de resumen automático: basado en Análisis de Grafos, TF-IDF y Gensim, lo que permite tener quince

[1] https://github.com/ShristiK/Text-Summarization

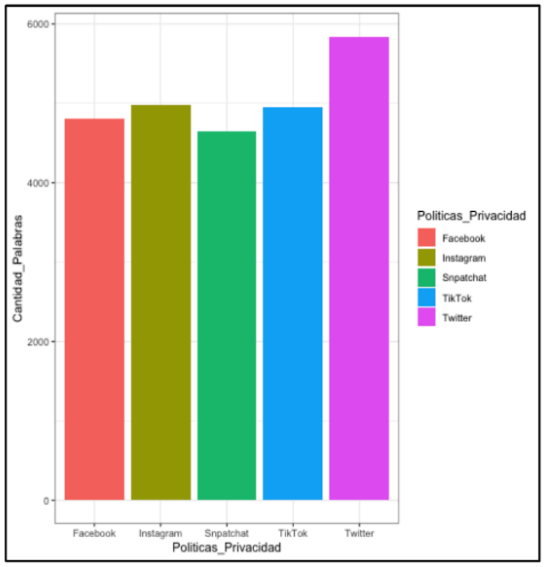

Figura 1: Cantidad de palabras de las 5 políticas de privacidad.

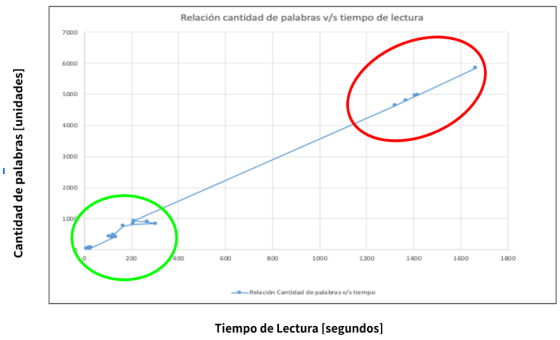

Figura 2: Relación entre cantidad de palabras y tiempo de lectura.

resúmenes distintos. Para el primero obtenemos en promedio una cantidad de 51,4 palabras y un tiempo de lectura de 21,4 segundos, para el segundo 420,6 palabras y 121,6 segundos respectivamente, y, por último, 845,6 palabras y 232 segundos, los que fueron evaluados de forma cualitativa con su complemento cuantitativo.

A partir del formulario respondido por el experto (abogado evaluador), se obtuvo una legibilidad de un 100 % para los quince resúmenes. Esto quiere decir, que en la lectura de los resúmenes no tuvo problemas respecto de la fluidez, gramaticalidad y coherencia de los mismos.

La Figura 2 presenta la relación entre la cantidad de palabras y el tiempo de lectura. Se observa una relación significativa alta, según el coeficiente de correlación de Pearson de $r = 0,955$ ($p < 0,001$), considerando las políticas de privacidad completa. Similar resultado se obtiene para la relación entre relevancia y la cantidad de palabras, $r = 0,878$ ($p < 0,001$), tal como se presenta en las Figuras 3 y 4 respectivamente.

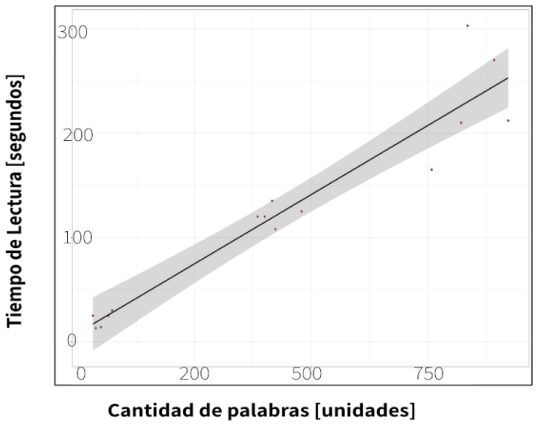

Figura 3: Correlación de Pearson: Cantidad de palabras (unidades) vs Tiempo de lectura (segundos) (promedios).

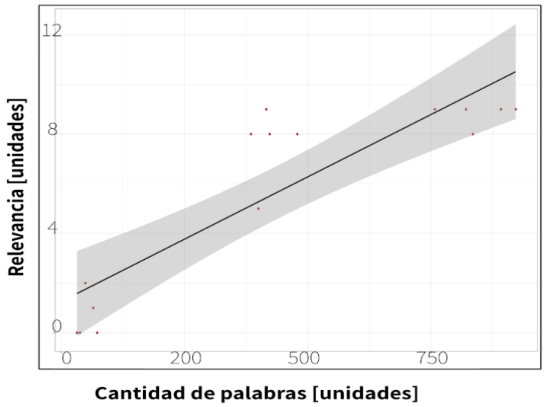

Figura 4: Correlación de Pearson: Cantidad de palabras (unidades) vs Relevancia (unidades) (promedios).

A continuación, se presentan las Figuras 5, 6, 7 y 8. En la primera se indica que las preguntas 4, 10 y 7 fueron las mejores evaluadas y, contrariamente, las preguntas 12, 14, 16, 17, 19 y 20 como las peor evaluadas.

La Figura 6 representa los mejores resúmenes, sin considerar el valor de la relevancia. Así el 015, 001 y el 011 se identifican como los mejor evaluados, por contraparte, los resúmenes 102, 104 y el 103 se identifican como los peor evaluados.

La Figura 7 nos indica qué resúmenes obtuvieron mejores y peores resultados solamente en términos de la relevancia, siendo los primeros el 003, 013, 014, 015, 011 y, por el otro lado, el 103, 105 y el 101.

Finalmente, la Figura 8 es la combinación de las Figuras 6 y 7, es decir, consideran las veinte preguntas con más relevancia en la que los con mejor calificación son el 015, 011 y 103. Para todos los gráficos el color verde representa el mejor 80 % y el color rojo el 20 % restante, según el análisis de Pareto.

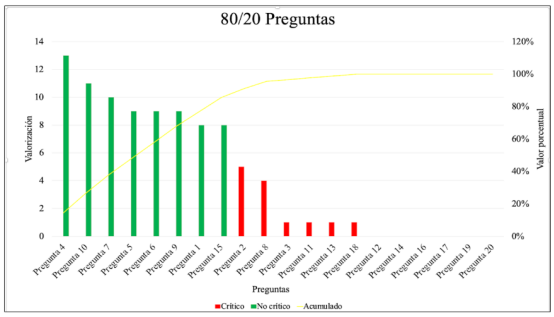

Figura 5: Análisis de Pareto de las 20 preguntas a partir de los resúmenes automáticos.

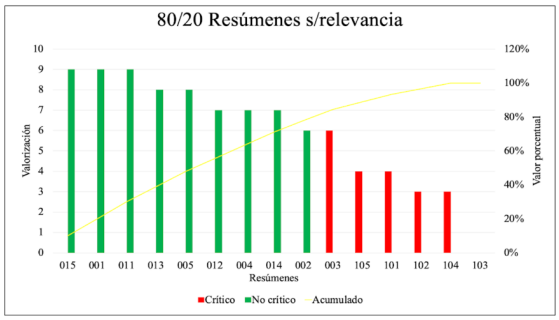

Figura 6: Análisis de Pareto de la Relevancia de las 20 preguntas, a partir de los resúmenes automáticos, sin considerar Relevancia.

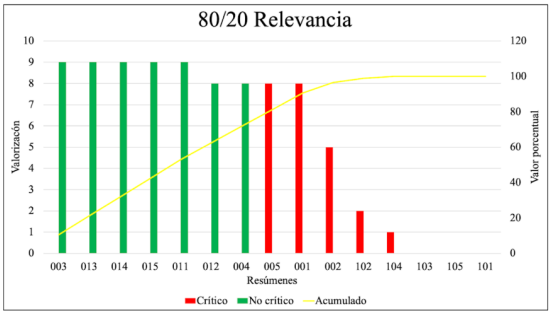

Figura 7: Análisis de Pareto de la Relevancia por resúmenes automáticos.

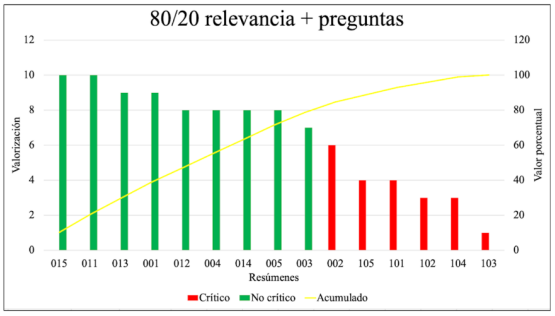

Figura 8: Análisis de Pareto de los resúmenes automáticos a partir de las 20 preguntas, más la Relevancia.

TF-IDF	Gensim	Grafos
001-Twitter	011-Twitter	101-Twitter
002-Facebook	012-Facebook	102-Facebook
003-Instagram	013-Instagram	103-Instagram
004-Snapchat	014-Snapchat	104-Snapchat
005-TikTok	015-TikTok	105-TikTok

Cuadro 2: Codificación de los resúmenes automáticos por métodos, categorizados a partir del Análisis de Pareto de la Figura 8.

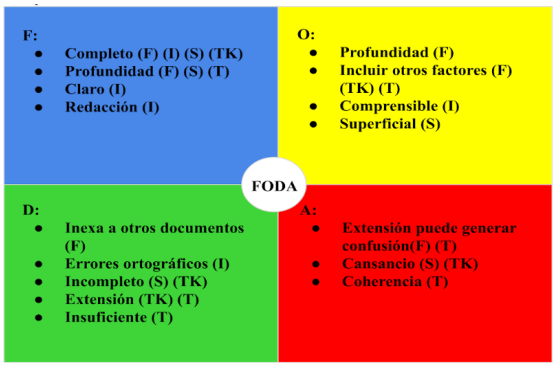

Figura 9: Análisis FODA del método Gensim de los resúmenes automáticos.

Todo lo anterior es resumido en el Cuadro 2, manteniendo la lógica de colores verde y rojo.

Tomando en cuenta el análisis FODA que desarrollamos para cada método y política de privacidad, los resultados son presentados en las Figuras 9, 10 y 11. Estos resulados están ordenados por método y cualidad, considerando cada política de privacidad, especificada entre paréntesis, a saber, **F**acebook, **T**witter, **T**ik**T**ok, **S**napchat e **I**nstagram. Para Gensim obtenemos que, de modo general, logra presentar información amplia de las veinte preguntas, es decir, es completo, abordando los temas con una profundidad tal que es identificada como aceptable para un resumen, además de ser claro y con buena redacción. No obstante, presenta algunas debilidades compartidas por las políticas de privacidad estudiadas, esto es, información insuficiente en el caso de Twitter o errores ortográficos. Además, se tiene en consideración que por su extensión, podrían requerir mayor esfuerzo cognitivo y sobrecarga de información al lector (Moy et al., 2018).

El método TF-IDF presenta características igual de aceptables, aunque muy segmentadas por tipo de empresa. Presenta menor extensión de los resúmenes y una redacción fácil de entender. Por lo mismo, no genera fatiga al lector, aunque si carecen de criterios o factores importantes. Asimismo, como principal amenaza no entrega información relevante para el lector meta.

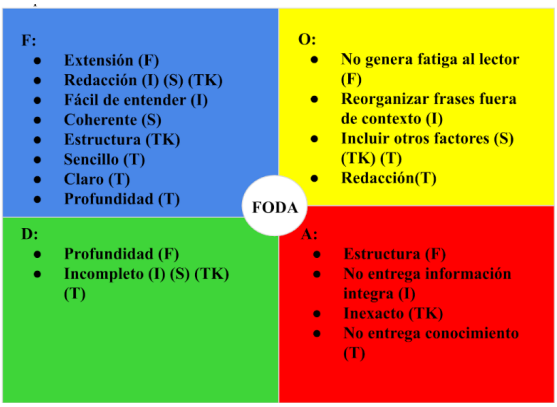

Figura 10: Análisis FODA del método TF-IDF de los resúmenes automáticos.

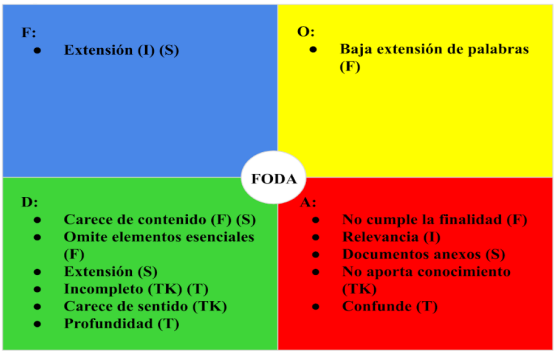

Figura 11: Análisis FODA del método basado en Grafos de los resúmenes automáticos.

Finalmente, el método grafo-analítico, tiene como principal fortaleza la extensión, pero no cumple la finalidad de informar o entregar conocimiento, carece de sentido y omite elementos esenciales. Cabe destacar, que a pesar de la poca cantidad de oraciones aún así permite generar resúmenes, aunque de menor calidad que con Gensim.

Luego de combinar el análisis Pareto con el FODA, obtenemos el Cuadro 3, el cual se presenta nuevamente con colores. Esta tabla agrega el amarillo a los posibles candidatos a ser un buen resumen, manteniendo el rojo para los peores evaluados y el verde para los mejores, que en este caso fueron el 014 y el 015, seguido levemente por el 012. A modo general, en verde encontramos resúmenes del método Gensim, en amarillo del método TF-IDF y en rojo el método grafo-analítico.

Respecto a los factores de riesgo de las PP, se obtiene un alto grado para todos los factores nombrados el Cuadro 4.

Como resultado concreto, a una velocidad de lectura de 3,5 palabras/segundo se logra reducir en un 91,29 % el tiempo de lectura general, un 83,84 % con el mejor modelo y se disminuyé la

TF-IDF	Gensim	Grafos
001-Twitter	011-Twitter	101-Twitter
002-Facebook	012-Facebook	102-Facebook
003-Instagram	013-Instagram	103-Instagram
004-Snapchat	014-Snapchat	104-Snapchat
005-TikTok	015-TikTok	105-TikTok

Cuadro 3: Codificación de los resúmenes automáticos por métodos, categorizados a partir del Análisis de Pareto de la Figura 8, sumado al Análisis FODA de los distintos métodos.

Datos bancarios	Finalidad/ Propósito para la recopilación de datos	Derecho al olvido
Datos sensibles	Seguro social	Editar o eliminar información personal del servicio
Información personal	Principios de protección de datos	Dirección de correo electrónico
Empresa proveedora del Servicio	Tratamiento de datos de menores de edad	Recopilación de datos
Consentimiento informado	Cambios en las Políticas de Privacidad	Violaciones de seguridad
Terceros involucrados		

Cuadro 4: Jerarquía de los factores de riesgo de las Políticas de Privacidad de aplicaciones infomáticas, identificando la mejor Política de Privacidad para Snapchat y Gensim.

cantidad de palabras en un 91,29 %, un 83,23 % con el mejor modelo, tal como muestra la Figura 12. A partir de lo anterior es posible proyectar que la utilización de resúmenes automáticos permitiría a los usuarios comprender mejor las PP que están aceptando, utilizando menos tiempo para su lectura.

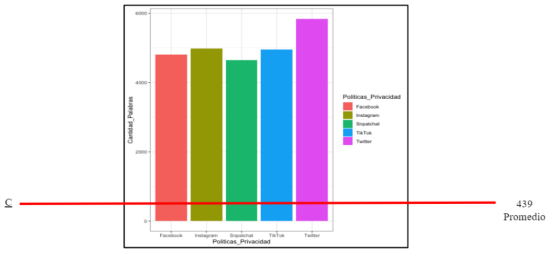

Figura 12: Cantidad de palabras de 5 Políticas de Privacidad: Facebook, Instagram, Snapchat, Twitter y TikTok, ya la curva $\bar{C}$ de cantidad de palabras promedio de los resúmenes automáticos.

5. Conclusiones y trabajo futuro

Como se aprecia en los resultados, los métodos evaluados tienen buenos índices en la construcción de los resúmenes, aunque TF-IDF, presenta una baja coherencia, dado su funcionamiento de repetición de palabras de contenido. Contrariamente, el método Gensim logra generar resúmenes más coherentes y con contenido más adecuado. Ello debido al procesamiento de textos con *Deep Learning*, lo que permite otros tipos de operaciones como el cálculo de similitud entre textos o palabras con mayor frecuencia.

Entre los resúmenes generados, el mejor evaluado es el 014 de Gensim, presentado en el Cuadro 5, relativo a las políticas de privacidad de Snapchat. No obstante, en el Cuadro 6 vemos que aún hay información que no es captada por los métodos de resumen. Este es el caso de la información relativa a los usuarios menores de 16 años, factor que se considera importante dentro de una PP. Además, se observa una correlación positiva alta entre la cantidad de palabras de un texto y su relevancia, así también entre la cantidad de palabras y el tiempo de lectura.

PREGUNTA 7: Snapchat

Podremos compartir la siguiente información con todos los Snapchatters, nuestros socios comerciales y el público en general: información pública, como tu nombre, nombre de usuario, Snapcódigo e imágenes de perfil; los envíos de Historia que se configuren para que todo el mundo pueda verlos y todo el contenido que envíes a un servicio inherentemente público tal como Nuestra Historia, y otros servicios de múltiples fuentes.

Cuadro 5: Extracto del resumen automático 014 a partir del método Gensim.

PREGUNTA 16: Snapchat

Niños

Nuestros servicios no están destinados a menores de 13 años y no los dirigimos a ellos. Ese es el motivo por el que no recabamos conscientemente información personal de ningún menor de 13 años. Por otra parte, podemos limitar el modo en que recabamos, usamos y guardamos parte de la información de los usuarios de la UE de entre 13 y 16 años. En algunos casos, esto significa que no podremos facilitar determinadas funciones a estos usuarios. Si tuviéramos que recurrir a un consentimiento como base jurídica para tratar esta información y tu país exige el consentimiento de uno de los padres, tal vez necesitemos el consentimiento de tu padre para recabar y utilizar dicha información.

Cuadro 6: Extracto de las políticas de privacidad de Snapchat.

El método TF-IDF sigue en la jerarquía de resultados, al ser un método de nivel intermedio, mientras que el de tipo grafo-analítico se queda en último lugar. No obstante, para trabajos futuros se planea potenciar el algoritmo de la Análisis de Grafos para ampliar y mejorar los contenidos de sus resúmenes.

Un aspecto interesante, en relación con los criterios de evaluación de la calidad de los resúmenes, es que no toda la información debe estar literalmente presentada en la política ni en el resumen. Así, por ejemplo, al evaluar Gensim (Cuadro 5) con la pregunta 7, se observa que esta no se responde literalmente y aun así se posiciona entre las tres mejor evaluadas. No obstante, de acuerdo con los expertos, esto tampoco es necesario. Ello porque al compartirse información con el "público en general" se hace evidente, utilizando la premisa del máximo-mínimo, que puede hacerse también con un destinatario específico como "la policía".

En suma, entre los beneficios de resumir las políticas de privacidad de las distintas aplicaciones informáticas con el método no supervisado Gensim, destacamos el sustancial ahorro de tiempo de lectura y una mejor comprensión por parte de los usuarios de los aspectos más relevantes declarados en las políticas de privacidad acerca del uso de sus datos privados.

En trabajos futuros se proyecta utilizar otros métodos de evaluación de resúmenes, como por ejemplo el paquete de medidas propuestas por (Lin, 2004), denominada ROUGE: Recall-Oriented Understudy for Gisting Evaluation, conformado por ROUGE-N, ROUGE-L, ROUGE-W y ROUGE-S, para evaluar la precisión (*precision*) y sensibilidad (*recall*) de los resúmenes automáticos respecto a resúmenes de referencia creados por humanos. No obstante, dado que no contamos con resúmenes de referencia creados por humanos o bien proporcionados por trabajos previos, el uso de esta métrica debe ser abordada independientemente con el fin de evaluar la calidad en términos computacionales. Así también, se proyecta explorar el uso de algoritmos de evaluación automatizada de resúmenes en los que no se requiere modelamiento por parte de escritores humanos (Louis & Nenkova, 2009; Saggion et al., 2010; Torres-Moreno et al., 2010).

Agradecimientos

Nuestros agradecimientos a la Pontificia Universidad Católica de Valparaíso por el financiamiento de esta investigación a través de los fondos Incentiva en el Aula (039.330/2020), 039.406/2021 y 039.344/2022. Así como al Núcleo de Procesamiento del Lenguaje Natural Aplicado (NIPLNA, https://www.niplna.com)

Referencias

Anastasopoulou, Kallia, Spyros Kokolakis & Panagiotis Andriotis. 2017. Privacy decision-making in the digital era: A game theoretic review. En *Human Aspects of Information Security, Privacy and Trust*, 589–603. Springer International Publishing. doi 10.1007/978-3-319-58460-7_41.

Barrios, Federico, Federico López, Luis Argerich & Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. ArXiv:1602.03606 [cs.CL].

Becerra, Xavier. 2020. California consumer privacy act (ccpa) fact sheet. https://oag.ca.gov/system/files/attachments/press_releases/CCPAFactSheet(00000002).pdf.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python*. O'Reilly.

Christian, Hans, Mikhael Pramodana Agus & Derwin Suhartono. 2016. Single document automatic text summarization using

term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications* 7(4). 285. 10.21512/comtech.v7i4.3746.

Corral, Hernán F. 2000. Configuración jurídica del derecho a la privacidad II: Concepto y delimitación. *Revista Chilena de Derecho* 27(2). 331–355.

Cucatto, Mariana. 2011. Algunas reflexiones sobre lenguaje jurídico como lenguaje de especialidad: más expresión que verdadera comunicación. *Revista virtual intercambios* 15. 1–19.

de la Cueva, Pablo Lucas M. & Alfonso Fernández-Miranda. 1990. *El derecho a la autodeterminación informativa: la protección de los datos personales frente al uso de la informática*. Tecnos.

da Cunha, Iria & M. Ángeles Escobar. 2021. Recomendaciones sobre lenguaje claro en español en el ámbito jurídico-administrativo: análisis y clasificación. *Pragmalingüística* 29. 129–148. 10.25267/Pragmalinguistica.2021.i29.07.

DataReportal. 2022. Digital 2022 global digital overview. https://datareportal.com/reports/digital-2022-global-overview-report.

Erkan, Günes & Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22(1). 457–479.

Gambhir, Mahak & Vishal Gupta. 2016. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47(1). 1–66. 10.1007/s10462-016-9475-9.

García Calderón, Jesús. 2019. Una retórica para la igualdad. *Revista del Ministerio Fiscal* 8. 41–57.

Hernández, Ángel Javier Alonso. 2017. *Deep learning aplicado al resumen de textos*: Universidad Complutense de Madrid. Trabajo de Fin de Máster.

Karmaker, Pradip Chandra & Md. Sharif Hossen. 2019. Performance analysis of frequency and graph theoretic based text summarization. En *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 10.1109/ecace.2019.8679452.

Ley Chile. 2014. Ley núm. 19.628, sobre protección de la vida privada (Chile). Ministerio Secretaría General de la Presidencia, http://www.leychile.cl/Navegar?idNorma=141599.

Lierde, Hadrien Van & Tommy W. S. Chow. 2019. Learning with fuzzy hypergraphs: A topical approach to query-oriented text summarization. *Information Sciences* 496. 212–224. 10.1016/j.ins.2019.05.020.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. En *Text Summarization Branches Out*, 74–81.

Louis, A. & A. Nenkova. 2009. Automatically evaluating content selection in summarization without human models. *Conference on Empirical Methods in Natural Language Processing* 306–314.

Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2). 159–165. 10.1147/rd.22.0159.

van de Luijtgaarden, Nick. 2019. *Automatic summarization of legal text*: Utrecht University. Trabajo de Fin de Máster.

Maybury, Mark T. 1995. Generating summaries from event data. *Information Processing & Management* 31(5). 735–751. 10.1016/0306-4573(95)00025-c.

McDonald, Aleecia M. & Lorrie Faith Cranor. 2009. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society* 4(3). 543–568.

Meier, Yannic, Johanna. Schäwel & Nicole C. Krämer. 2020. The shorter the better? effects of privacy policy length on online privacy decision-making. *Media and Communication* 8(2). 291–301.

Meza, Paulina, & Felipe González Catalán and. 2021. Un instrumento para evaluar la calidad lingüístico-discursiva de textos disciplinares producidos por estudiantes de derecho. *Onomázein Revista de lingüística filología y traducción* 51. 164–184. 10.7764/onomazein.51.08.

Montolío, Estrella & Anna López Samaniego. 2008. La escritura en el quehacer judicial: Estado de la cuestión y presentación de la propuesta aplicada en la escuela judicial de españa. *Revista Signos* 41(66). 33–64.

Montolío, Estrella & Mario Tascón. 2020. *El derecho a entender: la comunicación clara, la mejor defensa de la ciudadanía*. Catarata.

Moy, Naomi, Ho Fai Chan & Benno Torgler. 2018. How much is too much? the effects of information quantity on crowdfunding performance. *PLOS ONE* 13(3). e0192012. 10.1371/journal.pone.0192012.

OCDE. 2002. Resumen directrices de la OCDE sobre protección de la privacidad y flujos transfronterizos de datos personales. overview. https://www.oecd.org/sti/ieconomy/15590267.pdf.

Poblete, Claudia Andrea & Pablo Fuenzalida González. 2018. Una mirada al uso de lenguaje claro en el ámbito judicial latinoamericano. *Revista de Llengua i Dret* 69. 119–138. 10.2436/rld.i69.2018.3051.

Pérez Luño, Antonio-Enrique. 1984. *Derechos humanos, estado de derecho y constitución*. Tecnos.

RAE, Real Academia Española. 2021. Diccionario de la lengua española: privacidad. https://dle.rae.es/privacidad?m=form.

Rane, Neha & Sharvari Govilkar. 2019. Recent trends in deep learning based abstractive text summarization. *International Journal of Recent Technology and Engineering (IJRTE)* 8(3). 3108–3115. 10.35940/ijrte.c4996.098319.

Řehůřek, Radim. 2021. Gensim: Topic modelling for humans. https://radimrehurek.com/gensim/.

Rodrigo, Alfaro. & Héctor Allende. 2020. Clasificación de textos multi-etiquetados con modelo bernoulli multi-variado y representación dependiente de la etiqueta. *Revista signos* 53(104). 549–567. 10.4067/s0718-09342020000300549.

Ruidías, Héctor J., Karina Eckert, Juan M. Lezcano & Carolina V. Rosas. 2018. Tecnologías de la web semántica aplicadas al tratamiento de documentos jurídicos electrónicos. En *XX Workshop de Investigadores en Ciencias de la Computación (WICC)*, 290–294.

Saggion, Horacio, Juan-Manuel Torres-Moreno, Iria da Cunha, Eric SanJuan & Patricia Velázquez-Morales. 2010. Multilingual summarization evaluation without human models. En *23rd International Conference on Computational Linguistics (COLING)*, 1059–1067.

Saldaña, Maria Nieves. 2012. «the right to privacy»: la génesis de la protección de la privacidad en el sistema constitucional norteamericano, el centenario legado de warren y brandeis. *Revista de Derecho Político* 85. 195–239. 10.5944/rdp.85.2012.10723.

Salton, Gerard & Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5). 513–523. 10.1016/0306-4573(88)90021-0.

Schwabe, Jürgen. 2009. *Jurisprudencia del tribunal constitucional federal alemán*. Konrad Adenauer Stiftung.

TCCh. 1989. Sentencia nº rol 65 de tribunal constitucional, 16 de enero de 1989. Tribunal Constitucional de Chile, https://vlex.cl/vid/-58943056.

Torres-Moreno, Juan-Manuel, Horacio Saggion, Iria da Cunha, Eric SanJuan & Patricia Velazquez-Morales. 2010. Summary evaluation with and without references. *Polibits* 42. 13–19.

Valenzuela, Miguel, Stephanie Riff, Rodrigo Alfaro, Alan Bronfman & René Venegas. 2020. Análisis y resumen automático de políticas de privacidad. En *III Congreso Internacional de Lingüística Computacional y de Corpus (CILCC) y V Workshop en Procesamiento Automatizado de Textos y Corpus (WoPATeC)*, 286.

Wan, Changlin, Muhan Zhang, Wei Hao, Sha Cao, Pan Li & Chi Zhang. 2021. Principled hyperedge prediction with structural spectral features and neural networks. ArXiv:2106.04292 [cs.SI].

Warren, Samuel D. & Louis Brandeis. 1890. The right to privacy. *Harvard Law Review* 4(5). 193–219.

Widyassari, Adhika Pramita, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy & De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences* 34(4). 1029–1046. 10.1016/j.jksuci.2020.05.006.

Zaeem, Razieh Nokhbeh, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava & K. Suzanne Barber. 2020. PrivacyCheck v2: A tool that recaps privacy policies for you. En *29th ACM International Conference on Information & Knowledge Management*, 10.1145/3340531.3417469.

Zaeem, Razieh Nokhbeh, Rachel L. German & K. Suzanne Barber. 2018. PrivacyCheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology* 18(4). 1–18. 10.1145/3127519.

Zhang, Caiming & Yang Lu. 2021. Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration* 23. 100224. 10.1016/j.jii.2021.100224.

Propuesta recibida en marzo 2022 y aceptada para publicación en septiembre 2022.

Un estudio comparado de la traducción automática de eventos de movimiento de cruce de límites en inglés, español e italiano con Google Translate y DeepL

A comparative study of machine translations for boundary-crossing motion events in English, Spanish and Italian using Google Translate and DeepL

Nicola Florio
Universidad de Salamanca

Resumen

En este artículo se presenta un estudio comparado de las traducciones generadas por dos herramientas de traducción automática en línea (Google Translate y DeepL) para eventos de movimiento que implican un cruce de límites. Partiendo del inglés, una lengua tipológicamente opuesta al español y al italiano, se combinan verbos de movimiento que especifican la Manera en la que se produce el desplazamiento con complementos postverbales que expresan una Trayectoria de cruce de límites. El objetivo es analizar comparativamente las traducciones automáticas obtenidas en español e italiano con Google Translate y DeepL, y presentar los datos recopilados sobre las preferencias observadas en el patrón de lexicalización de los componentes semánticos de Trayectoria y Manera en ambas lenguas en comparación con el inglés.

Palabras clave

traducción automática; eventos de movimiento; cruce de límites; patrón de lexicalización; componentes semánticos

Abstract

This paper deals with a comparative study of translations obtained from two online machine translation tools (Google Translate and DeepL) for boundary-crossing motion events. Taking English as the source language, a typologically different language from Spanish and Italian, a list of Manner motion verbs has been combined with postverbal constructions encoding cross-boundary Paths. The aim is to comparatively analyze machine translations produced in Spanish and Italian using Google Translate and DeepL and present the data gathered about preferences in the lexicalization pattern for Path and Manner encoding in both languages compared to English.

Keywords

machine translation; motion events; boundary-crossing; lexicalization pattern; semantic components

1. Introducción

En los últimos años, gracias a los avances en las nuevas tecnologías y el desarrollo de un mercado globalizado a nivel mundial, hemos observado una proliferación de las herramientas de traducción automática disponibles en línea, muy demandadas por usuarios profesionales y no profesionales que precisan instrumentos rápidos, sencillos y económicos para la traducción de textos o la realización de consultas bilingües. Como explican Maldonado González & Liébana González (2021, p. 191), "la traducción automática es el proceso mediante el cual se utilizan herramientas software de computadora para traducir un texto de un lenguaje natural a otro". Fue a partir de la década de 1940, con la llegada de los primeros ordenadores, cuando las herramientas de traducción automática comenzaron a despegar y a convertirse en una de las aplicaciones más importantes dentro del mundo de la informática, y desde entonces, se han hecho grandes progresos e inversiones tanto por parte de instituciones como por parte de empresas privadas para perfeccionar las estrategias de traducción automática (Díaz Prieto, 2012). No es sorprendente, por tanto, que estas herramientas hayan despertado el interés de los investigadores sobre su utilidad en ámbitos tan dispares como la traducción, la enseñanza de lenguas o la medicina, como observamos en los estudios publicados por Maldonado González & Liébana González (2021), Ghasemi & Hashemian (2016), Lear et al. (2016), o Kol et al. (2018) entre muchos otros.

Aunque en la actualidad podemos encontrar numerosas herramientas de traducción automática en línea, en el presente artículo hemos centrado la atención en Google Translate y DeepL, y en las soluciones traductológicas que ofrecen de forma comparada en tres lenguas diferentes (inglés, español e italiano). Se han elegido estas

DOI: 10.21814/lm.14.2.368

dos herramientas por ser gratuitas, estar disponibles en línea y considerarse dos de las mejores y más frecuentemente utilizadas herramientas de traducción automática disponibles en la actualidad según numerosos especialistas y sitios web del campo de la traducción, como Fuentes-Luque et al. (2020), Loock & Léchauguette (2021), o la Asociación de Traductores Americanos[1], entre otros.

Google Translate es un instrumento desarrollado por Google para traducir textos y otros formatos de archivo (imagen, voz o vídeo) en 109 lenguas diferentes, aplicando en sus inicios un sistema de traducción automática estadística (Statistical Machine Translation), es decir, un sistema basado en el uso de corpus extensos de textos paralelos alineados por oraciones en los idiomas elegidos, que permite calcular las probabilidades de que una palabra en una lengua se corresponda con otra palabra en la otra lengua en función de las traducciones previas existentes (Maldonado González & Liébana González, 2021). No obstante, desde 2016 Google Translate ha adoptado un sistema de traducción automática neuronal (Neuronal Machine Translation) en la mayoría de los formatos traducidos para obtener mejores resultados. Aunque la información sobre los algoritmos empleados por Google Translate no es pública, esta herramienta utiliza todos los recursos disponibles en el motor de búsqueda de Google para recopilar y alinear textos bilingües (Adán Soriano, 2019).

Por su parte, DeepL es otra herramienta de traducción automática en línea lanzada al mercado en 2017, que trabaja con 26 lenguas diferentes y permite la traducción de textos escritos en varios formatos gracias a un sistema de traducción automática neuronal que utiliza una tecnología de inteligencia artificial a la vanguardia, basada en codificadores y decodificadores sofisticados que analizan las frases de origen y generan traducciones de forma automática aplicando complejos algoritmos y extensos corpus de entrenamiento (Casacuberta Nolla & Peris Abril, 2017). La traducción automática neuronal se basa en la creación de un sistema "capaz de analizar varios tipos de información a la vez, como la sintáctica, la semántica o el contexto comunicativo" (Adán Soriano, 2019). De este modo, se conectan los diferentes componentes del lenguaje con otro tipo de información subyacente (lingüística y no lingüística) para crear asociaciones y producir traducciones de forma automática mucho más fiables. Como explica Forcada (2017), la mayoría de los sistemas de traducción automática neuronal funcionan de forma similar al sistema de texto predictivo disponible en la mayoría de los dispositivos móviles, que permite completar un mensaje con diferentes opciones en función de las características de las palabras, su posición y el contexto de la oración de origen. DeepL permite la traducción de palabras, oraciones o incluso documentos completos, y ofrece la posibilidad de seleccionar la opción más adecuada mediante sugerencias de expresiones alternativas de acuerdo al contexto o al grado de formalidad requerido para la traducción (DeepL, 2022).

En este artículo se analizan de forma comparada en inglés, español e italiano las traducciones automáticas obtenidas a partir de Google Translate y DeepL para un tipo de construcción lingüística muy concreto: los eventos de movimiento. El estudio de la expresión del movimiento a través del lenguaje no es algo novedoso. De hecho, los lingüistas llevan décadas investigando acerca de la forma que tienen los hablantes de codificar la información sobre el espacio y el movimiento en las diferentes lenguas del mundo, pero hasta ahora no se había realizado ningún estudio que analizara de forma comparada en inglés, español e italiano las soluciones traductológicas generadas por herramientas de traducción automática para este tipo de estructuras.

2. La expresión del movimiento a través del lenguaje

En las últimas décadas, los conceptos de espacio y movimiento han generado gran debate dentro del ámbito de la lingüística, y han captado la atención de numerosos autores como Lakoff (1987, 2012), Langacker (2011), Fillmore & Baker (2012), Talmy (2000) o Slobin (1996), entre otros. Uno de los lingüistas más influyentes en el estudio de la expresión del movimiento a través del lenguaje es, sin duda, Leonard Talmy, que define los eventos de movimiento como aquellas situaciones en las que se produce un movimiento o en las que existe una continuación de un estado estacionario (Talmy, 2000, p. 25). Es importante diferenciar, por tanto, entre dos conceptos aparentemente similares pero en esencia diferentes: movimiento y desplazamiento. Como explica Morimoto (2001), para Talmy existe movimiento tanto en situaciones de desplazamiento como en contextos de estatismo o de movimiento interno de la Figura en los que no se produce ninguna modificación de la localización, pero solo se puede hablar de desplazamiento cuando se produce una alteración de

[1] ATA: American Translators Association (2022, 9 de agosto). Machine Translation. https://www.atanet.org/client-assistance/machine-translation/

la ubicación de la Figura como consecuencia del movimiento (p. 51). Conviene aclarar, por tanto, que en este artículo únicamente se han tenido en cuenta aquellos eventos de movimiento que hacen referencia a situaciones de desplazamiento, por lo que se han dejado a un lado los contextos de estatismo o movimiento interno.

Nuestro análisis parte de los estudios de Leonard Talmy, que a principios de los años setenta observó que las lenguas presentaban diferentes formas de construir eventos de movimiento en función del patrón de lexicalización empleado para codificar los componentes semánticos que los forman. Según el autor, los eventos de movimiento están compuestos por cuatro elementos básicos (Figura, Fondo, Movimiento y Trayectoria) y los denominados co-eventos (Manera y Causa). La Figura hace referencia al elemento que se desplaza con respecto a otro elemento que actúa como marco de referencia (Fondo); la Trayectoria indica el recorrido realizado por la Figura, y el Movimiento hace referencia al desplazamiento de la Figura en sí. La Manera indica la forma en la que se produce el movimiento y la Causa hace referencia a la acción que desencadena el movimiento de la Figura (Talmy, 2000, p. 26).

De acuerdo con Leonard Talmy, todas las lenguas lexicalizan el Movimiento en el propio verbo, pero este componente semántico puede aparecer acompañado de otros componentes adicionales como la Trayectoria o la Manera dentro de la raíz verbal. En función del componente semántico adicional lexicalizado junto al Movimiento dentro del verbo, Talmy (2000, p. 117) propone una división binaria de las lenguas del mundo en dos grandes categorías: las *lenguas de marco verbal* y las *lenguas de marco satélite*. El autor afirma que las lenguas de marco verbal tienden a lexicalizar en el propio verbo los componentes de Movimiento y Trayectoria, y suelen expresar los co-eventos como la Manera y la Causa a través de complementos externos al verbo, como construcciones adverbiales o formas de gerundio. Según Leonard Talmy, las lenguas románicas como el italiano o el español se englobarían dentro de esta categoría:

(1) *Il bambino* ***ha salito*** *le scale* ***di corsa****.*
(Movimiento + Trayectoria) (Manera)

(2) *El niño* ***subió*** *las escaleras* ***corriendo****.*
(Movimiento + Trayectoria) (Manera)

Por el contrario, las lenguas de marco satélite tienden a lexicalizar los co-eventos (Manera y Causa) dentro del propio verbo, junto al Movimiento, mientras que el componente de Trayectoria suele lexicalizarse de forma externa, a través de complementos que indican la meta del movimiento o la dirección del mismo. Según Talmy, algunas lenguas como el inglés se englobarían dentro del grupo de las lenguas de marco satélite.

(3) *The child* ***ran up*** *the stairs.*
(Movimiento + Manera) (Trayectoria)

(4) *The ball* ***rolled down*** *the street.*
(Movimiento + Manera) (Trayectoria)

Otro aspecto de gran relevancia que hemos tenido en cuenta en el presente artículo es la problemática del denominado 'cruce de límites', un fenómeno que analizan autores como Aske (1989), Slobin (2004), Hijazo Gascón (2011), Iacobini (2012), Cifuentes Férez (2013) o Ibarretxe-Antuñano (2017), entre otros. De acuerdo con Aske y Slobin, las lenguas de marco verbal no suelen permitir la construcción de eventos de movimiento con verbos que lexicalizan la Manera en el núcleo verbal y se acompañan de complementos que implican el cruce de un límite, es decir, complementos que indican que la Figura se desplaza hacia el interior o el exterior de un Fondo conceptualizado como un espacio cerrado o delimitado. De acuerdo con Slobin (2004), la tendencia de las lenguas de marco verbal consiste en lexicalizar la Trayectoria que implica un cruce de límites en el propio verbo (como *salir* o *entrar*) e incorporar el componente semántico de Manera a través de locuciones adverbiales o construcciones subordinadas o de gerundio (*a pie*, *corriendo* o *saltando*, por ejemplo). Sin embargo, en muchos casos, los hablantes de las lenguas de marco verbal parecen no hacer uso de estos complementos externos al verbo, o bien por considerarlo innecesario o bien por la carga cognitiva que añaden a la producción y comprensión de estas estructuras lingüísticas Slobin (2004). Por el contrario, este tipo de construcciones de cruce de límites con verbos que lexicalizan la Manera en el núcleo verbal (como *run, fly* o *walk*) parecen ser muy frecuentes en las lenguas de marco satélite como el inglés, puesto que la estructura de un evento de movimiento de estas características encaja perfectamente con el patrón de lexicalización habitual de estas lenguas, que tiende a indicar la Trayectoria de forma externa al verbo:

(5) *An owl* ***flew into*** *the room.*

(6) *A man* ***ran out*** *of the building.*

3. Objetivos

Teniendo en cuenta las diferencias en los patrones de lexicalización expuestas anteriormente, el propósito del presente artículo es evaluar las traducciones generadas por las herramientas de traducción automática en línea Google Translate y DeepL en italiano y español, dos lenguas englobadas tradicionalmente dentro de la categoría de las lenguas de marco verbal, tomando como punto de partida el inglés, una lengua considerada de marco satélite. El objetivo principal de nuestro estudio es observar si el patrón de lexicalización preferente asociado a las lenguas de marco verbal se ve reflejado en los resultados obtenidos con las herramientas de traducción automática Google Translate y DeepL para la construcción de eventos de movimiento de cruce de límites con verbos que lexicalizan la Manera en el propio núcleo verbal. La hipótesis de partida es que el inglés permite la construcción de eventos de movimiento de cruce de límites formados por verbos que lexicalizan los componentes semánticos de Movimiento y Manera en la propia raíz verbal, y la Trayectoria que implica el cruce de límites hacia el interior o el exterior del Fondo viene expresada de forma externa a través de complementos. Por el contrario, según el planteamiento expuesto anteriormente, el italiano y el español, dos lenguas de marco verbal, para la construcción de eventos de movimiento de cruce de límites deberían optar por verbos de movimiento que lexicalizan la Trayectoria de cruce de límites en la propia raíz verbal y expresar de forma externa, o incluso omitir, la información sobre la Manera en que se produce el movimiento.

Con el presente artículo se busca analizar si existen diferencias relevantes entre las opciones traductológicas ofrecidas por Google Translate y DeepL para este tipo de eventos, y se pretende asimismo observar si los resultados en italiano y español siguen en ambos casos el patrón de lexicalización propio de las lenguas de marco verbal o si se observan diferencias interesantes entre las dos lenguas. Para ello, hemos partido de estructuras en inglés como lengua de origen y hemos analizado el tipo de construcciones lingüísticas generadas por Google Translate y DeepL para este tipo de estructuras en italiano y español. Para poder evaluar de forma coherente los resultados obtenidos con las dos herramientas de traducción automática, hemos comparado las traducciones generadas por Google Translate y DeepL en italiano y español con las traducciones más frecuentes disponibles para esas mismas estructuras lingüísticas en la plataforma Glosbe, una herramienta en línea basada en memorias de traducción de gran tamaño elaboradas de forma colaborativa por más de 600.000 usuarios (Glosbe, 2022).

Hasta el momento, no hemos encontrado estudios previos que analicen la traducción automática de eventos de movimiento de forma comparada en tres lenguas que presentan patrones de lexicalización diferentes como el inglés, el italiano y el español. Por ello, considerábamos esencial la realización de un análisis de este tipo con el fin de extraer conclusiones relevantes sobre las opciones traductológicas ofrecidas por dos de las herramientas de traducción automática más utilizadas en línea en lo referente a la expresión del movimiento, sin entrar a cuestionar la corrección o la mayor o menor idoneidad de cada una de las opciones proporcionadas, sino las preferencias mostradas en el patrón de lexicalización para italiano y español.

4. Metodología

Para la realización de nuestro estudio, primeramente hemos llevado a cabo una tarea exhaustiva de recopilación de verbos de movimiento en inglés, italiano y español. Teniendo en cuenta las aportaciones de autores como Lamiroy (1991), Levin (1993), González Aranda (1998), De Miguel (1999), Cifuentes Honrubia (1999), Morimoto (2001), Cifuentes Férez (2013) y Stolova (2015), hemos seleccionado un total de 17 verbos de movimiento, presentes mayoritariamente en todos los listados de verbos consultados. Nuestra recopilación consta de 17 verbos de movimiento que en inglés lexicalizan simultáneamente dentro de la propia raíz verbal los componentes de Movimiento y Manera: *climb* (trepar, escalar), *crawl* (gatear), *creep* (arrastrarse), *cycle* (pedalear), *dance* (bailar), *drive* (conducir), *flutter* (revolotear), *fly* (volar), *jump* (saltar), *limp* (cojear), *roll* (rodar), *run* (correr), *sail* (navegar), *sneak* (colarse/escabullirse), *stroll* (pasear), *swim* (nadar) y *walk* (andar, caminar). Como podemos observar, son verbos de uso común que hacen referencia a un movimiento realizado de una Manera concreta, pero sin especificar la Trayectoria descrita por la Figura durante la ejecución del movimiento.

Los verbos seleccionados pueden agruparse en diferentes categorías en función del tipo de movimiento al que hacen referencia. En nuestro caso, hemos identificado las siguientes:

1. **Formas de desplazamiento mediante fricción con una superficie**: *climb, crawl, creep* y *roll*.

2. **Formas de caminar**: *limp, stroll* y *walk*.
3. textbfFormas de desplazamiento con un vehículo o medio de transporte: *cycle, drive* y *sail*.
4. **Formas de desplazamiento en medio acuático o aéreo**: *flutter, fly* y *swim*.
5. **Otros tipos de desplazamiento**: *dance, jump, run* y *sneak*.

Para la obtención de los resultados, los 17 verbos de movimiento se han utilizado en pasado y se han combinado con dos tipos de complementos externos que expresan una Trayectoria de cruce de límites en inglés: *into* + Fondo (desplazamiento hacia el interior del Fondo) y *out of* + Fondo (desplazamiento hacia el exterior del Fondo). Los elementos seleccionados como Figura y Fondo se han elegido de forma libre y las oraciones han sido construidas manualmente por parte de un hablante nativo de lengua inglesa, teniendo en cuenta las características de cada verbo de movimiento y sus posibilidades combinatorias. El objetivo de crear las oraciones manualmente era trabajar con estructuras sencillas compuestas por elementos básicos como Fondo (*habitación, cueva, iglesia, bahía, etc.*) y como Figura (*hombre, bebé, animal, pelota, mujer, etc.*) que facilitaran el análisis y la comparación de los resultados obtenidos con las herramientas de traducción automática. En cualquier caso, a la hora de seleccionar los complementos que actúan como Fondo, para construir eventos de movimiento de cruce de límites se han elegido elementos conceptualizados como espacios delimitados o cerrados en los que la Figura entra o de los que la Figura sale. Con cada verbo de movimiento se han construido dos oraciones diferentes: una con una Trayectoria de cruce de límites hacia el interior del Fondo y otra con una Trayectoria de cruce de límites hacia el exterior, manteniendo en ambos casos los mismos elementos como Figura y Fondo. Aunque el número de oraciones construidas con cada verbo no es muy elevado, la amplia variedad de verbos de movimiento analizados, combinados cada uno de ellos con dos complementos externos de cruce de límites (hacia el interior y hacia el exterior), permite obtener datos estadísticos relevantes sobre el panorama general de las traducciones automáticas obtenidas con las herramientas en línea analizadas en cuanto a eventos de movimiento que implican un cruce de límites. Paralelamente, las traducciones automáticas generadas por Google Translate y DeepL en italiano y español se han comparado con las traducciones más frecuentes disponibles en la plataforma Glosbe, que cuenta con memorias de traducción de gran extensión compuestas por millones de oraciones traducidas en más de 6.000 idiomas diferentes por su comunidad de usuarios (Glosbe, 2022). El fin último era evaluar las semejanzas y diferencias encontradas entre las traducciones automáticas de Google Translate y DeepL, y las memorias de traducción disponibles en Glosbe en cuanto a las preferencias en el patrón de lexicalización.

A continuación, se presentarán los resultados obtenidos y se analizarán las diferentes opciones de traducción ofrecidas en función de los siguientes parámetros:

(a) **Solo Trayectoria**: En la traducción se ha seleccionado un verbo de movimiento que lexicaliza la Trayectoria de cruce de límites en el núcleo verbal pero desaparece cualquier tipo de información sobre el componente de Manera.

(b) **Manera + Trayectoria externa**: En la traducción se ha seleccionado un verbo de movimiento que lexicaliza la Manera en el núcleo verbal y se especifica la Trayectoria de cruce de límites a través de complementos externos al verbo, manteniendo el patrón de lexicalización de la lengua original (inglés).

(c) **Trayectoria + Manera externa**: En la traducción se ha seleccionado un verbo de movimiento que lexicaliza la Trayectoria de cruce de límites en el núcleo verbal y se lexicaliza el componente de Manera a través de complementos externos (construcción de gerundio o locuciones adverbiales).

(d) **Alteración del significado**: La propuesta de traducción presenta un significado diferente con respecto al original expresado por la versión en inglés, o el significado no es exactamente el mismo.

(e) **Doble interpretación**: Para la traducción se ha elegido un verbo de movimiento que lexicaliza la Manera en el núcleo verbal y este se acompaña de complementos externos que especifican la Trayectoria de cruce de límites. En estos casos, la traducción ofrecida puede tener una doble interpretación en función del contexto. Se puede interpretar como un movimiento locativo, enmarcado dentro de los límites físicos del Fondo, o como un movimiento que implica un cruce de límites.

5. Análisis de los resultados

Las herramientas de traducción automática analizadas se han consultado en febrero de 2022. A continuación, se exponen los resultados extraídos haciendo distinción entre las muestras de Google Translate y DeepL, con el fin de analizar de forma conjunta y de forma comparada las traducciones automáticas encontradas. Los resultados se han agrupado en función de las diferentes categorías de verbos de movimiento presentadas anteriormente (desplazamiento mediante fricción con una superficie, formas de caminar, formas de desplazamiento con un vehículo o medio de transporte, desplazamiento en medio acuático o aéreo, y otros tipos de desplazamiento), con la intención de obtener también datos estadísticos sobre las traducciones obtenidas para cada categoría. Los resultados generados por las herramientas Google Translate y DeepL en las lenguas de destino han sido evaluadas por un hablante nativo de italiano y un hablante nativo de español, ambos especialistas en el campo de la lingüística y la traducción.

5.1. Desplazamiento mediante fricción con una superficie

Dentro de la primera categoría hemos analizado cuatro verbos en inglés (*climb, crawl, creep* y *roll*), combinados con complementos que implican un cruce de límites (hacia el interior del Fondo y hacia el exterior del Fondo). En total, hemos analizado 16 traducciones automáticas en cada lengua.

Climb

- **Inglés**

  The man climbed into the cave.
  The man climbed out of the cave.

- **Español**

  Google Translate:

  El hombre subió a la cueva. [d]
  El hombre salió de la cueva. [a]

  DeepL:

  El hombre subió a la cueva. [d]
  El hombre salió de la cueva. [a]

  Traducciones más frecuentes Glosbe:

  Subir a la cueva.[d]
  Salir de la cueva.[a]

Comprobamos que las traducciones más frecuentes encontradas en español en las memorias de traducción de Glosbe coinciden exactamente con las traducciones automáticas generadas por Google Translate y DeepL. En el caso del evento de movimiento de cruce de límites hacia el interior del Fondo, tanto en las traducciones automáticas de Google Translate y DeepL como en las memorias de traducción de Glosbe se ha alterado el significado de la oración original al utilizar el verbo de movimiento *subir*. Por lo que respecta al evento de movimiento de cruce de límites hacia el exterior del Fondo, en las traducciones automáticas y en las memorias de traducción se ha optado por utilizar un verbo de movimiento que expresa el cruce de límites dentro de la propia raíz verbal (*salir*) y se ha omitido cualquier tipo de información sobre la Manera.

- **Italiano**

  Google Translate:

  L'uomo si arrampicò nella grotta. [e]
  L'uomo è uscito dalla caverna. [a]

  DeepL:

  L'uomo entrò nella grotta. [a]
  L'uomo si arrampicò fuori dalla grotta. [e]

  Traducciones más frecuentes Glosbe:

  Arrampicarsi nella caverna. [e]
  Uscire dalla grotta. [a]

Como podemos observar, en el caso del italiano, las traducciones encontradas en Glosbe coinciden exactamente con las generadas por Google Translate, pero difieren de las opciones propuestas por DeepL. Tanto Google Translate como las memorias de traducción de Glosbe utilizan la construcción *arrampicarsi nella + Fondo* en el caso del evento de movimiento de cruce de límites hacia el interior del Fondo, que puede tener una doble interpretación locativa y de cruce de límites. En el caso del evento de movimiento de cruce de límites hacia el exterior del Fondo, Google Translate y las memorias de traducción de Glosbe optan por el verbo de movimiento *uscire*, que implica un cruce de límites pero no hace referencia en absoluto a la Manera en la que se produce el movimiento.

Crawl

- **Inglés**

  The baby crawled into the room.
  The baby crawled out of the room.

- **Español**

  Google Translate:

  El bebé se arrastró hasta la habitación.[b]
  El bebé salió gateando de la habitación. [c]

  DeepL:

  El bebé entró gateando en la habitación. [c]
  El bebé salió gateando de la habitación. [c]

  Traducciones más frecuentes Glosbe:

  Entrar en la habitación. [a]
  Salir a rastras de la habitación. [c]

En este caso, vemos que las traducciones más frecuentes en español encontradas en Glosbe no coinciden exactamente con las generadas por las herramientas de traducción automática. En el caso del cruce de límites hacia el interior del Fondo, en las memorias de traducción encontradas en Glosbe se opta por omitir cualquier tipo de información sobre la Manera en la que se produce el movimiento. Por el contrario, en el caso del movimiento de cruce de límites hacia el exterior del Fondo, la opción más frecuente en las memorias de traducción de Glosbe es lexicalizar la Trayectoria dentro del propio verbo y expresar la Manera de forma externa mediante una locución adverbial (*a rastras*).

- **Italiano**

  Google Translate:

  Il bambino è strisciato nella stanza. [e]
  Il bambino è strisciato fuori dalla stanza. [e]

  DeepL:

  Il bambino strisciò nella stanza. [e]
  Il bambino è strisciato fuori dalla stanza.[e]

  Traducciones más frecuentes Glosbe:

  Strisciare nella stanza. [e]
  Strisciare fuori dalla stanza. [e]

En el caso del italiano, las traducciones más frecuentes disponibles en las memorias de traducción de Glosbe coinciden exactamente con las generadas por Google Translate y DeepL, que optan por construcciones formadas por verbos de movimiento que lexicalizan la Manera dentro de la raíz verbal y que pueden tener una doble interpretación (locativa y de cruce de límites) dependiendo del contexto.

Creep

- **Inglés**

  The animal creeped into the room.
  The animal creeped out of the room.

- **Español**

  Google Translate:

  El animal se deslizó en la habitación. [d]
  El animal salió de la habitación. [a]

  DeepL:

  El animal entró sigilosamente en la habitación.[c]
  El animal salió sigilosamente de la habitación.[c]

  Traducciones más frecuentes Glosbe:

  Entrar en la habitación [a]
  Salir de la habitación sin hacer ruido [c]

En este caso, observamos que en español la traducción más frecuente que encontramos en las memorias de traducción de Glosbe para el evento de movimiento de cruce de límites hacia el interior del Fondo omite cualquier tipo de información sobre la Manera en la que se realiza el movimiento. Por lo que respecta al evento de movimiento de cruce de límites hacia el exterior del Fondo, la traducción más frecuente en Glosbe coincide con la propuesta de traducción automática generada por DeepL, que expresa la Manera de forma externa al verbo mediante una locución adverbial.

- **Italiano**

  Google Translate:

  L'animale si è insinuato nella stanza. [b]
  L'animale sgattaiolò fuori dalla stanza. [b]

  DeepL:

  L'animale entrò strisciando nella stanza.[c]
  L'animale è strisciato fuori dalla stanza.[e]

  Traducciones más frecuentes Glosbe:

  Insinuarsi nella stanza [b]
  Sgattaiolare fuori dalla stanza [b]

En el italiano, las traducciones más frecuentes identificadas en las memorias de traducción de Glosbe coinciden exactamente con las propuestas de traducción automática generadas por Google Translate, pero difieren de las traducciones de DeepL. La opción prioritaria en Glosbe para la traducción de ambos eventos de movimiento de cruce de límites consiste en elegir verbos que lexicalizan la Manera dentro de la raíz verbal y expresar la Trayectoria de forma externa.

Roll

- **Inglés**

  The ball rolled into the room.
  The ball rolled out of the room.

- **Español**

  Google Translate:

  La pelota rodó por la habitación.[d]
  La pelota salió rodando de la habitación.[c]

  DeepL:

  La pelota rodó hacia la habitación.[d]
  La pelota rodó fuera de la habitación.[d]

  Traducciones más frecuentes Glosbe:

  Entrar rodando en la habitación. [c]
  Salir de la habitación. [a]

De acuerdo con las memorias de traducción disponibles en Glosbe, las traducciones más frecuentes no coinciden en ningún caso con las traducciones automáticas generadas por Google Translate y DeepL. En el evento de movimiento de cruce de límites hacia el interior del Fondo, la opción prioritaria en Glosbe utiliza un verbo de movimiento que lexicaliza la Trayectoria en la propia raíz verbal (*entrar*) y expresa la Manera

en la que se produce el movimiento de forma externa, a través de una construcción de gerundio. En el caso del evento de movimiento de cruce de límites hacia el exterior, la traducción más frecuente en Glosbe omite cualquier tipo de información sobre la Manera del movimiento, haciendo hincapié únicamente en el cruce de límites con la selección del verbo *salir*.

- **Italiano**

  Google Translate:

  La palla è rotolata nella stanza.[e]
  La palla è rotolata fuori dalla stanza.[e]

  DeepL:

  La palla rotolò nella stanza.[e]
  La palla rotolò fuori dalla stanza.[e]

  Traducciones más frecuentes Glosbe:

  Entrare nella stanza [a]
  Uscire dalla stanza [a]
  y *rotolare fuori dalla stanza* [e]

Resulta interesante observar que las traducciones más frecuentes disponibles en las memorias de traducción de Glosbe para italiano optan por seleccionar un verbo de movimiento que implica un cruce de límites en la propia raíz verbal (*entrare* y *uscire*) y tienden a omitir mayoritariamente cualquier detalle sobre la Manera en la que se produce el movimiento. Solo encontramos traducciones que optan por utilizar un verbo que lexicaliza la Manera en la raíz verbal en el caso del evento de movimiento de cruce de límites hacia el exterior del Fondo (*rotolare fuori*), que presenta la misma frecuencia de uso que *uscire* en las traducciones disponibles en Glosbe.

Centrando la atención únicamente en las traducciones automáticas generadas por Google Translate y DeepL, comprobamos que, en el caso del español, hay dos patrones de lexicalización preferentes, con el 38 porciento de los casos cada uno de ellos. En 6 traducciones automáticas se selecciona un verbo que lexicaliza la Trayectoria de cruce de límites en el propio núcleo verbal y el componente de Manera aparece expresado de forma externa al verbo mediante construcciones de gerundio (*gateando* y *rodando*) o locuciones adverbiales (*sigilosamente*). En el otro 38 porciento de los casos se ha producido una alteración del significado original en inglés, o bien porque las traducciones automáticas no expresan un cruce de límites, sino simplemente un movimiento direccionado, como en *el hombre subió a la cueva* o *la pelota rodó hacia la habitación*, o bien porque el movimiento expresado es de carácter locativo, sin llegar a cruzar un umbral, como en *el animal se deslizó en la habitación*, *la pelota rodó por la habitación* y *la pelota rodó fuera de la habitación*.

Patrón de lexicalización	**Español**		**Italiano**	
a) Solo Trayectoria	3	18 %	2	13 %
b) Manera + Trayectoria (externa)	1	6 %	2	13 %
c) Trayectoria + Manera (externa)	6	38 %	1	6 %
d) Alteración del significado	6	38 %	—	—
e) Doble interpretación	—	—	11	68 %
TOTAL	16		16	

Cuadro 1: Desplazamiento mediante fricción con una superficie

En un 18 porciento de los casos se ha optado por un verbo de movimiento que indica en el propio núcleo verbal la Trayectoria de cruce de límites (*salir*), pero se omite cualquier tipo de referencia a la Manera en que se produce el movimiento. Por último, encontramos un caso en el que ha utilizado un verbo que lexicaliza la Manera en el propio núcleo verbal y la Trayectoria se ha expresado de forma externa para indicar el punto de llegada del movimiento, aunque no queda totalmente claro que llegue a cruzarse el umbral (*el bebé se arrastró hasta la habitación*).

En el caso del italiano, comprobamos que el patrón de lexicalización preferente, con el 68 porciento de los casos, es la selección de un verbo de movimiento que aporta información sobre la Manera, acompañado de complementos postverbales que indican una Trayectoria de cruce de límites. Como vemos, en todos estos casos se puede obtener una doble interpretación, tanto locativa como de cruce de límites, en función del contexto. Con un 13 porciento de casos, comprobamos que hay traducciones que usan verbos que lexicalizan la Trayectoria de cruce de límites en el propio núcleo verbal (*uscire* y *entrare*), pero omiten cualquier tipo de detalle sobre la Manera, y en otro 13 porciento de casos se eligen verbos que lexicalizan la Manera, mientras que la Trayectoria de cruce de límites se expresa de forma externa. Por último, encontramos una traducción automática en la que se ha utilizado un verbo de movimiento que lexicaliza la Trayectoria de cruce de límites en el núcleo verbal (*entrare*) y la Manera se expresa a través de una construcción de gerundio (*strisciando*).

5.2. Formas de caminar

En este segundo apartado analizamos los resultados de los verbos de movimiento que hacen referencia a distintas formas de caminar. Hemos seleccionado tres verbos (*limp, stroll* y *walk*), combinados con complementos de cruce de límites, por lo que en total hemos obtenido 12 traducciones automáticas en cada lengua.

Limp

- **Inglés**

  The woman limped into the room.
  The woman limped out of the room.

- **Español**

  Google Translate:

  La mujer entró cojeando en la habitación.[c]
  La mujer salió cojeando de la habitación.[c]

  DeepL:

  La mujer entró cojeando en la habitación.[c]
  La mujer salió cojeando de la habitación.[c]

  Traducciones más frecuentes Glosbe:

  Entrar cojeando en la habitación [c]
  Salir cojeando de la habitación. [c]

Comprobamos que las traducciones más frecuentes en las memorias de traducción de español disponibles en Glosbe coinciden exactamente con las traducciones automáticas generadas por Google Translate y DeepL, que optan por seleccionar un verbo de movimiento que expresa el cruce de límites en la propia raíz verbal (*entrar* y *salir*), y la Manera en la que se produce dicho movimiento se lexicaliza a través de complementos externos (en ambos casos, la forma de gerundio *cojeando*).

- **Italiano**

  Google Translate:

  La donna entrò zoppicando nella stanza.[c]
  La donna uscì zoppicando dalla stanza.[c]

  DeepL:

  La donna zoppicò nella stanza.[e]
  La donna zoppicò fuori dalla stanza.[e]

  Traducciones más frecuentes Glosbe:

  Entrare zoppicando nella stanza[c]
  Zoppicare fuori dalla stanza [e]
  y *uscire zoppicando dalla stanza* [c]

En el caso del italiano, la traducción más frecuente en Glosbe para el evento de movimiento de cruce de límites hacia el interior del Fondo coincide con la traducción automática generada por Google Translate, que selecciona un verbo de movimiento que lexicaliza el cruce de límites dentro de la propia raíz verbal (*entrare*) y la Manera se expresa de forma externa mediante una construcción de gerundio (*zoppicando*). Por lo que respecta al evento de movimiento de cruce de límites hacia el exterior del Fondo, en las memorias de traducción de Glosbe encontramos dos opciones con la misma frecuencia de uso: *zoppicare fuori* y *uscire zoppicando*, por lo que las traducciones automáticas proporcionadas por Google Translate y DeepL coinciden en ambos casos con las traducciones disponibles en Glosbe.

Stroll

- **Inglés**

  The bride strolled into the church.
  The bride strolled out of the church.

- **Español**

  Google Translate:

  La novia entró en la iglesia.[a]
  La novia salió de la iglesia.[a]

  DeepL:

  La novia entró en la iglesia.[a]
  La novia salió paseando de la iglesia.[c]

  Traducciones más frecuentes Glosbe:

  Entrar en la iglesia [a]
  Salir de la iglesia. [a]

En el caso del español, comprobamos que las traducciones más frecuentes disponibles en las memorias de traducción de Glosbe seleccionan verbos de movimiento que expresan el cruce de límites en la propia raíz verbal (*entrar* y *salir*) y se omite cualquier tipo de detalle sobre la Manera en la que se produce el movimiento. Por lo tanto, las traducciones automáticas generadas por Google Translate coinciden exactamente con las disponibles en Glosbe, mientras que DeepL ofrece una alternativa diferente para la traducción del evento de movimiento de cruce de límites hacia el exterior del Fondo, incorporando de forma externa información sobre la Manera del movimiento.

- **Italiano**

  Google Translate:

  La sposa entrò in chiesa.[a]
  La sposa uscì dalla chiesa.[a]

  DeepL:

  La sposa entrò in chiesa a piedi.[c]
  La sposa passeggiava fuori dalla chiesa.[d]

  Traducciones más frecuentes Glosbe:

  Entrare in chiesa [a]
  Uscire dalla chiesa [a]

En el caso del italiano, confirmamos una vez más que las traducciones más frecuentes disponibles en Glosbe coinciden exactamente con las proporcionadas por Google Translate, haciendo uso de verbos de movimiento que lexicalizan el cruce de límites en la raíz verbal (*entrare* y *uscire*) y omitiendo cualquier tipo de información sobre la Manera en la que se produce el movimiento. DeepL, por el contrario, incluye en las dos traducciones detalles sobre la Manera, o bien lexicalizando el componente semántico dentro del propio verbo o bien a través de complementos externos.

Walk

- **Inglés**

  The bride walked into the church.
  The bride walked out of the church.

- **Español**

  Google Translate:

  La novia entró en la iglesia.[a]
  La novia salió de la iglesia.[a]

  DeepL:

  La novia entró en la iglesia.[a]
  La novia salió de la iglesia.[a]

  Traducciones más frecuentes Glosbe:

  Entrar en la inglesa [a]
  Salir de la iglesia [a]

Como podemos observar, las traducciones más frecuentes encontradas en las memorias de traducción de Glosbe coinciden exactamente con las traducciones automáticas generadas por Google Translate y DeepL. En todos los casos se eligen verbos que expresan el cruce de límites en la propia raíz verbal (*entrar* y *salir*), sin incluir ningún tipo de detalle sobre la Manera del movimiento.

- **Italiano**

  Google Translate:

  La sposa entrò in chiesa.[a]
  La sposa è uscita dalla chiesa.[a]

  DeepL:

  La sposa è entrata in chiesa.[a]
  La sposa è uscita dalla chiesa.[a]

  Traducciones más frecuentes Glosbe:

  Entrare in chiesa [a]
  Uscire dalla chiesa [a]

En el caso del italiano, observamos el mismo fenómeno que en español. Las traducciones más frecuentes disponibles en Glosbe coinciden exactamente con las traducciones generadas por las dos herramientas de traducción automática analizadas. En todos los casos se ha elegido lexicalizar la Trayectoria de cruce de límites en el propio verbo y se ha omitido la información sobre la Manera.

A modo de resumen con cifras estadísticas, en el Cuadro 2 vemos que, en el caso del español, Google Translate y DeepL solo presentan dos patrones de lexicalización. En un 58 porciento de los casos, las traducciones automáticas utilizan un verbo de movimiento que lexicaliza la Trayectoria de cruce de límites en el propio núcleo verbal (*entrar* y *salir*), pero desaparece la información sobre la Manera del movimiento. En el 42 porciento de casos restantes, las traducciones automáticas utilizan un verbo de movimiento que lexicaliza la Trayectoria de cruce de límites en el propio núcleo (*entrar* y *salir*) y la Manera se especifica de forma externa a través de construcciones de gerundio (*cojeando* o *paseando*).

Patrón de lexicalización	**Español**		**Italiano**	
a) Solo Trayectoria	7	58 %	6	50 %
b) Manera + Trayectoria (externa)	—	—	—	—
c) Trayectoria + Manera (externa)	5	42 %	3	25 %
d) Alteración del significado	—	—	1	8 %
e) Doble interpretación	—	—	2	17 %
TOTAL	12		12	

Cuadro 2: Formas de caminar

Por lo que respecta al italiano, en las traducciones automáticas se ha elegido en el 50 porciento de los casos el mismo patrón de lexicalización predominante que en español: verbos de movimiento que lexicalizan la Trayectoria de cruce de límites en el propio núcleo verbal (*entrare* y *uscire*) y omisión de cualquier tipo de información sobre la Manera del movimiento. En un 25 porciento de los casos, las traducciones automáticas contienen un verbo que lexicaliza la Trayectoria de cruce de límites en el núcleo verbal y la Manera se indica de forma externa a través de construcciones de gerundio (*zoppicando*) o locuciones adverbiales (*a piedi*). En dos casos las traducciones automáticas permiten una doble interpretación del mensaje dependiendo del contexto (movimiento locativo o de cruce de límites), y en un caso se ha alterado el significado original del texto en inglés (*la sposa passeggiava fuori dalla chiesa*), puesto que desaparece el movimiento de cruce de límites y en su lugar se indica un movimiento prolongado en el tiempo y de forma locativa.

5.3. Desplazamiento con un vehículo o medio de transporte

En la tercera categoría analizamos tres verbos que hacen referencia a un desplazamiento mediante un vehículo o medio de transporte (*cycle, drive* y *sail*), combinados una vez más con complementos que implican la superación de un límite para acceder al interior del Fondo o para salir al exterior del Fondo.

Cycle

- **Inglés**

  The man cycled into the town.
  The man cycled out of the town.

- **Español**

  Google Translate:

  El hombre entró en bicicleta al pueblo.[c]
  El hombre salió en bicicleta de la ciudad.[c]

DeepL:

El hombre entró en la ciudad en bicicleta.[c]
El hombre salió de la ciudad en bicicleta.[c]

Traducciones más frecuentes Glosbe:

Ir en bicicleta a la ciudad [d]
Alejarse en bicicleta de la ciudad [d]

En el caso del español, las traducciones más frecuentes disponibles en las memorias de traducción de Glosbe no coinciden con las propuestas automáticas generadas por Google Translate ni DeepL. En las traducciones encontradas en Glosbe no se utilizan estructuras lingüísticas que expresan un cruce de límites, ni hacia el interior del Fondo ni hacia el exterior. Se ha optado en ambos casos por utilizar verbos de movimiento alternativos, como *ir* y *alejarse*, acompañados por complementos externos que especifican la Manera en la que se produce el desplazamiento (*en bicicleta*).

- **Italiano**

Google Translate:

L'uomo è entrato in bicicletta nella città.[c]
L'uomo è uscito dalla città in bicicletta.[c]

DeepL:

L'uomo è entrato in città in bicicletta. [c]
L'uomo è uscito in bicicletta dalla città. [c]

Traducciones más frecuentes Glosbe:

Pedalare fino in città.[d]
Pedalare fuori dalla città.[e]

En el caso del italiano, las traducciones más frecuentes encontradas en las memorias de traducción disponibles en Glosbe tampoco coinciden con las proporcionadas por Google Translate y DeepL. Observamos que las traducciones automáticas utilizan verbos de movimiento que implican un cruce de límites (*entrare* y *uscire*) y expresan la Manera de forma externa a través de complementos (*in bicicletta*). Por el contrario, en las memorias de traducción consultadas en Glosbe encontramos una alteración del significado en el primer caso, puesto que *pedalare fino in città* no implica necesariamente un cruce de límites, sino simplemente la llegada a una meta. En el segundo caso, las traducciones encontradas en Glosbe presentan una doble interpretación, puesto que pueden hacer referencia a un movimiento locativo (dentro del espacio delimitado por el Fondo) o a un movimiento de cruce de límites hacia el interior del Fondo.

Drive

- **Inglés**

The man drove into the parking lot.
The man drove out of the parking lot.

- **Español**

Google Translate:

El hombre entró en el estacionamiento.[a]
El hombre salió del estacionamiento.[a]

DeepL:

El hombre entró en el aparcamiento.[a]
El hombre salió del aparcamiento.[a]

Traducciones más frecuentes Glosbe:

Entrar en el aparcamiento. [a]
Salir del aparcamiento. [a]

Observamos que las traducciones más frecuentes en Glosbe para español coinciden exactamente con las traducciones automáticas proporcionadas por Google Translate y DeepL. En todos los casos, se han seleccionado verbos que lexicalizan el cruce de límites en la propia raíz verbal (*entrar* y *salir*) y no se ha incluido ningún tipo de información sobre la Manera en la que se produce el movimiento.

- **Italiano**

Google Translate:

L'uomo è entrato nel parcheggio.[a]
L'uomo è uscito dal parcheggio.[a]

DeepL:

L'uomo ha guidato nel parcheggio.[d]
L'uomo è uscito dal parcheggio.[a]

Traducciones más frecuentes Glosbe:

Entrare nel parcheggio [a]
Uscire dal parcheggio [a]

En el caso del italiano, las traducciones más frecuentes disponibles en las memorias de traducción de Glosbe seleccionan el mismo tipo de verbos que en español, es decir, verbos de movimiento que lexicalizan el cruce de límites en la propia raíz verbal (*entrare* y *uscire*) y se omiten los detalles sobre la Manera en la que se produce dicho movimiento. Coinciden, por tanto, con las traducciones automáticas ofrecidas por Google Translate y difieren parcialmente de las proporcionadas por DeepL.

Sail

- **Inglés**

The sailor sailed into the bay.
The sailor sailed out of the bay.

- **Español**

Google Translate:

El marinero navegó hacia la bahía.[d]
El marinero salió de la bahía.[a]

DeepL:

El marinero navegó hacia la bahía.[d]
El marinero salió de la bahía.[a]

Traducciones más frecuentes Glosbe:

Entrar en la bahía [a]
Navegar alejándose de la bahía [d]

Vemos que las traducciones automáticas generadas por Google Translate y DeepL coinciden entre sí, mientras que las traducciones más frecuentes encontradas en las memorias de traducción de Glosbe invierten las estructuras lingüísticas seleccionadas. Para el evento de cruce de límites hacia el interior del Fondo, las traducciones más frecuentes identificadas en Glosbe seleccionan un verbo de movimiento que lexicaliza el cruce de límites en la raíz verbal (*entrar*), sin hacer referencia a la Manera en la que se produce el movimiento. Por el contrario, la alteración del significado con respecto a la oración original en inglés la encontramos en el evento de movimiento de cruce de límites hacia el exterior, para el que se selecciona un verbo de movimiento que especifica la Manera (*navegar*), pero los complementos postverbales no implican un cruce de límites desde el interior del Fondo hacia el exterior, sino simplemente la dirección del movimiento, que se aleja del punto de partida.

- **Italiano**

Google Translate:

Il marinaio salpò nella baia.[d]
Il marinaio salpò dalla baia.[d]

DeepL:

Il marinaio entrò nella baia.[a]
Il marinaio salpò dalla baia.[d]

Traducciones más frecuentes Glosbe:

Entrare nella baia [a]
Uscire dalla baia [a]

En italiano, comprobamos que las traducciones más frecuentes encontradas en las memorias de traducción de Glosbe no coinciden en ningún caso con las propuestas generadas por Google Translate, pero se asemejan parcialmente a las traducciones automáticas de DeepL. En las traducciones disponibles en Glosbe se han seleccionado en los dos casos verbos de movimiento que lexicalizan el cruce de límites dentro de la propia raíz verbal (*entrare* y *uscire*), pero se ha omitido cualquier tipo de detalle sobre la Manera del movimiento.

Como podemos observar en el Cuadro 3, que recoge los datos estadísticos sobre las traducciones automáticas analizadas, en el caso del español, el patrón de lexicalización prioritario (en un 50 porciento de los casos) selecciona verbos de movimiento que expresan la Trayectoria de cruce de límites en el propio núcleo verbal (*salir* y *entrar*), pero se suprime cualquier tipo de detalle sobre la Manera en la que se produce el movimiento. El segundo patrón de lexicalización más frecuente (con un 33 porciento de los casos) utiliza verbos de movimiento que indican la Trayectoria de cruce de límites en el propio núcleo verbal y la Manera se expresa de forma externa, con locuciones adverbiales (*en bicicleta*). Por último, encontramos dos casos en los que se altera el significado original del texto en inglés, porque el verbo se acompaña de un complemento direccional que no implica la superación de un umbral (*el marinero navegó hacia la bahía*).

Patrón de lexicalización	**Español**		**Italiano**	
a) Solo Trayectoria	6	50 %	4	33 %
b) Manera + Trayectoria (externa)	—	—	—	—
c) Trayectoria + Manera (externa)	4	33 %	4	33 %
d) Alteración del significado	2	17 %	4	33 %
e) Doble interpretación	—	—	—	—
TOTAL	12		12	

Cuadro 3: Formas de desplazamiento con un vehículo o medio de transporte

Por lo que respecta al italiano, vemos que hay tres patrones de lexicalización preferentes, con el mismo porcentaje de frecuencia de uso. En un 33 porciento de los casos se produce una alteración del significado original del texto en inglés. En el caso de *l'uomo ha guidato nel parcheggio*, la interpretación es locativa, por lo que el movimiento de cruce de límites desaparece. Por su parte, en los ejemplos *il marinaio salpò nella baia* e *il marinaio salpò dalla baia*, el verbo seleccionado en las traducciones automáticas de italiano indica el punto de partida del movimiento (*salpare*), pero desaparece cualquier matiz de cruce de límites, por lo que se altera el significado original. En un 33 porciento de los casos, el patrón de lexicalización selecciona un verbo que indica la Trayectoria de cruce de límites en la raíz verbal y la Manera se expresa de forma externa mediante locuciones adverbiales (*in bicicletta*), y en otro 33 porciento de los casos las traducciones automáticas utilizan un verbo de movimiento que expresa la Trayectoria de cruce de límites en el propio núcleo verbal (*uscire* y *entrare*), pero se suprime cualquier tipo de información sobre la Manera.

5.4. Desplazamiento en medio acuático o aéreo

En esta cuarta categoría se analizan tres verbos que indican formas de desplazamiento en un medio acuático o aéreo (*flutter, fly* y *swim*), por

lo que se han obtenido 12 traducciones automáticas en cada lengua.

Flutter

- **Inglés**

  The bird fluttered into the room.
  The bird fluttered out of the room.

- **Español**

  Google Translate:

  El pájaro entró revoloteando en la habitación.[c]
  El pájaro salió volando de la habitación.[c]

  DeepL:

  El pájaro revoloteó hacia la habitación.[d]
  El pájaro salió revoloteando de la habitación.[c]

  Traducciones más frecuentes Glosbe:

  Entrar revoloteando en la habitación [c]
  Salir revoloteando de la habitación [c]

Comprobamos que las traducciones automáticas de Google Translate coinciden exactamente con las traducciones más frecuentes encontradas en las memorias de traducción de Glosbe, que seleccionan un verbo de movimiento que expresa el cruce de límites en la propia raíz verbal (*entrar* y *salir*) y lo acompañan de complementos externos que aportan detalles sobre el componente semántico de Manera. Por el contrario, vemos que las propuestas de traducción automática proporcionadas por DeepL difieren parcialmente de las identificadas en las memorias de traducción de Glosbe.

- **Italiano**

  Google Translate:

  L'uccello svolazzò nella stanza.[e]
  L'uccello svolazzò fuori dalla stanza.[e]

  DeepL:

  L'uccello volò nella stanza.[d]
  L'uccello è volato fuori dalla stanza.[d]

  Traducciones más frecuentes Glosbe:

  Entrare svolazzando nella stanza [c]
  Volare fuori dalla stanza [d]

En el caso del italiano, comprobamos que las traducciones más frecuentes encontradas en las memorias de traducción de Glosbe no coinciden exactamente con ninguna de las traducciones automáticas generadas por Google Translate ni DeepL. Solo coincide parcialmente la traducción del evento de cruce de límites hacia el exterior del Fondo con la traducción automática de DeepL, aunque en ambos casos se ha producido una alteración del significado con respecto a la oración original en inglés, puesto que se ha utilizado el verbo *volare* en lugar de *svolazzare*.

Fly

- **Inglés**

  The bird flew into the room.
  The bird flew out of the room.

- **Español**

  Google Translate:

  El pájaro voló a la habitación.[b]
  El pájaro salió volando de la habitación.[c]

  DeepL:

  El pájaro entró volando en la habitación.[c]
  El pájaro salió volando de la habitación.[c]

  Traducciones más frecuentes Glosbe:

  Entrar volando en la habitación. [c]
  Salir volando de la habitación. [c]

En este caso, las traducciones automáticas de DeepL coinciden exactamente con las traducciones más frecuentes identificadas en las memorias de traducción de Glosbe en español, con verbos de movimiento que expresan el cruce de límites en la propia raíz verbal e indican la Manera a través de complementos externos (la forma de gerundio del verbo *volar*). Las traducciones automáticas de Google Translate solo coinciden parcialmente con las memorias de traducción disponibles en Glosbe.

- **Italiano**

  Google Translate:

  L'uccello volò nella stanza.[e]
  L'uccello è volato fuori dalla stanza.[b]

  DeepL:

  L'uccello volò nella stanza.[e]
  L'uccello volò fuori dalla stanza.[e]

  Traducciones más frecuentes Glosbe:

  Volare nella stanza. [e]
  Volare via dalla stanza [b]
  y *volare fuori dalla stanza* [e]

En el caso del italiano, las traducciones más frecuentes disponibles en las memorias de traducción de Glosbe coinciden en ambos casos con las propuestas generadas tanto por Google Translate como por DeepL.

Swim

- **Inglés**

  The sailor swam into the bay.
  The sailor swam out of the bay.

- **Español**

  Google Translate:

  El marinero nadó hacia la bahía.[d]
  El marinero salió nadando de la bahía.[c]

  DeepL:

  El marinero nadó hasta la bahía.[b]
  El marinero salió nadando de la bahía.[c]

Traducciones más frecuentes Glosbe:
Nadar hacia la bahía. [d]
Salir nadando de la bahía. [c]

En este caso, las traducciones automáticas de Google Translate vuelven a coincidir exactamente con las traducciones más frecuentes disponibles en las memorias de traducción de Glosbe, y solo coinciden parcialmente con las generadas por DeepL para el evento de cruce de límites hacia el exterior del Fondo, al seleccionar un verbo que expresa el cruce de límites en la propia raíz verbal y especifica la Manera de forma externa.

- **Italiano**

Google Translate:
Il marinaio nuotò nella baia.[e]
Il marinaio nuotò fuori dalla baia.[e]
DeepL:
Il marinaio nuotò nella baia.[e]
Il marinaio nuotò fuori dalla baia.[e]
Traducciones más frecuentes Glosbe:
Entrare nella baia [a]
Nuotare fuori dalla baia [e]

Por último, en el caso del italiano, observamos que las traducciones automáticas de Google Translate y DeepL coinciden con las traducciones más frecuentes disponibles en Glosbe para el evento de cruce de límites hacia el exterior del Fondo, pero difieren en el otro caso. Para el evento de cruce de límites hacia el interior del Fondo, tanto Google Translate como DeepL han seleccionado estructuras lingüísticas que pueden presentar una doble interpretación (locativa o de cruce de límites), pero en las memorias de traducción de Glosbe la traducción más frecuente selecciona simplemente el verbo de movimiento *entrare* para expresar el cruce de límites y omite cualquier tipo de detalle sobre la Manera.

En el Cuadro 4 vemos los datos estadísticos de las traducciones automáticas y comprobamos que, en el caso del español, el patrón de lexicalización preferente (58 porciento de los casos) selecciona un verbo que expresa la Trayectoria de cruce de límites dentro del propio núcleo verbal y la Manera aparece lexicalizada de forma externa mediante construcciones de gerundio (*revoloteando, volando* y *nadando*). Encontramos un 25 porciento de casos en los que observamos una alteración del significado original del texto en inglés, o bien porque se utiliza un verbo en español que no coincide exactamente con el tipo de movimiento indicado en inglés (*volar* en lugar de *revolotear* como traducción de *flutter*), o bien porque las traducciones en español no implican un cruce de límites, sino simplemente un movimiento direccionado, como vemos en *el pájaro revoloteó hacia la habitación* o en *el marinero nadó hacia la bahía*. Por último, en las traducciones automáticas en español encontramos dos casos en los que se ha optado por un verbo que lexicaliza información sobre la Manera en el propio núcleo verbal y la Trayectoria de cruce de límites viene expresada de forma externa a través de complementos postverbales, aunque no quede claro por completo que la Figura atraviese el umbral del Fondo con su desplazamiento (*el pájaro voló a la habitación* y *el marinero nadó hasta la orilla*).

Patrón de lexicalización	**Español**		**Italiano**	
a) Solo Trayectoria	—	—	—	—
b) Manera + Trayectoria (externa)	2	17 %	1	8 %
c) Trayectoria + Manera (externa)	7	58 %	—	—
d) Alteración del significado	3	25 %	2	17 %
e) Doble interpretación	—	—	9	75 %
TOTAL	12		12	

Cuadro 4: Formas de desplazamiento en medio acuático o aéreo

En cuanto a los resultados en italiano, observamos que el patrón de lexicalización predominante (75 porciento de los casos) es el del uso de verbos de movimiento que especifican la Manera en la propia raíz verbal, combinados con complementos que lexicalizan la Trayectoria de cruce de límites de forma externa, aunque pueden recibir una doble interpretación, locativa o de cruce de límites, dependiendo del contexto. Encontramos dos ejemplos en los que la traducción automática en italiano no coincide exactamente con el significado original del texto en inglés, puesto que se ha elegido el verbo *volare* en lugar de *svolazzare* como traducción del verbo en inglés *flutter*. Por último, encontramos un ejemplo en italiano en el que se ha utilizado el verbo de movimiento *volare*, que especifica la Manera, y la Trayectoria de cruce de límites se ha lexicalizado de forma externa mediante los complementos postverbales. Conviene aclarar que, en italiano, el verbo *volare* puede formar los tiempos de pasado con dos auxiliares diferentes: *avere*, en general, o *essere* cuando el verbo aparece acompañado por un complemento que indica el origen o la meta del movimiento. Por ese motivo, la traducción *l'uccello è volato fuori dalla stanza* únicamente puede interpretarse como un desplazamiento dirigido hacia una meta, con cruce de límites, y nunca con valor locativo.

5.5. Otros tipos de desplazamiento

En esta última sección analizamos cuatro verbos que indican desplazamientos de difícil clasificación (*dance, jump, run* y *sneak*) acompañados

por complementos de cruce de límites en inglés. Hemos obtenido un total de 16 traducciones automáticas en cada una de las dos lenguas de estudio.

Dance

- **Inglés**

  The man danced into the room.
  The man danced out of the room.

- **Español**

  Google Translate:

  El hombre entró bailando en la habitación.[c]
  El hombre salió bailando de la habitación.[c]

  DeepL:

  El hombre bailó dentro de la habitación.[d]
  El hombre bailó fuera de la habitación.[d]

  Traducciones más frecuentes Glosbe:

  Entrar bailando en la habitación. [c]
  Salir bailando de la habitación [c]

En el caso del español, las traducciones más frecuentes encontradas en las memorias de traducción de Glosbe coinciden exactamente con las propuestas generadas por Google Translate, al seleccionar un verbo de movimiento que implica el cruce de límites en la raíz verbal y especificar la Manera de forma externa mediante construcciones de gerundio. Difieren, por tanto, de las traducciones automáticas ofrecidas por DeepL, que optan por verbos de movimiento que lexicalizan la Manera dentro de la raíz verbal e indican la Trayectoria de cruce de límites a través de complementos postverbales (*dentro* y *fuera*).

- **Italiano**

  Google Translate:

  L'uomo ballò nella stanza.[d]
  L'uomo ballò fuori dalla stanza.[d]

  DeepL:

  L'uomo danzò nella stanza.[d]
  L'uomo danzò fuori dalla stanza.[d]

  Traducciones más frecuentes Glosbe:

  Entrare danzando o ballando nella stanza [c]
  Uscire danzando dalla stanza [c]

En el caso del italiano, observamos que las traducciones más frecuentes encontradas en las memorias de traducción de Glosbe coinciden con las del español en el patrón de lexicalización utilizado, pero difieren por completo de las traducciones automáticas generadas por Google Translate y DeepL. Como podemos observar, las traducciones automáticas presentan una alteración del significado con respecto a las oraciones originales en inglés, puesto que expresan un movimiento locativo, no de cruce de límites, mientras que las traducciones más frecuentes identificadas en Glosbe optan por la lexicalización del cruce de límites dentro del propio verbo de movimiento (*entrare* y *uscire*) y expresan la Manera de forma externa, mediante construcciones de gerundio.

Jump

- **Inglés**

  The rabbit jumped into the den.
  The rabbit jumped out of the den.

- **Español**

  Google Translate:

  El conejo saltó a la guarida.[b]
  El conejo saltó de la guarida.[b]

  DeepL:

  El conejo saltó a la guarida.[b]
  El conejo saltó de la guarida.[b]

  Traducciones más frecuentes Glosbe:

  Saltar a la guarida. [b]
  Saltar de la guarida. [b]

En el caso del español, las traducciones automáticas proporcionadas por Google Translate y DeepL coinciden exactamente con las traducciones más frecuentes disponibles en las memorias de traducción de Glosbe, puesto que todas ellas seleccionan un verbo de movimiento que lexicaliza la Manera dentro de la propia raíz verbal y se indica la Trayectoria de forma externa, a través de los complementos postverbales.

- **Italiano**

  Google Translate:

  Il coniglio saltò nella tana.[e]
  Il coniglio saltò fuori dalla tana.[e]

  DeepL:

  Il coniglio saltò nella tana.[e]
  Il coniglio saltò fuori dalla tana.[e]

  Traducciones más frecuentes Glosbe:

  Saltare nella tana [e]
  Saltare fuori dalla tana [e]

Comprobamos igualmente que en el caso del italiano las traducciones más frecuentes identificadas en Glosbe coinciden exactamente también con las traducciones automáticas generadas por Google Translate y DeepL, que pueden recibir una doble interpretación (locativa o de cruce de límites) en función del contexto.

Run

- **Inglés**

  The man ran into the supermarket.
  The man ran out of the supermarket.

- **Español**

  Google Translate:

  El hombre entró corriendo al supermercado.[c]
  El hombre salió corriendo del supermercado.[c]

  DeepL:

  El hombre corrió hacia el supermercado.[d]
  El hombre salió corriendo del supermercado.[c]

  Traducciones más frecuentes Glosbe:

  Entrar corriendo al supermercado. [c]
  Salir corriendo del supermercado. [c]

Como podemos ver, las traducciones más frecuentes encontradas en las memorias de traducción de Glosbe coinciden una vez más con las traducciones automáticas generadas por Google Translate, puesto que seleccionan verbos de movimiento que lexicalizan el cruce de límites en la propia raíz verbal y se especifica la Manera del movimiento de forma externa, a través de construcciones de gerundio. Las traducciones más frecuentes encontradas en Glosbe solo coinciden parcialmente con las traducciones automáticas de DeepL, que en el primer caso altera ligeramente el significado original de la oración en inglés por no implicar de forma explícita un cruce de límites.

- **Italiano**

  Google Translate:

  L'uomo è corso al supermercato.[d]
  L'uomo è scappato dal supermercato.[b]

  DeepL:

  L'uomo corse nel supermercato.[e]
  L'uomo corse fuori dal supermercato.[e]

  Traducciones más frecuentes Glosbe:

  Correre nel supermercato [e]
  Correre fuori dal supermercato [e]

Por el contrario, en el caso del italiano, vemos que las traducciones más frecuentes encontradas en las memorias de traducción de Glosbe coinciden exactamente con las traducciones automáticas generadas por DeepL, con construcciones lingüísticas que abren la puerta a una doble interpretación (locativa o de cruce de límites) según el contexto.

Sneak

- **Inglés**

  The thief sneaked into the room.
  The thief sneaked out of the room.

- **Español**

  Google Translate:

  El ladrón se coló en la habitación.[b]
  El ladrón se escapó de la habitación.[b]

  DeepL:

  El ladrón se coló en la habitación.[b]
  El ladrón salió a hurtadillas de la habitación.[c]

  Traducciones más frecuentes Glosbe:

  Colarse en la habitación. [b]
  Salir sigilosamente de la habitación. [c]

En este caso, las traducciones más frecuentes disponibles en las memorias de traducción de Glosbe coinciden con las opciones generadas por DeepL. Para el evento de cruce de límites hacia el interior del Fondo, se ha elegido un verbo de movimiento que lexicaliza la Manera dentro de la propia raíz verbal (*colarse*) y se ha indicado la Trayectoria de cruce de límites a través de complementos postverbales externos. En el caso del evento de cruce de límites hacia el exterior del Fondo, se ha seleccionado un verbo de movimiento que lexicaliza el cruce de límites en la raíz verbal (*salir*) y se ha especificado la Manera de forma externa (con locuciones adverbiales). Las traducciones automáticas de Google Translate difieren parcialmente de las traducciones más frecuentes disponibles en Glosbe para el evento de cruce de límites hacia el exterior del Fondo.

- **Italiano**

  Google Translate:

  Il ladro si è intrufolato nella stanza.[b]
  Il ladro è uscito di soppiatto dalla stanza.[c]

  DeepL:

  Il ladro entrò di nascosto nella stanza.[c]
  Il ladro è uscito di nascosto dalla stanza.[c]

  Traducciones más frecuentes Glosbe:

  Entrare di nascosto nella stanza [c]
  Sgattaiolare fuori dalla stanza [b]

Por su parte, en italiano, la traducción más frecuente encontrada en las memorias de traducción de Glosbe para el evento de movimiento de cruce de límites hacia el interior del Fondo coincide con la traducción automática generada por DeepL, al lexicalizar la Trayectoria de cruce de límites dentro del propio verbo y especificar la Manera a través de complementos postverbales (locución adverbial). En el caso del evento de cruce de límites hacia el exterior del Fondo, las traducciones más frecuentes identificadas en Glosbe optan por la selección de un verbo de movimiento que lexicaliza la Manera dentro de la raíz verbal y se indica la Trayectoria de cruce de límites de forma externa, mientras que las traducciones automáticas de DeepL y Google Translate han repetido el mismo patrón de lexicalización que en el primer caso, para el evento de cruce de límites hacia el interior del Fondo.

Patrón de lexicalización	**Español**		**Italiano**	
a) Solo Trayectoria	—	—	—	—
b) Manera + Trayectoria (externa)	7	44 %	3	18 %
c) Trayectoria + Manera (externa)	6	38 %	3	18 %
d) Alteración del significado	3	18 %	4	25 %
e) Doble interpretación	—	—	6	38 %
TOTAL	16		16	

Cuadro 5: Otros tipos de desplazamiento

En el Cuadro 5 encontramos un resumen con los datos comentados en esta sección sobre el patrón de lexicalización prioritario encontrado en las traducciones automáticas analizadas para este tipo de verbos. Como podemos observar, en el caso del español, el patrón de lexicalización preferente (44 porciento) utiliza un verbo de movimiento que especifica la Manera en el núcleo verbal y la Trayectoria de cruce de límites aparece lexicalizada de forma externa, a través de complementos. En un 38 porciento de los casos, en las traducciones en español aparece un verbo de movimiento que indica la Trayectoria de cruce de límites en el propio núcleo verbal (*entrar* y *salir*), y se lexicaliza el componente de Manera de forma externa, a través de construcciones de gerundio (*bailando* o *corriendo*) o de locuciones adverbiales (*a hurtadillas*). Por último, encontramos tres casos (18 porciento del total) en los que detectamos una alteración del significado original del texto en inglés, una vez más, o bien porque la interpretación de las traducciones automáticas tiene un valor locativo (e*l hombre bailó dentro de la habitación* o *el hombre bailó fuera de la habitación*), o bien porque los complementos que acompañan al verbo son direccionales y no implican un cruce de límites (*el hombre corrió hacia el supermercado*).

En cuanto a las traducciones automáticas en italiano, encontramos un 38 porciento de casos con una doble interpretación (locativa o de cruce de límites) en función del contexto, con verbos de movimiento que indican la Manera en el núcleo verbal y aparecen acompañados de complementos externos que especifican un cruce de límites. Encontramos cuatro casos (25 porciento del total) en los que se ha producido una alteración del significado original del texto en inglés, puesto que las traducciones automáticas en italiano presentan un valor locativo y, por tanto, no pueden interpretarse como eventos de cruce de límites (*l'uomo ballò nella stanza* y *l'uomo ballò fuori dalla stanza*). Encontramos tres casos en los que se ha seleccionado un verbo de movimiento que lexicaliza la Trayectoria de cruce de límites en el núcleo verbal (*uscire* y *entrare*) y la Manera se especifica de forma externa, a través de locuciones adverbiales (*di nascosto* o *di soppiatto*), así como otros tres casos en los que se ha utilizado un verbo que contiene información sobre la Manera en el núcleo verbal y se ha lexicalizado la Trayectoria de cruce de límites de forma externa al verbo (*l'uomo è scappato dal supermercato*). Merece la pena aclarar que el verbo *correre*, al igual que *volare*, forma los tiempos de pasado con el auxiliar *avere* cuando se hace referencia a la acción en sí, y con el auxiliar *essere* cuando se indica una meta o punto de llegada del movimiento. Por lo tanto, el ejemplo *l'uomo è corso al supermercato* se interpreta como un evento de movimiento con un punto de llegada específico, que puede imaginarse en el interior del Fondo al que hace referencia, aunque no aparezca explícitamente en la traducción.

	Español			**Italiano**		
Patrón de lexicalización	Google Translate	DeepL	Total	Google Translate	DeepL	Total
a) Solo Trayectoria	9	7	16 (23 %)	7	5	12 (18 %)
b) Manera + Trayectoria (ext.)	6	4	10 (15 %)	5	—	5 (7 %)
c) Trayectoria + Manera (ext.)	14	15	29 (43 %)	5	6	11 (16 %)
d) Alteración del significado	5	8	13 (19 %)	5	7	12 (18 %)
e) Doble interpretación	—	—	—	12	16	28 (41 %)
			68			68

Cuadro 6: Resultados obtenidos en español e italiano con Google Translate y DeepL

5.6. Datos estadísticos totales

En el Cuadro 6 presentamos una recopilación de todos los datos obtenidos con nuestro estudio de traducciones automáticas generadas por las herramientas en línea Google Translate y DeepL en español e italiano para eventos de movimiento de cruce de límites. Como podemos observar, el patrón de lexicalización prioritario en las traducciones en español se basa en la preferencia por verbos de movimiento que lexicalizan la Trayectoria de cruce de límites en la propia raíz verbal (*entrar* o *salir*) acompañados de complementos externos que especifican la Manera en la que se produce el movimiento, o bien mediante construcciones de gerundio o bien a través de locuciones adverbiales. El segundo patrón de lexicalización más frecuente en las traducciones en español vuelve a seleccionar preferentemente verbos de movimiento que incorporan en la propia raíz verbal la Trayectoria de cruce de límites (*entrar* o *salir*), pero en este caso se elimina cualquier tipo de referencia a la Manera en la que se produce el

movimiento. Con estos datos, comprobamos que en el 66 porciento de las traducciones automáticas de español se eligen verbos de movimiento que lexicalizan la Trayectoria de cruce de límites en el propio núcleo verbal, frente al 15 porciento de casos en los que se ha optado por un verbo de movimiento que especifica la Manera en el núcleo verbal y externaliza la lexicalización de la Trayectoria de cruce de límites a través de complementos.

Por el contrario, en las traducciones automáticas en italiano vemos que el patrón de lexicalización predominante selecciona preferentemente verbos de movimiento que especifican la Manera en el propio núcleo verbal y estos se acompañan de complementos externos que indican una Trayectoria de cruce de límites. Teniendo en cuenta que el patrón de lexicalización [b] (Manera + Trayectoria externa) y el patrón de lexicalización [e] (Doble interpretación: locativa y de cruce de límites) seleccionan los mismos tipos de verbos y complementos (verbos de movimiento que especifican la Manera en la raíz verbal y complementos que lexicalizan la Trayectoria de cruce de límites de forma externa), podemos afirmar que en el 48 porciento de las traducciones automáticas en italiano el verbo lexicaliza la Manera y la Trayectoria de cruce de límites viene expresada de forma externa. Estas cifras destacan frente al 16 porciento de los casos en los que las traducciones automáticas en italiano seleccionan verbos de movimiento que lexicalizan la Trayectoria de cruce de límites dentro del propio núcleo verbal (*entrare* o *uscire*) y se acompañan de complementos externos que especifican la Manera, y frente al 18 porciento de casos en los que se omite cualquier tipo de información sobre la Manera y únicamente se lexicaliza la Trayectoria de cruce de límites en el verbo de movimiento.

Por lo que respecta a las alteraciones del significado original del texto en inglés, observamos que las cifras son muy similares en italiano y español. En el caso del italiano, en un 18 porciento de los casos las traducciones automáticas no han conservado el mismo significado que el texto original, y en el caso del español, el porcentaje es muy similar (19 porciento). En lo referente a las traducciones automáticas que pueden generar una doble interpretación (locativa o de cruce de límites), merece la pena comentar que no hemos identificado ningún caso en las traducciones automáticas de español, mientras que en italiano este tipo de construcciones representa la opción más frecuente (41 porciento de los casos).

Si comparamos los datos de las dos herramientas de traducción automática en línea analizadas, vemos que no existen grandes diferencias entre las opciones traductológicas proporcionadas por Google Translate y DeepL. El único dato llamativo lo encontramos en el caso del italiano con las construcciones formadas por verbos de movimiento que lexicalizan la Manera en el núcleo verbal e indican la Trayectoria de cruce de límites a través de complementos externos. Todos los casos identificados en italiano pertenecen a la herramienta Google Translate, por lo que no hemos encontrado ningún ejemplo de este tipo entre las traducciones automáticas proporcionadas por la herramienta DeepL.

	Español		**Italiano**	
Patrón de lexicalización	Glosbe	Total	Glosbe	Total
a) Solo Trayectoria	11	32 %	12	32 %
b) Manera + Trayectoria (externa)	3	9 %	4	11 %
c) Trayectoria + Manera (externa)	15	44 %	6	16 %
d) Alteración del significado	5	15 %	2	6 %
e) Doble interpretación (locativa y cruce de limites)	—	—	13	35 %
Total	34		37 (doble opción en 3 casos)	

Cuadro 7: Patrón de lexicalización más frecuente en las memorias de traducción de Glosbe

Por último, merece la pena comparar el patrón de lexicalización preferente encontrado en las traducciones automáticas de Google Translate y DeepL con el patrón de lexicalización más frecuente identificado en las memorias de traducción disponibles en Glosbe para español e italiano. En el Cuadro 7 observamos que, en el caso del español, las preferencias coinciden. El patrón de lexicalización más frecuente (44 porciento de los casos) para la traducción de eventos de movimiento de cruce de límites en español selecciona verbos que lexicalizan la Trayectoria de cruce de límites dentro de la propia raíz verbal y la Manera se especifica de forma externa, a través de complementos. También en las memorias de traducción de Glosbe ocupa el segundo lugar el patrón de lexicalización que omite cualquier tipo de información sobre la Manera y únicamente lexicaliza la Trayectoria de cruce de límites a través del verbo (32 porciento de los casos).

En el caso del italiano, comprobamos que en las memorias de traducción de Glosbe el patrón de lexicalización preferente para la traducción de este tipo de eventos coincide con el observado en las traducciones automáticas de Google Translate y DeepL, aunque el porcentaje de casos es inferior (35 porciento de los casos). Resulta más llamativo, sin embargo, ver que en las memorias de traducción de Glosbe se posiciona en segundo lugar con claridad el patrón de lexicalización que selecciona verbos de movimiento que expre-

san la Trayectoria de cruce de límites en la raíz verbal y omiten detalles adicionales sobre la Manera (32 porciento de los casos), mientras que en las traducciones automáticas solo presentaban este patrón de lexicalización el 18 porciento de los ejemplos analizados. Otro aspecto que merece la pena comentar es la presencia considerablemente más reducida en las memorias de traducción de Glosbe de casos en los que se produce una alteración del significado original de las oraciones en inglés. Vemos que, en las traducciones automáticas de Google Translate y DeepL, estos casos representaban un 18 porciento del total, mientras que en las memorias de traducción de Glosbe solo suponen un 6 porciento.

Si nos detenemos a analizar el grado de coincidencia en el patrón de lexicalización utilizado preferentemente en las memorias de traducción disponibles en Glosbe y el identificado en las traducciones automáticas de Google Translate y DeepL, observamos que, en el caso del español, Google Translate y Glosbe coinciden en un 70 porciento de los casos, frente al 62 porciento de DeepL. Por su parte, en italiano encontramos coincidencias más bajas. El patrón de lexicalización preferente en las memorias de traducción de Glosbe solo coincide en un 56 porciento de los casos con Google Translate, y en menos de la mitad de los casos con DeepL (un 47 porciento del total).

6. Conclusiones

El objetivo del presente artículo consistía en analizar de forma comparada las preferencias encontradas en el patrón de lexicalización de tres lenguas diferentes (inglés, español e italiano) en la construcción de eventos de movimiento de cruce de límites. Partiendo del inglés, una lengua englobada dentro de la categoría de las lenguas de marco satélite, hemos analizado las traducciones automáticas ofrecidas por dos herramientas en línea (Google Translate y DeepL) en español e italiano, dos lenguas consideradas de marco verbal según la clasificación binaria de Leonard Talmy de las lenguas del mundo. Para ello, se han seleccionado 17 verbos de movimiento en inglés que incorporan dentro del núcleo verbal información sobre la Manera en la que la Figura realiza el movimiento, y se han combinado con complementos postverbales que expresan una Trayectoria de cruce de límites de forma externa al verbo. A continuación, se ha analizado el patrón de lexicalización presente en las traducciones automáticas generadas por Google Translate y DeepL, y se han comparado los resultados con las traducciones más frecuentes encontradas en las memorias de traducción disponibles en la herramienta en línea Glosbe, con el fin de observar coincidencias y diferencias en los patrones de lexicalización.

De acuerdo con los datos obtenidos a partir de las traducciones automáticas generadas por Google Translate y DeepL, hemos observado que en la lengua española se muestra una clara preferencia por el patrón de lexicalización propio de las lenguas de marco verbal, es decir, por la selección de verbos de movimiento que expresan la Trayectoria de cruce de límites dentro del propio núcleo verbal (66 porciento de los casos). Por lo que respecta al componente de Manera, en un 43 porciento de los casos esta se lexicaliza de forma externa al verbo, a través de construcciones de gerundio o locuciones adverbiales, o simplemente se omite cualquier tipo de referencia a la Manera en la que se produce el movimiento (23 porciento de los casos).

Por el contrario, según los datos analizados en el presente artículo, en la lengua italiana se observa un patrón de lexicalización híbrido, a caballo entre las lenguas de marco verbal y las lenguas de marco satélite. Estas afirmaciones se basan en el hecho de que el 48 porciento de las traducciones automáticas obtenidas con Google Translate y DeepL en italiano están compuestas por verbos de movimiento que lexicalizan el componente de Manera en el propio núcleo verbal y externalizan la lexicalización de la Trayectoria de cruce de límites a través de complementos postverbales, una estructura asociada tradicionalmente a las lenguas de marco satélite. No obstante, en las traducciones automáticas generadas en italiano también encontramos un 16 porciento de casos en los que se opta por verbos de movimiento que lexicalizan la Trayectoria de cruce de límites dentro del propio verbo y la información sobre la Manera se expresa a través de complementos externos como construcciones de gerundio o locuciones adverbiales, así como otro 18 porciento de casos en los que simplemente se seleccionan verbos de movimiento que lexicalizan la Trayectoria de cruce de límites dentro del núcleo verbal y se omite cualquier información relacionada con la Manera, ambas estructuras tradicionalmente vinculadas a las lenguas de marco verbal.

Con el presente estudio también hemos comprobado que el patrón de lexicalización observado en las traducciones automáticas generadas por Google Translate y DeepL coincide en gran medida con las preferencias encontradas en las memorias de traducción disponibles en la herramienta en línea Glosbe, lo que nos permite corroborar las afirmaciones anteriores acerca del patrón

de lexicalización más frecuentemente utilizado en italiano y español para la traduccion de eventos de movimiento de cruce de límites. Por último, según los datos analizados, hemos observado que la coincidencia en la selección del patrón de lexicalización entre las herramientas de traducción automática y las memorias de traducción consultadas en Glosbe es superior en el caso de Google Translate en comparación con DeepL.

Confiamos en que el presente artículo sirva como punto de partida para investigaciones futuras sobre la expresión del movimiento a través del lenguaje y las estrategias de traducción de estas estructuras, con el objetivo de seguir arrojando luz sobre estas cuestiones y sobre las posibilidades de uso de herramientas de traducción automática como Google Translate y DeepL en campos como la enseñanza de lenguas y la traducción.

Referencias

Adán Soriano, Mónica. 2019. *Estudio comparativo de tres traductores automáticos en línea: DeepL, Yandex y Apertium*: Universidad Pontificia Comillas, Facultad de Ciencias Humanas y Sociales. Trabajo de Fin de Máster.

Aske, Jon. 1989. Path predicates in English and Spanish: A closer look. *Annual Meeting of the Berkeley Linguistics Society* 15. 1–14. 10.3765/bls.v15i0.1753.

Casacuberta Nolla, Francisco & Álvaro Peris Abril. 2017. Traducción automática neuronal. *Tradumàtica, Tecnologies de la Traducció* 15. 66–74. 10.5565/rev/tradumatica.203.

Cifuentes Férez, Paula. 2013. El tratamiento de los verbos de manera de movimiento y de los caminos en la traducción inglés-español de textos narrativos. *Miscelánea: A Journal of English and American Studies* 47. 53–80.

Cifuentes Honrubia, José Luis. 1999. *Sintaxis y semántica del movimiento. aspectos de gramática cognitiva.* Instituto de Cultura Juan Gil-Albert.

De Miguel, Elena. 1999. El aspecto léxico. En *Gramática descriptiva de la lengua española*, vol. 2, 2977–3060. Espasa Calpe.

DeepL. 2022. ¿por qué deepl? la traducción automática más precisa y sutil del mundo. (7 de febrero). https://www.deepl.com/es/whydeepl.

Díaz Prieto, Petra. 2012. Luces y sombras en los 75 años de traducción automática. En *Lengua, traducción, recepción: en honor de Julio César Santoyo*, 139–175. Universidad de León.

Fillmore, Charles J. & Collin Baker. 2012. A frames approach to semantic analysis. En *The Oxford Handbook of Linguistic Analysis*, 791–816. Oxford University Press.

Forcada, Mikel L. 2017. Making sense of neural machine translation. *Translation Spaces* 6(2). 291–309. 10.1075/ts.6.2.06for.

Fuentes-Luque, Adrián, & Alexandra Santamaría Urbieta. 2020. Machine translation systems and guidebooks: an approach to the importance of the role of the human translator. *Onomázein Revista de lingüística filología y traducción* 7. 63–82. 10.7764/onomazein.ne7.04.

Ghasemi, Hadis & Mahmood Hashemian. 2016. A comparative study of Google Translate translations: An error analysis of English-to-Persian and Persian-to-English translations. *English Language Teaching* 9(3). 13–17.

Glosbe. 2022. El mayor diccionario en línea. (10 de agosto). https://es.glosbe.com/.

González Aranda, Yolanda. 1998. *Forma y estructura de un campo semántico. A propósito de la sustancia de contenido 'moverse' en español.* Universidad de Almería, Servicio de Publicaciones.

Hijazo Gascón, Alberto. 2011. *La expresión de eventos de movimiento y su adquisición en segundas lenguas*: Universidad de Zaragoza, Departamento de Lingüística General e Hispánica. Tesis Doctoral.

Iacobini, Claudio. 2012. Grammaticalization and innovation in the encoding of motion events. *Folia Linguistica* 46(2). 359–386. 10.1515/flin.2012.013.

Ibarretxe-Antuñano, Iraide. 2017. Introduction. motion and semantic typology: A hot old topic with exciting caveats. En *Human Cognitive Processing*, 13–36. John Benjamins Publishing Company. 10.1075/hcp.59.02iba.

Kol, Sara, Miriam Schcolnik & Elana Spector-Cohen. 2018. Google Translate in academic writing courses? *The EuroCALL Review* 26(2). 50–57. 10.4995/eurocall.2018.10140.

Lakoff, George. 1987. *Women, fire and dangerous things. what categories reveal about the mind.* University of Chicago Press.

Lakoff, George. 2012. Explaining embodied cognition results. *Topics in Cognitive Science* 4(4). 773–785. 10.1111/j.1756-8765.2012.01222.x.

Lamiroy, Béatrice. 1991. *Léxico y gramática del español: Estructuras verbales de espacio y de tiempo*. Anthropos.

Langacker, Ronald W. 2011. Conceptual semantics, symbolic grammar, and the day after day construction. En *Mechanisms of Linguistic Behavior*, 3–24. Lacus.

Lear, A., L. Oke, C. Forsythe & A. Richards. 2016. "why can't i just use Google Translate?" a study on the effectiveness of online translation tools in translation of coas. *Value in Health* 19(7). A387. 10.1016/j.jval.2016.09.232.

Levin, Beth. 1993. *English verb classes and alternations. A preliminary investigation*. The University of Chicago Press.

Loock, Rudy & Sophie Léchauguette. 2021. Machine translation literacy and undergraduate students in applied languages: Report on an exploratory study. *Tradumàtica: tecnologies de la traducció* 19. 204–225. 10.5565/rev/tradumatica.281.

Maldonado González, María Concepción & María Liébana González. 2021. Los motores de traducción automática y su uso como herramienta lexicográfica en la traducción de unidades léxicas aisladas. *Círculo de Lingüística Aplicada a la Comunicación* 88. 189–211.

Morimoto, Yuko. 2001. *Los verbos de movimiento*. Visor Libros.

Slobin, Dan I. 1996. Two ways to travel: Verbs of motion in english and spanish. En *En Grammatical constructions: Their form and meaning*, 195–219. Oxford University Press.

Slobin, Dan I. 2004. The many ways to search for a frog: Linguistic typology and the expression of motion events. En *Relating events in narrative: Typological and contextual perspectives*, vol. 2, 219–257. Lawrence Erlbaum Associates.

Stolova, Natalya I. 2015. *Cognitive linguistics and lexical change. Motion verbs from latin to romance*. John Benjamins.

Talmy, Leonard. 2000. *Toward a cognitive semantics*. The MIT Press. 10.7551/mitpress/6847.001.0001.

Propuesta recibida en junio 2022 y aceptada para publicación en noviembre 2022.

Corpus de Falacias por Apelación a las Emociones: una Aproximación a la Identificación Automática de Falacias

Fallacies of Appeal to Emotions Corpus: an Approach to Automatic Fallacies Identification

Kenia Nieto-Benitez
Centro Nacional de Investigación y Desarrollo Tecnológico

Noé Alejandro Castro-Sánchez
Centro Nacional de Investigación y Desarrollo Tecnológico

Héctor Jiménez Salazar
Universidad Autónoma Metropolitana

Gemma Bel-Enguix
Universidad Nacional Autónoma de México

Resumen

Los discursos políticos en campañas electorales están orientados a movilizar y atraer con mensajes persuasivos al electorado y principalmente se argumenta apelando a las emociones incurriendo en falacias. Este artículo presenta un *corpus* de falacias en discursos políticos elaborados por candidatos a la presidencia de México, con el objetivo de obtener un recurso lingüístico en español que permita desarrollar sistemas computacionales para su minería. Hasta ahora no se conoce un *corpus* de falacias para el idioma español y los *corpus* de argumentos elaborados en el área de Minería de Argumentos se limitan a un etiquetado de la estructura argumentativa y no están elaborados a partir de discursos políticos. El *corpus* se elaboró con argumentos extraídos de los discursos y se realizó una anotación manual de premisas y conclusiones. Se obtuvo un acuerdo entre anotadores de 0.692 utilizando el índice *kappa* de Cohen. Posteriormente, se identificaron los argumentos válidos y las falacias, y como resultado se obtuvo un acuerdo de 0.442 con el mismo índice. Como contribución adicional, se presenta una línea base para la identificación de falacias utilizando los métodos de similitud coseno, *support vector machine*, *logistic regression* y *decision trees*, y la extracción de términos afectivos en los argumentos. En esta línea base se obtuvo un *F1-score* de 0.62 y es un resultado de comparación para futuras investigaciones.

Palabras clave

falacias; corpus; argumentos; emociones

Abstract

Political speeches in electoral campaigns are aimed at mobilizing and attracting the electorate with persuasive messages, and are mainly argued appealing to emotions, committing fallacies. This article presents a fallacies corpus of political speeches made by presidency candidates of Mexico, with the aim obtaining a linguistic resource in Spanish that allows computer development systems for its mining. Until now, there is no known fallacies corpus for Spanish language and arguments corpus elaborated in Argument Mining area are limited to argumentative structure tagging and are not elaborated from political speeches. The corpus was elaborated with arguments extracted from the speeches and a manual annotation of premises and conclusions was made. Inter-annotator agreement of 0.692 was obtained using Cohen's kappa index. Subsequently, valid arguments and fallacies were identified, and 0.442 agreement was obtained with the same index as a result. As an additional contribution, a fallacies identification baseline is presented using cosine similarity, support vector machine, logistic regression and decision trees methods, and effective extraction terms in the arguments.In this baseline, an 0.62 F1-score was obtained and it is a comparison result for future research.

Keywords

fallacies; corpus; arguments; emotions

1. Introducción

Las falacias son una estrategia para transformar el discurso y presentar posiciones en apariencia coherentes y sólidas con el objetivo de ganar la aprobación y simpatía de la audiencia, en especial en el discurso político emitido durante las campañas electorales.

La identificación automática de falacias es una tarea que está surgiendo en el área de Procesamiento de Lenguaje Natural. El automatizar la identificación de falacias permitiría analizar gran-

DOI: 10.21814/lm.14.2.376

des conjuntos de datos y se podrían desarrollar sistemas computacionales que permitan evaluar los discursos de candidatos políticos para facilitar al electorado la toma de decisiones en la jornada electoral; o en educación para apoyar a métodos que mejoran las capacidades del pensamiento y la evaluación crítica de los estudiantes, entre otros.

Con el fin de llegar a la minería de falacias[1] son necesarios recursos lingüísticos como *corpus* textuales que permitan el desarrollo de sistemas para discernir de manera confiable el razonamiento válido del inválido. Sin embargo, se carece de *corpus* diseñados ex profeso para abordar este problema. Las investigaciones de falacias existentes utilizaron principalmente textos escritos en inglés (Blassnig et al., 2019; Wells, 2018; Cabrejas-Peñuelas, 2015; Zurloni & Anolli, 2013; Tobolka Castro, 2007; Jason, 1986). Si bien, consisten en un conjunto de textos debidamente recopilados para el análisis, no presentan la información sobre el etiquetado de los datos y no ofrecen el acceso para su uso. En este mismo sentido, los *corpus* de argumentos elaborados en el área de Minería de Argumentos se limitan a un etiquetado de la estructura argumentativa y no están elaborados a partir de discursos políticos (Lippi & Torroni, 2016; García Gorrostieta & López López, 2019; Lawrence & Reed, 2020). La falta de estos recursos dificulta el proceso de automatizar la identificación de falacias. Por lo tanto, el interés de centrarnos en la elaboración de un *corpus* radica en la importancia de contar con recursos lingüísticos que apoyen las investigaciones en el área de Procesamiento de Lenguaje Natural.

En este artículo se reporta la elaboración de un *corpus* de falacias en español basado en discursos políticos. Se compone de argumentos realizados por los candidatos presidenciales de México en los comicios de los años 2006, 2012 y 2018. El etiquetado del *corpus* se enfoca en falacias por apelación a las emociones, las cuales aparecen con frecuencia en el contexto de la argumentación política (Jason, 1986; Cabrejas-Peñuelas, 2015; Blassnig et al., 2019; Al-Hindawi et al., 2015; Morales Gutiérrez, 2016). Además del etiquetado de falacias, cada argumento cuenta con un etiquetado de su estructura argumentativa. Esta información se considera útil para los investigadores que toman en cuenta los componentes del argumento para identificar falacias, así como para los investigadores en el área de Minería de argumentos.

[1] "Identificar y caracterizar aquellos elementos de un argumento falaz que lo distinguen de un argumento válido" y desarrollar métodos computacionales que permitan su identificación con esos elementos (Wells, 2018).

Por último, se presenta una línea base para la identificación de falacias mediante el uso de modelos de aprendizaje automático. Se considera que la conceptualización del término falacia, las características para su análisis e identificación y la relación entre los componentes argumentales podrían conducir a la implementación de técnicas basadas en aprendizaje automático que ayuden a detectar falacias. Considerando que la ocurrencia de términos afectivos y la similitud textual de las premisas con la conclusión del argumento puede ser indicativo de una posible falacia. Ya que, como veremos, una falacia presenta varios términos afectivos y una baja similitud entre los componentes de su argumento.

Las secciones del artículo se encuentran estructuradas de la siguiente forma. En la sección 2, se describe el concepto de falacia y se presentan las características para su identificación. En la sección 3, se describe el proceso y los resultados del *corpus* elaborado. En la sección 4, se presentan los resultados de la identificación de falacias. Y en la sección 5 se presentan las conclusiones del trabajo.

2. Falacias

En la literatura existe una variedad de conceptos para determinar qué es una falacia y esto ha llevado a la aparición de varios enfoques para su definición. En un sentido general del término, las falacias son tipos de argumento que parecen correctos, pero contienen un error de razonamiento por el mal manejo de sus proposiciones (Copi & Cohen, 2013); son errores comunes al argumentar (Govier, 2013) que resultan en una argumentación fallida o fraudulenta (Vega Reñón, 2013). También se puede definir como argumentos que tienen errores en su forma al infringir alguna de las estructuras deductivamente válidas (Copi & Cohen, 2013; Capaldi, 2011) o instancias de formas lógicas inválidas identificables (Hansen, 2019).

Otros investigadores hacen referencia a reglas o criterios que se deben seguir en el discurso o en la construcción de argumentos. En este contexto las falacias se caracterizan por infringir las reglas de una discusión crítica e interrumpir el proceso de resolución de una disputa (Van Eemeren & Grootendorst, 1987, 2002); o bien, por infringir las reglas o criterios para construir buenos argumentos (Weston, 2006; Damer, 2009). En su minoría las falacias se consideran como algo no plausible y se hace referencia a la intención persuasiva del argumento y los efectos que produce (Vega Reñón, 2013; Zurloni & Anolli, 2013).

2.1. Falacias por apelación a las emociones

Para fines de este artículo, una falacia es un argumento incorrecto que presenta un error de razonamiento cuyo patrón común puede detectarse en su contenido por el mal manejo de sus proposiciones (cuando las premisas no consiguen apoyar a la conclusión). El apoyo entre las proposiciones se ve afectado por diversos factores, por ejemplo, en argumentos que dependen de emociones en lugar de evidencia o justificaciones racionales para inferir la conclusión. A este tipo de falacia se le conoce como falacia por apelación a las emociones o *Ad populum* e incorpora otros tipos de falacia como *Ad misericordiam* (Copi & Cohen, 2013). Por ejemplo, en el siguiente argumento se tienen expresiones afectivas como *etapa sombría y de oscuridad* y *etapa de luz y esperanza* para contrastar la situación actual de México con lo que quiere la sociedad. Los ejemplos del 1 al 5 tienen una anotación propia sobre su estructura argumentativa.

1. (Premisa) México está muy claro en lo que quiere y está cierto que ya no quiere más de lo mismo; (Conclusión) quiere pasar de esta etapa sombría y de oscuridad a una nueva etapa de luz y esperanza (Morales Gutiérrez, 2016).

Las falacias por apelación a las emociones, como su nombre lo indica, recurren al lenguaje expresivo como recurso para conseguir un objetivo (Weston, 2006). El lenguaje expresivo en los argumentos son todas aquellas palabras cuya única función sea influir en las emociones (Copi & Cohen, 2013; Weston, 2006). Este lenguaje se presenta en la argumentación mediante un "conjunto de factores o circunstancias que provocan alguna reacción ya sea positiva o negativa en aquellos que escuchan dicha emisión" e incluye un contexto extralingüístico que va conformando la emoción en la argumentación (Camargo López, 2018). En la Tabla 1 se observa una lista de los recursos del lenguaje (patrones argumentativos y rasgos emotivos) usados en las falacias por apelación a las emociones y un conjunto de preguntas para su identificación.

El lenguaje expresivo busca provocar emociones como generosidad, altruismo, piedad y compasión por medio de palabras que provoquen estás emociones. Este lenguaje se utiliza por medio de diferentes patrones argumentativos (Tabla 1) que permiten justificar la opinión en el argumento (Copi & Cohen, 2013). Por ejemplo, en algunos argumentos se privilegia la cifra por encima de la calidad ciudadana, se usan apelativos con los que se denomina al pueblo (Morales Gutiérrez, 2016) o se caricaturiza al oponente (Weston, 2006). Por ejemplo:

2. (Conclusión) Estamos muy entusiasmados, (Premisa) porque ustedes y millones de mexicanos más, saben que tenemos el mejor proyecto, el proyecto para que México esté mejor (Morales Gutiérrez, 2016).

En algunos casos el lenguaje expresivo se puede identificar mediante rasgos emotivos (Tabla 1) que permitan la reconstrucción de la emoción utilizando los siguientes ejes: 1) personas involucradas; 2) intensidad/cantidad; y 3) tiempo (Camargo López, 2018).

En las personas involucradas el discurso se centra en el orador, donde el político encargado de emitir el discurso se autodescribe para justificar que cumple con sus compromisos, o involucra al auditorio, donde el político incorpora al oyente en el proyecto político para hacerlo sentir parte del equipo:

3. (Premisa) Cada vez menos gente cree en las promesas que hacen los políticos. (Conclusión) Por eso, allá en lo que fue mi campaña para Gobernador hace siete años, decidí innovar la forma de hacer política, asumiendo compromisos. En esta campaña, la que hoy inicio aquí en Guadalajara, volveré a firmar compromisos con todo México (Camargo López, 2018).

La intensidad/cantidad afecta a categorías como la distancia o la calidad de las personas mediante una modulación cuantitativa que impacta en la percepción de la audiencia:

4. (Conclusión) Duele reconocerlo, pero México no vive un buen momento. (Premisa) Muchos mexicanos atraviesan tiempos difíciles, sienten incertidumbre y desesperación. (Premisa) Muchos mexicanos viven angustiados y, lo que es peor, viven con miedo. (Premisa) Hay un México con enorme pobreza, con millones de familias a quienes no les alcanza ni lo más mínimo para comer... (Camargo López, 2018).

Finalmente, el tiempo se centra en la descripción del lapso en el que suceden los hechos que se narran. Es utilizado para hablar de la situación del país, en este caso se puede hablar bien o mal del presente, o hacer referencia al pasado para mencionar errores o logros cometidos por los partidos políticos:

Tipo	
Patrones argumentativos (Copi & Cohen, 2013) (Morales Gutiérrez, 2016)	– Apelativos con los que se denomina al pueblo: *los habitantes*, *los mexicanos*, *nuestros hermanos*, etc. – Contar con el apoyo mayoritario del pueblo: *millones de mexicanos*, *exigencias de los ciudadanos*, etc. – Privilegiar la cifra por encima del prestigio o circunstancias personales de los individuos: *50 mil veracruzanos*, *todo Veracruz*, etc. – Palabras dirigidas a generosidad, altruismo, piedad y compasión: *el joven es huérfano.*
Rasgos emotivos (Camargo López, 2018)	– Personas involucradas: *como latinoamericana* y *como suramericana*, *católico* y *cristiano*, etc. – Intensidad/cantidad: *muchos mexicanos*, *enorme pobreza*, *miles de muertes*, etc. – Tiempo: *nuestros orígenes*, *lo que hemos sido*, *nostalgia del ayer*, etc.
Preguntas críticas de evaluación (Tindale, 2007)	– ¿El argumentador apela a la lástima para sustentar la verdad de una afirmación o recomendar alguna acción? – ¿Este es un contexto en el que las apelaciones emocionales son relevantes? – ¿La premisa de lástima es relevante para la conclusión propuesta? – ¿La popularidad es relevante para la afirmación hecha en la conclusión?

Tabla 1: Características utilizadas para la identificación de falacias por apelación a las emociones.

5. (Concusión) El PRI tiene que asumir el papel que le corresponde, (Premisa) no inspirado en la nostalgia del ayer, sino en los retos del presente, para ganar el futuro (Camargo López, 2018).

Otros criterios utilizados en la identificación de falacias por apelación a las emociones son las preguntas críticas de evaluación y la relevancia de la información. Las preguntas presentadas en la Tabla 1 evalúan si el argumentador se ha apoyado con algún tipo de evidencia y si la apelación es relevante para la conclusión en el contexto del argumento, es decir, se evalúan las apelaciones en el contexto del argumento, así como la relación entre la premisa y conclusión (Tindale, 2007). Y la relevancia se enfoca en la aceptabilidad de la información en el argumento (Gilbert, 2004), en específico si la información dada en la premisa es relevante para la conclusión (Al-Hindawi et al., 2015; Damer, 2009).

2.2. Similitud textual en las falacias

Desde un punto de vista computacional, el lenguaje afectivo y la relevancia de la información se podrían evaluar mediante la similitud textual entre las proposiciones del argumento.

La similitud textual es la base de las tareas de procesamiento del lenguaje natural como la recuperación de información, la traducción automática, sistemas de diálogo, entre otros (Wang & Dong, 2020). Se encarga de comparar unidades textuales (documento, párrafo, oración o palabra) y expresa el grado de semejanza entre las unidades. Las unidades pueden ser similares léxicamente o semánticamente. Por ejemplo, dos textos son léxicamente similares si las oraciones contienen secuencias de palabras semejantes; y son semánticamente similares si las oraciones se refieren a lo mismo dentro del contexto aún sin tener necesariamente secuencias con palabras similares (Álvarez Carmona, 2014).

Considerando la similitud textual en los argumentos, podríamos esperar que las proposiciones de un argumento válido carezcan de un lenguaje afectivo y compartan información permitiendo establecer una relación entre ellas. En cambio, si en un argumento se encuentran diversas palabras afectivas podría decrecer la similitud entre las proposiciones y la información se consideraría irrelevante para establecer la conclusión del argumento, dando como resultado un argumento falaz.

3. Elaboración del corpusde falacias

El *corpus* consta de un conjunto de argumentos obtenidos a partir de discursos políticos escritos en español. Los argumentos se obtuvieron de los discursos generados por los candidatos presidenciales de México en los años de 2006, 2012 y 2018.

Se obtuvieron 80 discursos[2] de forma escrita. Estos discursos fueron realizados por los candidatos de los partidos políticos del PRD, PRI, PAN, MORENA y políticos independientes. La extensión de cada discurso es de 4 a 10 páginas y en total en los discursos se tienen 3,493 párrafos.

El etiquetado de los textos se realizó con apoyo de tres anotadores. Todos los anotadores son estudiantes en el área de Lingüística: dos de séptimo semestre y uno de décimo trimestre; y la lengua materna de los estudiantes es el español.

El proceso descrito en la Sección 3.1 forma parte de la guía proporcionada a los anotadores que incluye además, una descripción general de Minería de argumentos (qué es un argumento y cómo identificarlo), una lista de marcadores del discurso y palabras clave para identificar los componentes de un argumento, así como la descripción de la taxonomía de falacias utilizada, la definición de falacia por apelación a las emociones y las características para su identificación. La guía y el corpus están disponibles al público y podrá ser consultado en linea[3].

3.1. Proceso de anotación

El *corpus* contiene un conjunto de argumentos clasificados en falacias por apelación a las emociones y argumentos válidos. Además, cada argumento tiene identificada su estructura argumentativa mediante la clasificación de sus proposiciones (conclusión y premisa).

La extracción de los argumentos se realizó mediante la identificación de las oraciones argumentativas en cada uno de los párrafos de los discursos. Posteriormente, estas oraciones se clasificaron como premisa o conclusión y, en algunos casos, se identificaron segmentos no argumentativos. Estos segmentos acompañan al argumento para contextualizar o enfatizar su afirmación y pueden enunciarse al inicio, en medio o al final del argumento. Por ejemplo:

6. (Conclusión) Vamos a ganar esta elección (Premisa) porque es importante, (Premisa) y porque cuando algo va a ser importante para nuestras familias, las mujeres están dispuestas a dar la lucha hasta ganar, (Segmento no argumental) ¡esta elección la vamos a ganar con las mujeres!

Aun cuando estos segmentos no proporcionan información para determinar si un argumento es una falacia y no forman parte de los componentes del argumento, son importantes para las investigaciones en el área de Minería de argumentos y donde se podría utilizar este corpus.

Se obtuvieron uno a más argumentos dependiendo del tamaño del párrafo. En otros párrafos se obtuvo solo parte del argumento, una premisa o una conclusión. En este caso, la extracción de los argumentos se realizó considerando la unión de dos párrafos (párrafo actual + párrafo siguiente o párrafo anterior + párrafo actual).

La identificación de falacias por apelación a las emociones consistió en distinguir los argumentos válidos de falacias por medio de la inferencia que se realiza entre las premisas y la conclusión del argumento y el lenguaje expresivo presente en los componentes del argumento. La inferencia entre los componentes argumentativos se evaluó mediante las preguntas críticas de evaluación presentadas por Tindale (2007) (Tabla 1) y la relevancia de la información contenida en las premisas (Gilbert, 2004; Al-Hindawi et al., 2015; Damer, 2009), en este caso una premisa se considera relevante cuando la información presentada es aceptable para el tema tratado en el argumento. En el caso del lenguaje expresivo se consideraron todos aquellos segmentos de texto (palabras o frases) que apelan a las emociones y los rasgos emotivos presentados por Camargo López (2018) (Tabla 1).

Considerando lo anterior, los argumentos obtenidos se clasificaron como válidos o como falacias. Por ejemplo, el argumento 7 se considera una falacia por contener términos afectivos como miedo y estancamiento; y rasgos emotivos de intensidad (grandes cambios) y de tiempo (nuestro origen). Mientras que el argumento 8 no contiene algunas de estas características y mediante su evaluación con las preguntas críticas se determinó que es un argumento válido.

7. (Falacia) (Premisa) México no se resigna a vivir bajo una estela de miedo, estancamiento y falta de oportunidades. (Conclusión) Por eso, hoy regresamos a nuestro origen; hoy venimos a reafirmar que los grandes cambios sí son posibles cuando así lo decide la gente.

[2] Los discursos del 2006 se obtuvieron de Ojeda Cruz (2016) y los discursos del 2012 y 2018 de páginas oficiales de internet.

[3] http://corpus_falacias.tecln.cenidet.tecnm.mx/

8. (VÁLIDO) (CONCLUSIÓN) Creo en el principio de la autodeterminación de los pueblos, (PREMISA) porque guarda relación estrecha con la soberanía nacional, pero además, se remonta a nuestra proclamación de Independencia y a la voluntad del pueblo mexicano de determinar libremente su destino.

3.2. Resultados de la anotación

En el proceso de anotación se utilizaron los 3,493 párrafos contenidos en los 80 discursos. Se encontraron 629 párrafos con argumentos, esto representa solo el 18 % del total de párrafos que conforman los discursos políticos analizados.

Por el tipo de discurso utilizado (discurso político) los párrafos de los discursos contienen figuras retóricas o se encuentran opiniones sin ningún sustento. En el caso donde los párrafos contienen saludos, agradecimientos, convocatorias, diálogos o descripción de los problemas no se presentaron argumentos.

El acuerdo entre los anotadores se realizó considerando la estructura argumentativa y la identificación de falacias realizada por tres anotadores (A1, A2 y A3). El acuerdo se obtuvo mediante el índice *Kappa* de Cohen (k_c) (Cohen, 1960) y el índice *Kappa* de Fleiss (k_f) (Gordillo Torres & Rodríguez Perera, 2009). La interpretación para valorar el grado de acuerdo en función del índice de *kappa* (Landis & Koch, 1977) se muestra en la Tabla 2.

k_c/k_f	**Grado de acuerdo**
< 0.00	Sin acuerdo
0.00–0.20	Insignificante
0.21–0.40	Mediano
0.41–0.60	Moderado
0.61–0.80	Sustancial
0.81–1.00	Casi perfecto

Tabla 2: Interpretación del grado de acuerdo entre anotadores.

En la extracción de argumentos el anotador A1 obtuvo 647 argumentos; el anotador A2, 605 argumentos; y el anotador A3, 638 argumentos. En este proceso se obtuvo un acuerdo k_f de 0.456, equivalente a un total de 601 argumentos. Cada anotador (A) obtuvo un conjunto diferente de falacias (FA), argumentos válidos (AV), componentes argumentativos (premisas (P) y conclusiones (C)) y segmentos no argumentativos (S-NA) (Tabla 3).

Nº	**A**	**C**	**P**	**S-NA**	**FA**	**AV**
601	A1	601	1123	28	434	167
	A2	601	1036	29	388	213
	A3	601	1225	31	428	173

Tabla 3: Resultados obtenidos por anotador. Se etiquetaron los componentes del argumento y se determinó cuáles de los argumentos son falacias por apelación a las emociones.

Los argumentos no cuentan con una estructura específica. Es decir, un argumento puede tener una conclusión (C) y una premisa (P), mientras que otro anotador puede establecer una conclusión (C) y dos o más premisas ($P_1, P_2, \ldots, P_n$) para el mismo argumento. Además, se debe considerar el límite del texto seleccionado para cada componente del argumento. Por ello, la evaluación de la estructura argumentativa se estableció de acuerdo con la secuencia de los componentes argumentativos (SCA) identificados en los argumentos. Por ejemplo, en la Tabla 4 los tres anotadores coincidieron en la clasificación de componentes en los argumentos uno, dos y cinco, mientras que en los argumentos tres y cuatro sólo coincidieron los anotadores A1 y A3, y en el 601 cada anotador realizó una clasificación distinta. La secuencia con mayor frecuencia fue CP, es decir, los argumentos tienen por lo regular una conclusión seguida de una premisa (Tabla 5).

Nº	**A1**	**A2**	**A3**	**SCA**
1	PC	PC	PC	✓
2	PPC	PPC	PPC	✓
3	PCP	PPCP	PCP	
4	CP	CPP	CP	
5	CP	CP	CP	✓
...	...	...	...	
601	CP	PC	CPP	

Tabla 4: Coincidencias y discrepancia en la identificación de componentes.

Secuencia	**Frecuencia**	**Frecuencia en %**
CP	141	39.49
PC	86	24.08
CPP	39	10.92
PPC	23	6.44
PCP	14	3.92

Tabla 5: Frecuencias en el etiquetado de la estructura argumentativa.

En la estructura argumentativa se obtuvo un acuerdo k_c de 0.692 y un acuerdo k_f de 0.648, ambos resultados con un grado de acuerdo sustancial (Tabla 6). Las diferencias en el etiquetado fueron principalmente en el número de premisas identificadas en cada argumento.

Grupos	SCA	k	Grado
A1-A2	448	**0.692**	Sustancial
A1-A3	424	0.657	Sustancial
A2-A3	393	0.591	Moderado
A1-A2-A3	357	0.648	Sustancial

Tabla 6: Acuerdo en la estructura argumentativa.

La identificación de falacias fue una tarea compleja de realizar. Si bien, los anotadores siguieron una guía detallada para distinguir argumentos válidos (AV) de falacias (FA), se estima que predominó la opinión subjetiva respecto al tema tratado en el argumento o de su estructura. Esto influyó en cómo se evaluaron los argumentos, dando como resultado un índice k_c de 0.442 y un índice k_f de 0.282 (Tabla 7), valores por debajo de los resultados obtenidos en la estructura argumentativa.

Grupos	FA	AV	k	Grado
A1-A2	298	77	0.136	Insignificante
A1-A3	368	102	**0.442**	Moderado
A2-A3	313	98	0.278	Mediano
A1-A2-A3	262	63	0.282	Mediano

Tabla 7: Acuerdo en la identificación de falacias.

3.3. Descripción del corpus obtenido

El *corpus* contiene 34,739 *tokens*, 4,103 formas de palabra y 2,799 lemas. La palabra con mayor frecuencia absoluta (FA) y relativa (FR)[4] en el *corpus* fue el verbo *ir* con una frecuencia absoluta de 412 y una frecuencia relativa de 319.132 por cada 10,000 palabras (Tabla 8). Respecto a los bigramas y trigramas, los términos con mayor frecuencia en el corpus fueron *ir_ganar* con una FA de 41 y una FR de 30.490 e *ir_volver_ganar* con una FA de 17 y una FR de 12.853 (Tabla 9).

En el cálculo de la FA y FR se eliminaron las *stopwords* (signos de puntuación, pronombres, determinantes, cifras y numerales, preposiciones, conjunciones y fechas) y los verbos más comunes (ser, estar, tener y haber) mediante el procesamiento del texto con la bilbioteca de FreeLing.[5] Esto nos permitió obtener las palabras más significativas del *corpus*. Y en el caso de los bigramas y trigramas, se dejaron los verbos comunes para conocer su relación con otras palabras.

Palabras	FA	FR
Ir	412	319.132
México	232	179.705
Querer	181	140.201
Poder	176	136.328
País	168	130.131

Tabla 8: Palabras con mayor frecuencia en el *corpus*.

El lenguaje expresivo en los argumentos varía de acuerdo a su polaridad y categoría emocional. Estas características se obtuvieron de acuerdo con los diccionarios iSOL (Molina González et al., 2013) y SEL (Díaz Rangel et al., 2014). El diccionario iSOL[6] contiene 8,135 términos clasificados con una polaridad positiva y negativa; y el diccionario SEL[7] contiene 2,026 términos clasificados en seis categorías: sorpresa, tristeza, repulsión, miedo, enojo y alegría.

El *corpus* contiene 240 términos afectivos con polaridad positiva y 269 con polaridad negativa. Como se observa en la Figura 1, en los argumentos se encuentran con mayor frecuencia los términos positivos y contienen entre 0 y 5 términos afectivos. Con relación a la polaridad (Figura 2), los argumentos tienen principalmente una polaridad positiva y algunos tienen ambas polaridades (positivo/negativo), es decir contienen el mismo número de términos positivos como negativos. Y en el caso donde se tienen ambas polaridades, pero una es mayor que la otra, se consideró la polaridad predominante para clasificar el argumento.

[4] La frecuencia absoluta es el recuento real de todas las apariciones de una palabra o frase en el corpus y la frecuencia relativa es la frecuencia de palabras o frases en el corpus en un determinado número de palabras (Brezina, 2018).

[5] FreeLing contiene un conjunto de herramientas de código abierto para el análisis de lenguaje como análisis morfológico, detección de entidades nombradas, *PoS-tagging* y *parsing*. https://nlp.lsi.upc.edu/freeling/

[6] El diccionario iSOL contiene 2,509 palabras con una polaridad positiva y 5,626 palabras con una polaridad negativa.

[7] El diccionario SEL contiene en la categoría alegria 668 términos; en repulsión, 209; en enojo, 382; en miedo, 211; en sorpresa, 175; y en tristeza, 391. Algunos términos en el diccionario están clasificados con más de un tipo de emoción. Por ejemplo, el término "angustia" se encuentra en la categoría tristeza con una frecuencia de uso de 0.763 y en la categoría de miedo con una frecuencia de 0.797.

Bigramas	**FA**	**FR**	**Trigramas**	**FA**	**FR**
ir_ganar	41	30.490	ir_volver_ganar	17	12.853
haber_ser	27	20.078	actual_política_económico	10	7.560
política_económico	24	17.847	política_económico_haber	9	6.804
amigo_amigo	24	17.847	gente_estar_nosotros	8	6.048
ir_ser	19	14.129	cambiar_actual_política	7	5.292

Tabla 9: Bigramas y Trigramas con mayor frecuencia en el *corpus*.

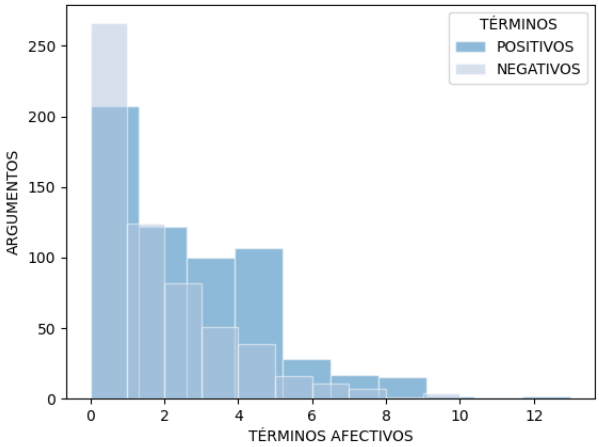

Figura 1: Frecuencia de términos afectivos en los argumentos.

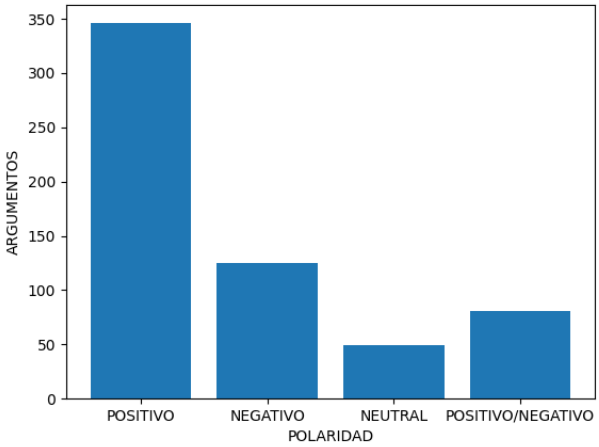

Figura 2: Polaridad de los argumentos.

De acuerdo con el tipo de emoción en los argumentos, el *corpus* contiene 411 términos afectivos clasificados con un tipo de emoción y los argumentos apelan principalmente a la alegría (Figura 3). Cabe mencionar que un argumento apela a más de un tipo de emoción. Por ejemplo, de los 431 argumentos que apelan a la alegría, 174 apelan a la alegría junto a otro tipo de emoción.

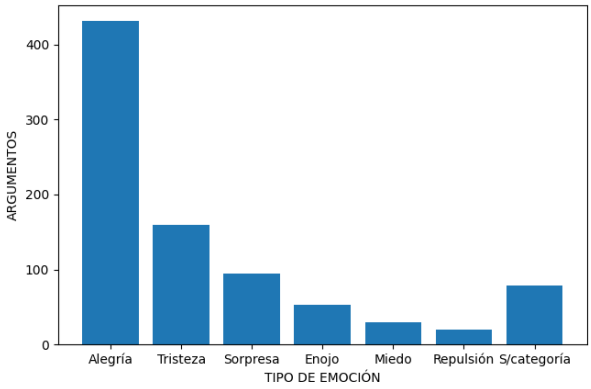

Figura 3: Argumentos por tipo de emoción.

Considerando el mejor acuerdo entre el grupo de dos anotadores, en el *corpus* se tienen 363 falacias y 102 argumentos válidos. De este conjunto, como se muestra en la Figura 4, algunos argumentos con polaridad neutra se encuentran clasificados como falacias y argumentos con polaridad positiva o negativa como válidos. Ambos tipos de argumento tienen entre cero y cinco términos afectivos y a diferencia de las falacias, los argumentos válidos contienen menos términos negativos. Por otra parte, los argumentos apelan al menos a un tipo de emoción y algunos no apelan a las emociones pero se consideran como falacias. Y en ambos tipos de argumento predomina la apelación a la alegría seguida de la tristeza (Figura 5).

La diversidad léxica[8] en los argumentos (considerando los segmentos no argumentativos que acompañan a los argumentos) se encuentra por debajo de 0.8 (Figura 6) y por tipo de argumento se tiene una frecuencia de la diversidad entre 0.5 y 0.7 (Figura 7), esto indica que los argumentos utilizan un limitado vocabulario y reciclan elementos léxicos. Con relación a la distribución de los datos por términos afectivos, diversidad y tipo de argumento, los argumentos tienden a ser

[8]La diversidad léxica mide si en general un texto o corpus utiliza una amplia gama de términos o se limita a elementos léxicos que se reciclan. Se utilizó el método *type/token* ratio para calcular la diversidad léxica en cada argumento (Brezina, 2018).

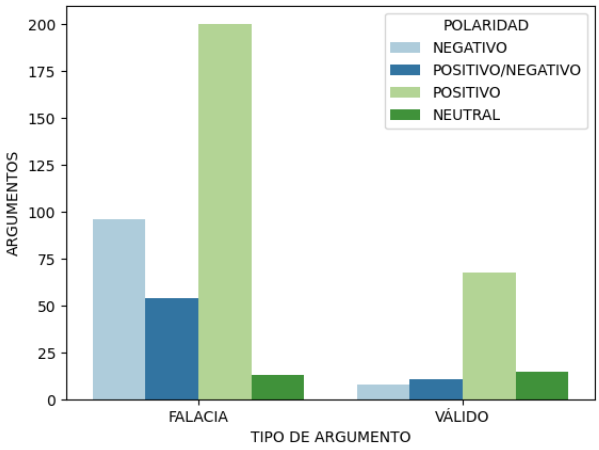

Figura 4: Polaridad por tipo de argumento.

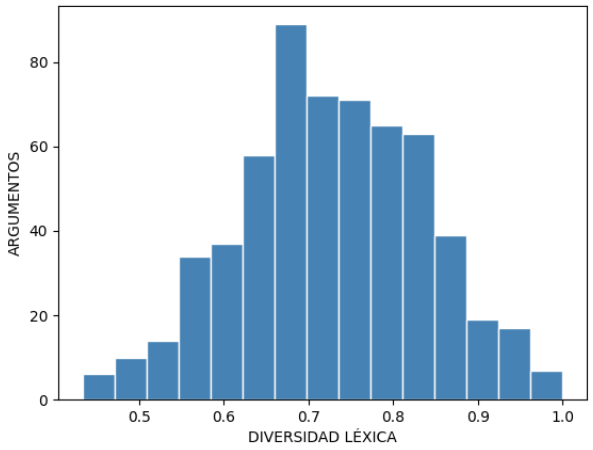

Figura 6: Cantidad de argumentos por valores de diversidad léxica.

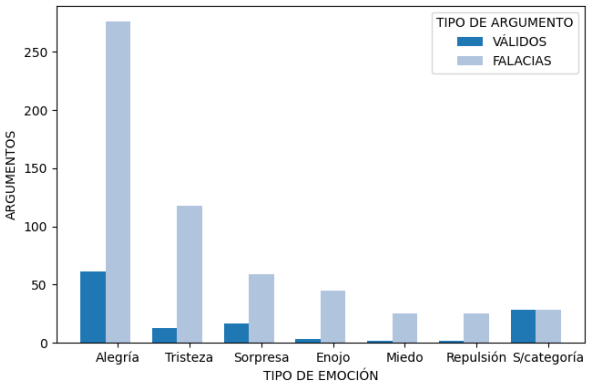

Figura 5: Cantidad de falacias y argumentos válidos por categoría emocional.

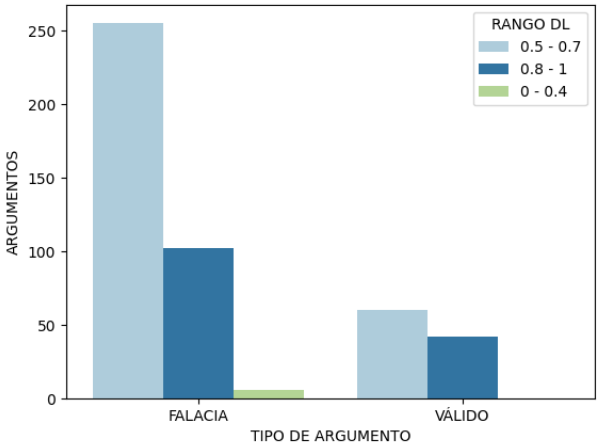

Figura 7: Cantidad de falacias y argumentos válidos por valores de diversidad Léxica (DL).

falacias mientras más número de términos afectivos contengan. Sin embargo, algunos argumentos son falaces incluso si tienen 0 términos afectivos. Por último, se observa una tendencia a disminuir la diversidad al aumentar el uso del lenguaje expresivo en los argumentos (Figura 8).

4. Clasificación de argumentos: línea base

El objetivo de la identificación de falacias por medio de un modelo computacional es proporcionar una línea base para fines de comparación con métodos futuros. Se utilizaron dos tipos de características para distinguir argumentos válidos de falacias y tres métodos comúnmente utilizados en la clasificación de textos basado en aprendizaje automático.

4.1. Características utilizadas en la clasificación de argumentos

Las características existentes para la identificación de falacias varían de acuerdo al tipo de falacia a identificar. Esta investigación se centra en la identificación de falacias por apelación a las emociones y su principal característica es el uso de lenguaje afectivo para justificar la opinión o postura en el argumento.

Como veremos, la similitud entre las proposiciones (premisa y conclusión) es un indicador del tipo de argumento. Las proposiciones de un argumento válido comparten palabras (sustantivos y verbos) que permiten establecer la idea y tema tratado en él y con ello se establece una relación (similitud textual) entre sus componentes. Las falacias utilizan palabras relacionadas a emociones y carecen de esta similitud, dando co-

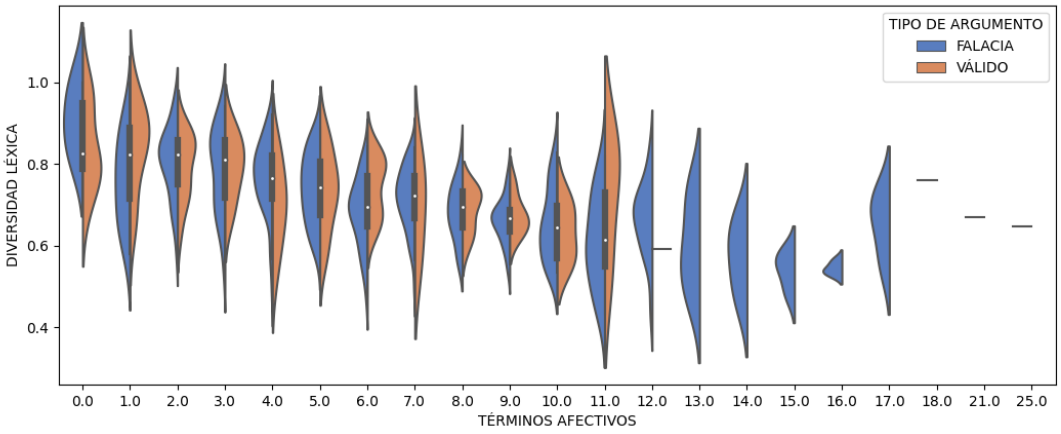

Figura 8: Dispersión de los valores de la diversidad léxica con respecto al número de términos afectivos contenidos en los argumentos.

mo resultado la irrelevancia de las premisas para inferir la conclusión. Se considera que cuanto mayor sea la similitud textual entre los componentes y menos palabras emocionales apuntaría a argumentos válidos.

4.2. Experimentación y discusión de los resultados

La clasificación de argumentos se realizó con los métodos de *Support vector machine* (SVM), *Logistic Regression* (LR) y *Decision Trees* (DT) y se utilizaron como características la similitud textual entre los componentes del argumento y los términos afectivos en los argumentos. Si bien, existen diversos métodos en tendencia como *Artificial Neural Network*, consideramos que estos métodos pueden utilizarse en trabajos con una orientación más específica a la identificación de falacias.

La anotación de los datos se acotó a los segmentos argumentativos, premisa y conclusión, y al etiquetado de falacias. El objetivo principal es identificar cuáles argumentos son falacias por apelación a las emociones mediante el contenido del argumento. El etiquetado de los segmentos no argumentales se excluyen ya que su función es acompañar al argumento y no aportan información para determinar si un argumento es una falacia.

El procesamiento del texto para la obtención de las características se realizó con la bilbioteca *spaCy*. Además, se utilizó el método de *Part-of-speech* de la misma librería para la lematización del texto y eliminación de las *stopwords*. Las *stopwords* eliminadas fueron los artículos, signos de puntuación, números, símbolos, conjunciones, preposiciones y pronombres.

La similitud entre los componentes del argumento se obtuvo mediante el método de similitud coseno[9] considerando dos procedimientos basados en similitud textual semántica. En el primer procedimiento se calcularon similitudes entre pares de componentes (una premisa y una conclusión) y se promediaron las similitudes obtenidas entre el número de combinaciones a partir de los componentes de cada argumento (SC–P). Y en el segundo procedimiento se obtuvo una similitud entre el conjunto de premisas y la conclusión (SC–A).

Argumento	SC-P	SC-A	TA
1	0.702	0.797	14
2	0.595	0.708	6
3	0.865	0.865	7
4	0.755	0.850	7
5	0.828	0.828	5

Tabla 10: Muestra de los valores de similitudes calculadas y la cantidad de términos afectivos en los primeros cinco argumentos del *corpus*.

En la identificación de los términos afectivos (TA) en los argumentos, estos se obtuvieron mediante el diccionario iSOL y SEL descri-

[9]Es un método ampliamente utilizado para obtener el grado de semejanza entre dos segmentos de textos. El método se implementó con la biblioteca de *spaCy* con el método *similarity()* y dentro de su configuración se utilizó el paquete de vectores *es_core_news_lg*. https://spacy.io/models/es#es_core_news_lg. La similitud se determina comparando vectores de palabras o *"word embeddings"* (https://spacy.io/usage/linguistic-features), en la cual se obtiene el vector promedio de los tokens de cada texto para después calcular el coseno entre ellos. https://spacy.io/usage/linguistic-features#vectors-similarity.

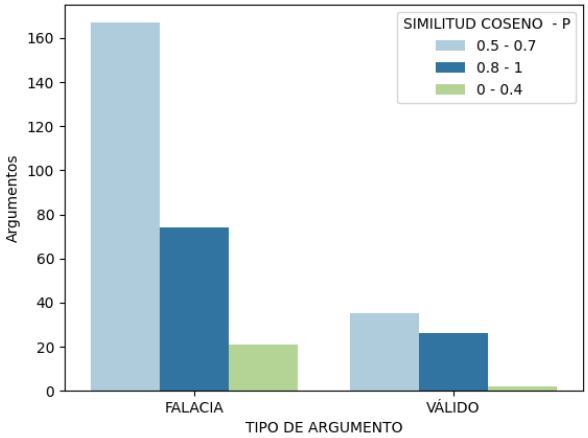

Figura 9: Cantidad de falacias y argumentos válidos por SC–P

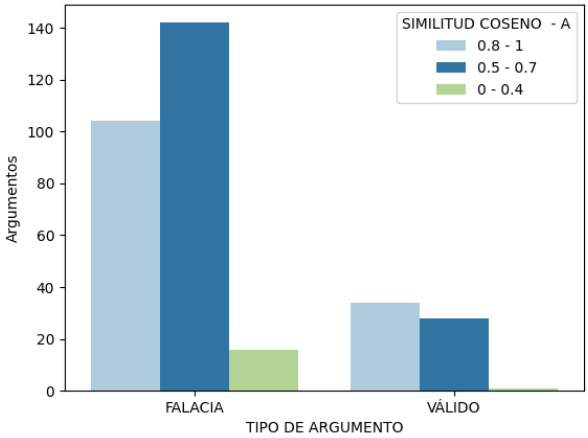

Figura 10: Cantidad de falacias y argumentos válidos por SC–A

tos en la Sección 3.3. La Tabla 10 muestra algunos de los resultados obtenidos a partir de estas características de una muestra de argumentos. Por ejemplo, el argumento 1 tiene una similitud entre sus proposiciones de 0.702 con el procedimiento SC–P, 0.797 con SC–A y contiene 14 términos afectivos. La dispersión de estas características y el tipo de argumento se muestran en las figuras 11 y 12.

El grado de similitud SC–P está entre 0.5 y 0.7. en los dos tipos de argumentos (Figura 9), en SC–A se encuentra el mismo rango en las falacias (Figura 10) y a diferencia de los argumentos válidos las falacias llegan a tener un grado de similitud entre 0 y 0.4 (Figura 9 y 10). Con el procedimiento SC–A se obtiene un conjunto mayor de argumentos válidos con un grado de similitud a 0.8 (Figura 10). La diferencia entre los tipos de argumento se presenta en el número de términos afectivos utilizados. Mientras mayor sea el número de términos afectivos, los argumentos se consideran falacias, sin embargo tienen un grado de similitud inicial de 0.5 en ambos procedimientos (Figura 11 y 12).

Se realizaron dos experimentos con el *corpus* presentado en este artículo. El primero, con el total de argumentos en el corpus. Y el segundo, con el conjunto de argumentos donde se obtuvo un mejor acuerdo entre los anotadores: 465 argumentos. El rendimiento de los métodos se evaluó con la métrica *F1-score* siguiendo una validación cruzada de 10 pliegues con el 70 % de los datos para el entrenamiento en cada método implementado.[10]

[10]Los métodos se implementaron utilizando la bilbioteca de *scikit-learn* con los valores de configuración predeterminados.

En la clasificación de los argumentos se obtuvo un rendimiento de 0.58 con el método SVM al procesar la similitud SC–P con los términos afectivos. Y el rendimiento disminuye a 0.50 al utilizar una sola característica: SC–P (Tabla 11). Como se observa en en la Tabla 11, los resultados alcanzan un *F1-score* de 0.50. Sin embargo, al tener un conjunto menor de datos con un mejor acuerdo entre los anotadores, el rendimiento incrementa a 0.62 al procesar las características SC–P y TA; y a 0.55 con la similitud SC–P (Tabla 12). Justamente con los atributos de similitud SC-P y términos afectivos se obtiene, para todos los clasificadores, el mayor desempeño.

Características	**SVM**	**LR**	**DT**
SC–P	**0.50**	**0.50**	**0.50**
SC–A	0.45	0.42	0.42
TA	0.42	0.42	0.42
SC–A+TA	0.55	0.50	0.50
SC–P+TA	**0.58**	0.56	0.55

Tabla 11: Resultado de la clasificación con 601 argumentos. Se utilizaron características individuales y la combinación de las mismas. Los resultados se obtuvieron con la métrica *F1-score*.

Algunos argumentos aunque tienen una similitud mayor a 0.8 y cero términos afectivos se consideran falacias. Del mismo modo sucede con los argumentos válidos, se consideran válidos aquellos argumentos que tienen una baja similitud y varios términos afectivos (Figura 9 y 10). Por ello, los resultados con el método de clasificación fueron menores a lo esperado, sin embargo son una base y hay un margen importante de mejora para llevar a cabo la minería de falacias en discursos políticos.

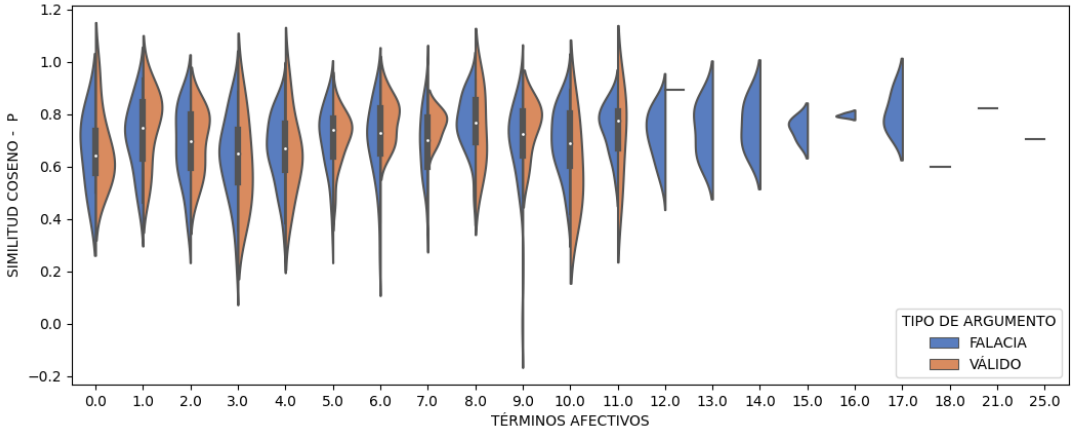

Figura 11: Dispersión de los argumentos con respecto a los resultados obtenidos por SC–P.

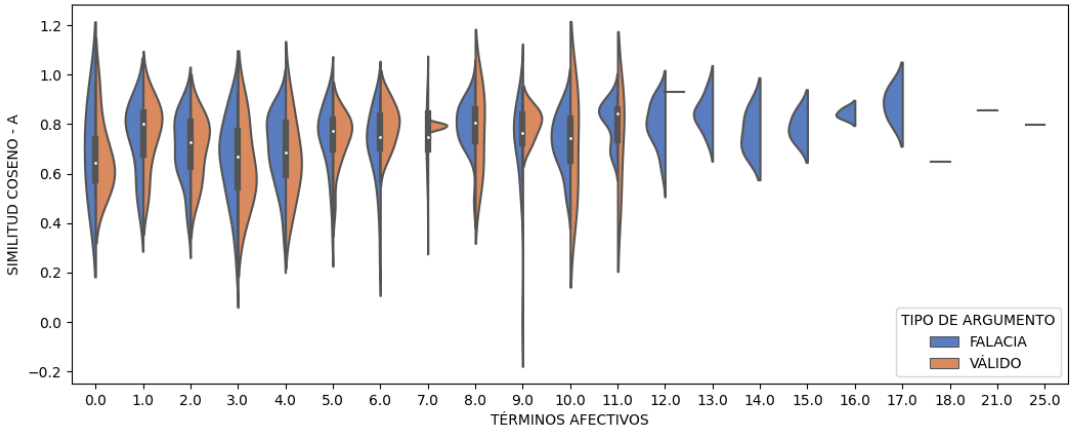

Figura 12: Dispersión de los argumentos con respecto a los resultados obtenidos por SC–A.

Características	SVM	LR	DT
SC–P	**0.55**	**0.55**	0.50
SC–A	0.48	0.45	0.45
TA	0.46	0.45	0.42
SC–A+TA	0.55	0.57	0.55
SC–P+TA	**0.62**	0.59	0.55

Tabla 12: Resultado de la clasificación con 465 argumentos. Se utilizaron características individuales y la combinación de las mismas. Los resultados se obtuvieron con la métrica *F1-score*.

5. Conclusión

En este artículo se presentó un *corpus* de argumentos en español donde se realizó una anotación de falacias por apelación a las emociones, argumentos válidos y estructura argumentativa. El *corpus* contiene 601 argumentos obtenidos de discursos políticos y fue descrito en términos del número de palabras, estructura argumentativa, falacias, lenguaje expresivo y diversidad léxica de los argumentos. El acuerdo entre anotadores fue de 0.442 para la identificación de falacias y 0.692 en la estructura argumentativa, ambos con el índice *Kappa* de Cohen.

En la elaboración del *corpus* predominó la opinión subjetiva de los anotadores para diferenciar argumentaciones racionales del uso de emociones en los argumentos. Al analizar el *corpus* no se encontró una relación entre el número de palabras y el tipo de argumento. Los argumentos válidos contienen términos afectivos y las falacias pueden no presentarlos. Sin embargo, mientras más número de palabras y menos diversidad léxica los argumentos tienden a ser falaces.

Como una contribución adicional, se realizó una línea base para proporcionar un punto de comparación para futuras investigaciones en la identificación de falacias por apelación en las emociones. En estos experimentos se utilizaron dos diccionarios de léxicos afectivos, y los méto-

dos de similitud coseno, *Support vector machine*, *Logistic Regression* y *Decision Trees* para la clasificación de argumentos en falacias o argumentos válidos. Como resultado se obtuvo un *F-score* de 0.55 utilizando la similitud textual entre los componentes del argumento y 0.62 combinando la similitud con los términos afectivos utilizados en los argumentos. No se encontró una relación entre estas dos características para distinguir falacias de argumentos válidos, debido a que las falacias pueden tener buena similitud y no utilizar palabras que provoquen una emoción.

Finalmente, se considera que este recurso representa una contribución importante para la comunidad dedicada a la identificación de falacias y, en general, para el análisis del discurso. Además, este *corpus* permite abordar la problemática de identificación de falacias mediante sistemas computacionales. Y con la información adicional en los argumentos se pueden desarrollar sistemas para la identificación de componentes argumentativos en el área de Minería de argumentos.

Agradecimientos

Este trabajo fue apoyado parcialmente por el gobierno de México (beca CONACYT con número de proyecto 653661, SNI).

Referencias

Al-Hindawi, Fareed H. H., Alkhazaali Musaab A. & Al-Awadi Duaa. 2015. A pragmatic study of fallacy in David Cameron's political speeches. *Journal of Social Science Studies* 2(2). 214–239.

Blassnig, Sina, Büchel Florin, Ernst Nicole & Engesser Sven. 2019. Populism and informal fallacies: An analysis of right-wing populist rhetoric in election campaigns. *Argumentation* 33(1). 107–136. 10.1007/s10503-018-9461-2.

Brezina, Vaclav. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Cabrejas-Peñuelas, Ana B. 2015. Manipulation in Spanish and American pre-election political debates: The Rajoy–Rubalcaba vs. Obama–McCain debates. *Intercultural Pragmatics* 12(4). 515–546. 10.1515/ip-2015-0025.

Camargo López, Sonia Patricia. 2018. La construcción de la emoción en los discursos políticos de campaña. *Pragmalingüística* 26. 199–220.

Capaldi, Nicholas. 2011. *Como ganar una discusión: El arte de la argumentación*. Gedisa.

Álvarez Carmona, Miguel Ángel. 2014. *Detección de similitud en textos cortos considerando traslape, orden y relación semántica de palabras*: Instituto Nacional de Astrofísica, Óptica y Electrónica. Trabajo de Fin de Máster.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1). 37–46. 10.1177/001316446002000.

Copi, Irving M. & Carl Cohen. 2013. *Introducción a la lógica*. Limusa.

Damer, T. Edward. 2009. *Attacking faulty reasoning: A practical guide to fallacy-free arguments*. Wadsworth Pub Co.

Díaz Rangel, Ismael, Grigori Sidorov & Sergio Suárez-Guerra. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein* 29. 31–46. 10.7764/onomazein.29.5.

García Gorrostieta, Jesús Miguel & Aurelio López López. 2019. A corpus for argument analysis of academic writing: argumentative paragraph detection. *Intelligent and Fuzzy Systems* 36(5). 4565–4577. 10.3233/JIFS-179008.

Gilbert, Michael A. 2004. Emotion, argumentation and informal logic. *Informal Logic* 24(3). 245–264. 10.22329/il.v24i3.2147.

Gordillo Torres, Juan Jesús & Víctor Hugo Rodríguez Perera. 2009. Cálculo de la fiabilidad y concordancia entre codificadores de un sistema de categorías para el estudio del foro online en e-learning. *Revista de Investigación Educativa* 27(1). 89–103.

Govier, Trudy. 2013. *A practical study of argument*. Cengage Learning.

Hansen, Hans. 2019. *Fallacies*. Stanford Encyclopedia of Philosophy Archive. https://plato.stanford.edu/archives/fall2019/entries/fallacies/.

Jason, Gary. 1986. Are fallacies common? a look at two debates. *Informal Logic* 8(2). 81–92. 10.22329/il.v8i2.2685.

Landis, Richard & Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1). 159–174.

Lawrence, John & Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics* 45(4). 765–818. 10.1162/coli_a_00364.

Lippi, Marco & Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology* 16(2). 1–25. 10.1145/2850417.

Molina González, M. Dolores, Eugenio Martínez Cámara, María Teresa Martín Valdivia & José M. Perea Ortega. 2013. Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications* 40(18). 7250–7257. 10.1016/j.eswa.2013.06.076.

Morales Gutiérrez, Irma Mariana. 2016. Falacias en los discursos de los candidatos presidenciales en méxico (2012). *Revista latinoamericana de estudios del discurso* 12(2). 11–32.

Ojeda Cruz, Felipe J. 2016. *Un método para la identificación automática de argumentos en discursos políticos escritos en español*: Tecnológico Nacional de México campus CENIDET. Trabajo de Fin de Máster.

Tindale, Christopher W. 2007. *Fallacies and argument appraisal*. Cambridge University Press.

Tobolka Castro, Sonia Jarmila. 2007. Competencia de los hablantes en la identificación de falacias: una perspectiva pragma-dialéctica. *Onomázein* 15. 129–155. 10.7764/onomazein.15.05.

Van Eemeren, Frans H. & Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation* 1(3). 283–301. 10.1007/BF00136779.

Van Eemeren, Frans H. & Rob Grootendorst. 2002. *Argumentación, comunicación y falacias: una perspectiva pragma-dialéctica*. Ediciones universidad católica de chile.

Vega Reñón, Luis. 2013. *La fauna de las falacias*. Trotta.

Wang, Jiapeng & Yihong Dong. 2020. Measurement of text similarity: a survey. *Information* 11(9). 421. 10.3390/info11090421.

Wells, Gordon. 2018. *Minería de falacias en el discurso político*: Universidad de Barcelona. Trabajo de Fin de Máster.

Weston, Anthony. 2006. *Las claves de la argumentación*. Ariel.

Zurloni, Valentino & Luigi Anolli. 2013. Fallacies as argumentative devices in political debates. En *International Workshop on Political Speech*, 245–257. 10.1007/978-3-642-41545-6_18.

Artigos Técnicos

Proposta rebuda el de juliol 2022 i acceptada per a publicació el d'octubre 2022.

El corpus paral·lel del *Diari Oficial de la Generalitat de Catalunya*

The parallel corpus of the Official Journal of the Catalan Government

Antoni Oliver
Universitat Oberta de Catalunya (UOC)

Resum

En aquest article presentem el procés de compilació de la nova versió del corpus paral·lel català–castellà creat a partir dels textos del *Diari Oficial de la Generalitat de Catalunya* (DOGC). Es descriuen els processos de descàrrega, conversió a text, segmentació i alineació automàtica. Tots els programes que s'han desenvolupat per dur a terme aquests processos es distribueixen amb una llicència lliure i el corpus compilat es pot descarregar lliurement. A més, es descriu el procés d'entrenament i avaluació de dos motors de traducció automàtica neuronal català–castellà i castellà–català que s'ha dut a terme fent servir aquest corpus paral·lel.

Paraules clau

corpus paral·lel; traducció automàtica neuronal

Abstract

In this paper, the process of compilation of the new version of the Catalan–Spanish parallel corpus of the Official Journal of the Catalan Government (DOGC) is presented. The processes of downloading, conversion to text, segmentation and automatic alignment are described. All the programs that have been developed to perform these processes are distributed under a free license and the compiled corpus can be freely downloaded. Furthermore, the process of training and evaluation of two neural machine translation systems, Catalan–Spanish and Spanish–Catalan, using this corpus is presented.

Keywords

parallel corpus; neural machine translation

1. Introducció

El *Diari Oficial de la Generalitat de Catalunya*[1] (DOGC) és el mitjà de publicació oficial de les lleis, normes, acords, resolucions, edictes, notificacions i anuncis de l'Administració i del Govern de Catalunya. El *Diari* té els seus inicis en el *Butlletí de la Generalitat de Catalunya*, el número 1 del qual aparegué el 3 de maig de 1931. El DOGC apareixia originàriament en paper i des del 2007 es va substituir íntegrament per una versió electrònica d'accés lliure. De la web del DOGC es poden descarregar textos en format electrònic des del número 2456 (18/08/1997), tot i que en aquests primers anys alguns textos no estan encara disponibles en HTML i només es poden obtenir en format PDF. Els textos provinents dels documents del DOGC tenen una llicència lliure i es poden distribuir i processar sense cap limitació legal. El tipus de llicència és la CC0 de Creative Commons[2].

El DOGC té dues edicions separades en català i en castellà. Molts dels textos del DOGC apareixen publicats tant en català com en castellà. Les normes amb rang de llei s'hi publiquen també en occità, així com les disposicions i els actes que afecten exclusivament l'Aran. L'aranès, llengua pròpia de la Vall d'Aran, és una variant de l'occità gascó. L'aranès és llengua oficial, junt amb el català i el castellà, a tot el territori de Catalunya.

En aquest article presentem el procés de creació de la nova versió del corpus DOGC, que comprèn els textos del 18/08/1997 al 31/12/2021. Aquesta nova versió actualitza la versió existent del corpus DOGC (Oliver, 2017), que comprenia els textos fins al 31/12/2015, i que, per tant, afegeix cinc anys més de la publicació. El resultat és un corpus català–castellà de temàtica administrativa i legislativa de gran mida. La versió anterior tenia un total de 5.026.847 segments paral·lels únics i la versió que presentem en aquest article assoleix els 8.472.786 segments paral·lels únics.

L'any 2016 l'Institut d'Estudis Catalans (IEC) va publicar una nova ortografia.[3]

[1] https://dogc.gencat.cat/

[2] https://creativecommons.org/publicdomain/zero/1.0/deed.ca

[3] https://www.iec.cat/llengua/documents/ortografia_catalana_versio_digital.pdf

DOI: 10.21814/lm.14.2.380

Els canvis principals es poden resumir en 6 punts:[4] (1) es modifica l'ús del guionet en alguns casos; (2) es redueix a quinze paraules la llista de mots amb accent diacrític: bé, déu, és, mà, més, món, pèl, què, sé, sí, sòl, són, té, ús i vós; (3) es recull la tradició de l'àmbit lingüístic valencià de representar amb accent agut els mots que es pronuncien amb e tancada en aquest parlar (que és oberta o neutra en altres parlars): café, comprén, francés, admés, etc.; (4) se suprimeix la dièresi en alguns derivats amb el sufix -al: laical, trapezoidal, etc.; (5) algunes paraules passen a doblar la r: arrítmia, cefalorraquidi, corresponsable, erradicar, otorrinolaringòleg, etc.; (6) algunes paraules compostes passen a escriure's amb e inicial del segon formant quan va seguida de s + consonant: arterioesclerosi, cirroestrat, electroestàtic, termoestable, etc. El canvi d'ortografia afecta els textos del DOGC des de la seva entrada en vigor, però amb dates canviants, ja que hi havia un període de transició on totes dues normatives estaven en vigor. Com veurem en l'article, s'ha desenvolupat un algorisme que adapta automàticament els textos a la nova normativa. Així doncs, els textos catalans del corpus resultant estan adaptats a la nova normativa ortogràfica.

Un corpus d'aquestes característiques pot tenir diverses utilitats, com per exemple estudis de llenguatges d'especialitat, treballs terminològics o l'entrenament de sistemes de traducció automàtica. En aquest treball presentem el procés d'entrenament de dos sistemes de traducció automàtica neuronal, un de català–castellà i un altre de castellà–català, i l'avaluació d'aquests sistemes, així com la comparació amb un sistema de transferència sintàctica superficial i amb un altre de neuronal.

Aquest corpus, per àmbit temàtic, es pot comparar amb els corpus paral·lels de la Unió Europea, com el JRC-Acquis[5] o el DGT.[6] Aquests corpus, però, estan disponibles per a molts parells de llengües, incloses el castellà i el portuguès. No tenim constància de la disponibilitat de corpus similars per a altres llengües de la península Ibèrica.

[4] https://www.uoc.edu/portal/ca/servei-linguistic/criteris/ortografia/resum-novetats/index.html

[5] https://joint-research-centre.ec.europa.eu/language-technology-resources/jrc-acquis_en

[6] https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en

2. Compilació del corpus

Tots els programes i arxius necessaris per a descarregar, convertir a text, segmentar i alinear el corpus DOGC es distribueixen sota una llicència lliure (GNU-GPL v.3) i es poden descarregar del lloc GitHub del projecte.[7]

2.1. Descàrrega dels arxius HTML

Els diferents arxius HTML del DOGC es poden descarregar fent servir el programa `downloadDOGC.py`, que demana quatre paràmetres d'entrada: el número d'article inicial i el número d'article final a descarregar, el directori de sortida per als arxius HTML en català i el directori de sortida per als arxius HTML en castellà. Per saber com fer servir el programa es pot fer servir l'opció `-h`. Per poder fer servir el programa de descàrrega és necessari disposar del *chromedriver*[8] adequat per a la versió de Chrome instal·lat en el sistema.

Per exemple, per descarregar de l'article 90000 al 91000 i deixar-los als directoris `html-ca` i `html-es` podem escriure.

```
python3 dowloadDOGC.py
   --nummin 90000 --nummax 910000
   --outdirCAT ./html-ca --outdirSPA ./html-es
```

El programa introdueix unes pauses aleatòries entre descàrregues d'arxius d'entre 3 i 6 segons per a evitar que el servidor denegui el servei per descàrrega massiva. Per aquest motiu el procés de descàrrega pot durar molta estona. La velocitat mitjana que hem obtingut ha estat de 625 parells d'articles (article en català i el mateix article en castellà) per hora.

2.2. Classificació dels arxius per any de publicació

Aquest pas és opcional i s'ha d'executar únicament si estem interessats a tenir els arxius HTML organitzats per anys. Aquesta classificació es pot dur a terme amb els programes `classifyByYear-ca.py` (per als HTML en català) i `classifyByYear-es.py` (per als HTML en castellà). Els programes disposen de l'opció `-h` que mostra l'ajuda. Per classificar els HTML catalans que són al directori `html-ca` podem escriure:

```
python3 classifyByYear-ca.py --dirin html-ca
```

[7] https://github.com/aoliverg/corpusDOGC

[8] https://chromedriver.chromium.org/downloads

El programa crearà un directori `html-pre1995-ca` per a copiar tots els HTML amb data de publicació anterior a 1995 (si n'hi ha cap) i directoris `html-1996-ca`, `html-1197-ca...` per a cada any (a mesura que siguin necessaris). D'aquesta manera, una vegada acabada l'execució del programa disposarem de tots els HTML endreçats en directoris per any de publicació.

2.3. Conversió d'HTML a text

Una vegada tenim els arxius HTML descarregats, i opcionalment classificats per anys, cal convertir aquests arxius a text. Ens interessa molt convertir únicament el text de l'article, sense incloure tot el text addicional que pot presentar la plana web, com pot ser els menús superiors i els enllaços destacats de la part inferior. Per aquest motiu s'ha desenvolupat un algorisme ad-hoc que detecta quan comença i quan acaba l'article i únicament converteix aquesta secció a text. Aquesta conversió es pot dur a terme amb el programa `DOGC2textDIR.py`, que converteix a text tots els arxius HTML d'un determinat directori. El programa disposa de l'opció `-h` que mostra l'ajuda. Per convertir tots els arxius HTML del directori `html-2015-ca` a text i posar els arxius al directori `text-2015-ca`, podem escriure:

```
python3 DOCG2textDIR.py
     --dirin html-2015-ca
     --dirout text-2015-ca
```

2.4. Segmentació

Una vegada tenim els arxius convertits a text ens interessarà segmentar-los per a poder fer una posterior alineació automàtica segment a segment. Per a segmentar els arxius es pot fer servir el programa `text2segmentedtextDIR.py`, que segmenta tots els arxius de text d'un determinat directori i guarda els arxius segmentats en un altre directori. El programa disposa de l'opció `-h` que mostra l'ajuda. El programa fa servir el mòdul `srx_segmenter.py`,[9] que utilitza arxius de definició de regles de segmentació en el format estàndard SRX[10] (*Segmentation Rules eXchange*) (Miłkowski & Lipski, 2009).

Per segmentar tots els arxius de text del directori `text-2015-ca` i guardar els arxius segmentats al directori `seg-2015-ca`, fent servir l'arxiu SRX segment.srx per la llengua *Catalan*, podem escriure.

```
python3 txt2segmentedtextDIR.py
     -i text-2015-ca/ -o seg-2015-ca/
     -s segment.srx -l Catalan
```

2.5. Alineació automàtica

Per dur a terme l'alineació automàtica dels arxius de text segmentat hem fet servir Hunalign[11] (Varga et al., 2007). Per facilitar l'ús d'aquest programa hem desenvolupat una sèrie de programes auxiliars. El primer programa auxiliar és el `createAlignScript.py`, que permet la creació d'un script d'alineació. El programa disposa de l'opció `-h` que mostra l'ajuda del programa. Seguint els exemples, si volem crear l'*script* d'alineació dels arxius segmentats de l'any 2015, podríem escriure:

```
python3 createAlignScript.py
  --dirSL seg-2015-ca/ --dirTL seg-2015-es/
  --dirALI ali-2015-ca-es/
  --dictionary hunapertium-ca-es.dic
  --script align2015.sh
  --r1 ca.txt --r2 es.txt
```

Fixem-nos que podem indicar un diccionari d'alineació. Hem creat tota una sèrie de diccionaris d'alineació en format Hunalign a partir dels diccionaris de transferència del sistema de traducció automàtica Apertium[12] (Forcada et al., 2011). Aquests diccionaris es poden descarregar de Github.[13] Una vegada executem aquest programa obtenim un script que conté una línia per a cada parell d'arxius català–castellà, com en el següent exemple:

```
timeout 5m ./hunalign hunapertium-ca-es.dic
    -utf -realign -text
    "seg-2015-ca/700000-ca.txt"
    "seg-2015-es/700000-es.txt"
    > "ali-2015-ca-es/ali-700000-ca.txt"
```

Cada comanda comença definint un *timeout* de 5 minuts per evitar que el procés es detingui si en algun procés d'alineació es deté per algun error. Si donem permisos d'execució a aquest *script* i l'executem, s'alinearan automàticament tots els arxius indicats. Una vegada finalitza l'alineació obtenim tota una sèrie d'arxius que contenen el segment en català, el segment en castellà i un valor de confiança, que indica la seguretat de l'alineació. El programa `selectAlignments.py` permet seleccionar tots els segments dels arxius alineats d'un determinat directori que tingui un valor de confiança superior a l'indicat. El valor

[9] https://github.com/narusemotoki/srx_segmenter
[10] https://www.gala-global.org/srx-20-april-7-2008
[11] https://github.com/danielvarga/hunalign
[12] https://www.apertium.org/
[13] https://github.com/aoliverg/hunapertium

de confiança acostuma a estar comprès entre -0,3 i 10, i els valors superiors a 0 acostumen a indicar alineacions vàlides. Per exemple, per seleccionar totes les alineacions amb confiança superior a 0 podem escriure:

```
python3 selectAlignments.py
  --inDir ali-2015-ca-es/
  --outFile alineacio-2015-cat-spa.txt
  --confidence 0
```

Una vegada seleccionades les alineacions s'eliminen els segments repetits amb instruccions estàndard de Unix[14]. Cal tenir en compte que l'eliminació de segments repetits suposa una pèrdua de l'ordre d'aparició dels segments.

2.6. Conversió a la nova ortografia catalana

Per assegurar-nos que tot el corpus està en la nova normativa ortogràfica catalana de l'IEC hem adaptat el programa `MTUOC-novaIEC`,[15] perquè funcioni amb un corpus paral·lel en format de text tabulat on el català està en el primer camp. El programa per dur a terme aquesta conversió és el `modificaIEC-PC1.py`, que disposa de l'opció -h per mostrar l'ajuda. En l'exemple següent adaptem l'alineació de l'any 2015 a la nova normativa ortogràfica catalana.

```
python3 modificaIEC-PC1.py
  --infile alineacio-2015-cat-spa.txt
  --outfile DOGC-2015-cat-spa.txt
```

2.7. Neteja del corpus

Un cop convertit el corpus a la nova ortografia catalana es du a terme un procés de neteja fent servir el programa `MTUOC-clean-parallel-corpus`[16]. La neteja ha consistit en les següents accions: (1) normalització de l'apòstrof; (2) eliminació d'etiquetes HTML/XML; (3) conversió de les entitats HTML, si n'hi ha, en el seu caràcter corresponent; (4) correcció dels errors en la codificació de caràcters, si n'hi ha; (5) eliminació dels segments buits; (6) eliminació dels segments curts, de menys de 10 caràcters; (7) eliminació dels segments que continguin el 60% o més de caràcters numèrics; (8) eliminació dels segments que siguin iguals en les dues llengües; (9) verificació automàtica de les llengües. Un parell de segments català–castellà s'elimina si un dels dos components compleix la condició d'eliminació.

[14] `cat`, `sort`, `uniq`, `shuf`

[15] https://github.com/aoliverg/MTUOC-novaIEC

[16] https://github.com/aoliverg/MTUOC-clean-parallel-corpus

2.8. Corpus resultant

Hem dut a terme el procés de compilació del corpus DOGC fins als articles corresponents a 31 de desembre de 2021. Hem fet una classificació per anys, i a la taula 1 podem veure el nombre d'articles, segments i segments sense repeticions per a cada any (els anys anteriors al 2.000 estan agrupats, ja que el nombre d'articles disponibles electrònicament és baix per als anys anteriors). El corpus global sense repeticions es pot descarregar lliurement[17] i està publicat sota la mateixa llicència que el DOGC (CC0 de Creative Commons[18]). Aquest corpus també està disponible a la col·lecció Opus Corpus[19] (Tiedemann, 2012). A la taula 2 es poden observar les dades de mida total en segments del corpus resultant.

Any	Articles	Segments	Seg. únics
< 2000	34.766	1.818.109	992.442
2000	23.475	1.352.777	635.971
2001	27.145	1.828.328	810.281
2002	29.281	1.923.290	893.101
2003	28.587	2.093.726	979.864
2004	28.805	1.985.084	957.778
2005	32.294	1.919.271	873.322
2006	38.115	2.049.614	932.173
2007	34.926	2.244.405	1.008.292
2008	33.177	2.276.233	1.027.718
2009	32.328	2.167.934	986.409
2010	31.852	2.046.071	941.081
2011	25.475	1.135.272	562.417
2012	23.239	838.820	468.802
2013	23.018	1.220.403	606.889
2014	22.981	1.275.181	615.327
2015	22.514	2.041.743	855.160
2016	22.810	3.391.425	1.202.426
2017	24.445	3.367.881	1.249.182
2018	23.716	2.723.208	1.083.260
2019	22.583	3.178.228	1.471.438
2020	19.212	2.521.235	1.194.678
2021	21.185	1.767.391	773.634

Taula 1: Nombre d'articles, segments i segments únics per a cada any.

El corpus DOGC pot tenir diverses utilitats. Per una banda, pot ser una bona base per a l'estudi del llenguatge juridicoadministratiu en català i castellà. També pot resultar interessant per a estudis terminològics i per a la confecció de glossaris terminològics català–castellà mit-

[17] http://lpg.uoc.edu/corpusDOGC/DOGC-2021-cat-spa.zip

[18] https://creativecommons.org/publicdomain/zero/1.0/deed.ca

[19] https://opus.nlpl.eu/DOGC.php

	Segments
Total	47.165.629
Únic brut	16.787.380
Únic net	8.472.786

Taula 2: Nombre de segments totals, totals únics i totals únics una vegada fet el procés de neteja.

jançant tècniques d'extracció automàtica de terminologia i de cerca automàtica d'equivalents de traducció. També pot resultar un corpus molt adequat per a entrenar sistemes de traducció automàtica per a textos juridicoadministratius. En la pròxima secció presentem l'entrenament i avaluació de sistemes de traducció automàtica neuronal català–castellà i castellà–català fent servir el corpus DOGC.

3. Entrenament de sistemes de traducció automàtica neuronal

En aquest apartant es descriu el procés d'entrenament de dos sistemes de traducció automàtica neuronal amb el corpus DOGC. S'ha entrenat un sistema català–castellà i un altre castellà-català. Tots els passos de preprocessament i entrenament s'han dut a terme fent servir els programes i *scripts* del projecte MTUOC[20] (Oliver, 2020).

3.1. Preprocessament del corpus

El corpus s'ha preprocessat fent servir l'*script* `MTUOC-corpus-preprocessing`.[21] La segmentació en unitats lèxiques i el càlcul de les unitats subparaules s'han dut a terme amb SentencePiece[22] (Kudo & Richardson, 2018), amb les següents característiques:

- Joining languages: True
- Model type: bpe
- Vocabulary size: 64000
- Vocabulary threshold: 50

El primer paràmetre (*joining languages*) significa que el càlcul de les unitats subparaules es fa a la vegada per a les dues llengües. Aquesta és la pràctica habitual i recomanada per a llengües que, com en el nostre cas, comparteixen alfabet. La resta de paràmetres s'han fixat en valors habituals en aquest tipus d'entrenaments.

Aquest pas de segmentació en unitats lèxiques i l'ús d'unitats subparaules és necessari en els sistemes de traducció automàtica neuronal, ja que cada una d'aquestes unitats es codificarà com a un vector. Les unitats subparaules, és a dir, unitats més petites que una paraula, es fan servir per evitar el problema de la limitació d'elements en el vocabulari del sistema, ja que les paraules poc freqüents es representen mitjançant fragments de paraules, que són més freqüents i tenen una representació vectorial pròpia. El corpus s'ha dividit en un fragment de 5.000 segments de validació, 5.000 segments per avaluació i la resta per a l'entrenament.

3.2. Entrenament

L'entrenament s'ha dut a terme fent servir `marian-nmt`[23] (Junczys-Dowmunt et al., 2018), amb una configuració de tipus *transformer*. S'han fet servir dues mètriques de validació: *bleu-detok* (el valor de BLEU calculat un cop desfeta la divisió en unitats lèxiques i subparaules) i l'entropia creuada (*cross-entropy*). El criteri de parada (*early-stopping*) s'ha fixat en 5 per les dues mesures, és a dir, es deté l'entrenament quan les dues mesures no milloren en 5 o més validacions. La freqüència de validació s'ha fixat en 5.000. L'entrenament català–castellà ha necessitat 4 èpoques i un total de 46 validacions. En canvi, l'entrenament castellà-català ha necessitat 5 èpoques i un total de 64 validacions. L'entrenament s'ha executat en un ordinador amb una unitat GPU Nvidia Quadro P6000 de 24 GB. L'entrenament català–castellà ha durat 2 dies, 2 hores, 40 minuts i 29 segons; i l'entrenament castellà–català 3 dies, 1 hora, 2 minuts i 3 segons.

3.3. Avaluació

Per a avaluar els motors entrenats s'ha fet servir el programa `MTUOC-eval`[24], que pot proporcionar les següents mètriques automàtiques:

- BLEU (Papineni et al., 2002)
- NIST (Doddington, 2002)
- WER (*Word Error Rate*) (Nießen et al., 2000)
- Distància d'edició percentual (DE%)
- TER (*Translation Error Rate* o taxa d'error per mot) (Klakow & Peters, 2002)

[20] https://github.com/aoliverg/MTUOC-server
[21] https://github.com/aoliverg/MTUOC-corpus-preprocessing
[22] https://github.com/google/sentencepiece
[23] https://marian-nmt.github.io/
[24] https://github.com/aoliverg/MTUOC-eval

Com més altes siguin les mesures BLEU i NIST indiquen millor qualitat; mentre que les mesures WER, distància d'edició percentual i TER, com més baixes siguin, millor. Per a aplicar aquestes mesures s'han fet servir 1.000 segments dels 5.000 segments totals del corpus d'avaluació, assegurant que aquests mateixos segments no apareguin en els corpus d'entrenament ni de validació. Aquests segments d'avaluació s'han traduït amb els motors entrenats, amb Apertium[25] (Forcada et al., 2011) i amb Google Translate[26].

A la taula 3 es poden observar els valors d'avaluació dels diferents sistemes per a les dues direccions. En tots dos casos, els valors són millors per al nostre sistema entrenat amb Marian, amb unes diferències prou importants. Per exemple, en la direcció cat–spa el nostre sistema Marian supera a Apertium en 4,9 punts de BLEU i a Google Translate en 2,3 punts de BLEU. En la direcció spa–cat les diferències són del mateix ordre, superant el nostre sistema Marian a Apertium en 3,8 punts de BLEU i a Google Translate en 6,3 punts de BLEU.

3.4. Sistemes resultants

Els sistemes de traducció automàtica resultants es poden descarregar dels enllaços que s'indiquen al GitHub del projecte.[27] Els sistemes estan configurats mitjançant un servidor MTUOC-server[28] (Oliver, 2020). Funcionen en un entorn GNU/Linux (o un Windows Subsystem for GNU/Linux). Es distribueix també amb un *marian-server* compilat per a CPU, però pot funcionar també sota GPU si proporcioneu un *marian-server* compilat per a la vostra GPU.

Editant l'arxiu `config-server.yaml` es poden indicar els ports a fer servir i també quin tipus de servidor posar en marxa: MTUOC, Moses, OpenNMT, NMTWizard o ModernMT. MTUOC-server pot emular tots aquests protocols de manera que és compatible amb diversos programes clients (com eines de traducció assistida per ordinador, per exemple).

4. Conclusions i treball futur

En aquest article hem presentat el procés de compilació de la nova versió del corpus del *Diari Oficial de la Generalitat de Catalunya* (DOGC). Tant el corpus, com els algorismes i programes que s'han fet servir per a la descàrrega, conversió a text, segmentació, alineació i adaptació a la nova normativa ortogràfica catalana estan disponibles per a la descàrrega i ús.[29] Aquests programes poden servir d'exemple per a la compilació d'altres corpus a partir de llocs web.

En aquest article també s'ha presentat un cas d'ús del corpus DOGC: l'entrenament de dos sistemes de traducció automàtica neuronal. Els sistemes entrenats s'han avaluat automàticament i comparat amb dos sistemes molt emprats. Totes les mètriques automàtiques analitzades indiquen que la qualitat dels sistemes entrenats, per a la traducció de textos juridicoadministratius, és superior a la de dos sistemes molt emprats per a aquest parell de llengües.

Com a treball futur volem detectar els textos publicats en aranès al DOGC i alinear-los amb el català i el castellà. Tot i que es preveu que el nombre de textos disponibles en aranès sigui petit, es podran fer noves descàrregues i detteccions futures per anar ampliant aquests corpus. També pretenem fer la compilació d'altres corpus similars per al català i el castellà, concretament un corpus a partir del *Butlletí Oficial de les Illes Balears* (BOIB)[30] i del *Diari Oficial de la Generalitat Valenciana* (DOGV)[31]. En el cas del BOIB només seran necessari uns canvis mínims en els programes de descàrrega per indicar les adreces de descàrrega corresponents. En canvi, el DOGV difereix molt en la manera de mostrar els articles al navegador, per la qual cosa serà necessari una modificació més profunda, que encara no hem estudiat.

També tenim la intenció d'estudiar la viabilitat de l'ús del sistema neuronal castellà-català entrenat amb el DOGC per a crear corpus sintètics (Park et al., 2017) entre les llengües oficials de les Nacions Unides (anglès, francès, rus, àrab i xinès, a més de l'espanyol) i el català, traduint automàticament la part espanyola dels corpus paral·lels entre aquestes llengües i l'espanyol. Crearíem també aquests corpus sintètics fent servir Apertium. Amb els corpus resultants entrenaríem sistemes de traducció automàtica i els avaluaríem amb mètriques automàtiques. D'aquesta manera podrem determinar si l'ús del sistema neuronal per a crear els corpus sintètics aporta millores respecte a l'ús d'un sistema basat en regles per a l'entrenament de nous sistemes de traducció automàtica neuronal. Si l'experiència fos satisfactòria s'ampliaria l'estudi a la creació de corpus sintètics entre les llengües oficials de

[25] https://www.apertium.org/

[26] La traducció s'ha fet mitjançant l'API de Google en data 18/07/2022)

[27] https://github.com/aoliverg/corpusDOGC

[28] https://github.com/aoliverg/MTUOC-server

[29] https://github.com/aoliverg/corpusDOGC

[30] http://www.caib.es/eboibfront/

[31] https://dogv.gva.es/va/

	BLEU	NIST	WER	DE(%)	TER
cat–spa					
Aquest treball	87,3	13,396	0,078	5,040	0,076
Apertium	82,4	12,866	0,114	5,923	0,100
Google T.	85,0	13,216	0,095	5,158	0,081
spa–cat					
Aquest treball	88,7	13,51	0.083	4,950	0,068
Apertium	84,9	13,117	0,102	5,435	0,087
Google T.	82,4	13,038	0,115	5,701	0,099

Taula 3: Valors de les mètriques d'avaluació dels sistemes analitzats.

la Unió Europea i el català i entrenar també sistemes per a aquests parells de llengües. Actualment, només existeix un sistema comercial d'àmbit general que inclogui aquestes llengües, Google Translate.

Referències

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. En *2nd International Conference on Human Language Technology Research (HLT)*, 138–145.

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. doi 10.1007/s10590-011-9090-0.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F.T. Martins & Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. En *56th Annual Meeting of the Association for Computational Linguistics (ACL: System Demonstrations)*, 116–121. doi 10.18653/v1/P18-4020.

Klakow, Dietrich & Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication* 38(1-2). 19–28. doi 10.1016/s0167-6393(01)00041-3.

Kudo, Taku & John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. En *2018 Conference on Empirical Methods in Natural Language Processing (EMNLP: System Demonstrations)*, 66–71. doi 10.18653/v1/D18-2012.

Miłkowski, Marcin & Jarosław Lipski. 2009. Using SRX standard for sentence segmentation. En *Language and Technology Conference (LTC)*, 172–182. doi 10.1007/978-3-642-20095-3_16.

Nießen, Sonja, Franz Josef Och, Gregor Leusch & Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. En *2nd International Conference on Language Resources and Evaluation (LREC)*, .

Oliver, Antoni. 2017. El corpus paral·lel del Diari Oficial de la Generalitat de Catalunya: compilació, anàlisi i exemples d'ús. *Zeitschrift für Katalanistik* 30. 269–291.

Oliver, Antoni. 2020. MTUOC: easy and free integration of NMT systems in professional translation environments. En *22nd Annual Conference of the European Association for Machine Translation*, 467–468.

Papineni, Kishore, Salim Roukos, Todd Ward & Wj Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. En *40th annual meeting of the Association for Computational Linguistics (ACL)*, 311–318. doi 10.3115/1073083.1073135.

Park, Jaehong, Jongyoon Song & Sungroh Yoon. 2017. Building a neural machine translation system using only synthetic parallel data. *ArXiv* abs/1704.00253.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. En *8th International Conference on Language Resources and Evaluation (LREC)*, 2214–2218.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh & Viktor Trón. 2007. Parallel corpora for medium density languages. En *Recent Advances in Natural Language Processing IV*, 247–258. John Benjamins. doi 10.1075/cilt.292.32var.

http://www.linguamatica.com/

linguaMATICA