Uso de uma ferramenta de processamento de linguagem natural como auxílio à coleta de exemplos para o estudo de propriedades sintático-semânticas de verbos

Larissa Picoli, Juliana Pinheiro Campos Pirovani, Elias Silva de Oliveira, Eric Laporte

Uso de uma Ferramenta de Processamento de Linguagem Natural como Auxílio à Coleta de Exemplos para o Estudo de Propriedades Sintático-Semânticas de Verbos

Using a Natural Language Processing Tool to Assist the Collection of Samples for the Study of Syntactic-Semantic Properties of Verbs

Larissa Picoli
Universidade Federal do Espírito Santo
larissa_picoli@hotmail.com

Juliana Campos Pirovani
Universidade Federal do Espírito Santo
juliana.campos@ufes.br

Elias de Oliveira
Universidade Federal do Espírito Santo
elias@lcad.inf.ufes.br

Éric Laporte
Université Paris-Est
eric.laporte@univ-paris-est.fr

Resumo

A análise e descrição de propriedades sintático-semânticas de verbos são importantes para a compreensão do funcionamento de uma língua e fundamentais para o processamento automático de linguagem natural, uma vez que a codificação dessa descrição pode ser explorada por ferramentas que realizam esse tipo de processamento. Esse trabalho experimenta o uso do Unitex, uma ferramenta de processamento de linguagem natural, para coletar uma lista de verbos que podem ser analisados e descritos por um linguista. Isso contribui significativamente para esse tipo de estudo linguístico, diminuindo o esforço manual humano na busca de verbos. Foi realizado um estudo de caso para automatizar parcialmente a coleta de verbos de base adjetiva com sufixo *-ecer* em um *corpus* de 47 milhões de palavras. A abordagem proposta é comparada com a coleta manual e a extração a partir de um dicionário para o PLN.

Palavras chave

Processamento de linguagem natural, Unitex, Lista de verbos, Propriedades sintático-semânticas.

Abstract

The analysis and description of syntactic-semantic properties of verbs are fundamental to both the knowledge of the grammar of a language and to the automatic processing of natural language, as an encoded form of this description can be exploited by automatic tools. This paper experiments with the use of Unitex, a natural language processing tool, to collect a list of verbs that can be analysed and described by a linguist. This work contributes significantly to linguistics, by decreasing the human manual effort in the search for verbs. A case study is performed to partially automate the collection of verbs in *-ecer* with adjectival bases in a *corpus* of 47 million words. The proposed approach is compared with manual collection and with extraction from an NLP dictionary.

Keywords

Natural language processing, Unitex, List of verbs, Syntactic-semantic properties

1 Introdução

O Processamento de Linguagem Natural (PLN) é uma área interdisciplinar que estuda a geração, representação e compreensão automática de fala e textos em línguas naturais. As aplicações do PLN incluem tradução automática, reconhecimento automático de voz, geração automática de resumos, recuperação de informação, correção ortográfica e outras ferramentas que auxiliam a escrita. De acordo com Vieira & Lima (2001), o PLN busca a construção de programas capazes de interpretar e/ou gerar informação fornecida em linguagem natural. Contudo, Laporte (2009) destaca que dicionários (léxicos) e gramáticas para o PLN, que podem ser construídos artesanalmente por linguistas, são também fundamentais para a implementação de ferramentas de qualidade. Os linguistas são responsáveis pela escolha de modelos de análise oriundos da teoria linguística, pela análise e descrição da língua e pela construção e atualização dos dicionários e gramáticas.

Em prática, nesse processo, são coletadas palavras e expressões, por exemplo expressões multipalavra em geral (RANCHHOD, 2005), nomes de pessoas (BAYRAKTAR & TEMIZEL, 2008), entidades nomeadas (TRABOULSI, 2009),

expressões jurídicas (CHIEZE *et al.*, 2010), expressões de sentimento (DURAN & RAMISCH, 2011), expressões metafóricas (MÜLLER, 2014), expressões com verbo-suporte[1] (*Vsup*) (BARROS, 2014; RASSI *et al.*, 2015)... Os linguistas analisam os itens encontrados, tendo em vista uma descrição sintático-semântica em dicionários para o PLN.

Este artigo compara vários métodos de coleta de palavras e expressões. O contexto da coleta é uma descrição das propriedades sintático-semânticas de verbos de base adjetiva com os sufixos *-ecer e -izar* (SMARSARO & PICOLI, 2013; PICOLI, 2015).[2]

Três abordagens são focadas: a coleta manual por introspeção ou com auxílio da *web*, como no estudo inicial de Picoli (2015); a extração a partir de um dicionário para PLN existente; e a extração a partir de um *corpus* de textos com o Unitex (PAUMIER, 2015), uma ferramenta de PLN.

A descrição de Picoli (2015) visa as derivações que apresentam equivalência semântica entre a frase de base (1) e a frase transformada (2):

(1)*O sol aqueceu a areia*
Fórmula*:* N_0 *Adj-v* N_1

(2) *O sol tornou a areia quente*
Fórmula: N_0 *tornar* N_1 *Adj*

A frase de base (1) contém o verbo com o sufixo *-ecer*, no caso *aquecer*. Na fórmula sintática correspondente, N_0 significa nome ou grupo nominal que ocupa a posição de sujeito na frase base, *Adj-v* denota verbo de base adjetiva e N_1 significa nome ou grupo nominal que ocupa a posição de complemento do predicado na frase base. A frase transformada (2) contém o verbo *tornar* e o adjetivo, no caso *quente*.

A autora selecionou, para a descrição sintático-semântica, apenas verbos que admitem a correspondência semântica, como em (1) e (2). Foi construída uma lista de 88 verbos de base adjetiva com sufixo *-ecer*. A partir dessa seleção de verbos, a autora analisou e descreveu suas propriedades sintático-semânticas formais (estruturais, distribucionais e transformacionais), que são aquelas relacionadas à natureza dos argumentos admitidos pelo verbo e às transformações que o verbo pode sofrer. As propriedades foram formalizadas em uma tabela do Léxico-gramática, segundo o formato proposto por Gross (1975).

Essa tabela, que foi obtida observando o comportamento sintático-semântico dos verbos numa estrutura frasal e descreve construções sintáticas, pode ser importante para aplicações como a tradução de um texto em português para outra língua. Em português, por exemplo, *brutalizar* pode equivaler semanticamente a *tornar brutal*, mas em francês, a tradução de *tornar brutal*, que é *rendre brutal*, não equivale semanticamente ao verbo *brutaliser* "tratar com brutalidade". Com os resultados desse estudo, o verbo *brutalizar* poderia ser traduzido por *rendre brutal*. Além das aplicações para o PLN, esse tipo de base de dados permite um melhor conhecimento do uso de frases e de suas transformações, contribuindo para o ensino da língua.

Esse artigo está estruturado em 6 seções. Na Seção 2, são apresentados alguns trabalhos correlatos que examinam métodos de coleta de listas de palavras e expressões. A Seção 3 apresenta a metodologia utilizada no desenvolvimento desse trabalho. Os grafos construídos no Unitex para coleta dos verbos e exemplos desejados são apresentados na Seção 4. Na Seção 5 são apresentados e discutidos os resultados do trabalho realizado e a Seção 6 apresenta as conclusões e trabalhos futuros.

2 Revisão de literatura

A literatura científica descreve as três abordagens de coleta focadas neste artigo, mas não encontramos publicações que comparassem várias abordagens.

Em várias pesquisas linguísticas recentes na área da descrição lexical, a coleta de itens lexicais e/ou exemplos para análise é realizada com auxílio da *web* e/ou por introspecção (RODRIGUES, 2009; DAVEL, 2009; PACHECO & LAPORTE, 2013; PICOLI, 2015).

Listas de palavras ou expressões são também extraídas de dicionários preexistentes. Por exemplo, Barros (2014) extrai do trabalho de Chacoto (2005) uma parte de sua lista de predicados nominais com o verbo-suporte *fazer*, e Rassi *et al.* (2014) aproveitam as listas de Vaza

[1] Uma expressão com verbo-suporte comporta um verbo, mas o papel de núcleo do predicado e a seleção dos argumentos não são cumpridos pelo verbo, e sim por um item de outra categoria gramatical, geralmente um nome ou um adjetivo, como em *ter inveja* ou *ser invejoso* (NEVES, 1999). Sendo essa definição semântica pouco precisa, critérios sintáticos foram estabelecidos para cada idioma, sempre verificando a existência de uma construção em que o verbo-suporte pode ser removido sem mudança de sentido, como em *a inveja que João tem/a inveja de João* (GROSS, 1981; LANGER, 2005; RASSI *et al.*, 2015). As construções com verbo-suporte são um dos principais tipos de expressões multipalavra.

[2] O objetivo da coleta é a construção de um dicionário, mas o objetivo do artigo é a comparação de métodos de coleta.

(1988) e Baptista (1997) para constituir uma lista de predicados nominais com os verbos-suporte *ter* e *dar*.

A terceira abordagem consiste no uso de uma ferramenta de PLN, o Unitex ou outra, para coletar itens lexicais ou exemplos automaticamente em um *corpus* de textos. Essas ferramentas realizam processamentos que incluem a segmentação em frases, a segmentação em palavras ou tokenização, a classificação gramatical de palavras e a busca em textos. Dessa forma, os linguistas podem se beneficiar do PLN por meio das ferramentas construídas pelos profissionais da computação, da mesma forma em que, simetricamente, a qualidade do PLN pode depender da descrição da língua pelos linguistas.

Assim, Ranchhod (2005), Bayraktar & Temizel (2008), Traboulsi (2009), Chieze *et al.* (2010), Barros (2014), Rassi *et al.* (2015) utilizam o Unitex para coletar diversos tipos de expressões em diversos *corpora*. Rassi *et al.* (2015) extraem uma lista de 4.668 construções com *Vsup*, que totalizam 45 variantes de *Vsup* e 3.200 nomes diferentes. Para criar essa lista, os autores elaboram grafos no Unitex e aplicam esses grafos a um *corpus* de 103.080 textos do jornal *Folha de São Paulo*. Müller (2014) coleta expressões metafóricas com o NooJ, uma ferramenta historicamente relacionada com o Unitex e com funcionalidades e funcionamento próximos.

O Unitex[3] (PAUMIER, 2015) é um sistema *open-source* para o PLN, desenvolvido inicialmente na universidade Paris-Est Marne-La-Vallée (França), disponível gratuitamente e utilizado por empresas de PLN. Os alicerces do Unitex (SILBERZTEIN, 1994) foram elaborados no Laboratoire d'Automatique Documentaire et Linguistique (LADL), dirigido por Maurice Gross, que guiou um trabalho de análise e descrição sintático-semântica do francês. O Unitex é distribuído com dicionários desenvolvidos para vários idiomas pela rede de laboratórios RELEX (RELEX, 2015), incluindo um dicionário do português do Brasil (MUNIZ *et al.*, 2005).

A ferramenta aplica dicionários a *corpora* não anotados e extrai informações através de expressões regulares e redes de transições recursivas representadas como grafos. Um exemplo de grafo no Unitex é apresentado na Figura 1. Esse grafo reconhece *o menino bonito* e *o garoto inteligente*. O código <A> reconhece um adjetivo: qualquer símbolo que aparecer entre "<" e ">" é interpretado pelo sistema como código de propriedade lexical nos dicionários ou como lema. O padrão simples da Figura 1 poderia ser representado na forma de uma expressão regular igualmente legível, mas a representação de padrões complexos por grafos é mais conveniente do que por expressões regulares para o leitor humano.[4] A função dos padrões não se limita ao reconhecimento dos próprios itens a extrair, e pode também se estender ao reconhecimento do contexto desses itens, em caso de ambiguidade.

Além do Unitex, concebido por linguistas com experiência na descrição lexical e gramatical em grande escala, outras ferramentas são utilizadas para coleta de itens lexicais e exemplos. Por exemplo, Duran & Ramisch (2011) extraem expressões de sentimento com o auxílio do sistema mwetoolkit. Arranz *et al.* (2005) extraem ocorrências de expressões verbais cristalizadas morfologicamente diferentes da forma registrada no dicionário WordNet 1.6. A tecnologia dessas duas experiências tem pontos comuns com os sistemas de NLP padrão como GATE[5] e Stanford NLP Software[6], oriundos de uma inspiração voltada para a computação, e é menos adaptada à coleta do que o Unitex, por duas razões.

Primeiro, a única forma de definir padrões é a das expressões regulares, que convém para padrões simples como NA (nome seguido de adjetivo) e NAA, mas passa a ser menos legível do que os grafos quando os padrões se tornam mais complexos. Para extrair os padrões desse trabalho, as expressões se tornariam muito complexas e de difícil leitura.

A segunda razão é uma consequência da ausência de dicionários nesses sistemas. O reconhecimento das categorias gramaticais especificadas nos padrões depende da anotação das palavras do *corpus*. Essa anotação é realizada por etiquetadores automáticos, que cometem erros. Se as anotações estão revisadas, o custo da revisão limita o tamanho dos *corpora* disponíveis em português, e a diversidade dos textos. Se as anotações não foram revisadas, os erros de etiquetagem podem impedir a extração de ocorrências. Nos dois casos, a abrangência da extração é prejudicada. Com o Unitex, o uso de um

[3] http://www-igm.univ-mlv.fr/~unitex/

[4] Por exemplo, no grafo da Figura 5, as linhas que saem do nó inicial indicam visualmente quais nós correspondem ao início de cada variante do padrão. Numa expressão regular equivalente, mesmo apresentada com uma indentação cuidadosa, a identificação do início de cada variante do padrão necessita que o leitor distinga os operadores união e concatenação, o que é menos evidente visualmente do que a oposição entre a presença/ausência de uma linha entre dois nós.

[5] https://gate.ac.uk/

[6] http://nlp.stanford.edu/software/

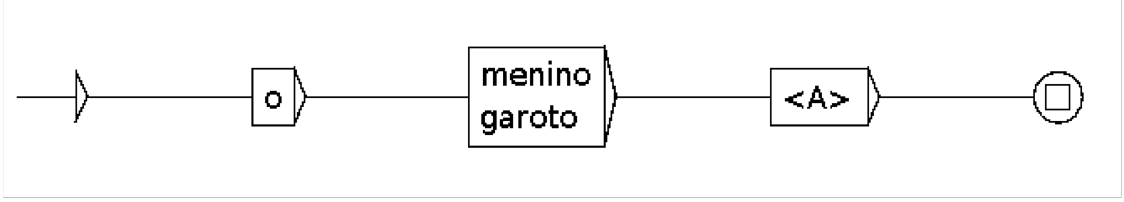

Figura 1 - Exemplo de grafo no Unitex.

dicionário de grande cobertura garante uma abrangência elevada.

A Sketch Engine (KILGARRIFF *et al.*, 2014) poderia ser utilizada para a coleta de palavras e expressões, com as duas dificuldades que acabamos de descrever e outras: não permite a definição de padrões, não pode ser utilizada em linha de comando e o uso é cobrado.

Cook *et al.* (2008) coletam expressões verbais cristalizadas no BNC, um *corpus* anotado e revisado. No caso do português, o custo da revisão limita o tamanho dos *corpora* disponíveis e a diversidade dos textos, prejudicando a abrangência da extração.

3 Metodologia

No estudo inicial, Picoli (2015) coletou os itens desejados por introspecção e manualmente, em diversos materiais como a *web* e dicionários. A descrição sintático-semântica necessitava selecionar apenas verbos que admitissem a correspondência semântica, como em (1) e (2). Essa operação, necessitada nas três abordagens, produziu uma lista de 88 verbos em *-ecer*.

A segunda abordagem, a extração dos verbos em *-ecer* no dicionário de lemas do Unitex (MUNIZ *et al*., 2005) com o comando *grep*, produz uma lista de 298 verbos que necessita uma revisão.

A contribuição principal deste artigo está na terceira abordagem, em que os verbos derivados com sufixo *-ecer* e os exemplos práticos reais de frases com esses verbos são tirados de um *corpus* de textos publicados.

As fórmulas (1) e (2) podem representar frases com outros verbos. Além dos verbos de base adjetiva com sufixo *-ecer*, como *enriquecer*, os verbos de base adjetiva com sufixo *-izar* também podem ser inseridos em frases com as mesmas fórmulas, por exemplo:

(3)*O adubo fertilizou a terra*
Fórmula*:* N_0 *Adj-v* N_1

(4) *O adubo tornou a terra fértil*
Fórmula: N_0 *tornar* N_1 *Adj*

Ferramentas de PLN podem ser usadas para buscar automaticamente frases que possuam determinadas estruturas, como as apresentadas em (1) e (2). A ferramenta de PLN utilizada neste trabalho foi o Unitex (UNITEX, 2015).

Dadas as construções sintáticas descritas em Picoli (2015), representadas pelas fórmulas (1) e (2), grafos foram construídos no Unitex para reconhecer essas estruturas. Em seguida, o Unitex foi utilizado para buscar frases a partir desses grafos no *corpus* da *Tribuna*, um *corpus* de 45.908 textos jornalísticos escritos em português e publicados pelo jornal do Espírito Santo *A Tribuna*[7]. Os arquivos do *corpus* possuem, em média, 1.032 palavras. Esse *corpus*[8] possui textos publicados nos anos de 2002 a 2006. Os textos abordam assuntos diversos como Economia, Política, Família, Ciência e Tecnologia, Concursos, TV, etc.

A aplicação de grafos do Unitex a cada arquivo do *corpus* gera um arquivo de concordância que lista as ocorrências de frases identificadas pelos grafos. O Unitex permite que *tags* sejam adicionadas aos arquivos de concordância. Assim, os verbos foram colocados entre as *tags* <verbo> e </verbo> e os adjetivos entre as *tags* <adj> e </adj>. Os arquivos de concordância foram concatenados gerando um único arquivo com todas as ocorrências identificadas no *corpus* para cada grafo utilizado.

Após essa primeira etapa, os verbos e adjetivos identificados pelos dois grafos construídos por meio das fórmulas (1) e (2), respectivamente, são extraídos dos arquivos de concordância de cada grafo e dois novos arquivos são criados contendo esses verbos e adjetivos, sem repetição. Assim, para cada grafo, dois arquivos são gerados para análise de um especialista: um arquivo com os itens lexicais (Verbos.txt ou Adjetivos.txt) identificados e um arquivo com todas as frases do *corpus* utilizado onde esses itens aparecem (ExemplosVerbos.txt e ExemplosAdjetivos.txt).

[7] http://www.redetribuna.com.br/jornal

[8] http://www.inf.ufes.br/~elias/dataSets/aTribuna-21dir.tar.gz

Grafos = {G_1, G_2} um repositório de grafos onde G_1 é o grafo construído para fórmula (1) e G_2 o grafo construído para fórmula (2)
Textos = {D_1, D_2, …, D_N} um repositório com os arquivos de entrada do *corpus A Tribuna* onde N = 45908

```
1.  for G in Grafos
2.  do
3.   for D in Textos
4.   do
5.     Aplique G em D gerando C
6.     cat C >> ConcordGeral.txt
7.   done
8.   while read linha
9.   do
10.    if [ G == G1 ];
11.    then
12.        verbo = palavra entre as tags <verbo> e </verbo>
13.        vFCanonica=`grep ^$verbo dlf | cut -d"," -f2 | cut -d "." -f1 | head -1`
14.       Se vFCanonica finaliza com -ecer, coloque o verbo no arquivo Verbos.txt e a linha no arquivo
    ExemplosVerbos.txt
15.    fi
16.    if [ G == G2 ];
17.    then
18.        adjetivo = palavra que aparece entre as tags <adj> e </adj>
19.       Coloque o adjetivo no arquivo Adjetivos.txt e a linha no arquivo ExemplosAdjetivos.txt
20.   fi
21.  done < ConcordGeral.txt
22. done
```

Quadro 1 – Pseudocódigo do programa em shell script

Apenas a construção dos grafos foi realizada manualmente. Todas as etapas seguintes foram realizadas por um programa de computador[9], implementado em *shell script*, com essa finalidade, que utiliza as ferramentas do Unitex. O pseudocódigo do programa é apresentado no Quadro 1.

As linhas de 3 a 7 mostram a aplicação de um grafo G a cada arquivo de entrada D gerando a concordância C. A linha 6 concatena o conteúdo do arquivo C gerado em um arquivo chamado ConcordGeral.txt. Nas linhas de 8 a 21, esse arquivo é lido linha a linha e, dependendo do grafo que foi aplicado, é tomada uma decisão. Se o grafo aplicado é o que reconhece a fórmula (1), deve-se obter o verbo entre as *tags* <verbo> e </verbo> e verificar se o lema (ou forma canônica) dele termina com sufixo *-ecer.* Em caso afirmativo, o verbo e a linha são colocados nos arquivos correspondentes. Se o grafo é o que reconhece a fórmula (2), basta obter o adjetivo que aparece entre as *tags* <adj> e </adj> e colocar o adjetivo e a linha nos arquivos correspondentes.

9 http://www.inf.ufes.br/~elias/codes/linguamatica2015.tar.gz

O lema do verbo é obtido a partir do arquivo chamado *dlf* gerado pelo Unitex ao aplicar os dicionários durante o pré-processamento de um texto (linha 13). O *dlf* é um dicionário gerado para as palavras simples do texto. Um exemplo de linha no arquivo *dlf* é:

entristeceu,**entristecer**.V:J3s

O lema está em negrito no exemplo. Ele aparece após a palavra que ocorre no texto original (forma flexionada) e é separado dela por uma vírgula. A classificação gramatical da palavra aparece após um ponto que segue o lema. No exemplo, a classe verbo é representada pelo código V no dicionário do Unitex. O que segue a classificação gramatical após “:” são as informações flexionais. Nesse caso, J representa passado, 3 indica terceira pessoa e S singular.

Feito isto, o linguista realiza uma análise minuciosa dos arquivos gerados para verificar se os verbos encontrados fazem parte do grupo de verbos que se pretende descrever e se os adjetivos encontrados derivam verbos relevantes. Em uma busca para selecionar verbos com final *-ecer* por exemplo, é possível que a ferramenta encontre alguns tipos de verbos: os de base adjetiva com

sufixo *-ecer*, como *apodrecer*; os de base substantiva com o sufixo *-ecer*, como *amanhecer*; e ainda os que não são formados por derivação, como *acontecer*. O trabalho do linguista é selecionar, a partir da busca feita com o Unitex, o grupo de itens lexicais que pretende descrever.

4 Descrição dos grafos construídos no Unitex

O grafo responsável por reconhecer a estrutura (1) é apresentado na Figura 2. SN é utilizado para reconhecer um sintagma nominal. O reconhecimento de sintagmas nominais foi detalhado previamente em um outro grafo, que será apresentado posteriormente, e incluído nesse como subgrafo. Referências a subgrafos são representados em nós com fundo cinza pelo Unitex. O código <V> no nó do grafo é utilizado durante o processamento pelo Unitex para reconhecer verbos.

O Unitex permite inserir saídas (texto em negrito sob setas) no grafo. Existem três modos de utilizar as saídas ao aplicar um grafo para identificar padrões em um texto. As saídas podem ser ignoradas (opção "are not taken into account"), podem ser usadas para substituir a sequência reconhecida no arquivo de concordância (opção "REPLACE recognized sequences") ou podem ser inseridas no arquivo de concordância (opção "MERGE with input text"). Na Figura 2, saídas são utilizadas para inserir as *tags* <verbo> e </verbo> no arquivo de concordância. Assim, aplicando esse grafo no modo "MERGE with input text", o verbo identificado será apresentado entre essas *tags* no arquivo de concordância.

Esse grafo reconhece qualquer verbo e não apenas os verbos de base adjetiva derivados com sufixo *-ecer*. Um filtro foi realizado posteriormente pelo programa implementado usando os resultados da aplicação dos dicionários do Unitex, como apresentado no pseudocódigo.

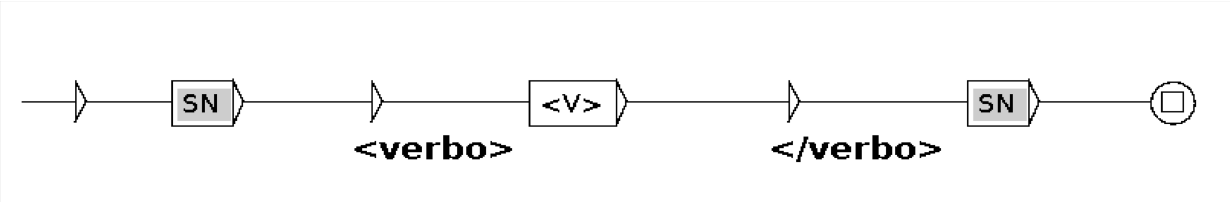

Figura 2 – Grafo que reconhece frases que possuem a estrutura da Fórmula (1) criado no Unitex.

{S} A empresa<verbo> oferece</verbo> cursos para formação
{S} Este recurso<verbo> fornece</verbo> som excepcional
{S} a, na Praça do Papa, e já tem gente<verbo> aquecendo</verbo> as turbinas para quebrar tu
{S} Eu<verbo> agradecia</verbo> a Deus 24 horas por dia principalmente, me<verbo> entristeceu</verbo> muito
{S} Isso me<verbo> emagreceu</verbo> e me deixou mais saudável.{S}
A pele<verbo> apodreceu</verbo> e , se a infecção se genera

Figura 3 - Parte da concordância obtida em modo merge com o grafo da Figura 2.

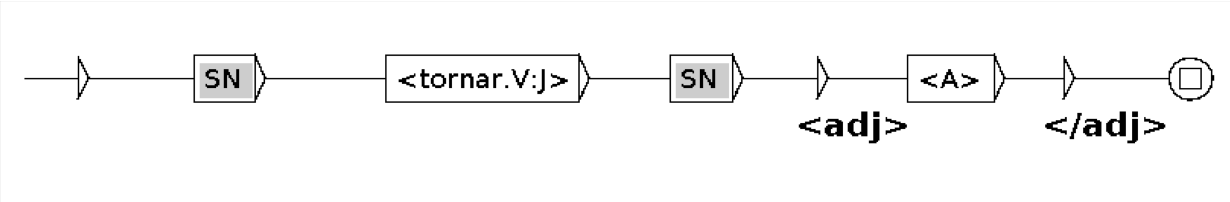

Figura 4 – Grafo que reconhece frases que possuem a estrutura da Fórmula (2) criado no Unitex.

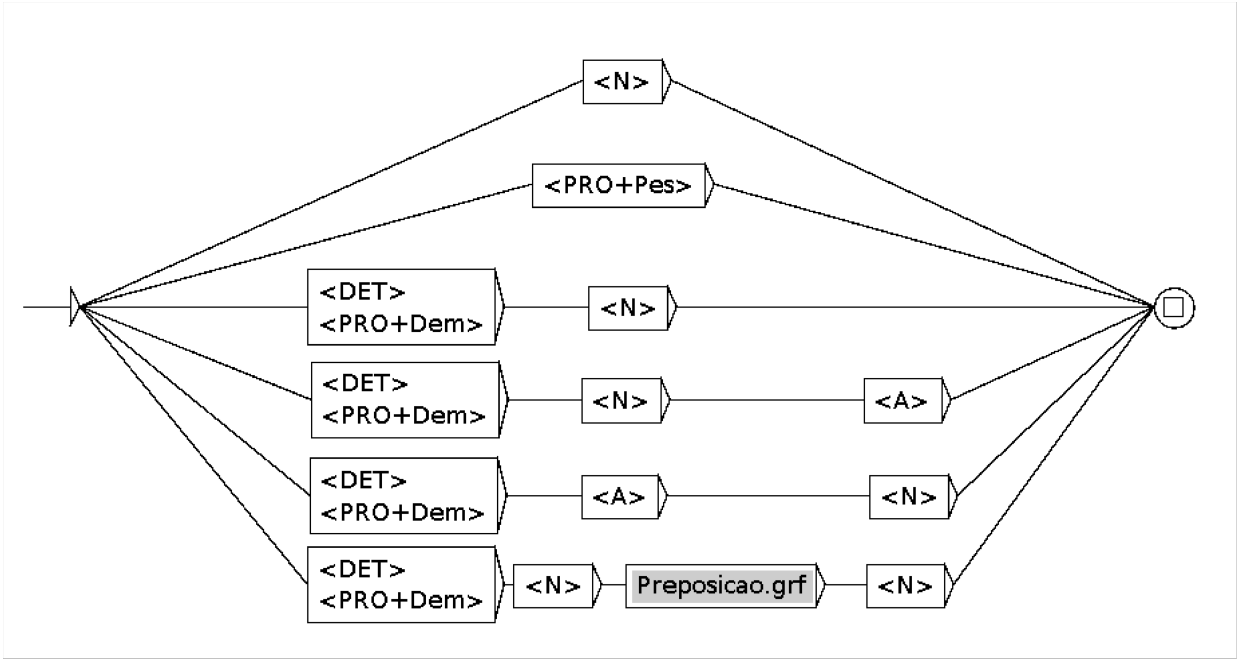

Figura 5 – Subgrafo que reconhece sintagmas nominais (SN.grf) criado no Unitex

A Figura 3 apresenta uma amostra do arquivo de concordância obtido com a aplicação do grafo da Figura 2. Alguns dos exemplos listados não contêm o verbo entre dois sintagmas nominais, mas foram extraídos assim mesmo porque a descrição dos sintagmas nominais no grafo SN inclui poucas restrições e por causa das homonímias entre entradas lexicais no dicionário do Unitex. O {S} é um símbolo separador de frases inserido no Unitex, durante o pré-processamento do texto.

Para reconhecer a estrutura (2), foi criado o grafo apresentado na Figura 4. O código <tornar.V:J> é utilizado durante processamento pelo Unitex para reconhecer verbos no passado que tem *tornar* como lema. O código <A> reconhece um adjetivo. As saídas <adj> e </adj> são utilizadas para taguear o adjetivo identificado no arquivo de concordância. O grafo SN utilizado como subgrafo nas Figuras 2 e 4 para reconhecer sintagmas nominais é apresentado na Figura 5. Esse grafo reconhece as principais estruturas de sintagmas nominais que devem ser encontradas nas frases buscadas. Observe na Figura 5 que ele também inclui um outro subgrafo, Preposicao.grf, usado para reconhecer preposições.

A Tabela 1 mostra o significado dos códigos usados nesse grafo. Observe que o símbolo + pode ser usado para reconhecer sequências de informações gramaticais ou semânticas, isto é, sequências mais específicas. Por exemplo, o código <PRO> é usado para reconhecer quaisquer pronomes. Já o código <PRO+Pes>, usado no grafo SN, reconhecerá apenas os pronomes pessoais.

Alguns exemplos de sintagmas nominais que podem ser reconhecidos pelo grafo da Figura 5 são:

(a) *João.*
(b) *O menino bonito.*
(c) *Aquele belo sapato.*

Código Unitex	Função
<N>	Reconhecimento lexical de substantivos
<PRO + Pes>	Reconhecimento lexical de pronomes pessoais
<DET>	Reconhecimento lexical de determinantes
<PRO + Dem>	Reconhecimento lexical de pronomes demonstrativos
<A>	Reconhecimento lexical de adjetivos

Tabela 1 – Códigos do Unitex usados na Figura 5 e seu significado.

5 Resultados e discussão

As três abordagens focadas neste artigo podem ser comparadas em termos de abrangência, esforço e tempo.

A abrangência é o critério mais importante porque mede a qualidade e a quantidade do resultado. A abordagem manual produziu 88 entradas em *-ecer* depois da revisão. A extração automática a partir do dicionário Unitex-PB forneceu 298 entradas em *-ecer* que necessitavam revisão, mas a revisão não foi feita para evitar o efeito de repetição em que a mesma operação sobre as mesmas entradas pelo mesmo linguista se torna mais fácil e rápida a cada iteração.

A extração a partir do *corpus* pelo grafo da Figura 2 apresentou nos arquivos de concordância 79 verbos diferentes com sufixo *-ecer* e 10.693 exemplos de frases com esses verbos. Dos 79 verbos, foi verificado manualmente por um linguista que 27 são de base adjetiva e possuem correspondência semântica quando inseridos em frases como (1) e (2): vários dos verbos com sufixo *-ecer* não eram de base adjetiva e não foram selecionados, como o verbo *oferecer* apresentado na Figura 3. Comparando a lista obtida com a de Picoli (2015), foi observado que um dos verbos identificados que possui correspondência semântica, *enraivecer*, não consta na lista de verbos descritos pela autora.

Já o grafo da Figura 4 deu 177 adjetivos e 234 exemplos de frases com esses adjetivos. Dos 177 adjetivos, foi verificado manualmente que 9 têm derivados em *-ecer* que admitem a correspondência semântica em frases como (1) e (2). Nessa lista de adjetivos, apenas 3 formam verbos que não foram coletados a partir do grafo da Figura 2. Assim, utilizando a extração a partir do *corpus*, foi construída uma lista final com 30 verbos (27 coletados pelo grafo da Figura 2 e 3 outros coletados a partir do grafo da Figura 4).[10] Esse resultado é comparável com os 88 verbos apresentados em Picoli (2015). Assim, a extração a partir de um grande *corpus* homogêneo é eficiente para coleta de verbos e adjetivos, mas menos do que a coleta manual, que leva em conta as várias comunidades em que o pesquisador está inserido.

Todavia, os dois métodos de coleta são complementares. O Unitex extraiu do *corpus* um grande número de exemplos de frases com esses verbos que podem ser utilizados no estudo e na atestação de suas propriedades. Analisando exemplos de frases com o mesmo verbo, podem-se observar propriedades distintas. Tal riqueza de exemplos é complementar à análise por introspecção, indispensável para identificar, entre outras, as propriedades que os verbos não possuem.

Enfim, a comparação entre a abordagem por dicionário e as duas outras sugere que o dicionário contém entradas pouco usadas, que podem dificilmente ser submetidas a um estudo aprofundado de suas propriedades sintático-semânticas, pois tal estudo pressupõe que o linguista domine o uso das entradas em todas suas construções.

O segundo critério, o do esforço humano desempenhado, é difícil de avaliar, mas durante esta experiência, achamos mais trabalhosa a busca lexical sem ferramentas computacionais.

O terceiro critério é a duração do processo de coleta.

No estudo inicial, Picoli (2015) não registrou uma estimação do tempo gasto para a busca manual de itens lexicais, nem para a seleção dos itens que possuem correspondência semântica quando inseridos em frases como (1) e (2). A busca manual não consistiu em ler grandes quantidades de textos; contudo, considerando que aproximadamente 4 minutos são necessários, em média, para ler um dos arquivos do *corpus*, seria necessário aproximadamente 3.000 horas para apenas ler todos os 45.908 arquivos desse *corpus*.

Na segunda abordagem, a resposta do sistema é imediata.

Na terceira abordagem, o programa implementado para coleta dos verbos e adjetivos foi executado em um computador com as seguintes características: processador Intel core i5, memória de 4GB, sistema operacional Ubuntu 14.04. O tempo de execução para coleta dos adjetivos foi 1h35min e para coleta dos verbos 3h54min. Essa diferença de tempo é consequência do tempo adicional necessário para gerar outro arquivo de concordância mantendo apenas os verbos com sufixo *-ecer*. Portanto, esses verbos foram identificados em poucas horas, e o processo pode ser agilizado, concatenando todos os textos do *corpus* em um único arquivo e aplicando cada grafo ao *corpus* completo. Para verificar se os verbos são de base adjetiva e se possuem correspondência semântica em frases como (1) e (2), analisando cada verbo e adjetivo extraído para verificar se eles ocorrem na estrutura buscada e se possuem as propriedades de interesse, o linguista gastou cerca de 2 minutos por verbo.

[10] Esse mesmo grafo poderia ser usado para coletar adjetivos que derivam verbos com outros sufixos, como *-izar.* Analisando rapidamente as listas, verificou-se que 13 adjetivos extraídos do *corpus* têm derivados em *-izar* e 272 verbos têm o sufixo *-izar.* Portanto, esse sufixo é mais frequente e mais produtivo neste *corpus* do que *-ecer*.

Assim, com a utilização de uma ferramenta de PLN, o linguista pode, em um curto período de tempo, extrair uma parte do seu objeto de análise, que são neste estudo os verbos que aparecem na estrutura sintática de interesse.

6 Conclusão

Nesse trabalho foi utilizada uma ferramenta de processamento de linguagem natural, o Unitex, para auxiliar o recenseamento de verbos de base adjetiva com sufixo *-ecer* e de exemplos de frases com estes verbos, para análise e descrição de propriedades sintático-semânticas.

Foram construídos dois grafos no Unitex: um para coleta de frases contendo verbos e outro para coleta de frases contendo adjetivos que podem derivar os verbos de interesse. Esses grafos foram construídos a partir das fórmulas (1) e (2) que descrevem as estruturas de frase cuja equivalência semântica poderia ser estudada como em Picoli (2015). Foi implementado um programa que aplica esses grafos no *corpus* da *Tribuna* e gera arquivos para análise de um linguista.

O linguista deve selecionar para descrição apenas aqueles verbos com as características desejadas: base adjetiva e correspondência semântica em frases como (1) e (2). Foram identificados 30 verbos com essas características, sendo que um deles não fora identificado pelo estudo de Picoli (2015).

A coleta dos verbos por este método produz resultados menos abrangentes, mas também é menos trabalhosa para os linguistas que buscam descrever a língua portuguesa. O linguista pode ainda poupar esforço e tempo com o Unitex por meio de outros métodos: por exemplo, extraindo todos os 298 verbos com sufixo *-ecer* do dicionário de lemas do Unitex, antes de revisar a lista obtida.

Como trabalho futuro, pretende-se realizar o mesmo experimento em outros *corpus*, maiores e de gêneros textuais diferentes do apresentado no jornal *A Tribuna* e realizar a coleta de itens em dicionários existentes. Pretende-se também repetir a experiência com versões mais completas dos grafos. Por exemplo, pode ser incluído no grafo da Figura 2 o reconhecimento de frases com o verbo na primeira pessoa do singular, sem pronome sujeito explícito, como em *Fortaleço meus braços*. Da mesma forma, é possível relaxar as restrições do grafo da Figura 4, para aceitar o verbo em outros tempos que não o perfeito (*O sol torna a areia quente*), ou aceitar os complementos na ordem inversa (*O sol tornou quente a parte superficial*). Grafos adicionais podem ser construídos para representar construções frequentes associadas a (1), por exemplo, N_1 *Adj-v =: A areia aqueceu.*

Além disso, pretende-se também criar uma interface *web* onde o linguista poderá informar as fórmulas sintáticas de interesse, tais como as apresentadas nas fórmulas (1) e (2). Essas estruturas serão convertidas automaticamente em grafos do Unitex que serão usados para buscá-las em alguns corpora internos predefinidos. Assim, uma lista de verbos será criada automaticamente para que o linguista continue seu trabalho.

Referências

ARRANZ, Victoria; ATSERIAS, Jordi; CASTILLO, Mauro. Multiwords and word sense disambiguation. In Alexander Gelbukh, editor, *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING), volume 3406 of Lecture Notes in Computer Science (LNCS), 2005, p. 250–262, Mexico City, Mexico. Springer-Verlag.

BAPTISTA, Jorge. Sermão, tareia e facada. Uma classificação das construções conversas dar-levar. *Seminários de Linguística*, volume 1, 1997,.p. 5–37, Faro. Universidade do Algarve.

BARROS, Cláudia D. *Descrição de classificação de predicados nominais com verbo-suporte fazer: especificidades do Português do Brasil.* Tese de doutorado, Universidade Federal de São Carlos, 2014.

BAYRAKTAR, Őzkan; TEMIZEL, Tuğba Taşkaya. Person Name Extraction From Turkish Financial News Text Using Local Grammar Based Approach. *23rd International Symposium on Computer and Information Sciences* (ISCIS'08). Istanbul. 2008.

CHACOTO, L. *O verbo 'fazer' em construções nominais predicativas*. Tese de Doutorado, Universidade de Algarve, Faro, 2005.

CHIEZE, Emmanuel; FARZINDAR, Atefeh; LAPALME, Guy. An Automatic System for Summarization and Information Extraction of Legal Information; In *Semantic Processing of Legal Texts*, 2010, p. 216-234, Springer-Verlag.

COOK, Paul; FAZLY, Afsaneh; STEVENSON, Suzanne. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions* (MWE), 2008, p. 19–22. Marrakech, Morocco.

DAVEL, Alzira da P. C. *Um estudo sobre o verbo-suporte na construção Dar+SN*. Dissertação de Mestrado. Vitória, UFES, 2009.

DURAN, Magali Sanches; RAMISCH, Carlos. "How do you feel? Investigating lexical-syntactic patterns in sentiment expression", *Proceedings of Corpus Linguistics 2011: Discourse and Corpus Linguistics Conference*, Birmingham, UK, July 2011.

GROSS, Maurice. *Méthodes en syntaxe. Régime des constructions complétives*. Paris: Hermann. 1975.

GROSS, Maurice. Les bases empiriques de la notion de prédicat sémantique, *Langages* 63, 1981, p. 7-52 e 127-128, Paris: Larousse.

KILGARRIFF, Adam; BAISA, Vít; BUSTA, Jan; JAKUBÍČEK, Miloš; KOVÁŘ, Vojtěch; MICHELFEIT, Jan; RYCHLÝ, Pavel; SUCHOMEL, Vít. The Sketch Engine: ten years on. *Lexicography ASIALEX* 1, 2014, p. 7-36.

LANGER, Stefan. A linguistic test battery for support verb constructions. *Lingvisticae Investigationes* 27 (2). 2005, p. 171–184.

LAPORTE, Éric. Lexicons and Grammar for language processing: industrial or handcrafted products? In: *Léxico e gramática*: dos sentidos à descrição da significação. REZENDE, L. DIAS DA SILVA,B. C.; BARBOSA, J. B. (Org). São Paulo: Cultura Acadêmica, 2009, p. 51-84.

MÜLLER, Ralph. "NooJ as concordancer in computer-assisted textual analysis. The case of the German module" In *Formalising Natural Languages with NooJ 2013: Selected papers*. 2014, p. 203-214.

MUNIZ, Marcelo C.; NUNES, Maria das Graças V.; LAPORTE, Éric. UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. In *Proceedings of the Workshop on Technology on Information and Human Language (TIL)*, 2005, p. 2059-2068.

NEVES, Maria H. M. *Gramática de usos do português*. São Paulo: UNESP, 1999.

PACHECO, Wagner L; LAPORTE, Éric. Descrição do verbo cortar para o processamento automático de linguagem natural. In: *Dialogar é preciso. Linguística para o processamento de línguas.* Vitória PPGEL/UFES, 2013, p. 165-175.

PAUMIER, S. (2015). Unitex 3.1 user manual. Disponível em: http://igm.univ-mlv.fr/~unitex/ UnitexManual3.1.pdf. Acesso em: 29/11/2015

PICOLI, Larissa. *Descrição de verbos de base adjetiva derivados com os sufixos -ecer e -izar, para o Processamento Automático de Linguagem Natural***.** Dissertação de Mestrado. Vitória, UFES, 2015.

RANCHHOD, Elisabete. Using Corpora to Increase Portuguese MWU Dictionaries. Tagging MWU in a Portuguese Corpus. In Pernilla DANIELSSON, Martijn WAGENMAKERS (eds.), *Corpus Linguistics* (CL'05), 2005.

RASSI, Amanda P.; SANTOS-TURATI, Cristina; BAPTISTA, Jorge; MAMEDE, Nuno; VALE, Oto A. The fuzzy boundaries of operator verb and support verb constructions with dar "give" and ter "have" in Brazilian Portuguese, In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, Coling, 2014, p. 92–101.

RASSI, Amanda P.; BAPTISTA, Jorge; VALE, Oto A. Um corpus anotado de construções com verbo-suporte em Português. *Gragoatá* 38, 2015, p. 207-230, Niterói.

RELEX. *The RELEX network*. Disponível em: http://infolingu.univ-mlv.fr/Relex/introduction.html. Acesso em: 13/07/2015.

RODRIGUES, Carlos R. de S. *Descrição e formalização de estruturas com verbos de ação-processo para a elaboração de um parser*. Dissertação de Mestrado. Vitória, UFES, 2009.

SILBERZTEIN, Max D., "INTEX: a corpus processing system", in *COLING 94 Proceedings*, Kyoto, Japan, vol. 1, 1994, p.579-583.

SMARSARO, Aucione; PICOLI, Larissa. Propriedades sintático-semânticas de verbos *Adj-ecer*. *Cadernos do CNLF (CiFEFil)*, Rio de Janeiro, Vol. XVII, nº 02, 2013.

TRABOULSI, Hayssam. Arabic Named Entity Extraction: A Local Grammar-Based Approach. *International Multiconference on Computer*

Science and Information Technology (IMCSIT'09), vol. 4, 2009, p. 139–143.

UNITEX. Disponível em: http://www-igm.univ-mlv.fr/~unitex/. Acesso em: 15/10/2015.

VAZA, Aldina. *Estruturas com nomes predicativos e o verbo-suporte dar.* Dissertação de mestrado, Faculdade de Letras, Universidade de Lisboa. 1988.

VIEIRA, Renata; LIMA, Vera L. S. Linguística computacional: princípios e aplicações**.** In: *IX Escola de Informática da SBC-Sul*. Luciana Nedel (Ed.) Passo Fundo, Maringá, São José. SBC-Sul, 2001.