VII SIMELP
SIMPÓSIO MUNDIAL DE ESTUDOS DE LÍNGUA PORTUGUESA
Porto de Galinhas - Pernambuco, Brasil
20 a 24 de agosto de 2019

ISBN 978-85-7946-353-2

SIMPÓSIO AT019

ELABORAÇÃO DE CORPUS ANOTADO PARA O PROCESSAMENTO DE SUBSTANTIVOS PREDICATIVOS

MARTINEZ, Ryan Marçal Saldanha Magaña
UFSCAR - Universidade Federal de São Carlos
ryan.saldanha.martinez@gmail.com

Resumo: Apresentamos parte de um trabalho de mestrado que tem como objetivo descrever os padrões de seleção de argumento dos substantivos predicativos da porção brasileira do corpus Bosque (AFONSO, 2002), correspondente a 4213 frases. Na primeira parte deste texto, descrevemos os testes utilizados com o apoio de exemplos do corpus. Suas definições de predicado e, mais especificamente, dos substantivos predicativos, são baseadas em Gross (1981), em que as diferentes manifestações do substantivo predicativo são resultado de transformações sobre uma frase elementar caracterizada por um verbo que adiciona marcas de tempo, modo e aspecto ao substantivo - isto é, o verbo suporte - selecionando um dos argumentos do substantivo como seu sujeito. As relações predicado-argumento constantes no corpus são estabelecidas por meio de testes formais de reconstrução da frase elementar e suas formas intermediárias. Uma vez definidos esses testes, passamos a uma descrição dos procedimentos de anotação e sua concordância Kappa, cujos valores variaram entre 0,495 e 0,713 para diferentes características. Por fim, apresentamos as sequências encontradas no corpus, nas quais se verifica a prevalência, primeiramente, grupos nominais e, secundariamente, substantivos com elisão completa dos argumentos e suporte, seguido de diversas manifestações envolvendo verbos suporte e causativos. Espera-se que essa descrição possa contribuir para o processamento automático dos substantivos predicativos, seja indicando caminhos para diretrizes de anotação ou lançando bases para abordagens automáticas ou semiautomáticas para a expansão de tais recursos.

Palavras-chave: predicação; substantivos predicativos; processamento de língua natural; linguística de corpus

Abstract: We present part of a masters research whose objective was describing the patterns of argument selection for predicative nouns in the Brazilian portion of corpus Bosque (AFONSO, 2000), which corresponds to 4213 sentences. In the first part of this text, we describe the different tests used, supported by corpus samples. The definition of predicate - and, more specifically, predicative nouns - used for such tests are based on Gross (1981), to whom different schemes of predicative nouns are a result of transformations over elementary sentences, characterized by a verb that adds tense, mood and aspect marks to the noun - i.e., support verbs - and selects one of the

ISBN 978-85-7946-353-2

noun's arguments as its subject. The predicate-argument relations in the corpus are established using formal tests that rely on elementary sentence and intermediate forms reconstruction. Once the tests are defined, we proceed to a description of annotation procedures and Kappa agreement, which ranged from 0.495 to 0.713 in different factors. Finally, we present the sequences found, where one can observe, firstly, the prevalence of noun groups and, secondarily, nouns with complete elision of arguments and support, followed by strings containing different manifestations of support verbs and causatives. We expect this description to contribute to predicative noun processing by signalling annotation directives or laying foundations for automatic and semiautomatic approaches to the expansion of such resources.

Keywords: predication; predicative nouns; natural language processing; corpus linguistics

Introdução

Lidando com unidades de sentido que se expressam por mais de um segmento separado por espaços, a tarefa de identificação de expressões multi-palavras (MWEs, do inglês *multi-word expressions*) é uma parte do Processamento de Língua Natural (PLN) que tem recebido atenção considerável em anos recentes. O conceito de MWE inclui uma série de fenômenos linguísticos com características distintas, dentre os quais se incluem os verbos suporte, ou "verbos leves", os quais que se distinguem das outras MWEs por sua flexibilidade sintática e a idiossincrasia do par verbo-substantivo (SAG et al, 2002, p. 7). Essas características do fenômeno vêm sendo tratadas para processamento computacional em projetos como o Parse.me (SAVARY et al, 2017), que conta com *corpora* multilíngues de construções com "verbo leve", e a construção de léxicos exaustivos (RASSI, 2015; BARROS, 2014; SANTOS, 2014; entre outros). Algumas empreitadas exploram também o processamento de sua estrutura argumental (a exemplo de MEYERS et al, 2004 e BONIAL, 2014, ambos para língua inglesa), junto a substantivos isolados.

Ocorre que o sujeito da CVS ativa é argumento do substantivo que acompanha o verbo, fazendo com que verbo e substantivo funcionem como um único predicado (GROSS, 1981).

Mais comumente, o verbo da CVS se reduz e seu sujeito se torna um segmento introduzido por preposição, formando uma sintagma nominal

ISBN 978-85-7946-353-2

acompanhado de seus complementos nominais. Vê-se, assim, que há uma série de esquemas envolvendo esses substantivos que correspondem a um mesmo grupo de formas parafrásticas, das quais a MWE corresponde apenas a uma fração.

Vislumbrando a automatização da tarefa, anotamos os substantivos predicativos das 4213 frases da porção brasileira do corpus Bosque (AFONSO, 2002), trecho do corpus jornalístico CETENFolha que conta ainda com anotações morfossintáticas, de relações sintáticas e de constituintes anotadas automaticamente pelo *parser* PALAVRAS (BICK, 2000) e revisada por linguistas. São apresentados a concordância entre anotadores e a frequência das sequências. Concluímos comentando a eficácia do processo de anotação e a relevância dos dados obtidos como recurso para o processamento dessas estruturas.

1. Definição de substantivo predicativo

Nossa visão de linguagem se baseia em Harris (1968), para quem a sintaxe pode ser caracterizada por: 1- um conjunto de frases elementares, as quais incluem o predicado e os elementos selecionados por ele (isto é, seus argumentos); e 2- um conjunto de transformações, que podem se dar sobre uma única frase elementar (unárias) ou sobre duas frases elementares (binárias). Assim, as frases analisadas são frases complexas, formadas pela sobreposição desses dois conjuntos, e cuja descrição se dá pela identificação de seus estágios anteriores em uma cadeia transformacional. Esse processo de identificação culmina na frase elementar.

Para a identificação dos substantivos predicativos, utilizamos os critérios delineados em Gross (1981). Entendemos que haja substantivos cuja presença licencia alguns de seus elementos vizinhos, nomeadamente sujeito, objetos, preposições e um verbo que adiciona marcas de tempo, modo e aspecto ao substantivo. O verbo em questão é denominado verbo suporte. Para Gross (1981), a relativização do substantivo predicativo é uma etapa na formação dos grupos nominais. Por isso, as formas abaixo são paráfrases:

ISBN 978-85-7946-353-2

"Outra pesquisa do Datafolha mostra que estes alimentos recuaram 0,47% na última semana de novembro." (CETENFolha, 366)

"Outra pesquisa (que o Datafolha fez + feita pelo Datafolha) mostra que estes alimentos recuraram 0,47% na última semana de novembro" (frases construídas)

Constatamos que ambas incluem alguma variação de uma frase elementar, que é nosso critério para marcar o objeto de análise:

"O Datafolha fez uma pesquisa sobre alguma coisa" (frase construída)

Esses substantivos podem também surgir como adjunto adverbial, tomando o sujeito da oração principal como argumento. Trata-se de uma redução de um verbo suporte preposicionado. O verbo suporte pode sofrer alterações aspectuais e aceita substituição por uma gama de variantes que não acarretam qualquer mudança de significado. Muitas vezes, essas variantes são idiossincráticas. Por fim, há o fenômeno do verbo suporte converso, que inverte as posições de sujeito e objeto (para uma descrição mais detalhada, ver CALCIA, 2016). No teste abaixo, criamos uma paráfrase de duas construções conversas utilizando verbos suporte padrão, demonstrando que são equivalentes.

Assim, substantivos capazes de participar de algum número dos tipos de sequência aqui descritos são considerados substantivos predicativos, e seus elementos coocorrentes, caso possam remontar a uma frase elementar, são considerados argumentos ou verbos suporte.

2. Anotação do *corpus* e concordância

A anotação do *corpus* consistiu, para todas as frases, de um anotador (Anotador 1) utilizando julgamentos de aceitabilidade sobre formas construídas decorrentes das frases originais. Em caso de frases construídas de aceitabilidade duvidosa, houve pesquisa empírica de sua ocorrência em outros

ISBN 978-85-7946-353-2

corpora pela ferramenta AC/DC e na *web*. O Anotador 1 foi o único a anotar todas as 4213 frases.

A validação desses dados foi realizada pelo Anotador 2, que recebeu 200 frases selecionadas aleatoriamente por um algoritmo, sem estar ciente desse critério de seleção. Pediu-se que utilizasse os critérios do léxico-gramática (isto é, Gross 1981).

Apresentamos abaixo o coeficiente Kappa entre o Anotador 1 e o Anotador 2 para os elementos constitutivos da frase elementar, além dos resultados para cada um dos anotadores. Para cada substantivo das frases utilizadas, a concordância do substantivo é calculada em relação a se os substantivos analisados foram ou não considerados predicativos. As linhas Suporte, N0 (Sujeito), N1 e N2 (Objetos) mostram a concordância desses fatores para os substantivos considerados positivos por ambos os anotadores. O coeficiente Kappa foi calculado utilizando-se a biblioteca SciKit-Learn para Python.

Segmento	**Anotador 1 (Positivos/ Total)**	**Anotador 2 (Positivos/ Total)**	**Concordância Kappa**
Npred	231/776	230/776	0,713
Vsup	46/231	84/230	0,495
N0	48/231	59/230	0,693
N1	51/231	52/230	0,640
N2	7/231	8/230	0,405

Tabela 1: Comparação e concordância entre anotadores

3. Segmentos e sua prevalência

Apresentamos abaixo os números relativos às anotações do Anotador 1 para o *corpus* Bosque completo, validadas por sua alta concordância com o Anotador 2. *Tokens* se refere ao número de ocorrências de substantivo predicativo, enquanto os tipos são a contagem de uma mesma classe de substantivo predicativo.

ISBN 978-85-7946-353-2

Categoria 1 / Categoria 2	***Tokens***	**Tipo**	**Frases**	**TOTAL**
Tokens	1	2,347	1,1263	4745
Tipos	0,310	1	0,479	2021
Frases	0,887	2,084	1	4213

Tabela 2: Relações de frequência entre tipos, *tokens* e frases

Vê-se que o fenômeno é bastante difundido em textos em língua portuguesa, ainda que a extensão do *corpus* utilizado não apresente um número substancial de representantes de cada tipo, se considerada a variedade de formas que podem obter. Na tabela abaixo, apresentamos as formas que esses predicados tomam.

Sequência	**Número**	**Frequência por tipo**	**Frequência por *token***
Npred s/ argumentos	1791	0,886	0,377
Grupos Nominais	1974	0,976	0,416
Alternâncias de Verbo suporte	878	0,434	0,185
Verbo Operador Causativo	102	0,050	0,021

Tabela 3: Frequência das sequências do corpus analisado

Considerações finais

Ao longo deste trabalho, discutimos as manifestações do substantivo predicativo como MWE (na CVS) e como item isolado, a qual correspondem diferentes estratégias de formalização para seu processamento, seja registrando as idiossincrasias dos pares verbo suporte + substantivo ou estabelecendo seus relações de predicado-argumento. Em seguida, delineamos uma abordagem (derivada de Gross, 1981) baseada em testes transformacionais para a caracterização desses fenômenos em frases complexas, visando a atender o segundo tipo de formalização. A abordagem consiste em remontar a manifestação do substantivo predicativo a uma frase elementar, aplicando sucessivas transformações que demonstrem sua equivalência ao que, em nosso entender, é a unidade sintática mínima. Para isso, o anotador deve estabelecer paráfrases que possuam aceitabilidade.

ISBN 978-85-7946-353-2

Como resultado dos esforços de quatro anotadores orientados por esses preceitos, obtivemos uma concordância moderada a alta, o que demonstra a confiabilidade do modelo para PLN. O *corpus* anotado produzido ao longo desse processo cobre uma gama de construções ampla, mas o número de exemplos para cada uma ainda é reduzida. Dessa forma, o recurso construído é um passo em direção à automatização da tarefa, que pode ser ainda complementadas por outros esforços de anotação humana, automática ou semiautomática. Para os primeiros, as observações aqui registradas podem servir para demonstrar as qualidades e limitações da abordagem utilizada. Para os dois últimos, o recurso produzido pode ser unido a outras propostas, de maneira a sanar a escassez de dados.

Referências

AFONSO, Susana et al. Floresta Sintá(c)tica: A treebank for Portuguese. In: Rodrigues, M. G. & Araujo, C. P. S. (Org) **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)**. Paris: ELRA, 2002

BARROS, Cláudia Dias. **Descrição e classificação de predicados nominais com o verbo-suporte fazer no Português do Brasil.** Tese de doutorado. São Carlos: Universidade Federal de São Carlos, 2014.

BICK, Eckhard. The parsing system Palavras. In: **Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework,** 2000.

BONIAL, Claire. **Take a Look at This! Form, Function and Productivity of English Light Verb Constructions**. Tese de doutorado. Colorado: University of Colorado at Boulder, 2014.

CALCIA, Nathalia Perussi. **Descrição e classificação das construções conversas do português do Brasil**. Dissertação de mestrado. São Carlos: Universidade Federal de São Carlos, 2016.

GROSS, Maurice. Les bases empiriques de la notion de prédicat sémantique. In: **Langages**, p. 7–52. 1981.

HARRIS, Zellig. **Mathematical structures of language.** Interscience publishers, 1968.

MEYERS, Adam et al. The NomBank project: An interim report. In: **HLT-NAACL 2004 workshop: Frontiers in corpus annotation**, v. 24. 2004.

ISBN 978-85-7946-353-2

SAG et al, SAG, Ivan A. et al. Multiword expressions: A pain in the neck for NLP. In: **International Conference on Intelligent Text Processing and Computational Linguistics**. Springer, Berlin, Heidelberg, 2002. p. 1-15.

SAVARY, A. et al. The PARSEME shared task on automatic identification of verbal multiword expressions. In: **Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)**, Valência, p. 31-47, 2017.

RASSI, Amanda. **Descrição, classificação e processamento automático das construções com o verbo 'dar' em Português do Brasil. Tese de Doutorado.** São Carlos: Universidade Federal de São Carlos, 2015.

SANTOS, M. C. A. **Descrição dos predicados nominais com o verbo-suporte 'ter'.** Tese de Doutorado. São Carlos: Universidade Federal de São Carlos, 2015.