Proposta recebida em Outubro 2017 e aceite para publicação em Dezembro 2017.

Processamento automático de expressões idiomáticas do português europeu

Enhancing Parsing of European Portuguese Verbal Idioms

David Antunes
HLT, INESC ID Lisboa
IST, Univ. Lisboa

Jorge Baptista
HLT, INESC ID Lisboa
FCHS, Univ. Algarve

Nuno Mamede
HLT, INESC ID Lisboa
IST, Univ. Lisboa

Resumo

Expressões idiomáticas verbais são expressões multipalavra em que o verbo principal é distribucionalmente fixo com um ou mais dos seus argumentos. O significado global destas expressões é, geralmente, não composicional, isto é, não pode ser regularmente inferido a partir do significado individual dos seus constituintes, quando usados separadamente. O principal objetivo deste trabalho é a construção de um sistema capaz de processar expressões idiomáticas do português europeu, integrado de forma fluida numa cadeia (pipeline) de processamento de linguagem natural. Para tal, foram desenvolvidas duas componentes fundamentais: (i) a criação de um corpus anotado com instâncias de expressões idiomáticas verbais do português europeu, e (ii) o desenvolvimento de um sistema que gera regras de análise de dependência para identificar automaticamente expressões idiomáticas com base nas restrições linguísticas representadas numa matriz léxico-gramatical. O sistema foi avaliado com frases produzidas manualmente, frases geradas artificialmente (por um módulo específico do sistema) e usando documentos selecionados a partir de dois *corpora* e onde foram anotadas instâncias de expressões idiomáticas. Os resultados globais indicam que a Precisão do sistema é bastante satisfatória, enquanto a Abrangência (Recall) é menos favorável. Estes dados destacam a necessidade de direcionar esforços para melhorar o desempenho do sistema, nomeadamente das regras que permitem identificar automaticamente as expressões fixas em textos. Uma parte do *corpus* utilizado bem como das anotações de expressões idiomáticas são disponibilizados para a comunidade científica.

Palavras chave

expressões multipalavra; português europeu; expressões idiomáticas; expressões fixas; léxico-gramática

Abstract

Verbal idioms (or verbal idiomatic expressions) are multiword expressions in which the main verb is distributionally frozen with one or more of its arguments (subject or complements). For the most part, they convey a non-compositional meaning that cannot be inferred from the individual meanings of their constituents when used separately. The primary goal of this project is the creation of a system capable of processing verbal idioms from European Portuguese, seamlessly integrated into a natural language processing pipeline. To this end, two key components were developed: (i) the creation of a corpus annotated for instances of verbal idiomatic expressions in European Portuguese, and (ii) the development of a system that generates dependency parsing rules to identify proper instances of specific verbal idioms based on the linguistic restrictions outlined in a lexicon-grammar matrix. The system is evaluated using manually produced sentences, artificially generated ones, and real sentences from two selected *corpora* annotated for verbal idioms. The overall results indicate that the Precision of the system is very satisfactory; whilst the Recall is less favorable, highlighting the need for further efforts to enhance the automatic identification of verbal idioms in texts. A portion of the annotated *corpora* is made available to the scientific community.

Keywords

multiword expressions; European Portuguese; verbal idioms; idiomatic expressions; lexicon-grammar

1. Introdução

Expressões idiomáticas (ou expressões fixas de núcleo verbal) podem ser definidas como um tipo especial de *expressões multipalavra* (EMP), nas quais o verbo principal e um ou mais de seus argumentos são distribucionalmente fixos entre si, ou seja, apresentam restrições combinatórias distribucionalmente imprevisíveis (Gross, 1982; Baptista et al., 2004; Constant et al., 2017). Haverá outros tipos de expressões idiomáticas, mas este artigo foca-se nas construções idiomáticas de

DOI: 10.21814/lm.17.1.464

núcleo verbal (em ing., *verbal idioms*), como, por exemplo, *O Pedro* **bateu** *a* **bota** 'O Pedro morreu', *A Joana* **não bate bem da bola** 'A Joana é maluca', etc. O significado geral dessas expressões muitas vezes não pode ser derivado do significado que cada elemento apresenta quando usado separadamente; em outras palavras, o significado dessas construções é *não-composicional*. Por exemplo, em *A Ana* **perdeu** *a* **cabeça** 'A Ana descontrolou-se', o sentido da expressão não resulta do sentido das palavras 'perder' e 'cabeça', quando usadas de forma independente.

Considerando também que a frequência de expressões idiomáticas em textos excritos é habitualmente muito baixa, fica claro que o processamento dessas expressões é uma tarefa desafiante por vários motivos: a imprevisibilidade das restrições distribuicionais; a possível flexão dos complementos fixos, principalmente do verbo; a estrutura sintática que apresentam, muitas vezes permitindo inserções, permutações de constituintes e outras transformações sintáticas; e o significado não-composicional destas expressões.

Embora se possa pensar que a frequência de EMP em textos, seja na oralidade, seja na escrita, é baixa o suficiente para que se possa ignorar a sua peculiaridade, o número estimado dessas expressões no léxico de um falante é considerável. As estimativas variam: desde serem da mesma ordem de grandeza que o número de verbos simples (Jackendoff, 1997), até várias vezes o número de verbos simples com construções distribuicionalmente livres —por exemplo, Gross (1996) apresenta um léxico-gramática do francês com 20.340 expressões fixas, em contraste com 13.225 verbos simples, distribuicionalmente livres. Nesta perspetiva teórica, o léxico e a gramática de uma língua natural estão indissociavelmente interligados, sendo muitas propriedades sintático-semânticas e transformacionais das expressões linguísticas determinadas pelos elementos lexicais, e neste caso, pelas combinatórias desses elementos. Daí o termo *léxico-gramática*, que aqui também utilizamos. Fica, portanto, claro que não se pode negligenciar essas expressões idiomáticas no processamento das línguas naturais, pois elas contêm informações essenciais para a compreensão dos textos.

Um grande volume de trabalho tem sido realizado para integrar a análise de expressões idiomáticas em sistemas de Processamento de Linguagem Natural (PLN). O projeto PARSEME (Savary et al., 2017)[1], uma iniciativa de uma rede de pesquisa europeia focada no papel das EMP na análise sintática, produziu resultados interessantes, como um *corpus* multilíngue de 5 milhões de palavras, que inclui o português brasileiro, anotado com EMP de vários tipos. Em particular, a segunda edição desta tarefa (Ramisch et al., 2018b) indica que cerca de 20% das EMP anotadas correspondiam a expressões idiomáticas do tipo aqui tratado. Este *corpus* foi, inclusive, utilizado (Baptista et al., 2022) no âmbito de um estudo que envolveu a utilização do sistema que aqui pretendemos desenvolver, a STRING (Mamede et al., 2012)[2].

A STRING é numa cadeia de processamento (uma *pipeline*) modular de PLN desenvolvida especificamente para o português europeu, que executa todas as tarefas básicas de processamento, incluindo segmentação de texto (em frases e tokens), marcação da classe gramatical e traços morfossintáticos e extração de dependências sintáticas, entre outras operações. O módulo responsável pela análise sintática é o *XIP* (Xerox Incremental Parser; Aït-Mokhtar et al. (2002)). Este analisador utiliza uma gramática baseada em regras do português para segmentar frases em constituintes sintáticos elementares (ing. *chunks*, e.g., sintagmas nominais, *NP*) e extrair relações de dependência sintática entre os núcleos desses *chunks* (e.g., a relação de *sujeito* entre um verbo e a cabeça de um NP). O XIP inclui um módulo denominado Xipificator que permite a extração de dependências sintáticas que representam expressões idiomáticas (dependência `FIXED`[3]). Este sistema tem vindo a ser desenvolvido ao longo dos anos (Baptista et al., 2014, 2015; Galvão et al., 2019) e é o foco deste trabalho, cujas principais contribuições face a versões anteriores do sistema são: (i) o aumento da automação na geração de regras e exemplos de expressões idiomáticas; (ii) a definição de 3 modos diferentes de geração de regras; (iii) a criação de ferramentas de suporte ao refinamento do sistema, que auxiliam na manutenção e expansão do léxico-gramática; (iv) a introdução de procedimento de avaliação com textos reais, propiciando a expansão do léxico-gramática; e (v) a criação de um corpus anotado com expressões idiomáticas verbais do português europeu, cuja versão parcial é disponibilizada à comunidade científica como

[1] https://typo.uni-konstanz.de/parseme/ (último acesso: 2025-07-24; todos os URL neste artigo foram verificados nesta data).

[2] https://string.hlt.inesc-id.pt/

[3] A dependência `FIXED` usada pelo nosso sistema não corresponde à dependência homónima do modelo das *Universal Dependencies* (UD). O conceito de 'fixed' das UD abrange muitas outras construções que não só as expressões idiomáticas (como se pode consultar em https://universaldependencies.org/treebanks/pt_pud/pt_pud-dep-fixed.html).

recurso linguístico[4]. A arquitetura e funcionamento deste sistema são descritos em pormenor na Secção 3.

2. Estado da Arte

A estratégia prevalecente em sistemas de PLN que visam processar expressões idiomáticas (e expressões multipalavra em geral), na sua complexidade, envolve a utilização de um léxico ou dicionário, que contém as várias expressões idiomáticas de uma dada língua (Savary et al., 2019b). Subsequentemente, este recurso: ou é diretamente incorporado num sistema de PLN, que aproveita a riqueza desta informação lexical para conseguir identificar em textos instâncias das expressões idiomáticas conhecidas; ou é utilizado para a extração de exemplos de expressões idiomáticas que são depois utilizados para treinar modelos de aprendizagem automática.

2.1. Recursos Lexicais

Geralmente, o desenvolvimento do léxico é uma tarefa manual que envolve a compilação de expressões relevantes, acompanhadas da definição das restrições essenciais necessárias para as identificar de forma precisa nos textos, preservando o significado convencionalizado dessas expressões. Infelizmente, a área de PLN tem testemunhado uma exploração limitada deste campo (Sag et al., 2002; Savary et al., 2019b), sendo este o principal impedimento para o avanço dos sistemas no tratamento de expressões idiomáticas.

Podemos encontrar léxicos abrangentes de expressões idiomáticas para várias línguas, tais como o árabe (Kourtin et al., 2021), o espanhol (Mogorrón Huerta, 2020), o grego moderno (Fotopoulou, 1993), o italiano (Vietri, 2014), o russo (Fukova, 2016), e o português, nomeadamente, a variedade do português brasileiro, como exemplificado em Vale (2001).[5]

Os trabalhos de Baptista (2004), Baptista et al. (2015) e de Galvão et al. (2019) representam avanços significativos na descrição linguística das expressões fixas do português europeu. Deles resultou a construção de um léxico-gramática das expressões idiomáticas nesta variedade do português, o qual já foi integrado no sistema de PLN STRING (Baptista et al., 2014).

Os trabalhos acima referidos apresentam semelhanças com as descrições linguísticas já feitas para as expressões fixas em várias outras línguas românicas ou europeias, concretamente das propriedades das expressões idiomáticas de núcleo verbal. No entanto, embora as duas maiores variedades da língua sejam bastantes semelhantes, um estudo Baptista (2008) mostrou que o português europeu e o português brasileiro partilham apenas uma quantidade muito reduzida de expressões idiomáticas (cerca de 10%) e, ainda assim, com muitas diferenças formais.

2.2. Identificação Automática de Expressões Idiomáticas

Há numerosos trabalhos que visam a identificação automática de expressões idiomáticas em textos, alguns dos quais seram apresentados nesta Secção. Um trabalho ilustrativo pode ser visto em de Uzeda Garrão & Dias (2001), que tentou aprimorar a tradução entre o português brasileiro e o inglês, com foco nas construções idiomáticas da forma *verbo + sintagma nominal*, sem recorrer a dados estatísticos do mundo real. Para alcançar esse objetivo, um conjunto de expressões idiomáticas que apresentam o verbo 'bater' foi introduzido na base de dados (léxico) do sistema, juntamente com seus significados correspondentes em inglês. As traduções assim produzidas teriam permitido manter o significado idiomático destas expressões, evitando traduções literais e inadequadas. O estudo não divulgou, porém, quaisquer resultados quantitativos.

Um projeto semelhante (Salton et al., 2014) recorre a 3 recursos lexicais para traduzir expressões idiomáticas: um dicionário de expressões na língua fonte; um dicionário de expressões na língua destino; e um dicionário bílingue com a correspondência entre expressões idiomáticas das duas línguas. O sistema foi testado com a tradução entre inglês e português brasileiro. Para este efeito foram construídos dois *corpora* que em conjunto apresentavam 10 exemplos reais para cada uma das 28 construções diferentes consideradas (todas elas do tipo *verbo + sintagma nominal*).

No que diz respeito à aprendizagem profunda (do inglês *Deep Learning*) podemos encontrar trabalhos como a tarefa partilhada *Multilingual Idiomaticity Detection and Sentence Embedding* (Madabushi et al., 2022). Este esforço colaborativo contou com um total de 25 equipas, que desenvolveram os seus modelos de PLN para reali-

[4] A partição do corpus correspondente aos textos do *CETEMPúblico* está disponível em `https://portulanclarin.net/repository/search/?q=VIDiom-PT`. Porém, devido a restrições de licença, não nos é possível disponibilizar os documentos relativos ao Parlamento Português.

[5] Note-se, porém, que este trabalho, até onde sabemos, ainda não foi incorporado em nenhum sistema de PLN existente.

zar 2 tarefas independentes, usando um conjunto de dados (*dataset*) que apresentava frases em português, galego e inglês, juntamente com anotação humana do grau de composicionalidade das expressões. A primeira tarefa consistia no desenvolvimento de modelos de aprendizagem supervisionada capazes de distinguir instâncias não-idiomáticas de instâncias idiomáticas de EMP. Os resultados foram bastante satisfatórios, com a equipa mais bem-sucedida a alcançar uma F1-score de 0,9385 no total das 3 línguas e 0,8944 para o português especificamente. A segunda tarefa era focada no cálculo do nível de similaridade entre frases com uma EMP e frases com o significado respetivo escrito explicitamente. O desempenho foi avaliado com o coeficiente de correlação de *Spearman*, sendo o maior valor de correlação alcançado 0,6648, a que corresponde um grau de correlação "forte". Embora os resultados globais deste trabalho sejam positivos, é importante destacar que as conclusões que se podem tirar deles são superficiais: na primeira tarefa, o objetivo era identificar se uma expressão é idiomática ou não (ou seja, uma classificação binária); para avaliar a segunda tarefa, os modelos são fornecidos com duas frases a serem comparadas, o que significa que está a ser negligenciado todo o processo de identificar qual é a expressão idiomática que está em causa e de determinar o seu significado.

Para fechar esta secção, Pasquer et al. (2020) apresentam um sistema capaz de identificar EMP verbais com base na combinação de filtros morfossintáticos, evitando técnicas de aprendizagem profunda ou aprendizagem automática complexa. Este sistema foca-se especialmente em expressões observadas na fase de treino para identificar novas instâncias das mesmas construções na fase de teste. O método começa com a extração de candidatos utilizando *multisets* (não ordenados e com repetições) dos lemas característicos das EMP verbais observadas. Segue-se a aplicação de um conjunto de filtros que consideram uma série de aspetos, entre os quais: a desambiguação dos componentes utilizando anotações morfossintáticas, a ordem dos componentes, a distância e relações sintáticas entre os componentes, etc. O desempenho foi avaliado com o dataset da edição 1.1 do *PARSEME* (Ramisch et al., 2018a), tendo sido obtida uma F-score a rondar os 0.8 para EMP verbais observadas. Apesar de ser desenvolvido para EMP verbais observadas, o sistema também apresentou um bom desempenho para construções não observadas, refletindo a alta proporção de expressões observadas face às não observadas. Este trabalho destaca-se pela sua intepretabilidade (pois é simples identificar os filtros responsáveis pelas falhas) e por permitir um desenvolvimento incremental e customizado (com a adição de novos filtros adequados). Finalmente, este trabalho salienta a importância da disponibilidade de corpora anotados com expressões idiomáticas para o desenvolvimento e a avaliação de sistemas de identificação automática destas construções em textos – o que é um dos objectivos deste trabalho.

Reconhecemos que uma comparação direta com sistemas existentes poderia ser feita com base num pequeno subconjunto de expressões em comum. No entanto, optamos por não a incluir por considerarmos que uma comparação abrangente entre abordagens tão distintas está fora do âmbito deste trabalho.

3. Arquitetura

Nesta Secção apresenta-se a arquitetura do sistema de processamento das expressões idiomáticas. O Xipificator tem como fonte de informação lexical uma matriz lexical, a que chamamos *léxico-gramática*, que está em desenvolvimento contínuo (Baptista, 2004; Baptista et al., 2015; Galvão et al., 2019) e que descreve as expressões idiomáticas verbais do português europeu. O sistema identifica as expressões ali representadas em textos, estabelecendo uma relação entre os elementos-chave das expressões através da extração da dependência `FIXED`, usando o módulo de análise sintática (*parser*) XIP Aït-Mokhtar et al. (2002).

3.1. XIP

O XIP realiza a análise sintática de unidades textuais através de um método incremental, que pode ser dividido em três etapas. Na primeira etapa, as regras de segmentação (ing. *chunking*) são aplicadas ao input e uma árvore de segmentação é gerada para cada frase, delimitando os constituintes sintáticos elementares (*chunks*), como grupos/sintagmas nominais (`NP`), adjetivais (`AP`), adverbiais (`ADVP`), preposicionais (`PP`), etc. Na etapa seguinte, as regras de extração de dependências sintáticas mais básicas entram em ação e extraem as relações mais simples entre os segmentos assim formados; por exemplo, a relação `DETD` entre a cabeça de um sintagma nominal, `NP`, e um determinante definido. Na última etapa é onde as dependências mais complexas são identificadas, aplicando as regras de dependência restantes às relações básicas extraídas na etapa anterior. Trata-se de dependências fundamentais como, por exemplo, as que ligam um verbo ao seu sujeito (`SUBJ`),

Class	N0 = Nhum	N0 = N-hum	Vse	NegObrig	V	Prep-link	Vc	ADV1	Prep1	Det1	Modif1-E	C1	Modif1-D	Prep2	Det2	Modif2-E	C2	Modif2-D
C1	+	-	-	-	<abanar>	-	-	-	-	o	<E:	capacete	<E>	-	-	-	-	-
C1PN	+	+	-	-	<abrir>	-	-	-	-	<D	<E:	<caminho>	<E>	a	-	-	-	-
C1PN	+	-	-	-	<abrir>	-	-	-	-	<D	<M	<porta>	<MOD>	a	-	-	-	-
CNP2	+	-	-	-	<apagar>	-	-	-	-	-	-	-	-	de	a	<E>	memória	<E>
C1	+	-	-	-	<apagar>	-	-	-	-	<D	<E:	<fogo>	<E>	-	-	-	-	-
C1	+	-	-	-	<arregalar>	-	-	-	-	os	<E:	olhos	<E>	-	-	-	-	-
C1PN	+	-	-	-	<arregalar>	-	-	-	-	os	<E:	olhos	<E>	a	-	-	-	-
C1	+	-	-	-	<bater>	-	-	-	-	a	<E:	<bota>	<E>	-	-	-	-	-
C1	+	-	-	-	<borrar>	-	-	-	-	a	<E:	pintura	<E>	-	-	-	-	-
CNP2	+	-	-	-	<chamar>	-	-	-	-	-	-	-	-	a	a	<E>	atenção	<E>
CAN	+	-	-	-	<chamar>	-	-	-	-	a	<E:	atenção	<E>	de	-	-	-	-
CNP2	+	-	-	-	<conhecer>	-	-	-	-	-	-	-	-	por	a	<E>	pinta	<E>
C1P2	+	-	-	+	<gastar>	-	-	-	-	<E:	<E:	cera	<E>	com	<DE	<MOD>	<defunto>	<MOD>
C1P2	+	-	-	+	<gastar>	-	-	-	-	<E:	<E:	cera	<E>	em	<E>	<E>	<defunto>	ruim
CADV	+	-	-	-	<ir>	-	-	ADV:embora	-	-	-	-	-	-	-	-	-	-
CADV	+	-	+	-	<ir>	-	-	ADV:embora	-	-	-	-	-	-	-	-	-	-
CVt	+	+	-	-	<ir>	<E:	dar	-	a	-	-	-	-	-	-	-	-	-
C1PN	+	-	-	-	<juntar>	-	-	-	-	os	<E:	<trapo>	<E>	com	-	-	-	-
CNP2	+	-	-	-	<levar>	-	-	-	-	-	-	-	-	a,"a	o	<E>	fim	<E>
C1	+	-	-	-	<perder>	-	-	-	-	a	<E:	cabeça	<E>	-	-	-	-	-

Figura 1: Aspeto geral da matriz léxico-sintática (léxico-gramática) das expressões idiomáticas do português europeu.

ao complemento direto (`CDIR`) ou aos diferentes tipos de modificador (`MOD`). É nesta etapa que são aplicadas um conjunto de regras de dependência especificamente criadas para identificar expressões idiomáticas: trata-se de dependências `FIXED`, em que se representam os constituintes essenciais da construção em questão e em que a classe da expressão (segundo a tipologia do léxico-gramática) aparece como traço (ing. *feature*) da dependência. Por exemplo, a expressão **abanar** *o* **capacete** 'dançar' (da classe *C1*) é representada pela dependência `FIXED_C1(abanar,capacete)`.

3.2. Formalização das Expressões Idiomáticas

O léxico-gramática das expressões idiomáticas do português europeu (em parte ilustrado na Figura 1) tem vindo a ser desenvolvido ao longo dos anos (Baptista, 2004; Baptista et al., 2004; Galvão et al., 2019). Trata-se de um processo contínuo e dinâmico, que envolve a adição de nova informação linguística, a identificação e classificação gradual de novas expressões fixas, bem como a revisão sistemática de aspetos linguísticos em função da observação de diversos fenómenos em *corpora*. A versão do léxico-gramática atualmente incorporado no sistema é a versão 9.26, datada de 13 de setembro de 2024.

Atualmente, o léxico-gramática consiste numa matriz que inclui um total de 2.574 expressões, descritas num ficheiro XLSX (uma expressão por linha), com as suas propriedades representadas nas colunas da matriz.

De um modo geral, há nesta matriz dois tipos de colunas: (i) colunas com valores *binários*, que podem ser preenchidas com '+' (positivo) ou '-' (negativo), indicando se uma restrição está presente ou não, respetivamente; estas colunas são tipicamente usadas para restrições relacionadas com a estrutura da frase ou para marcar transformações sintáticas aplicáveis à construção de base; e (ii) colunas *lexicais*, que geralmente contêm o *lema* ou a *forma* (ing. *surface*) de uma palavra (ou expressão), indicando a presença e restringindo o conteúdo de um elemento-chave de uma dada expressão idiomática. Ao todo, a matriz é constituída atualmente por 106 colunas, incluindo uma com um exemplo (elaborado manualmente) ilustrativo de cada uma das construções. Uma explicação pormenorizada das restrições representadas por cada coluna pode ser consultada em Antunes (2024).

Para tornar mais claro como as expressões são descritas pelas colunas da matriz, segue-se uma breve apresentação do conteúdo e significado das colunas relevantes da matriz para a expressão **juntar** *os* **trapos com** *alguém* 'casar'. Esta expressão faz parte da classe `C1PN` o que é indicado na coluna `Class`. O sujeito é distribucionalmente livre (coluna `C0` marcada com um '-') e pode ser definido pelo traço distribucional de nome humano (coluna `N0 = Nhum` marcada com um '+'). O verbo principal, que caracteriza esta expressão pode ser qualquer forma do verbo 'juntar' (coluna `V` preenchida com '<juntar>', onde os delimitadores '<>' indicam restrição sobre o lema da palavra). Este verbo não pode aparecer numa construção reflexa (coluna `Vse` marcada com valor negativo). A construção também não apresenta uma negação obrigatória (coluna `NegObrig` marcada com valor negativo). Quanto ao primeiro complemento (fixo): é um complemento direto, ou seja, não é introduzido por uma preposição (coluna `Prep1` marcada com um '-'); como a coluna `C1` está preenchida com '<trapo>', sabe-se que é um complemento fixo e que pode tomar qualquer forma flexionada associada ao lema do nome 'trapo' (e.g., 'trapinhos')[6]; O nome *trapos* tem de ser introduzido pelo determinante 'os' (coluna `Det1` preenchida com 'os'). Quanto ao segundo complemento: é introduzido pela preposição 'com' (coluna `Prep1` preenchida por 'com'); é um complemento livre (coluna `C2` marcada com um '-') e é distribucionalmente uma posição de um nome humano (coluna `N2 = Nhum` marcada com um '+'). Esta expressão não apresenta outros complementos essenciais (as colunas que descrevem um possível terceiro e quarto complementos estão marcadas com '-'). Em termos de transformações, a única que se aplica é a de *Simetria* (coluna `Sim2` marcada com valor positivo), caracterizada pela possibilidade de se trocar o sujeito e segundo complemento ou de os coordenar na posição de sujeito, mantendo o significado da expressão. Por fim, a última coluna a considerar é a coluna `Exemplo`, que contém uma frase produzida manualmente e onde se apresenta um exemplo típico da expressão em causa (neste caso o exemplo é *O Rui juntou os trapinhos com a Ana* 'O Rui casou com a Ana'.

[6]Contrastando com as 92 ocorrências da expressão com diminutivo (0,07 ocorrências por milhão de tokens), a expressão ocorre ainda assim 10 vezes sem sufixo diminutivo no corpus PtTenTen2023, acessível pelo Sketch Engine `https://app.sketchengine.eu/`: *O casal* juntara os trapos *há cerca de um ano*; ... *formaria um belo casal,* juntando os trapos*, num cerimonial excepcional*; decidem *juntar os trapos.*

3.3. Identificação Automática de Expressões Idiomáticas

O diagrama da Figura 2 mostra a arquitetura do sistema. Este está implementado para: (i) validar formalmente o conteúdo das entradas da matriz (detalhado na Secção 3.3.1); (ii) gerar regras de dependência XIP para cada expressão idiomática (detalhado na Secção 3.3.2); e, paralelamente, (iii) gerar exemplos dessas expressões (detalhado na Secção 3.3.4); tal é feito com base nas informações contidas na matriz de léxico-gramática; O sistema permite ainda (iv) incorporar as novas regras XIP na STRING e, em seguida, extrair as dependências para os exemplos gerados, (v) avaliando o desempenho do sistema ao comparar as dependências `FIXED` extraídas com o que é esperado para cada frase processada (detalhado na Secção 3.3.5); por fim, o sistema permite ainda (vi) analisar automaticamente as falhas do sistema e apontar possíveis causas para as mesmas (detalhado nas Secções 3.3.3 e 3.3.6). Os ficheiros sucessivamente processados pelo sistema foram numerados na Figura 2 para facilitar a referência nas Secções seguintes.

3.3.1. Validação da Matriz

O input do sistema é a matriz léxico-sintática, sob a forma de um ficheiro XLSX (Ficheiro 1), descrito na Secção 3.2. O desenvolvimento do léxico-gramática para expressões idiomáticas é uma tarefa complexa que é realizada manualmente, tornando-a suscetível à introdução de erros. Após a matriz ser convertida para um CSV (Ficheiro 2), entra em ação um validador automático que garante a adequação da informação lexical. Este validador faz três tipos de verificações diferentes: (i) verifica a consistência de cada valor com os valores possíveis na respetiva coluna; (ii) garante que as colunas preenchidas estão de acordo com as propriedades inerentes à classe da expressão; e (iii) verifica a consistência entre colunas cujos valores dependem dos valores de outras colunas. Na ausência de problemas, o ficheiro CSV está pronto para ser processado.

3.3.2. Geração de Regras

O Gerador de Regras faz uso da informação codificada no CSV para gerar regras de dependência XIP, que permitem a identificação precisa das expressões idiomáticas descritas na matriz do léxico-gramática. Estas regras apresentam uma estrutura `if()`, a qual engloba todas as dependências necessárias para que a expressão idiomática possa ser identificada, considerando

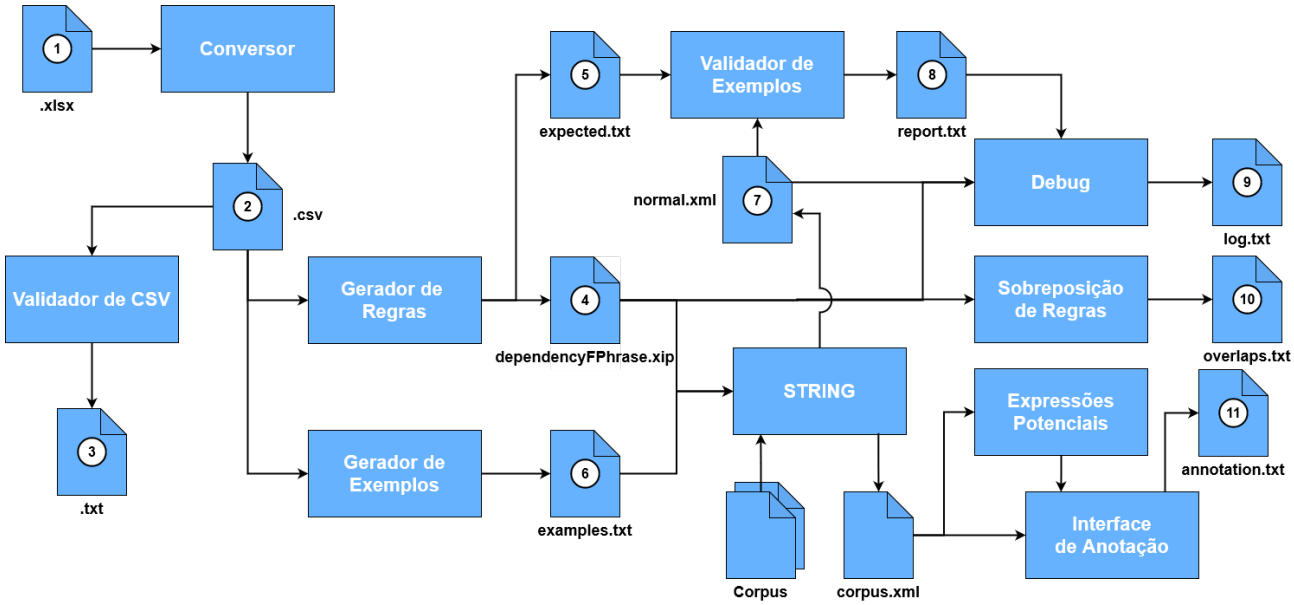

Figura 2: Arquitetura do sistema. As setas indicam o fluxo de informação no sistema. Os ícones com forma de ficheiro representam ficheiros de input ou output utilizados pelos módulos do sistema, enquanto estes são representados por retângulos. Os ficheiros são numerados para facilitar a sua referência no texto.

o seu significado não composicional. Estas dependências são interligadas com o operador lógico `AND`. Quando a condição da instrução `if()` é verificada, é gerada a dependência `FIXED` correspondente (ou seja, o consequente do `if()`).

A grande maioria das regras são geradas seguindo um processo sistemático, passo-a-passo, em que as restrições descritas na matriz são traduzidas sequencialmente para dependências XIP. Porém, há casos irregulares, que correspondem, neste momento, a cerca de 1,4% das expressões da matriz, e em que a geração automática da regra não é possível devido a incompatibilidades entre as dependências extraídas pela pipeline e as dependências geradas através deste processo sistemático. Estas expressões são marcadas com um '+' na coluna `AllManual`. Tal leva a que a regra XIP para a expressão em causa seja obtida diretamente a partir da coluna `Manual`, sem necessidade de qualquer passo suplementar.

Para as restantes expressões, a geração de regras segue um processo automatizado. Este processo consiste nos seguintes passos, para cada linha da matriz: (i) o sistema percorre cada coluna que representa uma restrição ou dependência sintática na expressão; (ii) se o valor na célula de cada coluna for *positivo* (no caso das colunas com valores binários) ou se a célula contiver uma palavra (no caso de campos lexicais) o sistema acrescenta a dependência correspondente à estrutura `if()`. Após a conclusão desse processo sequencial, a dependência `FIXED` adequada é colocada como o consequente da estrutura `if()`. Esta abordagem sistemática e abrangente garante a geração precisa e eficiente de regras para a maioria das expressões fixas da matriz. Para ilustrar como as restrições da matriz são traduzidas para dependências XIP, apresentamos os passos envolvidos na geração da instrução `if()` correspondente à regra da expressão idiomática *O Hugo* **apagou** *o* **fogo** 'O Hugo resolveu um problema urgente' (a regra resultante pode ser observada após o parágrafo seguinte): (i) o verbo principal que caracteriza esta expressão é obtido a partir da coluna `V`, que tem o valor '<apagar>'. É adicionado ao `if()` a dependência `VDOMAIN` (que captura esta função sintática) na qual o segundo argumento referencia o token do verbo e impõe a restrição de o lema ser 'apagar', ou seja, qualquer flexão deste verbo é aceite; (ii) o sujeito não é fixo (coluna `C0` marcada com '-') e tem de ter as propriedades de um nome humano (coluna `N0 = Nhum` marcada com '+'). É adicionada a dependência `SUBJ` (indicadora do sujeito de um verbo) e que tem o identificador do verbo como primeiro argumento, enquanto no segundo argumento se referencia o token do sujeito; esta mesma dependência impõe a restrição de o sujeito ter a feature `UMB-Human`, que representa qualquer grupo nominal humano; (iii) o primeiro complemento é fixo e o seu valor é obtido da coluna `C1` — '<fogo>'. Este complemento corresponde a um complemento direto (dependência `CDIR`), pois é o primeiro complemento e não é de tipo preposicional (coluna

`Prep1` marcada com '-'). O primeiro argumento da dependência `CDIR`) referencia o verbo e o segundo referencia o token do complemento, sendo imposta a restrição de o lema ser 'fogo', ou seja, qualquer flexão deste nome é aceite; (iv) para consequente do `if()` temos a dependência `FIXED` que tem como traço (ing. *feature*) a classe da expressão (obtida da coluna `CLASS`) — `C1`. Os argumentos do `FIXED` são os identificadores do verbo e do primeiro complemento, respetivamente.

Juntamente com a instrução `if()`, cada regra incorpora também um cabeçalho e um rodapé. O cabeçalho contém um exemplo da expressão idiomática, que é obtido da coluna `Exemplo` da matriz, e um identificador único atribuído à regra. O rodapé repete o exemplo e também inclui a dependência `FIXED` que deverá ser extraída para a expressão em questão. Abaixo, apresentamos a regra completa gerada para esta expressão idiomática:

```
//==================================
// Example: O Hugo apagou o fogo
// Rule ID: 159
//==================================
   if ( VDOMAIN(?,#2[lemma:apagar]) &
          SUBJ(#2,#1[UMB-Human]) &
        CDIR[post](#2,#3[lemma:fogo]) )
     FIXED[C1=+](#2,#3)
////ORIGINAL O Hugo apagou o fogo
////EXPECTED FIXED_C1(apagar,fogo)
```

Note-se, nesta regra, o uso de identificadores, de formato `#n`, que referenciam os elementos principais da expressão analisada. No exemplo apresentado acima, eles são: `Hugo (#1)`, `apagou (#2)` e `fogo (#3)`. Não há noção de ordem entre estes identificadores, nem na frase nem na dependência `FIXED`. O que importa é que se mantenha uma relação unívoca entre os nós (as palavras-chave ou os constituintes da construção) e os respetivos identificadores.

Note-se também que a expressão 'apagar fogo' é ambígua e esta regra pode ser ativada para usos literais da expressão. As regras não pretendem, pois, evitar a identificação de leituras literais, mas, pelo contrário, identificar potenciais instâncias da construção idiomática. Não obstante, Savary et al. (2019a) mostram que, quando os componentes lexicalizados de uma EMP coocorrem, estando presentes as condições morfossintáticas necessárias para uma leitura idiomática, o significado desta expressão também é quase sempre não-composicional.

Este sistema foi dotado de três modos diferentes de geração de regras: `Core`, `CoreDistS`, `CoreDistSC`, sucessivamente mais restritivos. Cada um destes modos gera regras que incorporam um conjunto de restrições diferentes na condição da instrução `if()`. O `Core` é o modo menos restritivo e gera regras que restringem apenas os consituintes essenciais da expressão. São eles a construção verbal da frase, o sujeito fixo, e os complementos fixos e as preposições fixas que os introduzem. As regras geradas pelo modo `CoreDistS`, apresentam todas as dependências consideradas pelo modo anterior, com a adição das restrições distribucionais sobre o sujeito livre da expressão. Finalmente, o modo mais restritivo é o `CoreDistSC`, fazendo uso de todas as restrições descritas na matriz. Isto consiste em usar todas as dependências consideradas pelo modo anterior, sendo ainda acrescentadas as restrições sobre os determinantes dos complementos fixos, bem como as restrições distribucionais sobre os complementos livres e preposições que os introduzem. Estes diferentes modos permitem testar que conjunto de restrições é mais adequado para identificar com precisão as expressões desejadas.

Há casos em que a tradução dos valores das colunas para as regras de dependência é mais complexa, nomeadamente quando alguma transformação pode ser aplicada a uma determinada expressão idiomática, sem que esta passe a ter um sentido literal. Para a maioria das transformações, a solução descritiva passa por incorporar uma restrição adicional, utilizando o operador lógico `OR`) à regra de base do constituinte afetado. Por exemplo, para a expressão *O Rodrigo* **apagou** *a Mónica* **da memória** 'O Rodrigo esqueceu intencionalmente a Mónica', o complemento direto ('Mónica') pode ser reduzido a um pronome, i.e., *O Rodrigo* **apagou-*a*** **da memória** 'O Rodrigo esqueceu-a intencionalmente'). Assim, a regra gerada deve permitir tanto um complemento direto preenchido por um grupo nominal, como um que seja expresso por um pronome clítico acusativo[7]. Para tal variação, é, pois, gerada a regra: `CDIR[post](#2,#3)) || CLITIC(#2,#3[acc])`, onde `||` representa o operador `OR`.

Contudo, para expressões que permitem a passivização, não é suficiente apenas modificar a regra base, pois estas transformações implicam uma reestruturação completa das frases. No caso

[7]Note-se que as restrições distribucionais sobre o complemento direto livre, quando expresso por um grupo nominal (concretamente a sua natureza humana), não são representadas quando este constituinte se encontra expresso por um pronome, daí a necessidade de incluir a dependência `CLITIC`.

```
if (VDOMAIN(?,#2[lemma:arregalar]) &
      SUBJ(#2,#1[UMB-Human]) &
    CDIR[post](#2,#3[surface:olhos]) &
      DETD(#3,?[surface:os]) )
  FIXED[C1=+](#2,#3)
```

```
if (VDOMAIN(?,#2[lemma:arregalar]) &
      SUBJ(#2,#1[UMB-Human]) &
    CDIR[post](#2,#3[surface:olhos]) &
      DETD(#3,?[surface:os]) &
    MOD[post](#2,#4[UMB-Human]) &
      PREPD(#4,?[surface:a]) )
  FIXED[C1PN=+](#2,#3)
```

Figura 3: Exemplo de duas regras com sobreposição.

da formação da frase passiva, o complemento direto torna-se o sujeito e é inserido um verbo auxiliar do particípio passado do verbo principal. Consequentemente, torna-se necessário gerar uma nova regra para as formas passivas. Para ilustrar esta diferença, apresentamos abaixo a estrutura `if()` gerada para a construção passiva da expressão idiomática apresentada anteriormente: *O* **fogo** *foi* **apagado** 'O problema foi resolvido'.

```
if ( VDOMAIN(?,#2[pass-ser,
             lemma:apagar]) &
       SUBJ(#2,#1[lemma:fogo]) )
  FIXED[C1=+](#2,#1)
```

O output deste módulo consiste num ficheiro denominado *dependencyFPhrase.xip* (Fig. 2, Ficheiro 4), que contém todas as regras XIP geradas, bem como um ficheiro *expected.txt* (Ficheiro 5), cujo conteúdo são as dependências `FIXED` que devem ser extraídas para cada par regra-exemplo.

3.3.3. Sobreposição de Regras

O Módulo de Sobreposição de Regras ajuda na identificação precisa de expressões idiomáticas ao detectar quando duas regras estão sobrepostas. Duas regras estão sobrepostas se todas as dependências contidas na condição de uma delas também estiverem presentes na condição da outra. Note-se que os argumentos dessas dependências também devem ser exatamente iguais para que uma sobreposição seja detetada.

Estes casos prejudicam o desempenho do sistema, pois isto significa que mais de uma regra pode ser ativada para uma dada expressão, resultando na extração de múltiplas dependências `FIXED` para a mesma expressão no texto. Embora tal não seja um problema para a descoberta de expressões idiomáticas, tem um impacto negativo na identificação adequada destas expressões, uma vez que se torna ambíguo qual delas está realmente presente no texto analisado. Assim, este módulo compila todas as sobreposições detetadas num ficheiro TXT (Fig. 2, Ficheiro 10), o qual pode ser consultado para determinar que regras precisam de ser ajustadas de modo a não entrarem em conflito com outras.

Para demonstrar o que é uma sobreposição e que tipo de vantagens a presença deste módulo traz, a Figura 3 apresenta duas regras diferentes que são geradas para as expressões idiomáticas presentes nas seguintes frases: *O Hugo* **arregalou** *os* **olhos** 'ficar surpreso' (regra à esquerda) e *O Hugo* **arregalou** *os* **olhos à** *Joana* 'censurar alguém só pela expressão facial' (regra à direita).

Ao comparar estas duas regras, é detetada uma sobreposição pois todas as 4 dependências que estão presentes na condição da primeira regra estão também presentes na condição da segunda. Além disto, as restrições sobre os argumentos de todas estas dependências são as mesmas. Isto significa que, neste estado, se a segunda regra for ativada para uma dada frase, a primeira irá necessariamente ser também ativada.

Face a esta situação, o Módulo de Sobreposição de Regras faz uma breve comparação da classe e argumentos da dependência `FIXED` correspondente a cada regra (neste caso `FIXED_C1(arregalar,olhos)` e `FIXED_C1PN(arregalar,olhos)`) e escreve o seguinte no seu ficheiro de output:

```
Rule 100 is contained in Rule 898
The Rules present a similar FIXED:
Rule 100: FIXED_C1(arregalar,olhos)
Rule 898: FIXED_C1PN(arregalar,olhos)
```

Para se lidar com sobreposições de regras em que os argumentos da dependência `FIXED` são os mesmos, decidiu-se remover a linha da matriz que descreve a expressão mais simples, quando tal não parece afetar o significado global da expressão, apenas se mantendo ambas as construções quando o significado é claramente distinto. Efetivamente, com a introdução dos modos de geração de regras, é possível controlar-se qual das regras se pretende utilizar a partir

da descrição da regra mais complexa. No nosso exemplo, gerando as regras no modo `CoreDistSC` obtém-se a segunda regra, enquanto que, usando um dos outros modos, se obtém a primeira regra.

Para casos em que há diferença nos argumentos do `FIXED`, é introduzida na regra mais simples uma condição negativa sobre uma das dependências relativas a um argumento suplementar da regra mais complexa, de forma a garantir que as regras se tornem mutuamente exclusivas.

3.3.4. Gerador de Exemplos

Além de ser utilizado para gerar as regras XIP, o ficheiro CSV também é necessário para obter e gerar exemplos de expressões idiomáticas que irão servir de *input* à *pipeline*, permitindo avaliar o desempenho das regras de dependência geradas. Para cada expressão idiomática, pelo menos uma frase manual é obtida a partir das colunas `Example` e `Other Example`. Esta última apresenta, em geral, algum aspeto de variação da construção, como, por exemplo, a variação da preposição que introduz um complemento, e.g. *O João* **não gasta cera com/em defuntos ruins** 'O João não desperdiça tempo, dinheiro, etc. com coisas que não têm importância/mérito'.

Além disso, para cada transformação que pode ser aplicada a uma expressão, é gerada uma frase com base nas restrições codificadas nas diversas colunas da matriz. Por exemplo, para a expressão idiomática **conhecer** *alguém* **pela pinta** 'reconhecer alguém pela aparência ou modos', a frase 'A Rosa conheceu o vizinho pela pinta' é obtida da coluna `Example`, e as frases 'A Rosa conheceu-o pela pinta' e 'O Filipe foi conhecido pela pinta' são geradas, respetivamente, para a pronominalização acusativa do primeiro complemento (livre) e para a construção passiva com o verbo 'ser'. O *output* deste módulo consiste num ficheiro denominado *examples.txt* (Fig.2, Ficheiro 6), que contém todas as frases-exemplo de base, bem como as frases geradas automaticamente.

Um tipo específico de frases, que apelidámos *expressões pseudo-idiomáticas*, é também gerado por este módulo. Estas expressões são criadas substituindo um dos constituintes fixos de uma dada expressão idiomática por uma palavra genérica ou não ambígua. A ideia por detrás da geração destes exemplos consiste em testar o sistema com estas frases para garantir que, do ponto de vista lexical, as regras estão a restringir corretamente todos os elementos-chave. Por exemplo, para a expressão idiomática *A Maria* **borrou** *a* **pintura** 'A Maria estragou algo que estava a correr bem', com dois elementos fixos, são geradas duas expressões pseudo-idiomáticas: uma em que o verbo 'borrar' é substituído por outro, 'A Maria mói a pintura', e outra em que o primeiro complemento fixo é substituído por um pronome indefinido, 'A Maria borrou isso'.

3.3.5. Validador de Exemplos

De forma a validar as regras geradas, o ficheiro de regras é integrado na gramática do analisador sintático XIP e a STRING processa os exemplos de expressões idiomáticas. As dependências extraídas resultantes deste processamento, incluindo a dependência `FIXED`, são, então, armazenadas num ficheiro XML (Fig. 2, Ficheiro 7). O Validador de Exemplos começa, pois, por analisar a informação contida nesse ficheiro, visando localizar e isolar a dependência `FIXED` para cada frase. Essas dependências são, depois, comparadas com as dependências `FIXED` esperadas. O módulo avalia quatro aspectos dos resultados obtidos , que podem ser utilizados posteriormente para avaliar o desempenho do sistema: (i) verifica se foi extraída alguma dependência `FIXED`; (ii)- verifica se foi extraída mais do que uma dependência `FIXED`; (iii) confirma que o número de argumentos da dependência `FIXED` extraída é o mesmo que o número de argumentos esperados; e (iv) verifica se os argumentos da dependência estão de acordo com os argumentos esperados. Os resultados desta análise são, finalmente, registados num ficheiro de saída *report.txt* (Fig. 2, Ficheiro 8) para análise posterior.

3.3.6. Debug

O Módulo de Debug foi desenvolvido sobretudo para lidar com falhas em que nenhuma dependência `FIXED` é extraída pela STRING. Melhorar o desempenho desses exemplos é prioritário, pois é o mais frequente tipo de falha. Para que um humano realize as tarefas envolvidas no processo de depuração, seria necessário voltar a correr o exemplo na pipeline e verificar visualmente o output, o que é um processo que consome muito tempo. No entanto, o Módulo de Debug não enfrenta tal problema, pois obtém estas informações a partir do ficheiro XML (Fig. 2, Ficheiro 7) que contém o output da STRING.

Para cada exemplo em que não foi identificada nenhuma expressão fixa, o programa compara as dependências extraídas pela STRING ao processar a frase com as dependências que constituem a regra que deveria ter sido ativada para essa mesma falha. São realizados 3 níveis sequenciais de verificações para cada dependência esperada:

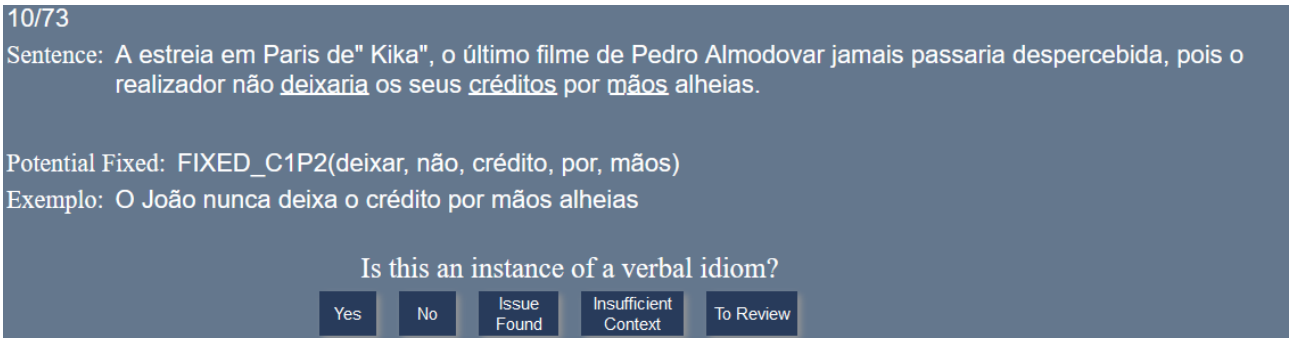

Figura 4: Aspeto geral da interface de anotação.

(i) é verificado se essa dependência foi extraída; (ii) é verificado se os argumentos da dependência extraída correspondem aos esperados; e (iii) é verificado se os traços da dependência extraída correspondem aos esperados.

O módulo gera o ficheiro *log.txt* (Fig. 2, Ficheiro 9), que apresenta pormenorizadamente, para cada falha de identificação: (i) a frase em que ocorreu; (ii) o ID da regra esperada; (iii) as dependências esperadas em falta (se as houver); (iv) as dependências incompatíveis (se as houver), com uma indicação dos argumentos que foram causa dessa incompatibilidade; e (v) as dependências com traços incorretos (se as houver), com uma indicação dos traços esperados e dos traços da dependência extraída correspondente.

3.4. Ferramentas de Anotação

Pretendendo-se avaliar o desempenho do sistema com dados do mundo real, foi necessário realizar uma anotação de *corpora*. Para otimizar o uso dos recursos humanos bem como o tempo necessário para realizar a tarefa de anotação das expressões fixas que ocorrem nos textos, foram desenvolvidos dois programas de suporte aos anotadores humanos.

3.4.1. Deteção de Expressões Idiomáticas Potenciais

O primeiro programa percorre o *corpus* de textos selecionados em busca de possíveis ocorrências de expressões idiomáticas em cada frase. Em seguida, compila as frases que encontrou num ficheiro formatado. Considera-se que uma frase contém uma expressão idiomática potencial se estiverem presentes o verbo e todos os seus complementos fixos, ou seja, os principais elementos lexicais que formam essa expressão. Além disso, e com base na análise de Manning (1999, pp. 157 ff.), a distância entre elementos consecutivos da expressão fixa na frase analisada não deve ser superior a 5 tokens.

3.4.2. Interface de Anotação

Após a execução do programa anterior, é possível utilizar a interface de anotação (Figura 4) para marcar as expressões potenciais que são, de facto, expressões idiomáticas. Este programa consiste numa interface gráfica onde é apresentada ao utilizador uma janela para cada frase em que foi detetado um candidato potencial. Em cada janela, é visível ao utilizador um conjunto de componentes informativos: (i) a frase do *corpus* onde a expressão idiomática potencial foi detetada, estando sublinhadas as palavras que correspondem aos argumentos da expressão; (ii) a dependência `FIXED` que corresponde à expressão em causa; e (iii) o exemplo manual contido na matriz para essa mesma expressão. Na parte inferior da janela, existem dois botões que permitem ao anotador decidir se marca esse caso como idiomático ou não. Outros três botões podem ser usados para registar problemas encontrados. Após a conclusão da anotação, é gerado um ficheiro TXT (Fig. 2, Ficheiro 11), contendo todas as expressões idiomáticas marcadas e as frases em que estas ocorrem.

Apesar da existência de ferramentas de anotação já desenvolvidas, optou-se por criar uma solução própria, à semelhança do que foi feito em Haagsma et al. (2020), devido à especificidade do método de identificação de expressões potenciais e à informação que se pretendia apresentar em cada janela da interface. Esta decisão foi motivada, em particular, pelos seguintes fatores: (i) a deteção das expressões baseia-se nos lemas das palavras; (ii) pretendia-se sublinhar essas palavras na interface; (iii) o output da ferramenta STRING, utilizada para obter os lemas, apresenta-se num formato XML específico. Estes aspetos dificultam a integração direta da informação proveniente dessa estrutura XML em programas de anotação já existentes, exigindo a construção de um conversor apropriado.

4. Avaliação

A avaliação do sistema foi organizada em duas fases complementares: (i) uma fase de testes controlados, que designamos por avaliação intrínseca, e (ii) uma avaliação extrínseca, baseada em dados reais.

A avaliação intrínseca não tem como objetivo aferir a capacidade de generalização do sistema, mas sim garantir que as regras geradas são internamente consistentes e capazes de reconhecer as expressões idiomáticas sempre que estas ocorrem nas suas formas canónicas ou esperadas. Trata-se, assim, de uma etapa fundamental para validar a adequação do sistema em condições ideais e assegurar que não há falhas sistemáticas na definição ou aplicação das regras.

Num primeiro momento, seguiu-se um processo iterativo, no qual cada iteração consistia na realização de testes intrínsecos, seguidos de uma fase de refinamento com base nos erros detetados. Uma vez atingido um desempenho satisfatório nestes testes, avançou-se para a avaliação extrínseca.

Nesta nova fase, o sistema foi aplicado a textos do mundo real, onde as expressões idiomáticas ocorrem de forma espontânea e não controlada. Esta avaliação é essencial para aferir a capacidade do sistema de generalizar para além dos exemplos usados na sua construção. A necessidade de lidar com a variabilidade natural das expressões motivou, por sua vez, a expansão da informação linguística do sistema, o que exigiu uma nova ronda de testes intrínsecos, agora com o objetivo de verificar a adequação dos novos dados introduzidos.

4.1. Avaliação Intrínseca

A avaliação intrínseca recorre a três tipos de frases: as frases de base, *i.e.* exemplos obtidos diretamente da matriz; as frases geradas através da aplicação de diferentes transformações sobre as frases de base; e as expressões que chamámos de pseudo-idiomáticas (v. 3.3.4). Após o processamento destas frases pela STRING, segue-se a análise das dependências `FIXED` extraídas. Para avaliar cada exemplo são considerados cinco critérios, nos quais a dependência `FIXED` extraída é comparada com a esperada. Observam-se as seguintes situações: (i) nenhuma dependência é `FIXED` extraída; (ii) são extraídas várias dependências `FIXED`; (iii) é extraída uma dependência `FIXED` mas com número incorreto de argumentos; (iv) é extraída uma dependência `FIXED` com o número correto de argumentos, mas com argumentos incorretos; e (v) é extraída uma dependência `FIXED` com os argumentos corretos. Para os exemplos de expressões pseudo-idiomáticas, o único critério considerado é se uma dependência `FIXED` foi ou não extraída.

4.1.1. Resultados

A Tabela 1 apresenta os resultados obtidos (para as frases de base e para os exemplos gerados automaticamente) antes e após o processo iterativo de refinamento do sistema, bem como após o refinamento da informação linguística introduzida como resultado da fase seguinte de avaliação (apresentado em detalhe na Secção 4.2). Podemos observar que, inicialmente, os resultados gerais foram relativamente satisfatórios: a tarefa de reconhecer tanto frases produzidas manualmente como geradas automaticamente foi bastante bem-sucedida, verificando-se a extração da dependência `FIXED` para cerca de 90% dos exemplos considerados. No entanto, numa parte significativa destes casos, verificou-se a extração de múltiplas dependências `FIXED`, indicando a sobreposição de várias regras, o que representa um entrave à identificação precisa das expressões idiomáticas. Os restantes tipos de falhas verificavam-se muito raramente e eram, na grande maioria, erros pontuais.

Com o apoio da informação fornecida pelo Módulo de Sobreposição de Regras e pelo Módulo de Debug, foi possível refinar significativamente os processos de geração de regras e de exemplos, eliminando muitos dos problemas detetados e reduzindo o número de falhas do sistema. Após o processo iterativo de refinamento, obteve-se um resultado totalmente correto, sendo extraída uma dependência `FIXED` adequada para cada exemplo.

Por fim, em relação às expressões pseudo-idiomáticas, como pode ser observado na Tabela 2, os resultados iniciais mostram que apenas 30 das 4 980 frases geradas foram incorretamente reconhecidas como expressões idiomáticas, número este que foi reduzido para apenas 2 de 4 984 frases no final do processo de refinamento. A variação no número de exemplos deve-se às várias mudanças do léxico-gramática durante o processo de refinamento. Uma análise dos erros revelou que estes aconteciam em expressões nas quais um dos componentes fixos coincidia com a palavra genérica que o substituia. Por exemplo, a expressão **não dar por isso** 'não notar algo' tem como complemento fixo a palavra 'isso', a qual, por design, é o elemento utilizado para substituir o primeiro complemento fixo ao gerar a ex-

Frases	#Total	% Não Detetado	% Múltiplos FIXED	% N. Args Incorreto	% Args Incorretos	% Args Corretos
Antes do Processo Iterativo de Refinamento						
Base	2 626	9,4%	5,2%	0,6%	0,3%	84,5%
Geradas	1 463	10,4%	9,4%	0,1%	0,0%	80,1%
Depois do Processo Iterativo de Refinamento						
Base	2 538	0,0%	0,0%	0,0%	0,0%	100,0%
Geradas	1 392	0,0%	0,0%	0,0%	0,0%	100,0%
Após o Refinamento da nova Informação Linguística						
Base	2 682	0,0%	0,0%	0,0%	0,0%	100,0%
Geradas	1 480	0,0%	0,0%	0,0%	0,0%	100,0%

Tabela 1: Resultados obtidos para os exemplos de expressões idiomáticas usados na avaliação intrínseca do sistema. A coluna **% Não Detetado** corresponde a expressões para as quais não foi extraída qualquer dependência FIXED. As últimas 3 colunas correspondem a expressões para as quais foi extraída um única dependência FIXED

#Total	#Não Detetado	#FIXED Extraído
Antes do Refinamento		
4 980	4 950	30
Após o Refinamento		
4 984	4 982	2

Tabela 2: Resultados obtidos para os exemplos de expressões pseudo-idiomáticas usados na avaliação intrínseca do sistema.

pressão pseudo-idiomática. Desta forma, a frase gerada consiste numa instância válida da construção idiomática. Estes casos resultam, portanto, de imprecisões do sistema de geração das expressões pseudo-idiomáticas, e não de falhas das regras que se pretendem testar.

4.2. Avaliação Extrínseca

De forma a se poder avaliar o desempenho do sistema de um ponto de vista extrínseco, foi constituído um *corpus* de textos reais, escritos em português europeu, coligido a partir de recursos textuais existentes e onde foram anotadas as expressões idiomáticas neles presentes. Em paralelo, o conteúdo desse *corpus* foi processado pela STRING para a identificação automática de instâncias destas expressões. As dependências FIXED geradas foram, então, comparadas com as obtidas a partir da anotação manual.

A avaliação do desempenho do sistema na avaliação extrínseca recorre às métricas *standard* de *Precisão*, *Recall* e *F1-score*. Considera-se como a referência para a avaliação as expressões marcadas manualmente no *corpus* como sendo expressões idiomáticas. Em contraste, as dependências FIXED extraídas pela STRING são tratadas como previsões de casos positivos. Assim, um *verdadeiro-positivo* é uma expressão do *corpus* anotada como idiomática e para a qual a STRING extraiu corretamente uma dependência FIXED. Um *falso-positivo* ocorre quando a STRING extrai uma dependência FIXED para uma expressão do *corpus* que não foi anotada como idiomática. Por fim, um *falso-negativo* é a uma expressão anotada no *corpus* como sendo idiomática, mas para a qual a STRING não extraiu uma dependência FIXED.

4.2.1. Descrição do Corpus e Processo de Anotação

O *corpus* utilizado na avaliação extrínseca pode ser dividido em dois grandes conjuntos: (1) o *corpus* de *treino* — utilizado para avaliar o desempenho do sistema, bem como expandir a informação linguística. A primeira avaliação com este *cor-*

pus motivou um processo de refinamento do sistema, com a correção de erros detetados e a introdução de novas expressões na matriz. Após se assegurar um certo nível de adequação da nova informação adicionada (através da realização de nova avaliação intrínseca), o desempenho do sistema foi mais uma vez avaliado com este *corpus*; e (2) o *corpus* de *teste* — utilizado para avaliar o desempenho da versão final do sistema. O *corpus* de treino conta com 75 documentos selecionados a partir de duas fontes: 49 documentos são transcrições de sessões do Parlamento Português, desde Maio de 2004 até Março de 2005; os restantes 26 documentos foram obtidos a partir da coleção do *CETEMPúblico*. A Tabela 3 apresenta uma análise dos documentos de ambas as fontes, especificando o número de documentos, de frases e de expressões potenciais detetadas.

Fonte	Parlamento Português	CETEMPúblico
# Documentos	49	26
# Frases	50 689	50 349
# Expressões Potenciais	3 075	2 466

Tabela 3: Descrição das fontes de documentos que constituem o *corpus* de treino.

O *corpus* de teste consiste em 103 documentos selecionados de duas fontes: 78 documentos são transcrições de sessões do Parlamento Português, desde Março de 2018 até Setembro de 2018; os restantes 25 documentos foram obtidos a partir da coleção do *CETEMPúblico*. A Tabela 4 apresenta uma descrição pormenorizada dos documentos de ambas as fontes, especificando o número total de documentos, frases e expressões idiomáticas potenciais detetadas pela ferramenta especificamente desenvolvida para este estudo.

Fonte	Parlamento Português	CETEMPúblico
# Documentos	78	25
# Frases	50 911	51 376
# Expressões Potenciais	2 749	2 331

Tabela 4: Descrição das fontes de documentos que constituem o *corpus* de teste.

Note-se que, embora o número de documentos provenientes de cada fonte seja distinto, o número de frases de cada subconjunto é notavelmente semelhante. Na prática, isto significa que as duas fontes contribuem de forma similar para a avaliação. Curiosamente, os documentos do corpus do Parlamento Português apresentam um maior número de expressões idiomáticas potenciais. Contudo, o número de expressões potenciais não reflete necessariamente o número de expressões idiomáticas presentes nos textos.

O *corpus* foi integralmente anotado por três anotadores, peritos em expressões idiomáticas do português europeu, utilizando as ferramentas de anotação descritas na Secção 3.4 e seguindo as diretivas de anotação (disponíveis no repositório onde também disponibilizamos o *corpus*). Um subconjunto inicial de documentos, composto por 7 textos e contendo aproximadamente 10% das expressões verbais idiomáticas potenciais detetadas no *corpus* de treino, foi anotado pelos três anotadores. Esta amostra permitiu calcular o acordo entre os anotadores e validar a adequação das diretrizes de anotação. Devido à natureza da tarefa, foi utilizado o índice Krippendorff-alfa para dados nominais (Krippendorff, 2008) como métrica de concordância entre anotadores.

Após a conclusão da anotação desses 7 documentos, os anotadores resolveram colaborativamente as discrepâncias de forma a produzir uma anotação consensual, sendo assim criada uma coleção dourada/de referência. A necessidade de discussão entre os três anotadores para chegar a um consenso destaca a complexidade na identificação das expressões verbais idiomáticas, uma vez que determinar o caráter idiomático de uma expressão se mostrou uma tarefa desafiadora, sobretudo com um contexto limitado. Não obstante, a maioria das diferenças na anotação foram atribuídas a faltas de atenção dos anotadores. Ainda assim, tendo-se atingido um acordo entre os anotadores de α=0,869 (acima, portanto, do limite de 0,8)[8], este valor foi considerado "satisfatório", sendo possível assumir que o desempenho dos anotadores nesta tarefa seria bastante consistente. Todo este processo permitiu reduzir a carga de trabalho para a anotação dos 68 documentos remanescentes, distribuindo-os proporcionalmente entre os anotadores e sendo cada documento anotado por apenas uma pessoa.

Note-se que, após a primeira avaliação com o *corpus* de treino e consequente introdução de expressões idiomáticas na matriz, foi necessário anotar as novas expressões potenciais detetadas como resultado da nova informação linguística.

[8] https://www.k-alpha.org/methodological-notes

4.2.2. Resultados

Passemos à análise dos resultados (Tabela 5). Em primeiro lugar, note-se que foram inicialmente anotadas no *corpus* de treino um total de 2.041 expressões idiomáticas. Número este que subiu (consideravelmente) para 2.839 com a introdução de novas construções no léxico-gramática. Já no *corpus* de teste foram anotadas um total de 2.339 expressões idiomáticas.

Comparando os resultados entre os diferentes modos de execução, é evidente que a *recall* do sistema aumenta à medida que as regras se tornam menos restritivas. Isto era esperado, pois regras menos restritivas são acionadas com mais frequência, levando a uma taxa de deteção mais alta de expressões idiomáticas. Em termos de *precisão*, já se esperava que modos de execução menos restritivos apresentassem valores mais baixos, dada a maior flexibilidade das regras. No entanto, isso não foi necessariamente o caso: o modo `Core` apresenta uma precisão mais alta do que o modo `CoreDistS`, sugerindo que a ausência de restrições distribucionais sobre o sujeito das expressões idiomáticas impacta positivamente o desempenho do sistema. Esta diferença destaca as restrições distribucionais sobre sujeitos livres como um foco potencial para futuros aprimoramentos do sistema. Observando a *F1-score*, que corresponde à média harmónica entre a *precisão* e a *recall*, o modo `Core` emerge como o modo de execução mais bem-sucedido e deve ser considerado o modo de referência daqui em diante.

Para examinar mais pormenorizadamente as falhas do sistema, foi desenvolvido um programa que compila, para cada expressão idiomática que foi erradamente detectada pela STRING, todas as frases nas quais a falha ocorre. Isto é, para cada expressão idiomática que é um falso-positivo ou falso-negativo, são recolhidas e apresentadas todas as frases onde a dependência `FIXED` correspondente foi extraída mas não anotada, ou anotada mas não extraída, respectivamente. Esta organização dos dados permite analisar os casos de falhas mais frequentes, encontrar padrões nessas falhas e definir claramente os problemas subjacentes.

Uma avaliação pormenorizada dos falsos-positivos recorrentes mostrou que a maioria dos casos eram explicados por falhas na anotação do *corpus*. Por exemplo, todas as falhas para as expressões representadas por `FIXED_CADV(ir,embora)` e `FIXED_CADV(ir,se,embora)` (ambas significando 'sair de um lugar, abandonar uma posição/cargo/função') vêm de frases que não tinham sido anotadas como expressões idiomáticas, sobretudo pela alternância entre a construção reflexa e a não pronominal; por exemplo, 'Os senhores **foram embora**, porque não eram capazes de fazer o que nós fizemos!').

Além disto, várias instâncias de falsos-positivos surgiram de informações imprecisas codificadas no léxico-gramática. Para certas expressões, os complementos fixos estavam limitados às suas formas no singular mas, na verdade, esses elementos podiam também aparecer no plural. Tal é o caso de 'caminho' em `FIXED_C1PN(abrir,caminho)`, no sentido de 'preparar o terreno' e 'porta' em `FIXED_C1PN(fechar,porta)`, no sentido de 'rejeitar'. Nesses casos, atualizar a matrizm indicando simplesmente trata-se não da forma mas do lema desse elemento fixo, bastaria para que o sistema passasse a aceitar as variantes no plural, como '**abrir caminhos**' e '**fechar** as **portas**'.

Em relação aos casos de falsos-negativos, a análise revelou que as falhas seguiam padrões identificáveis. De acordo com as diretrizes de anotação, expressões idiomáticas que ainda não estavam descritas na matriz mas que eram detetadas como expressões potenciais, devido a serem formadas pelos mesmos elementos lexicais que ocorrem noutra expressão já recenseada, deveriam ser marcadas como expressões idiomáticas. Um exemplo notável é o frequente falso-negativo `FIXED_CAN(chamar,atenção)`, por exemplo em *A Clara* **chamou** *a* **atenção** *da Marta para isso* 'A Clara alertou a Marta para isso'. Esta expressão era frequentemente marcada como um falso-negativo porque, na verdade, era uma instância de uma expressão idiomática diferente: **chamar** *alguém* **à atenção** 'repreender alguém', a qual ainda não constava na matriz.

Outros casos semelhantes surgiram nesta análise, como a variação de preposições, que não tinham sido consideradas na matriz, mas cujas expressões eram capturadas como potenciais, pois o processo ignorava as preposições. Assim, por exemplo, para a expressão representada por `FIXED_CNP2(levar,a,fim)` (significando 'concluir algo'), a preposição 'a' pode variar com 'até a' ('**levar até ao fim**') mas esta segunda preposição não tinha ainda sido registada no léxico-gramática, ainda que a sequência de 'levar' e 'fim' fosse capturada como uma expressão idiomática potencial.

Após a análise dos resultados obtidos para o *corpus* de treino, procedeu-se à alteração de informação lexical na matriz léxico-gramatical, tendo em conta os problemas identificados. Além disso, foram também adicionadas à matriz um

	# Expressões Idiomáticas Anotadas	Precisão	Recall	F1-score
	***corpus* de Treino Inicial**			
Core	2 041	0,807	0,570	0,668
CoreDistS		0,759	0,265	0,392
CoreDistSc		0,807	0,172	0,286
	***corpus* de Treino Final**			
Core	2 839	0,929	0,586	0,719
CoreDistS		0,940	0,295	0,449
CoreDistSc		0,968	0,236	0,380
	***corpus* de Teste**			
Core	2 339	0,902	0,605	0,725
CoreDistS		0,905	0,276	0,423
CoreDistSc		0,947	0,207	0,340

Tabela 5: Resultados obtidos em vários momentos para os *corpora* usados na avaliação extrínseca do sistema.

total de 142 novas expressões idiomáticas, que tinham sido encontradas durante a anotação do *corpus* de treino, tendo sido também definida uma nova classe de expressões. Com as alterações feitas ao léxico-gramática, foi necessário voltar a detetar as expressões idiomáticas potenciais e anotar as novas expressões descobertas. A Tabela 5 mostra os resultados obtidos para a última avaliação do *corpus* de treino.

Nesta fase, o número total de expressões anotadas nestes documentos totaliza 2 839 expressões (soma dos falsos-negativos com os verdadeiros-positivos). A introdução de novas expressões da matriz e deteção dos novos potenciais resultaram na anotação de 798 novas instâncias de expressões idiomáticas.

Quanto aos resultados propriamente ditos, é visível um aumento significativo no número de verdadeiros-positivos bem como uma diminuição significativa do número de falsos-positivos. Estas duas mudanças traduzem-se num aumento significativo da precisão do sistema, que, por sua vez, se reflete num ligeiro aumento da *F1-score*.

Considerando que foram introduzidas mudanças substanciais para aprimorar o desempenho do sistema, nomeadamente no refinamento do léxico-gramática das expressões idiomáticas, na otimização dos processos de geração de regras e no desenvolvimento da gramática da STRING; e tendo em conta que estas transformações se basearam nos dados do *corpus* de treino; surgiu a necessidade de tornar a avaliar o sistema com um novo conjunto de dados (designado por *corpus* de teste), para analisar o impacto dessas melhorias no desempenho da STRING. A Tabela 5 apresenta os resultados obtidos pelo sistema ao processar o *corpus* de teste.

É visível que não há grande diferença no desempenho do sistema para o *corpus* de treino final e para o *corpus* de teste, com uma ligeira redução da precisão e ligeiro aumento da recall, resultando numa F1-score muito semelhante. Isto indica que, apesar de as mudanças no sistema terem sido baseadas no *corpus* de treino, a melhoria de desempenho é geral.

5. Conclusões

Este trabalho teve como objetivo aprimorar o processamento de expressões idiomáticas na pipeline STRING, particularmente no módulo XIP, responsável pela deteção destas expressões em textos. O trabalho resultou em melhorias significativas em vários aspetos do sistema, contribuindo para um melhor desempenho, maior automação na geração de regras e exemplos, e maior capacidade de rastrear diferentes tipos de falhas.

Os resultados gerais da avaliação intrínseca foram muito positivos, refletindo os esforços investidos tanto no desenvolvimento do Gerador de

Regras quanto no Validador de Exemplos, bem como a eficácia das ferramentas de suporte desenvolvidas, que tiveram um impacto substancial no desempenho global do sistema.

O desempenho na avaliação extrínseca também foi bastante favorável. Em termos de Precisão, os resultados finais foram muito satisfatórios — cerca de 90% das dependências FIXED extraídas pela STRING foram corretamente capturadas. Quanto à deteção das expressões idiomáticas existentes no *corpus*, a capacidade do sistema para identificar as expressões-alvo foi um pouco mais limitada, com uma *recall* em torno de 60%, revelando aspectos do sistema que ainda ainda precisam de aperfeiçoamento, assim como algumas limitações deste trabalho.

Considerando as ferramentas de deteção e de anotação utilizadas, apenas instâncias de expressões idiomáticas descritas no léxico-gramática poderiam ter sido identificadas e anotadas. No entanto, este é, tanto quanto sabemos, o maior léxico formal de expressões idiomáticas de núcleo verbal do português europeu, que apresenta uma cobertura lexical bastante ampla. Como é óbvio, dificilmente algum recurso poderá algumas vez conter *todas* as expressões idiomáticas da língua e muitas variantes de expressões existentes e já recenseadas carecem ainda de uma descrição adequada e de integração no léxico-gramática. Além disso, como o processo de anotação não incluiu uma inspeção de todo o conteúdo textual, não podemos afirmar que o *corpus* foi integralmente anotado para expressões idiomáticas. Porém, este é, tanto quanto sabemos, o primeiro *corpus* anotado com expressões idiomáticas do português europeu (como referido na Secção 1 e que doravante estará disponível à comunidade científica).

Por fim, é importante destacar como este trabalho permitiu a deteção de erros na matriz de dados léxico-sintáticos, contribuindo para uma maior consistência do léxico-gramática. Esse processo não apenas esclareceu o significado preciso de várias propriedades codificadas, mas também promoveu a evolução da informação lexical contida na matriz. A importância de avaliar o desempenho do sistema em textos reais é inquestionável. Esta tarefa não apenas possibilitou a identificação de expressões idiomáticas ainda não descritas na matriz e, especialmente, de variantes inéditas de expressões já catalogadas, mas também evidenciou áreas onde as regras geradas ainda são insuficientes para identificar com precisão instâncias das expressões-alvo.

Como trabalho futuro, consideramos que uma das prioridades principais deverá ser o desenvolvimento da anotação e avaliação de textos reais, dada a importância desta abordagem para identificar falhas e lacunas no desempenho do sistema de identificação de expressões idiomáticas verbais. A avaliação com dados reais não apenas possibilita a descoberta de expressões ainda não integradas na matriz léxico-gramatical, mas também permite o refinamento contínuo das regras, assegurando a descoberta de novos fenómenos e uma cobertura mais abrangente. Além disso, sugere-se a realização de uma anotação sistemática de corpora de diferentes domínios, ampliando a base de dados para suportar análises mais diversificadas. Uma exploração de *corpora* específicos, como o de texto futebolístico (Correia et al., 2016), por exemplo, poderá oferecer *insights* sobre a frequência e a variação contextual das expressões fixas na linguagem coloquial e especializada. O alargamento a outros géneros de texto poderá, ainda, contribuir para uma descrição léxico-gramatical não limitada a um domínio específico.

Aproveitando os *corpora* anotados produzidos neste estudo, prevê-se a possibilidade de explorar métodos estatísticos para a deteção de candidatos a novas expressões idiomáticas verbais ainda não descritas no léxico-gramática. Esta abordagem permitiria identificar padrões frequentes e coocorrências relevantes que podem escapar à análise manual, promovendo uma expansão dinâmica do léxico-gramática. Ao aplicar técnicas como a análise de associações, redes semânticas ou modelagem de tópicos, tornar-se-ia possível reconhecer automaticamente estruturas verbais e combinações lexicais que, embora fixas ou semi-fixas, não foram anteriormente lexicalizadas nem formalizadas. Esta deteção assistida por algoritmos permitiria aumentar a cobertura do léxico-gramática, enriquecendo a descrição das expressões idiomáticas com novas entradas e possibilitando uma adaptação mais precisa e atualizada do sistema ao uso real da língua.

Agradecimentos

Este trabalho foi financiado por fundos nacionais portugueses através da Fundação para a Ciência e a Tecnologia (Referência: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) e pela Comissão Europeia (Projeto: iRead4Skills, Referência: 1010094837, Programa: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

Referências

Aït-Mokhtar, Salah, Jean-Pierre Chanod & Claude Roux. 2002. Robustness beyond shallowness: Incremental dependency parsing. *Natural Language Engineering* 8(2–3). 121–144. 10.1017/s1351324902002887

Antunes, David. 2024. *Enhancing parsing of European Portuguese verbal idioms*: Instituto Superior Técnico, Universidade de Lisboa. Tese de Mestrado

Baptista, Jorge. 2004. [compositional vs. frozen sequences] (em chinês). Em *Lexicon-Grammar Workshop*, 81–93

Baptista, Jorge. 2008. Structuring of cross-linguistic database of frozen sentences. Em Carmen González Royo & Pedro Mogorròn Huerta (eds.), *Estudios y análisis de fraseología contrastiva: lexicografía y traducción*, 37–46. Universidade de Alicante

Baptista, Jorge, Anabela Correia & Graça Fernandes. 2004. Frozen sentences of Portuguese: Formal descriptions for NLP. Em *Workshop on Multiword Expressions: Integrating Processing*, 72–79.

Baptista, Jorge, Graça Fernandes, Rui Talhadas, Francisco Dias & Nuno Mamede. 2015. Implementing European Portuguese verbal idioms in a natural language processing system. Em *Conference of the European Society of Phraseology (EuroPhras)*, 102–115

Baptista, Jorge, Nuno Mamede & Ilia Markov. 2014. Integrating verbal idioms into an NLP system. Em *Computational Processing of the Portuguese Language (PROPOR)*, 250–255. 10.1007/978-3-319-09761-9_28

Baptista, Jorge, Nuno Mamede & Sónia Reis. 2022. Support verb constructions across the ocean sea. Em *18th Workshop on Multiword Expressions*, 26–36.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner & Amalia Todirascu. 2017. Multiword expression processing: A Survey. *Computational Linguistics* 43(4). 837–892. 10.1162/COLI_a_00302

Correia, José, Jorge Baptista & Nuno Mamede. 2016. Syntax deep explorer. Em *Computational Processing of the Portuguese Language (PROPOR)*, 189–201. 10.1007/978-3-319-41552-9_19

Fotopoulou, Aggeliki. 1993. *Une classification des phrases a complèments figés en grec moderne: étude morphosyntaxique des phrases figées*: Université Paris VII. Tese de Doutoramento

Fukova, Tatyana. 2016. *Lexicon-grammar of Russian verbal idioms*. Faro, Portugal: Universidade do Algarve. Tese de Mestrado

Galvão, Ana, Jorge Baptista & Nuno Mamede. 2019. Processing European Portuguese verbal idioms: From the lexicon-grammar to a rule-based parser. Em *Computational and Corpus-based Phraseology (EuroPhras)*, 70–77. 10.26615/978-2-9701095-6-3_009

Gross, Maurice. 1982. Une classification des phrases «figées» du français. *Revue québécoise de linguistique* 11(2). 151–185. 10.7202/602492ar

Gross, Maurice. 1996. Lexicon-Grammar. Em Keith Brown & Jim Miller (eds.), *Concise Encyclopedia of Syntactic Theories*, 244–259. Pergamon

Haagsma, Hessel, Johan Bos & Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. Em *12th Language Resources and Evaluation Conference (LREC)*, 279–287.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. MIT Press

Kourtin, Asmaa, Asmaa Amzali, Mohammed Mourchid, Abdelaziz Mouloudi & Samir Mbarki. 2021. Lexicon-grammar tables for modern Arabic frozen expressions. Em *15th International NooJ Conference*, 28–38. 10.1007/978-3-030-92861-2_3

Krippendorff, Klaus. 2008. Systematic and random disagreement and the reliability of nominal data. *Communication Methods and Measures* 2(4). 323–338. 10.1080/19312450802467134

Madabushi, Harish Tayyar, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart & Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual idiomaticity detection and sentence embedding. Em *16th International Workshop on Semantic Evaluation (SemEval)*, 107–121. 10.18653/v1/2022.semeval-1.13

Mamede, Nuno, Jorge Baptista, Cláudio Diniz & Vera Cabarrão. 2012. STRING: A hybrid statistical and rule-based natural language processing chain for Portuguese. Em *Computational Processing of the Portuguese Language (PROPOR) — Demo Session*,

Manning, Christopher D. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press

Mogorrón Huerta, Pedro. 2020. La polysémie dans les constructions verbales figées de l'espagnol. *Lingvisticæ Investigationes* 43(2). 241–264. 10.1075/li.00048.mog

Pasquer, Caroline, Agata Savary, Carlos Ramisch & Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? Em *28th International Conference on Computational Linguistics*, 3333–3345. 10.18653/v1/2020.coling-main.296

Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya & Abigail Walsh. 2018a. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. Em *Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG)*, 222–240.

Ramisch, Carlos, Renata Ramisch, Leonardo Zilio, Aline Villavicencio & Silvio Cordeiro. 2018b. A corpus study of verbal multiword expressions in Brazilian Portuguese. Em *Computational Processing of the Portuguese Language (PROPOR)*, 24–34. 10.1007/978-3-319-99722-3_3

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. Em *Computational Linguistics and Intelligent Text Processing (CICLing)*, 1–15. 10.1007/3-540-45715-1_1

Salton, Giancarlo, Robert Ross & John Kelleher. 2014. Evaluation of a substitution method for idiom transformation in statistical machine translation. Em *10th Workshop on Multiword Expressions (MWE)*, 38–42. 10.3115/v1/W14-0806

Savary, Agata, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta & Voula Giouli. 2019a. Literal occurrences of multiword expressions: rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics* 112. 5–54. 10.2478/pralin-2019-0001

Savary, Agata, Silvio Ricardo Cordeiro & Carlos Ramisch. 2019b. Without lexicons, multiword expression identification will never fly: A position statement. Em *Joint Workshop on Multiword Expressions and WordNet (MWE-WN)*, 79–91. 10.18653/v1/W19-5110

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. Em *13th Workshop on Multiword Expressions (MWE)*, 31–47. 10.18653/v1/W17-1704

de Uzeda Garrão, Milena & Maria Carmelita P Dias. 2001. Um estudo de expressões cristalizadas do tipo V+SN e sua inclusão em um tradutor automático bilíngüe (português/inglês). *Cadernos de Tradução* 2(8). 165–182.

Vale, Oto Araújo. 2001. *Expressões cristalizadas do Português do Brasil: uma proposta de tipologia*: Universidade Estadual Paulista. Tese de Doutoramento

Vietri, Simonetta. 2014. The lexicon-grammar of Italian idioms. Em *Workshop on Lexical and Grammatical Resources for Language Processing*, 137–146. 10.3115/v1/W14-5817