UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS EXATAS, NATURAIS E DA SAÚDE

DEPARTAMENTO DE COMPUTAÇÃO
CIÊNCIA DA COMPUTAÇÃO

Daniel Hand Santiago

MELHORIAS DE GRAMÁTICAS LOCAIS PARA RECONHECIMENTO DE EXPRESSÕES CRISTALIZADAS

ALEGRE - ES

2023

MELHORIAS DE GRAMÁTICAS LOCAIS PARA RECONHECIMENTO DE EXPRESSÕES CRISTALIZADAS

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

por

Daniel Hand Santiago

Orientador
Juliana Pinheiro Campos Pirovani

Universidade Federal do Espírito Santo

ALEGRE - ES

2023

DANIEL HAND SANTIAGO

MELHORIA DE GRAMÁTICAS LOCAIS PARA RECONHECIMENTO DE EXPRESSÕES CRISTALIZADAS

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 03 de fevereiro de 2023.

Profa. Dra. Juliana Pinheiro Campos Pirovani
Universidade Federal do Espírito Santo
Orientadora

Prof. Dr. Bruno Vilela Oliveira
Universidade Federal do Espírito Santo

gov.br Documento assinado digitalmente
ELIAS SILVA DE OLIVEIRA
Data: 20/02/2023 17:16:32-0300
Verifique em https://verificador.iti.br

Prof. Dr. Elias Silva de Oliveira
Universidade Federal do Espírito Santo

ALEGRE - ES
FEVEREIRO 2023

SUMÁRIO

1 Introdução **9**

1.1 O problema e sua importância 11

1.2 Objetivos 13

1.2.1 Objetivos Gerais 13

1.2.2 Objetivos Específicos 13

2 Revisão de Literatura **14**

2.1 Tábuas do Léxico-Gramática 14

2.2 Unitex 15

2.3 Gramáticas Locais e Grafos Parametrizados 16

2.4 Trabalhos Correlatos 19

3 Metodologia **22**

3.1 Análise dos Resultados das GLs anteriores e das ECs 22

3.2 Adaptações no Grafo Parametrizado 25

4 Resultados e Discussão **29**

5 Considerações Finais **32**

Referências **33**

Apêndice A – Primeira versão do Grafo Parametrizado para tábua EC1 . . . . . **36**

Apêndice B – Segunda versão do Grafo Parametrizado para tábua EC1 . . . . . **37**

Apêndice C – Terceira versão do Grafo Parametrizado para tábua EC1 38

Apêndice D – Quarta versão do Grafo Parametrizado para tábua EC1 39

LISTA DE FIGURAS

Figura 1 Tábua do Léxico Gramática - EC2. Fonte: (PICOLI, 2020) 15

Figura 2 Interface Gráfica do Unitex 15

Figura 3 Exemplo de Grafo no Unitex 16

Figura 4 Grafo Parametrizado para a tabela EC2 18

Figura 5 Menu Léxico-Gramática no Unitex 18

Figura 6 Menu Compilar Léxico-Gramática para GRF 19

Figura 7 Grafo que reconhece sintagmas nominais. 24

Figura 8 Subgrafo de Sintagma Nominal para sujeitos *Nhum* 26

Figura 9 Subgrafo de Sintagma Nominal para sujeitos *Nnhum* 27

LISTA DE TABELAS

Tabela 1 Resultados Obtidos com as GLs originais. 26

Tabela 2 Resultados Obtidos com as diferentes versões da GL EC1. 28

Tabela 3 Resultados Obtidos com as GLs modificadas para este trabalho. . . . 29

LISTA DE SIGLAS

PLN	Processamento de Linguagem Natural
GL	Gramáticas Locais
PLN	Processamento de Linguagem Natural
EC	Expressões Cristalizadas
CVS	Construção com Verbo Suporte
GL	Gramáticas Locais
IDE	Integrated Development Environment
EMP	Expressões Multi-Palavra

RESUMO

O Processamento de Linguagem Natural (PLN) é uma subárea da Ciência da Computação, juntamente com a Linguística, que tem como objetivo o estudo da geração, representação e compreensão da língua natural por computadores. O PLN encontra, com a complexidade das linguagens humanas, diversas dificuldades e obstáculos. Entre estes, podemos citar as Expressões Cristalizadas. Expressões Cristalizadas são conjuntos de palavras cujo significado não pode ser extraído diretamente pela compreensão de suas partes de maneira individual como *Esta história não tem pé nem cabeça*. Uma máquina não consegue extrair seus significados diretamente e essas informações podem ser de grande utilidade para o PLN. Tendo isso em vista, reconhecer ECs de maneira automática em textos de escrita livre pode ser um passo importante para o PLN. Este trabalho teve como objetivo o estudo e a adaptação de Gramáticas Locais (GL) para reconhecimento de Expressões Cristalizadas em textos em Português e, como produto deste, as GLs adaptadas são capazes de reconhecer expressões com maior número de variações. O aprendizado deste pode ser utilizado para o reconhecimento de outros tipos de expressões multi-palavras, como Construções com Verbo Suporte.

Palavras-chave: processamento de linguagem natural; expressões cristalizadas; gramáticas locais; tábuas do léxico-gramática.

1 INTRODUÇÃO

Processamento de Linguagem Natural (PLN) é um campo de estudo multidisciplinar da Ciência da Computação e Linguística que estuda a geração e compreensão automática de linguagem natural por computadores. Essa área tem ganhado destaque cada vez maior no meio acadêmico e comercial devido à quantidade de benefícios gerados por suas aplicações.

A importância da pesquisa em PLN se tornou maior à medida que, com o avanço da internet, uma quantidade maior de informação se tornou disponível na forma textual. Grande parte desses textos encontra-se em escrita livre (ou seja, não estruturada) e precisa de tratamento para ser utilizada por aplicações que buscam informações específicas a partir deles. E eis que, neste ponto, as limitações dos computadores e da forma que eles processam dados ficam mais evidentes.

> O Processamento de Linguagem Natural tem sido um dos santos graais da ciência da computação. Como é evidente que os computadores são melhores que os humanos em dar sentido a informações altamente estruturadas, mas algumas áreas, por exemplo, entender o significado contextual da linguagem, os humanos são inegavelmente melhores que as máquinas. (KUMAR, 2017).

As máquinas encontram, dada a complexidade da linguagem natural humana, muitos desafios e barreiras que atrasam ou impedem o processamento de todas as informações presentes nos textos de escrita livre. Entre os principais problemas enfrentados, destacam-se a interpretação de contextos, ambiguidade no significado e sentido das palavras e a existências de expressões complexas, como é o caso das Expressões Cristalizadas (EC).

Segundo Picoli (2020), ECs podem ser entendidas como uma estrutura formada por mais de uma palavra que pode ser interpretada como uma entrada lexical. Por exemplo, a expressão *Sicrano não tem papas na língua*, onde *não ter papas na língua* pode ser interpretado como uma entrada lexical com o sentido de *ser tagarela*. Em outras

palavras, ECs são expressões das quais não é possível extrair seu significado apenas somando o significado individual das suas partes. Outro exemplo de EC é *ter o olho maior que a barriga*, como em *"Fulano tem o olho maior que a barriga"*. Essa expressão, se analisada apenas pela soma dos significados de seus termos e de maneira literal, leva a um entendimento diferente do que geralmente é atribuído à frase. Essa construção admite um significado metafórico de que Fulano tem gula ou de que Fulano se alimentou mais do que deveria. A extração do significado metafórico da expressão só é possível com pelo menos certo nível de conhecimento linguístico e/ou cultural específico da língua portuguesa brasileira e uma máquina não seria capaz de fazê-lo sem que haja o fornecimento das informações necessárias para tal.

É válido destacar que as ECs, além da não composicionalidade, se distinguem de outras construções livres por apresentar expressividade, não produtividade e pela existência de argumentos livres (CARNEIRO, 2016a). Além disso, elas se diferenciam também das Construções com Verbo Suporte (CVS) pelo fato de não admitirem construção sem verbos, o que as CVS admitem, como mostrado em Picoli (2020).

O reconhecimento de expressões como ECs (e também CVS) pode beneficiar muitas aplicações de PLN, como ferramentas que trabalham com tradução e geração de resumos de textos de maneira automática e afins. Em Oliveira, Spalenza e Pirovani (2020), os autores dissertam sobre o desenvolvimento de um *chatbot*, capaz de interpretar informações de um livro ou texto e gerar perguntas envolvendo entidades nomeadas encontradas nele, criando uma base de dados para responder tais perguntas. Muitas outras pesquisas na área de sistemas conversacionais inteligentes (RAPP; CURTI; BOLDI, 2021), (REDDY; CHEN; MANNING, 2019), (GRUDIN; JACQUES, 2019), (IO; LEE, 2017) para citar alguns, exploram construções mais complexas, indo além do casamento de padrões para a produção de sistemas que responde de maneira mais adequada às interações humanas.

Acrescentar as informações (em especial o seu significado) presentes nas ECs a estes sistemas poderia auxiliar em diversos aspectos, tanto como forma de acrescentar conhecimento existente àqueles presentes nas aplicações, quanto para permitir que essas mesmas aplicações consigam se aproximar mais da maneira como humanos se comunicam, utilizando e interpretando expressões mais próximas às falas reais entre humanos.

ECs podem ser descritas no formato de tábuas (ou tabelas) do Léxico-Gramática (GROSS, 1975). Uma tábua é um arquivo que descreve um determinado conjunto

de expressões (não somente cristalizadas). O Léxico-Gramática é tanto um método quanto uma prática. Ele prevê que os dados estejam dispostos em tabelas nas quais possam se relacionar os itens lexicais com as propriedades. A utilidade dessas tábuas se dá na constituição de uma base de informações sintático-semânticas e pelo fator cumulativo, o que significa que uma única tábua pode conter inúmeras expressões com construção semelhante, inclusive, se aplicadas formalizações e padronizações, por diferentes pesquisadores. Picoli (2020) apresenta as tábuas descrevendo propriedades sintático-semânticas de um conjunto de ECs. O formalismo presente nas tábuas permite que ferramentas como o Unitex (PAUMIER, 2021) extraia as informações presentes nelas e, com o auxílio de outras tecnologias (neste caso as Gramáticas Locais), gere dados úteis para outras aplicações e pesquisas.

As Gramáticas Locais (GL) (GROSS, 1997) são uma ferramenta de reconhecimento de expressões em texto. Segundo Gross (1999) "Gramáticas Locais são gramáticas de estados finitos ou autômatos de estados finitos que representam conjuntos de expressões de uma língua natural." Essas gramáticas são, em sua base, regras construídas manualmente para reconhecer expressões com características sintático-semânticas em comum. Seu uso não se limita a ECs, mas abrange várias outras áreas de pesquisa, como as de Reconhecimento de Entidades Nomeadas (PIROVANI; OLIVEIRA, 2015), (PIROVANI; OLIVEIRA, 2021). A construção dessas GLs pode ser realizada na ferramenta Unitex[1].

No trabalho de Santiago et al. (2022) foram construídas GLs para reconhecimento de ECs presentes em Picoli (2020). Essas gramáticas foram testadas em dois *corpora*, um composto por um conjunto de textos publicados pelo jornal do Espirito Santo A Tribuna[2] e outro composto por um conjunto de *tweets*, retirados da plataforma Twitter[3] obtidos usando *APIs* fornecidas pela própria empresa e com o tema "Olimpíadas". Os resultados obtidos mostraram que o uso do método aplicado é viável e que o estudo é promissor. Contudo, alguns problemas foram observados.

1.1 O problema e sua importância

Reconhecer ECs em *corpora* de maneira a produzir informações relevantes é uma tarefa difícil. Pela sua natureza de conter partes que são fixas mas outras que podem

[1] https://unitexgramlab.org/pt
[2] http://pdf.redetribuna.com.br/
[3] https://twitter.com.br

variar, uma busca pela expressão em textos pode, ao mesmo tempo, não encontrar todas as ocorrências e trazer muitos casos que tem estrutura de ECs mas não são. E a busca pela parte fixa não fornece todas as informações importantes para os casos em que o reconhecimento de ECs pode ser utilizado. Por exemplo, para extração de informações de textos não é suficiente apenas encontrar as ECs, mas também seu contexto e a composição da oração em que ela se encontra.

As etapas do processo que teve como resultado as gramáticas em Santiago (2022) demonstraram precisar de adaptações e melhorias. Alguns dos problemas encontrados, entretanto, não estavam ligados às gramáticas geradas, e sim à certas propriedades das ferramentas utilizadas. As tábuas do Léxico-Gramática permitem a seus autores destrinchar as sentenças que formam ECs e incluir nessas tabelas possíveis flexões nas expressões (por ex.: uma expressão pode ser descrita com mais de um verbo e ainda ter o mesmo significado) e o Unitex conta com recursos que permitem reconhecer vocábulos com certo nível de flexibilização (por ex.: reconhecer um mesmo verbo em diferentes tempos verbais). Esses recursos trazem benefícios para o reconhecimento de ECs, porém trazem também alguns pontos negativos. Dos resultados analisados pôde-se notar que a flexibilização de todos os termos, mesmo quando feita em suas devidas classificações (verbos, preposições, etc), faz com que muitos falso-positivos sejam reconhecidos, expressões que para o computador eram estruturadas da maneira das ECs, mas que na verdade não são ECs.

Um exemplo de falso-positivo encontrado diversas vezes pelas GLs é a expressão *fora de casa*, termo bastante utilizado em temas relacionados ao futebol, como em "Tal time enfrentou outro time *fora de casa*". Nesse exemplo, o Unitex apresenta para cada palavra todas as possíveis classificações gramaticais e, ao fazê-lo, a palavra *fora* pode ser classificada tanto como advérbio, mas também como a conjugação do verbo *ser* na 1ª ou 3ª pessoa do singular do mais-que-perfeito do indicativo. Essa segunda classificação faz com que a frase configure a estrutura da EC *ser de casa*, o que não está correto.

Como o material obtido pela execução das GLs em *corpora* são textos anotados com as informações desejadas das ECs, ter maior precisão quanto ao reconhecimento de mais expressões e ter menos falso-positivos garante uma melhor base de conhecimento na forma textual a ser utilizada por outras aplicações e pesquisas que se beneficiam delas. Além disso, melhorias nas ferramentas e modelos atuais podem facilitar a obtenção de mais ECs em trabalhos futuros, tendo como conhecimento os possíveis

pontos de falhas e as adaptações que trazem mais precisão ao processo, uma vez que para criação de outras GLs seria necessário apenas pequenas mudanças para adequação às tábuas.

1.2 Objetivos

1.2.1 Objetivos Gerais

O objetivo geral foi melhorar GLs geradas em Santiago (2022) para o reconhecimento de ECs em *corpora* em Português.

1.2.2 Objetivos Específicos

1. Analisar os resultados obtidos pelas GLs de Santiago (2022) identificando possibilidades de melhorias no processo
2. Adaptar os grafos parametrizados construídos em Santiago (2022) com o objetivo de melhorar o reconhecimento de ECs
3. Gerar novamente as GLs a partir dos grafos parametrizados adaptados e tábuas de Picoli (2020)
4. Avaliar o desempenho das novas GLs após aplicá-las aos *corpora*

2 REVISÃO DE LITERATURA

Esta seção apresenta a metodologia aplicada em Santiago (2022) no processo de geração de GLs para o reconhecimento de ECs e que foi alvo de melhorias neste trabalho.

2.1 Tábuas do Léxico-Gramática

O Léxico-Gramática utiliza tabelas (ou tábuas) para descrever propriedades de entrada em um conjunto de entradas lexicais. Elas podem ser usadas por linguistas e pesquisadores de outras áreas para descrever grande parte dos fenômenos linguísticos. As tábuas constituem um recurso linguístico que pode ser utilizado em aplicações de PLN. A descrição linguística para essa finalidade exige um formalismo, que no Léxico-Gramática é representado por fórmulas codificadas de acordo com convenções que são, portanto, reutilizáveis. O padrão adotado pelo Unitex, documentado em PAUMIER (2021), possui dois eixos: horizontal (linhas), que contém as entradas lexicais como verbos, ECs, entre outros; e vertical (colunas), que corresponde às propriedades sintático-semânticas (RASSI et al., 2008).

A Figura 1 representa uma das tábuas retiradas de Picoli (2020). Essa tábua é um arquivo no formato .csv, formato este usado para adequação aos padrões do Unitex. A primeira linha contém a descrição das propriedades que serão mostradas nas linhas abaixo. Cada linha, a partir da segunda, equivale a uma expressão. O que diferencia uma expressão de outra são as informações presentes em cada uma das colunas. Essas colunas podem ser divididas em dois grupos: o que contém apenas a informação se uma propriedade é utilizada e o que contém qual léxico é utilizado na estrutura da EC em questão. O primeiro funciona de forma binária: quando a entrada lexical admite determinada propriedade seu valor é + (positivo) e caso não admita seu valor é - (negativo). O segundo grupo de colunas admite termos que podem compor as expressões.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	N0=:Nhum	N0=:Nnhum	ser	estar	estar (prep)	ficar	ter	permanecer	continuar	continuar (prep	tornar (se)	virar	acabar	terminar	entrar	sair	DET=:E	ADJ/ADV	N	PREP/CONJ	N	Prep	N1=:Nhum	N1=: F	Comparação	Intensificação	Significado	Exemplo
2	+	-	-	+	-	+	-	-	-	-	-	-	+	+	+	+	+	<E>	cara	a	cara	com	+	-	-	-	perto	João ficou cara a cara com Ana
3	-	+	+	-	-	+	-	-	+	-	+	-	-	-	-	-	+	grego	<E>	<E>	<E>	para	+	-	+	-	difícil	Matemática é grego para mim
4	+	-	+	-	-	-	-	-	-	-	+	+	-	-	-	-	+	<E>	peixe	<E>	<E>	de	+	-	-	-	favorito	João é peixe de alguém
5	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	<E>	cara	<E>	<E>	para	-	+	+	+	ousadia	João tem cara para mentir

Figura 1: Tábua do Léxico Gramática - EC2. Fonte: (PICOLI, 2020)

2.2 Unitex

O Unitex (PAUMIER, 2021) engloba um conjunto de *software* livre de PLN que usa tecnologias baseadas em autômatos de estados finitos e recursos linguísticos (como dicionários e gramáticas eletrônicos) disponíveis em mais de 20 línguas. Ele fornece ao usuário uma *Integrated Development Environment* (IDE) que facilita a criação e aplicação de recursos linguísticos por meio de uma interface gráfica (Figura 2), facilitando o desenvolvimento de tais recursos. Vale destacar a possibilidade de execução das ferramentas via linha de comando, funcionalidade bastante útil quando se trata de aplicação dos recursos repetidas vezes e em arquivos de texto diferentes.

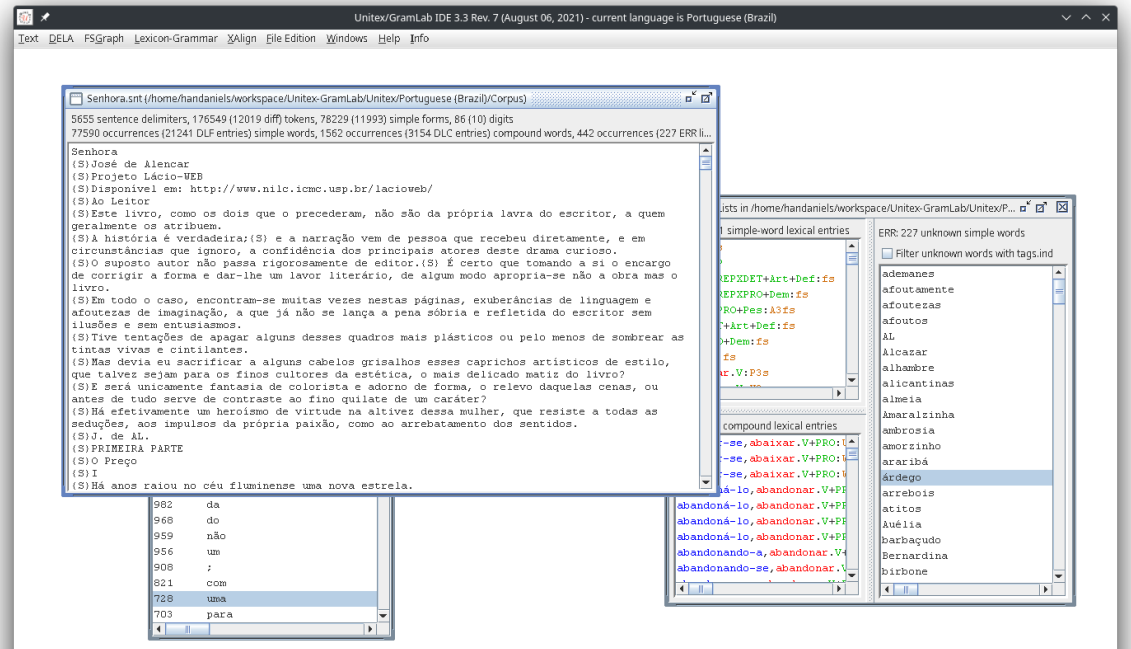


Figura 2: Interface Gráfica do Unitex

Para a geração de GLs a partir de tábuas do Léxico-Gramática são necessários dois recursos do Unitex: o de criação de grafos parametrizados para moldar o formato das GLs e o de compilar o grafo parametrizado junto a uma tábua e, assim, gerar uma gramática (um arquivo no formato *.grf*) com sub-gramáticas que reconhecem as expressões, especificamente para cada expressão (linha) presente na tábua.

2.3 Gramáticas Locais e Grafos Parametrizados

Introduzidas por Maurice Gross na década de 1990, as GLs são autômatos de estados finitos capazes de reconhecer expressões em textos. A representação no Unitex se dá visualmente como um grafo direcionado e pode ser demonstrado na Figura 3. As informações a serem reconhecidas estão presentes nos estados do autômato e esses estados podem ter uma ou mais palavras ou expressões a serem reconhecidas. Um estado pode admitir mais de uma entrada, sendo que é necessário apenas que uma das informações presentes no estado seja encontrada. Um grafo pode conter, também, outros subgrafos como estados internos e ainda recursos como anotação no texto em caso da localização do padrão desejado.

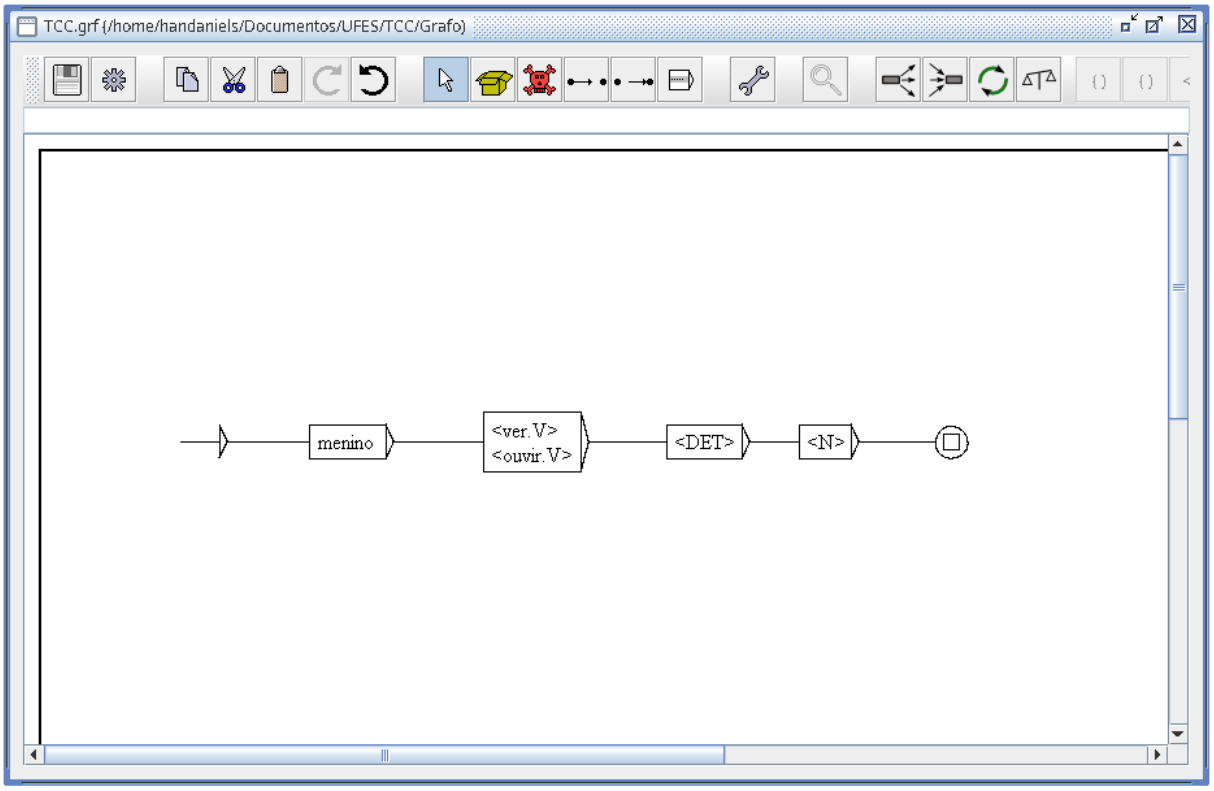


Figura 3: Exemplo de Grafo no Unitex

O Unitex, ao aplicar o grafo, percorre as palavras do texto e tenta combinar a palavra selecionada com o primeiro estado do(s) grafo(s) aplicados. Caso a combinação aconteça, a análise passa para um dos possíveis estados seguintes do grafo e a próxima palavra do texto, sucedendo até o último estado onde é confirmado ou, em caso de não haver estados combinados com a palavra analisada, o término da combinação para o trecho do texto. Neste último caso pode-se imaginar, a fim de se obter um maior entendimento da funcionalidade que, para cada estado do autômato, há uma transição que o leva para um estado final, com resultado da busca da expressão como sendo negativo.

A Figura 3 ilustra um simples grafo que reconhece expressões que configurem a seguinte estrutura: contenham a palavra menino, seguida ou do verbo *ouvir* ou do verbo *ver* (incluindo suas flexões de tempo verbal), um determinante e um substantivo. Um exemplo de expressão reconhecida por este grafo seria "O *menino viu o carro*". Caso esse grafo fosse aplicado a algum texto, o Unitex tentaria encontrar o casamento de cada palavra do texto com o primeiro estado do grafo, ou seja, a palavra *menino*. Ao encontrar a primeira ocorrência, a ferramenta passaria a analisar as seguintes palavras do texto e compará-las com os seguintes estados do grafo. Nota-se que existem diferentes formas de descrever a palavra a ser reconhecida nos estados. É possível apenas usar a palavra, como no estado *menino*. Neste caso, a única variação a ser reconhecida é se houver letras maiúsculas na palavra analisada no texto. O segundo estado apresenta os dois verbos no infinitivo, cercados pelos sinais < (menor) e > (maior) indicando um código lexical pré-definido no Unitex e seguida de "*.V*". Isso indica que o Unitex reconhecerá palavras que sejam classificadas como flexões dos verbos que aparecem no código. Os dois últimos estados do grafo representam códigos lexicais que reconhecem respectivamente determinantes ou substantivos.

Para unir as funcionalidade de reconhecimento das GLs e a base de dados das tábuas do Léxico-Gramática por meio do Unitex são utilizados grafos parametrizados. Estes funcionam exatamente como os grafos explicados anteriormente, mas seus estados podem ter conteúdo diferente. Esses grafos são criados de acordo com as tábuas do Léxico-Gramática e com as informações presentes nelas. As informações são extraídas usando os campos da tabela como variáveis, de modo que as colunas são enumeradas com letras maiúsculas precedidas por um caractere @, iniciando com *@A*. Para os estados nos quais as informações são binárias (como as colunas de A à Q mostrado na Figura 1, por exemplo) são verificados os seus conteúdos e, caso os valores sejam positivos, o estado é adicionado à GL, caso contrário, os estados não

são adicionados. Vale notar que, devido à natureza binária do conteúdo, é necessário fornecer a informação presente na tábua, como é o caso dos estados “@A” e “@B”, que são sucedidos de estados que contém as informações relativas a eles. Essa questão não ocorre nos estados referentes ao outro tipo de coluna, pois neles as informações estão dispostas diretamente na tabela e, caso não esteja, o símbolo $<E>$, reconhecido como um estado vazio, será ignorado na GL.

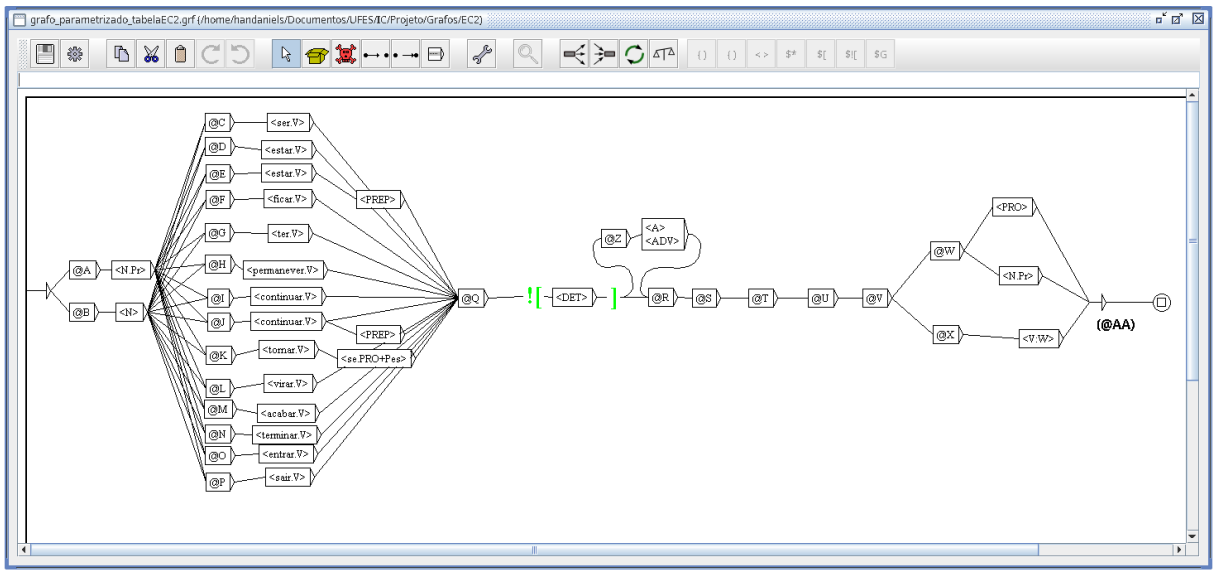


Figura 4: Grafo Parametrizado para a tabela EC2

O Unitex permite abrir as tábuas do Léxico-Gramática para visualização e também permite compilar as informações da tábua aberta para um grafo (ou seja, uma GL).

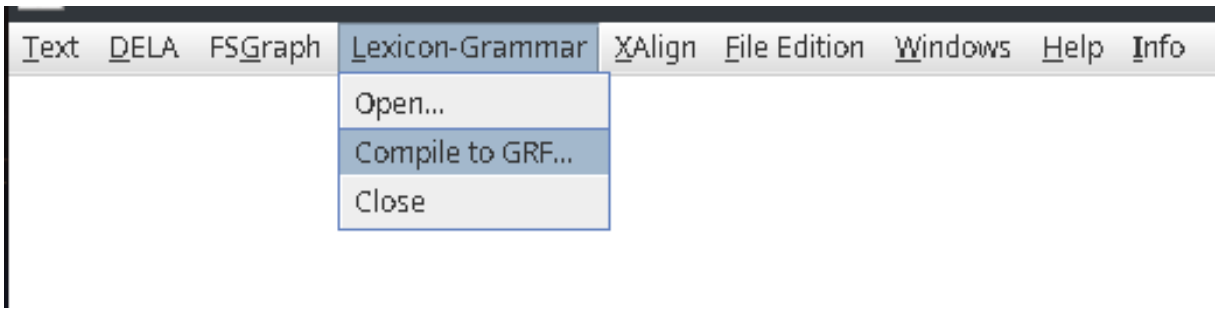


Figura 5: Menu Léxico-Gramática no Unitex

Ao acessar essa segunda opção, é mostrado um menu com três campos, exibido na Figura 6. No primeiro deles, chamado “Grafo de Referência”, deve ser apontado o arquivo do grafo parametrizado criado para a tábua em questão, no formato .grf. No segundo campo deve ser inserido o diretório e o nome da principal GL a ser criada.

Esta conterá um único estado, além do inicial e final, contendo a união de todas as sub-gramáticas geradas, a fim de que todas elas sejam testadas. O último campo, por fim, funciona de maneira semelhante ao anterior, com a exceção de que, após o nome dos subgrafos e antes do formato do arquivo, é necessário o uso dos caracteres *"@%"* para que a GL seja gerada de maneira adequada. Na hora da compilação, para cada entrada lexical presente na tábua, será criado um grafo que a reconhece. Esses grafos criados têm que ser salvos no formato reconhecido pelo Unitex, para isso, o campo *"@%"* é substituído pela ordem das entradas lexicais na tábua. Por exemplo, digamos que uma tábua contenha 10 entradas lexicais e sua GL seja nomeada "gramatica.grf". Neste caso, ao utilizar a formatação do nome de arquivo adequada, preenchendo o campo do nome dos subgrafos com "gramatica@%.grf" serão criados 10 arquivos com os nomes "gramatica001.grf" até "gramatica010.grf".

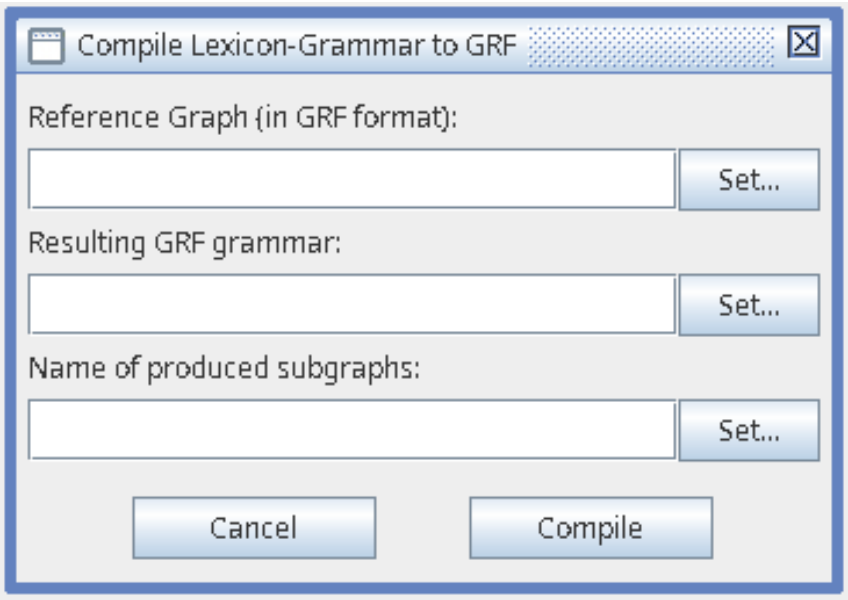


Figura 6: Menu Compilar Léxico-Gramática para GRF

2.4 Trabalhos Correlatos

Expressões multi-palavras (EMP) ou expressões de múltiplas palavras, segundo Constant et al. (2017) "[...] são uma classe de formas linguísticas que abrangem fronteiras convencionais de palavras que são idiossincráticas e abrangentes em diferentes idiomas." O autor disserta sobre como a estrutura de processamento linguístico precisa ser modificada para acomodar o processamento de EMPs e como EMPs levantam muitos desafios para diversas aplicações de PLN. O artigo ainda encara a ausência

de princípios orientadores e propõe a clarificação das interações entre o processamento de EMPs e outras aplicações de PLN.

Há diversas pesquisas com descrições sintático-semânticas de itens lexicais do Português, formalizadas em tábuas, que utilizam o modelo teórico-metodológico do Léxico-Gramática, como Picoli (2015), MALACOSKI (2017) e Barros (2014).

Gross (1993) introduziu exemplos de conjuntos de expressões fixas cujas formas não poderiam ser diretamente relacionadas seguindo regras formais de estrutura de frases ou transformacionais e mostra como o formalismo de autômatos finitos pode ser usado para representar essas expressões de maneira mais natural. O formalismo demonstrado se tornou as Gramáticas Locais (Local Grammar, no inglês) (GROSS, 1997; GROSS, 1999). Seu uso dentro de PLN é deveras amplo, utilizado em diversas aplicações, podendo estar ou não em conjunto com outros recursos computacionais como inteligência artificial, aprendizado de máquina e entre outros como Pirovani e Oliveira (2021) e Pirovani e Oliveira (2015) exemplificam.

Vale (2001) pode ser considerado o primeiro a unir os dois métodos quando criou as primeiras tábuas do Léxico-Gramática descrevendo as ECs, sendo também o principal difusor do termo. Outro trabalho correlato é o de Caminada (2008), que retrata as dificuldades encontradas na pesquisa de PLN na língua portuguesa devido à diferença de quantidade de ferramentas e pesquisadores existentes se comparadas a outras línguas como inglês e alemão. O autor disserta sobre um método de identificação de expressões candidatas à classificação como multi-palavras formados por sintagmas preposicionados. O método de identificação utiliza quatro fórmulas estatísticas classicamente utilizadas nesta tarefa.

É possível, todavia, encontrar trabalhos que, além de conter descrições em tábuas, também apresentam GLs, construídas a partir das tábuas, com a finalidade de identificar automaticamente as propriedades das construções em análise. Rassi (2015) criou grafos para a descrição de Construções com Verbo-Suporte (CVS) *dar*. Santos (2015), por sua vez, construiu grafos de busca no Unitex para encontrar construções com verbo *ter* seguido de predicado nominal. Para encontrar possíveis variantes dos verbos-suporte *receber* e *levar*, Calcia (2016) elaborou grafos de busca. Já Carneiro (2016b) confeccionou grafos no intuito de fazer um levantamento de expressões cristalizadas com *ser* e *estar*.

Diferente dos trabalhos anteriormente citados, as tábuas do Léxico-Gramática retiradas de Picoli (2020), utilizadas em Santiago (2022) e que são o alvo de adaptações

neste, contém ECs com os verbos *ser*, *estar*, *ficar* e *ter*.

3 METODOLOGIA

Esta seção aborda a metodologia utilizada neste trabalho para adaptar as GLs para o reconhecimento de ECs.

3.1 Análise dos Resultados das GLs anteriores e das ECs

O primeiro passo foi o estudo e análise dos resultados de Santiago (2022) e das ECs presentes nas tábuas quanto à sua construção. Os textos anotados em Santiago (2022) foram analisados e cada expressão reconhecida foi estudada no *corpus* original, buscando entendimento do contexto e da construção da expressão.

Nos próprios exemplos escritos nas tábuas, pôde-se notar que a construção das expressões não é completamente cristalizada (ou seja, fixa), podendo admitir outros vocábulos que não estavam presentes nas tábuas. A maioria das expressões admite advérbios em sua construção, como a expressão *Ela nunca teve papas na língua*, que configura corretamente uma EC mesmo não estando na estrutura descrita na tábua.

Outro ponto também avaliado foi a construção do sujeito da oração, onde foram acrescentadas regras para o reconhecimento de sujeitos além de substantivos comuns e próprios e isso se dá pelo entendimento de que os sujeitos humanos (*Nhum*) e não humanos (*Nnhum*) podem abranger outras construções, inclusive construções multipalavras.

A análise começou de maneira manual, apenas olhando o núcleo, ou seja, a parte cristalizada das expressões e buscando outras possíveis construções com ECs que tinham o mesmo sentido e significado das ECs descritas na tábua. Para isso, apenas a tábua EC1 foi analisada por ter maior quantidade de ECs. Em primeira instância, pôde-se notar alguns pontos como a existência de advérbios e outro vocábulos nas expressões, mas como a busca de maneira manual se limitava à uma quantidade

pequena, não foi possível observar uma grande quantidade de variações para as ECs.

Para auxiliar a busca dessas construções foram reunidos *tweets* contendo as ECs. Como a *API* do *Twitter* permitia buscar *tweets* utilizando palavras chaves, foram criados dois *scripts*: um em *shell script* que extraiu a parte imutável das expressões da tábua contendo as ECs e criou um arquivo com essas partes; e outro em *python* que buscou 100 *tweets* com cada uma dessas partes e imprimiu em um arquivo.

O resultado desses dois *scripts* foi um arquivo de texto contendo aproximadamente 100 exemplos de cada EC da tábua EC1. A busca é realizada pela *API* do *Twitter* e esta retorna resultados semelhantes, logo alguns dos *tweets* não continham exatamente as ECs desejadas e isso, apesar de parecer um fator impeditivo, auxiliou na confirmação do entendimento que uma expressão, por mais que esteja na estrutura de uma EC, não necessariamente pode ser classificada como uma EC, podendo ter o significado literal dos vocábulos.

Uma primeira observação, antes mesmo do tratamento das palavras que não estavam sendo consideradas nas expressões, foi o da variação do tipo e da estrutura do sujeito das orações. Em Santiago (2022), o tratamento de sujeitos nas expressões foi realizado de maneira muito limitada, devido ao pouco conhecimento na área e ao entendimento limitado do significado das colunas das tábuas. Na tábua EC1, por exemplo, os sujeitos poderiam ser do tipo *Nhum*, o que significa que seriam sujeitos ”humanos“ ou relacionados a humanas; ou sujeitos *Nnhum*, que são sujeitos não relacionados a coisas humanas diretamente. Esses dois tipos de sujeitos foram inicialmente traduzidos para as regras da GL apenas como substantivos próprios e comuns, respectivamente, e isto limitou os resultados encontrados pelas GLs. Nas variações das ECs encontradas, pôde-se observar sujeitos não necessariamente mais complexos, como pronomes pessoais, não estavam sendo reconhecidos. Sem levar isso em consideração, expressões como *Ela não tem papas na língua* não poderiam ser reconhecidas.

A partir do entendimento dessa limitação foi escolhido usar um subgrafo presente em Picoli et al. (2015), que reconhecia melhor sintagmas nominais. Sintagmas Nominais se referem às partes de uma oração que têm como núcleo um substantivo, como é o caso do sujeito, complemento verbal ou nominal e aposto. A Figura 7 mostra o grafo criado que é capaz de reconhecer sintagmas nominais. Além de reconhecer substantivos, outras opções foram levadas em conta, como pronomes pessoais, além das combinações de substantivos, determinantes, pronomes, advérbios e prepo-

sições. Sem a inclusão desse subgrafo, expressões como *Estar entre os oito primeiros não teria preço* não seriam reconhecidas (neste caso o verbo *estar* não foi reconhecido, apenas a partir da palavra *entre*, porém foi mencionado apenas para dar sentido à expressão, já que o reconhecimento pela GL foi correto).

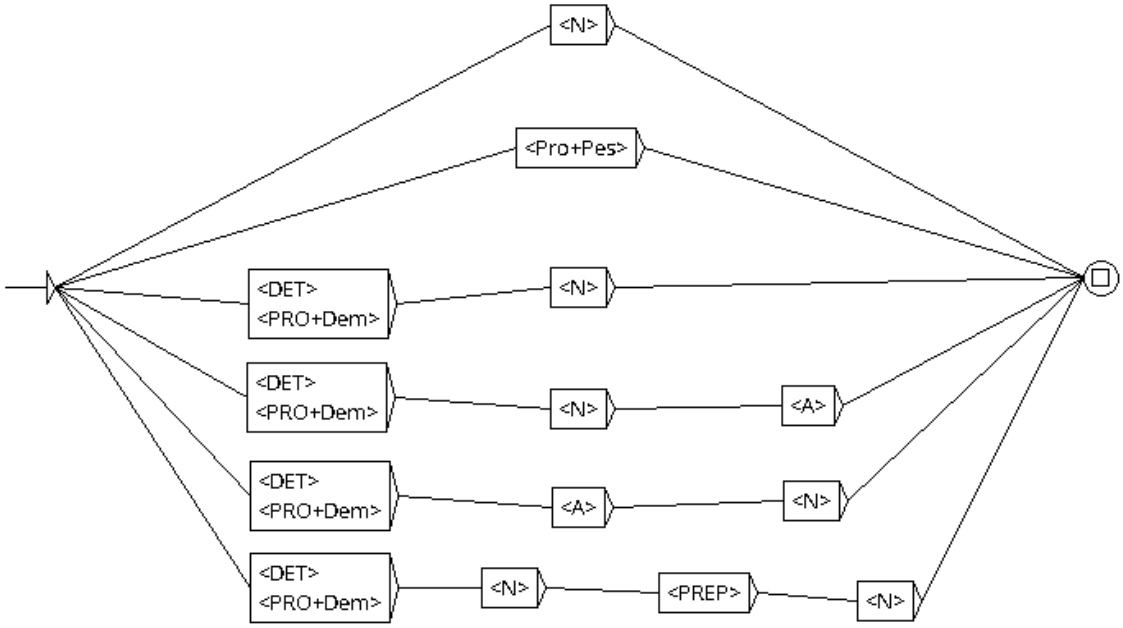


Figura 7: Grafo que reconhece sintagmas nominais.

O passo seguinte foi a análise de quais palavras poderiam aparecer antes do núcleo das ECs. Com a análise dos *tweets*, alguns pontos foram destacados. O primeiro é que os vocábulos mais comuns encontrados estavam relacionados aos verbos: advérbios e preposições. O mais comum foi a utilização dessas duas classes de palavras entre os sujeitos e os verbos, seguindo a sequência: sujeito da oração, podendo ser seguido ou não de uma preposição, que por sua vez pode ser seguido de um advérbio e então o verbo e o restante da oração. Os dois tipos de vocábulos não foram encontrados necessariamente em todas as frases anotadas, podendo aparecer apenas um ou outro. Exemplos desse comportamento seriam a expressão *dar uma colher de chá* que aparece em *Daniel também deu uma colher de chá*, onde o advérbio *também* aparece; ou a expressão *ser mato* em *Falsidade por aqui é mato* onde aparecem tanto a preposição *por* quanto o advérbio *aqui*.

Como a tábua EC1 contém uma coluna de negação obrigatória, pôde-se notar outra falha não percebida no trabalho original. Neste, o estado do grafo parametrizado correspondente à negação obrigatória continha apenas a palavra *não*. Nos exemplos de ECs obtidos que tinham o campo negação obrigatória marcados, percebeu-se que outros advérbios de negação foram utilizados, como *nunca* e *jamais*. A adição desses advérbios poderia aumentar a quantidade de expressões reconhecidas.

3.2 Adaptações no Grafo Parametrizado

Com as informações obtidas no estudo das ECs abordado na seção anterior, o passo seguinte foi a adaptação dos grafos parametrizados. No início, apenas um grafo parametrizado foi criado, com o pensamento de que ao se adicionar os campos que foram entendidos como necessários, a quantidade de expressões encontradas seria maior, sem prejudicar a precisão das GLs, que é calculada pela Fórmula 7. No entanto, viu-se que esse não foi o caso e modificações foram feitas para tentar equilibrar esses dois fatores.

$$\text{Previsão} = \frac{\text{Quantidade de reconhecimentos corretos}}{\text{Quantidade total de reconhecimentos}} \tag{7}$$

Com isto, foram criadas quatro versões do grafo parametrizado para a tábua EC1. Cada versão foi criada após a execução da versão anterior e análise dos falso positivos e de ECs que não foram reconhecidas, tentando corrigir quaisquer detalhes que atrapalharam a precisão das GLs. A Tabela 2 apresenta o resultado da execução das GLs adaptadas no corpus A Tribuna, com cada uma das versões do grafo parametrizado.

A primeira versão (nomeada V1 para simplificação) é apresentada no Apêndice A. Nela foram acrescentados o subgrafo com sintagmas nominais apresentado na Figura 7, o subgrafo contendo os advérbios de negação e estados para as preposições e advérbios antes do verbo. A GL construída a partir deste grafo foi capaz de reconhecer um número maior de expressões, no entanto a quantidade de falsos positivos foi superior, fazendo com que a precisão ficasse abaixo até da primeira versão da GL, construída em Santiago (2022), presentes na Tabela 1.

GL	ECs encontradas	ECs corretas	Falso positivos	Precisão
EC1	64	21	43	32,81%
EC2	4	4	0	100,00%
ECP1	41	11	30	26,83%
ECP2	1	1	0	100,00%

Tabela 1: Resultados Obtidos com as GLs originais.

Com os resultados da primeira versão e com a análise das expressões reconhecidas, foi criada a segunda versão do grafo parametrizado, apresentado no Apêndice B. Nela, o subgrafo de sintagmas nominais foi dividido em dois para acomodar-se aos dois tipos de sujeitos diferentes da tábua EC1 e a inclusão de algumas outras alterações.

O subgrafo apresentado na Figura 8, contém um estado para substantivos próprios, um para pronomes pessoais e um outro subgrafo, que reconhece palavras que se iniciam com letra maiúscula e que não podem ser classificadas pelos dicionários do Unitex. Essa última foi acrescentada devido a um número considerável de expressões que continham sujeitos como sobrenomes, nomes de empresas ou outras palavras que não eram reconhecidas pelo Unitex. A GL construída por este grafo parametrizado, embora tenha obtido uma quantidade total de expressões reconhecidas menor, obteve a melhor precisão, então o resultado foi entendido como melhor.

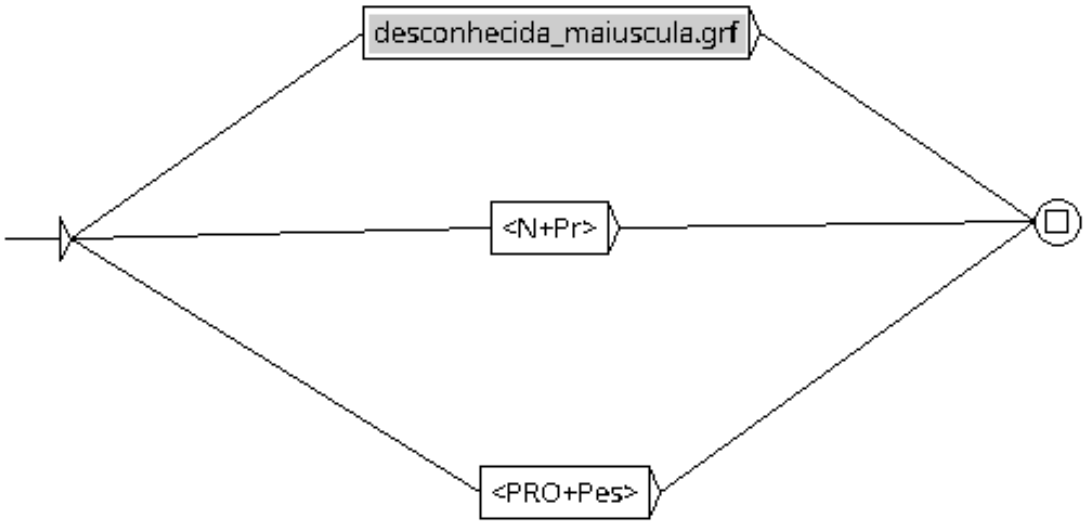


Figura 8: Subgrafo de Sintagma Nominal para sujeitos *Nhum*

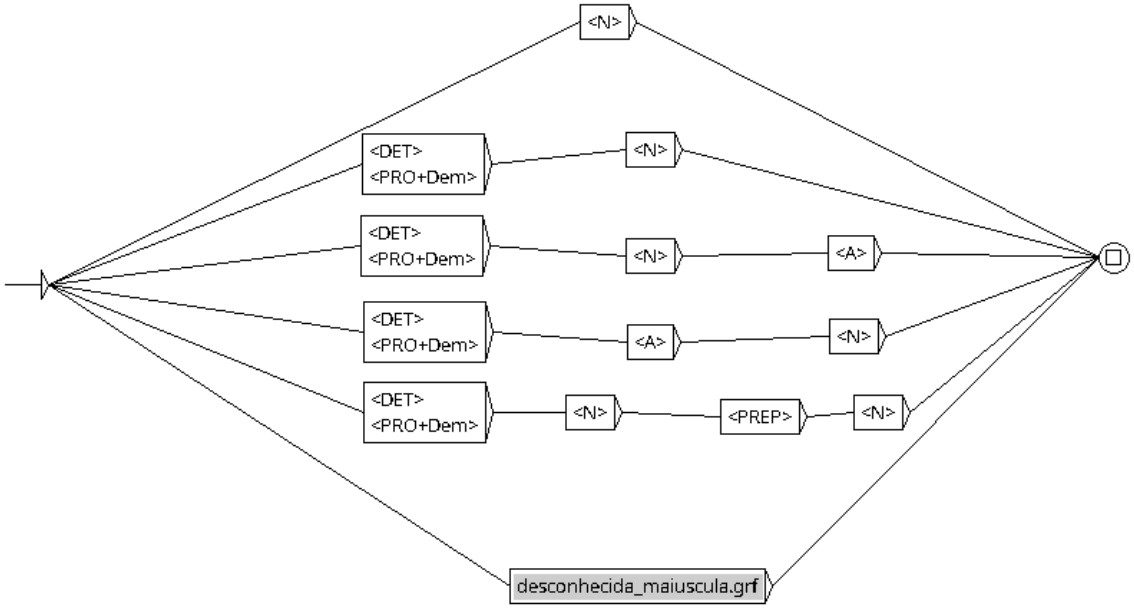


Figura 9: Subgrafo de Sintagma Nominal para sujeitos *Nnhum*

As duas versões posteriores foram feitas como complementos às suas respectivas versões anteriores. A terceira versão, apresentada o Apêndice C, teve como objetivo eliminar ainda mais a quantidade de falso positivos encontrados. Nela houve um retrocesso quanto ao sujeito das orações, tentando eliminar os outros falsos positivos reconhecidos por estes. Pode-se dizer que os subgrafos de sintagmas nominais foram reduzidos para apenas três estados, considerando estes os mais relevantes: substantivos próprios e pronomes pessoais para sujeitos do tipo *Nhum* e substantivos simples para sujeitos do tipo *Nnhum*. O resultado, porém, foi a diminuição de expressões reconhecidas de maneira não proporcional a de falso positivos, o que diminuiu a precisão da GL.

Por fim, com o intuito de alcançar, ou até mesmo superar, a precisão das segunda versão do grafo parametrizado, a quarta versão adaptada foi criada e é apresentada no Apêndice D. Foram adicionados mais alguns dos elementos dos subgrafos de sintagmas nominais de volta ao grafo, sendo eles o subgrafo de palavras desconhecidas iniciadas com letra maiúscula para os sujeitos *Nhum* e um estado que reconhecia apenas palavras desconhecidas ao Unitex para os sujeitos *Nnhum*.

Essa última alteração não teve o resultado esperado, ficando com precisão abaixo da versão anterior (V3). Como mostrado na Tabela 2, a versão que teve maior precisão foi a V2. A partir dessa informação, ela foi escolhida como a melhor, entre as criadas

e avaliadas, em termos de precisão, e então foi utilizada como base e exemplo para a adaptação das GLs para as demais tábuas de Picoli (2020).

GL	ECs encontradas	ECs corretas	Falso positivos	Precisão
V1	178	50	128	28,09%
V2	82	31	51	37,80%
V3	76	28	48	36,94%
V4	104	32	72	30,77%

Tabela 2: Resultados Obtidos com as diferentes versões da GL EC1.

Com a escolha do grafo parametrizado que gerou GLs com maior precisão para a tábua EC1, as mudanças feitas nela foram aplicadas às demais tábuas. O processo consistiu em acrescentar os campos de acordo com as necessidades das tábuas, gerar as GLs e aplicá-las no corpus A Tribuna. Nem todas as alterações foram aplicadas às tábuas, tendo em vista a diferença de estrutura das mesmas (por exemplo, nem todas as tábuas continham propriedade de negação obrigatória). Algumas pequenas modificações foram feitas nos grafos parametrizados após uma primeira verificação dos resultados, apenas a fim de retirar falso positivos muito repetidos ou que se distanciavam de maneira extrema das ECs descritas na tábua. A quantificação dos resultados e a precisão serão abordados na seção seguinte.

4 RESULTADOS E DISCUSSÃO

São apresentados nas Tabelas 1 e 3 os resultados da aplicação das GLs no corpus A Tribuna, considerando respectivamente as versões presentes em Santiago (2022) e neste trabalho. Vale destacar que apenas a precisão foi levada em consideração pelo fato de não existir uma versão desse *corpus* anotada com as ECs para o cálculo da abrangência.

GL	ECs encontradas	ECs corretas	Falso positivos	Precisão
EC1	82	31	51	37,80%
EC2	4	4	0	100,00%
ECP1	72	15	57	20,83%
ECP2	13	11	2	84,62%

Tabela 3: Resultados Obtidos com as GLs modificadas para este trabalho.

A diferença de quantidade de expressões encontradas e da precisão entre as versões original e a alterada variou para cada tábua do Léxico Gramática utilizada. Apenas a GL gerada para a tábua EC2 manteve o mesmo resultado, tanto na quantidade de expressões reconhecidas quanto na precisão. A GL da tábua EC1, que foi utilizada para as modificações (embora analisada em um corpus diferente), teve um aumento na quantidade de ECs encontradas e reconhecidas corretamente, chegando a uma melhora de aproximadamente 5% de precisão.

Passando para a GL gerada para a tábua ECP2, a aplicação da nova GL obteve um número maior de expressões encontradas e, mais importante, de ECs reconhecidas corretamente, saltando de apenas 1 expressão para 13 reconhecidas, sendo 11 destas reconhecidas corretamente. A precisão da GL foi reduzida, mas apenas pelo fato de antes haver uma única expressão reconhecida e esta estar correta. Pode-se entender

que, apesar da precisão ser inferior, o resultado das alterações foi melhor do que a versão anterior.

Considerando o exposto acima, a única tábua que teve o seu resultado pior do que na versão anterior foi a ECP1. A quantidade de expressões reconhecidas corretamente, embora tenha aumentado em comparação com o trabalho anterior, não foi proporcional à quantidade de expressões reconhecidas no geral. O número de ECs encontradas corretamente saiu de 11 para 15, porém a quantidade de falsos positivos foi de 30 para 57, o que reduziu a precisão de 26,83% para 20,83%. A GL da tábua ECP1 foi a com menor precisão entre as GLs do trabalho original e um dos principais fatores de tal foi a estruturação de ECs específicas que pode aparecer em textos e que não são ECs.

Analisando novamente as ECs que mais foram reconhecidas pelas GLs mas que não eram ECs, podemos observar que a que obteve mais quantidade de falso positivos, assim como na versão de Santiago (2022), foi a expressão *ser de casa*. Embora possa ser uma expressão comum no cotidiano, ela tem uma estrutura que pode abranger muitas frases que não têm o mesmo significado, especialmente no contexto jornalístico. Essa expressão, sozinha, foi responsável por 40 dos 57 falso positivos reconhecidos para as GL ECP1. A segunda expressão com mais falso positivos, *estar de cara*, tem 7 falso positivos.

A abordagem utilizada neste trabalho foi capaz de trazer alguns resultados melhores para as GLs na aplicação no corpus A Tribuna, embora os mesmos fiquem longe do esperado. É possível que a aplicação destas GLs em outros gêneros textuais seja diferente do contexto aplicado neste. Não há dúvidas de que muitos fatores podem ser alterados para chegar em resultados possivelmente melhores.

Durante todo o processo da construção e adaptação dos grafos parametrizados para as tábuas do Léxico-Gramática foram encontradas dificuldades. As tábuas, mesmo com adaptações para se adequar à maneira com que o Unitex as usa e para tentar obter resultados melhores, apresentam problemas. Um deles é a estruturação que, embora feita para agrupar quantidades diferentes de ECs em cada uma das quatro tábuas, não suporta todos os campos de todas as expressões. Isso não seria problema se houvesse mais campos, mesmo que vazios em algumas expressões, mas é um grande limitador quando não há campos suficiente para comportar mais variações de uma mesma EC.

A quantidade de falsos positivos encontradas também acentua a ideia de que uma

frase ter a estrutura que configura uma EC não quer dizer que ela seja uma EC. As palavras podem ter mais de uma significado e como o Unitex não consegue fazer a etiquetagem das palavras (ou seja, o *Part-of-Speech tagging* (VOUTILAINEN, 2003)) baseado em contexto, todas as possíveis classificações são utilizadas, o que abre as portas para falso-positivos serem encontrados. Utilizar outra ferramenta que seja capaz de fazer a etiquetagem baseada no contexto poderia resolver parcialmente este problema, embora provavelmente seja necessário realizar uma tradução das etiquetas, já nem todas usam o padrão do Unitex.

A busca e análise de cada expressão e suas variações, como feita para a tábua EC1, é muito custosa. Desde o passo de extrair o núcleo das expressões até a busca por elas nos *tweets* (que por si só já demanda tempo devido à política de limitação de requisições por dia do *Twitter*) é necessário muito tempo e análise. Realizar esta análise para as demais tábuas poderia trazer resultados melhores porém ainda sim essas melhorias estariam limitadas aos problemas anteriores das tábuas e do Unitex.

Buscar as partes fixas das ECs sem analisar a sua estrutura, embora possa parecer mais relevante, pode não ser o suficiente no reconhecimento de ECs. O reconhecimento feito pela metodologia apresentada em Santiago (2022) e adaptada neste trabalho pode ajudar não só aplicações do PLN, mas também ajudar o trabalho de descrição dos linguistas, fornecendo informações que podem ser usadas para a descoberta de novas ECs ou ECs com significados diferentes e ampliando a possibilidade de construções onde as ECs podem aparecer.

5 CONSIDERAÇÕES FINAIS

Este trabalho reafirma que é possível gerar GLs a partir de grafos parametrizados construídos para de tábuas do Léxico-Gramática para o reconhecimento de ECs, mas acredita-se que para evoluir a metodologia atual é necessário mais detalhamento na descrição das expressões, incluindo mais possíveis variações. Essa descrição mais detalhada é possível desde que exista um estudo aprofundado das ECs e, com esses estudos, mais adaptações podem ser realizadas nos grafos parametrizados e resultados ainda melhores podem ser alcançados.

Foram implementadas mudanças para GLs que reconhecem Expressões Cristalizadas, criadas a partir de tábuas do Léxico-Gramática em Santiago (2022). As mudanças trouxeram, mesmo que não em todos os casos, resultados melhores para o reconhecimento de ECs, e também proporcionaram conhecimentos importantes que podem ser aplicados em trabalhos futuros, não só apenas no reconhecimento de ECs. A possibilidade de alterações em outros fatores não abordados neste trabalho, como o pré-processamento e a etiquetagem do texto, também traz a perspectiva de que há espaço para melhorias. Espera-se que o que foi aprendido com este trabalho possa ser usado em outras aplicações, como o reconhecimento de CVSs, também apresentados em Picoli (2020), e em outros trabalhos em que linguistas utilizam as tábuas do Léxico-Gramática.

REFERÊNCIAS

BARROS, C. D. d. Descrição e classificação de predicados nominais com o verbo-suporte fazer no português do brasil. Universidade Federal de São Carlos, 2014. Citado na página 20.

CALCIA, N. P. Descrição e classificação das construções conversas do português do brasil. Universidade Federal de São Carlos, 2016. Citado na página 20.

CAMINADA, N. *Identificação Automática de Expressões Cristalizadas Preposicionais em Corpora da Língua Portuguesa*. 2008. Citado na página 20.

CARNEIRO, A. d. S. Descrição e classificação das epressões cristalizadas com ser e estar do português do brasil. Universidade Federal de São Carlos, 2016. Citado na página 10.

CARNEIRO, A. d. S. Descrição e classificação das expressões cristalizadas com ser e estar do português do brasil. Universidade Federal de São Carlos, 2016. Citado na página 20.

CONSTANT, M. et al. Multiword expression processing: A survey. *Computational Linguistics*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 43, n. 4, p. 837–892, 2017. Citado na página 19.

GROSS, M. *Méthodes en syntaxe: régime des constructions complétives*. [S.l.]: Hermann Paris, 1975. Citado na página 10.

GROSS, M. *Local grammars and their representation by finite automata*. [S.l.]: Harper-Collins, 1993. Citado na página 20.

GROSS, M. 1 the construction of local grammars. *Finite-state language processing*, p. 329, 1997. Citado 2 vezes nas páginas 11 e 20.

GROSS, M. A bootstrap method for constructing local grammars. In: UNIVERSITY OF BELGRAD. *Proceedings of the Symposium on Contemporary Mathematics*. [S.l.], 1999. p. 229–250. Citado 2 vezes nas páginas 11 e 20.

GRUDIN, J.; JACQUES, R. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2019. (CHI '19), p. 1–11. Citado na página 10.

IO, H.; LEE, C. Chatbots and Conversational Agents: A Bibliometric Analysis. In: IEEE. *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. [S.l.], 2017. p. 215–219. Citado na página 10.

KUMAR, V. *Data Mining Techniques for Information Retrieval*. Dissertação (monografia) — National University of Science And Tecnology "MISIS", Moscou, 2017. Citado na página 9.

MALACOSKI, L. *Descrição sintático-semântica de construções com o verbo-suporte ter e nomes humanos de relação para o Processamento Automático de Linguagem Natural. 2017*. Tese (Doutorado) — Dissertação (Mestrado)–Programa de Pós-Graduação em Estudos Linguísticos . . . , 2017. Citado na página 20.

OLIVEIRA, E. d.; SPALENZA, M.; PIROVANI, J. P. rava: A robot for virtual support of learning. In: SPRINGER. *International Conference on Intelligent Systems Design and Applications*. [S.l.], 2020. p. 1238–1247. Citado na página 10.

PAUMIER, S. *Unitex 3.3 User Manual*. 2021. Disponível em: <https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf>. Citado 3 vezes nas páginas 11, 14 e 15.

PICOLI, L. *Descriçao de verbos de base adjetiva derivados com os sufixos-efficere-izar, para o processamento automático de linguagem natural: UFES*. Tese (Doutorado) — Tese de Mestrado, 2015. Citado na página 20.

PICOLI, L. *Contínuo e limite entre expressão cristalizada e construção com verbo-suporte à luz do Léxico-Gramática*. Tese (Doutorado) — Universidade federal de São Carlos, 2020. Citado 10 vezes nas páginas , 9, 10, 11, 13, 14, 15, 20, 28 e 32.

PICOLI, L. et al. Uso de uma ferramenta de processamento de linguagem natural como auxílio à coleta de exemplos para o estudo de propriedades sintático-semânticas de verbos. *Linguamá tica*, v. 7, n. 2, p. 35–44, 2015. Citado na página 23.

PIROVANI, J. P.; OLIVEIRA, E. Studying the adaptation of portuguese ner for different textual genres. *The Journal of Supercomputing*, Springer, v. 77, n. 11, p. 13532–13548, 2021. Citado 2 vezes nas páginas 11 e 20.

PIROVANI, J. P. C.; OLIVEIRA, E. S. de. Extração de nomes de pessoas em textos em português: uma abordagem usando gramáticas locais. In: *Computer on the Beach*. [S.l.: s.n.], 2015. p. 1–10. Citado 2 vezes nas páginas 11 e 20.

RAPP, A.; CURTI, L.; BOLDI, A. The Human Side of Human-Chatbot Interaction: A Systematic Literature Review of Ten Years of Research on Text-Based Chatbots. *International Journal of Human-Computer Studies*, Elsevier, v. 151, p. 102630, 2021. Citado na página 10.

RASSI, A. P. *Descrição, classificação e processamento automático das construções com o verbo'dar'em português brasileiro*. Tese (Doutorado) — Universidade Federal de São Carlos, 2015. Citado na página 20.

RASSI, A. P. et al. Estatuto sintático-semântico do verbo fazer no português escrito do brasil. Universidade Federal de Goiás, 2008. Citado na página 14.

REDDY, S.; CHEN, D.; MANNING, C. D. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 7, p. 249–266, 2019. Citado na página 10.

SANTIAGO, D. et al. *Gramáticas Locais para o reconhecimento de Expressões Cristalizadas em Português*. 2022. Artigo não publicado. Citado na página 11.

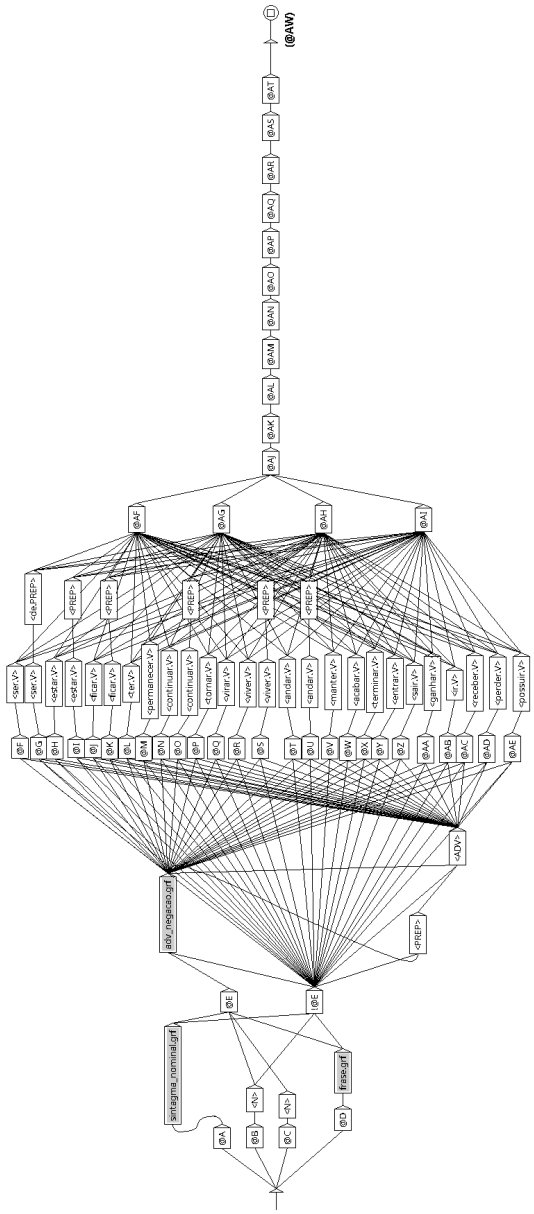
SANTIAGO, D. H. Gramáticas locais para reconhecimento de expressões cristalizadas em português. Relatório Parcial de Pesquisa, Programa Institucional de Iniciação Científica 2021/2022. 2022. Citado 11 vezes nas páginas 12, 13, 14, 20, 22, 23, 25, 29, 30, 31 e 32.

SANTOS, M. C. d. *Descrição dos predicados nominais com o verbo-suporte"ter"*. Tese (Doutorado) — UFSCar, 2015. Citado na página 20.

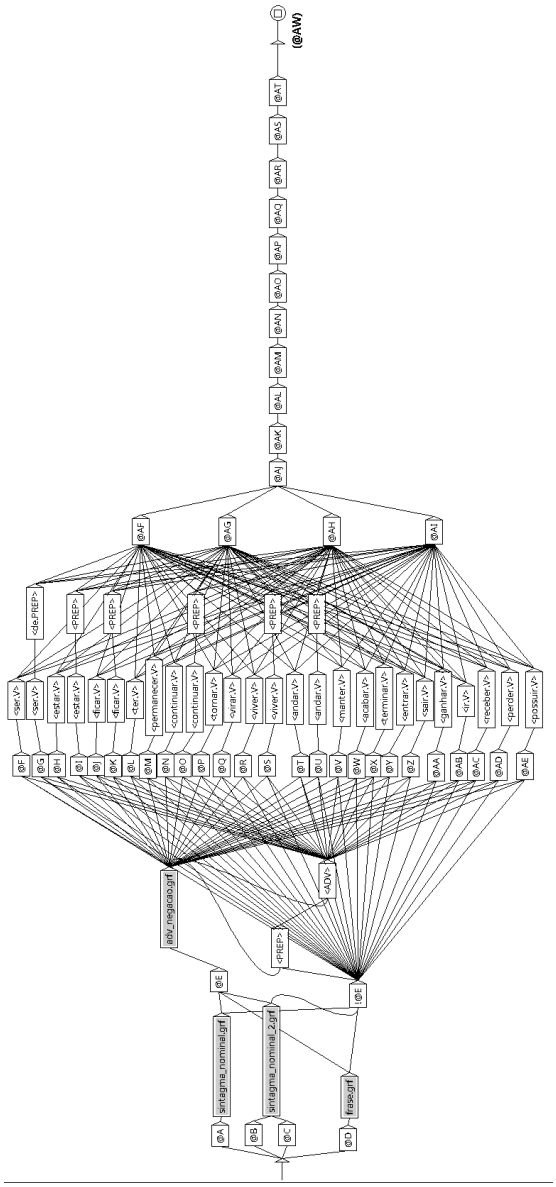
VALE, O. A. Expressões cristalizadas do português do brasil: uma proposta de tipologia. *Araraquara: UEJMF (PhD Thesis)*, 2001. Citado na página 20.

VOUTILAINEN, A. Part-of-speech tagging. *The Oxford handbook of computational linguistics*, University Press, Oxford, p. 219–232, 2003. Citado na página 31.

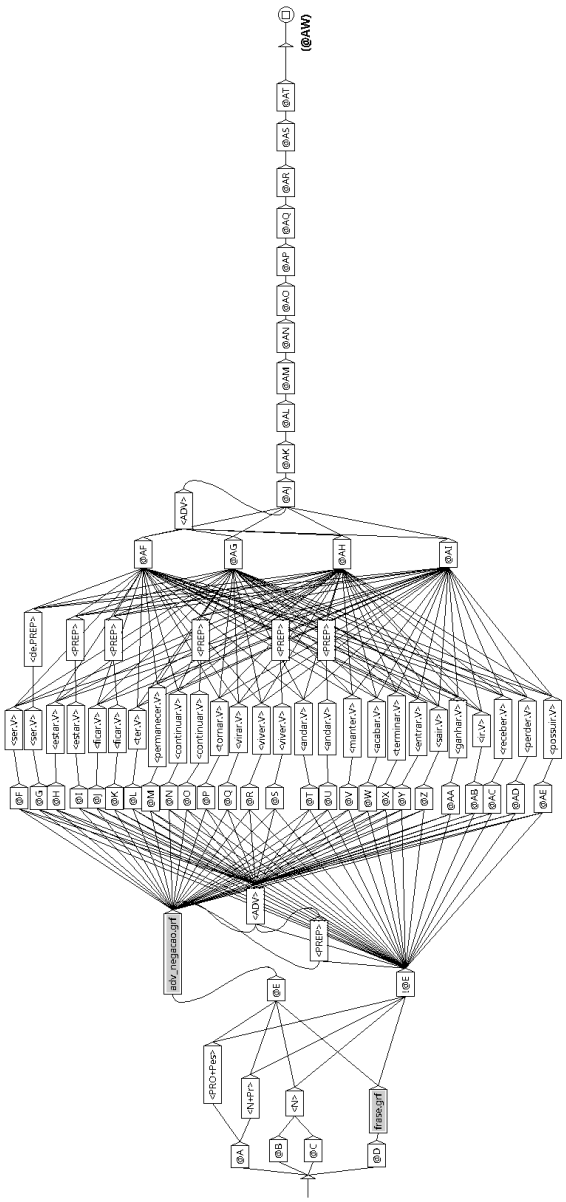
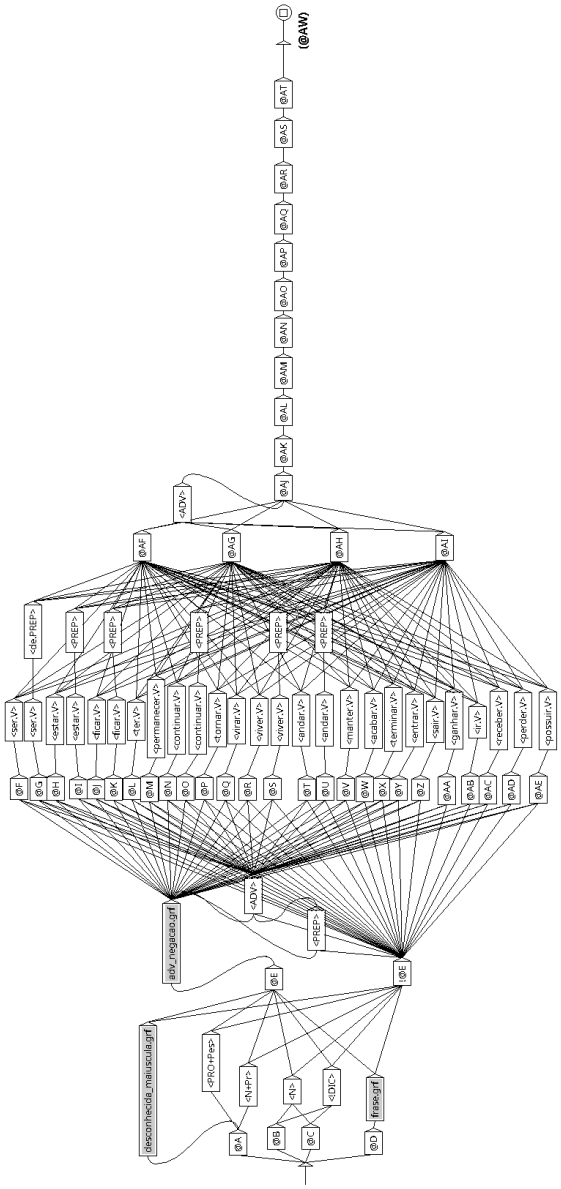
APÊNDICE A – PRIMEIRA VERSÃO DO GRAFO PARAMETRIZADO PARA TÁBUA EC1



APÊNDICE B – SEGUNDA VERSÃO DO GRAFO PARAMETRIZADO PARA TÁBUA EC1



APÊNDICE C – TERCEIRA VERSÃO DO GRAFO PARAMETRIZADO PARA TÁBUA EC1



APÊNDICE D – QUARTA VERSÃO DO GRAFO PARAMETRIZADO PARA TÁBUA EC1