Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 10, Número 2 (2018)

ISSN: 1647-0818

Volume 10, Número 2 – 2018

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

POP: Por Outras Palavras

POP – Por Outras Palavras

Este volume contém os trabalhos apresentados no POP – Por Outras Palavras, o 1º seminário sobre Ferramentas e Recursos Linguísticos para Parafraseamento em Português, realizado a 24 de Setembro de 2018 em Canela (RS), Brasil. O seminário teve como objetivo reunir investigadores linguistas e que trabalham na área do Processamento de Linguagem Natural interessados em discutir novas ideias sobre o desenvolvimento e uso de recursos linguísticos orientados para pararafraseamento em português com aplicações do mundo real.

As paráfrases são extremamente importantes na comunicação humana, tanto na produção como na compreensão da linguagem, e assumem um papel cada vez mais importante em atividades e projetos de investigação. Diversas experiências linguísticas mostraram a viabilidade de usar recursos parafrásticos numa ampla variedade de aplicações de software, pois permitem reconhecer e gerar formas equivalentes de expressar o mesmo conteúdo, permitindo que os sistemas forneçam ao utilizador sugestões para dizer e escrever a mesma coisa / ideia por outras palavras, aumentar a fluência, a criatividade e a diversidade estilística. No atual estágio de desenvolvimento, os sistemas de parafraseamento exigem conhecimento linguístico e "inteligência" sensível ao contexto para "compreender" e reconhecer uma ampla variedade de expressões. Para o português, a utilidade dos recursos parafrásticos já foi explorada em cenários aplicativos, como um sistema de diálogo, para aumentar o conhecimento linguístico de um agente virtual inteligente, em ferramentas de sumarização e simplificação e também em ferramentas que visam obter tradução automática de qualidade superior. No entanto, é necessária mais investigação para a viabilidade e sucesso de um sistema de parafraseamento a longo prazo nas áreas de produção e revisão de texto, nomeadamente no desenvolvimento e melhoria de plataformas de autoria online, desenvolvendo programas interativos para ajudar os estudantes de português como língua estrangeira a produzir frases diferentes mas equivalentes ou até para estudantes nativos, para os auxiliar nas tarefas de produção e revisão dos seus textos.

Ao propor o seminário POP, queríamos (i) reunir investigadores com interesse no campo das paráfrases, e com especial enfoque no português, para aprender e partilhar informação sobre o tema; (ii) reunir um conjunto de artigos de boa qualidade que discutam as últimas tendências na área e contribuam para melhorar o estado da arte das paráfrases em português; (iii) trocar ideias e disseminar as melhores práticas para ajudar a fomentar a investigação nesta área; (iv) fomentar uma convergência de esforços de investigação para uma definição consensual dos métodos científicos, e incentivar a cooperação internacional, a fim de alcançar estratégias comuns que respondam às necessidades tecnológicas atuais; (v) discutir novas metodologias, como redes neuronais, etc., e aprender a combinar essas metodologias com esforços linguísticos; (vi) discutir desafios futuros e trocar informação sobre aspetos científicos e tecnológicos; (vii) incentivar e reforçar a criação de corpora paralelos de paráfrases para o português como conjuntos de dados para a coleta de recursos de alinhamento

parafrástico para treino e teste de sistemas de parafraseamento; e (viii) localizar fontes de financiamento para impulsionar ainda mais a investigação, apoiar a inovação e desenvolver esta tecnologia capacitante essencial.

O Comité do Programa era composto por 22 membros de Portugal (8), Brasil (7), Espanha (4), França (2) e Noruega (1), e todos os membros são especialistas de renome em Processamento de Linguagem Natural, Linguística Computacional, Engenharia da Linguagem, e áreas afins, com ampla experiência no processamento da língua portuguesa e especificamente em tópicos relacionados à paráfrase.

Os organizadores do seminário POP gostariam de reconhecer publicamente várias instituições e pessoas cuja ajuda foi imprescindível para o sucesso do seminário: a Organização do PROPOR'2018, por aceitar a proposta de integrar o POP nos eventos satélite da principal conferência internacional sobre Processamento da Língua Portuguesa, bem como pelo seu apoio constante e colaboração; todos os membros do Comité de Programa, cuja colaboração inestimável foi fundamental para o sucesso do seminário e para a sua qualidade científica; as diferentes instituições que apoiaram, de diferentes formas, a participação de autores e organizadores na conferência.

Anabela Barreiro
Jorge Baptista
Renata Vieira
Paulo Quaresma

Comissão de Programa POP@PROPOR2018

Sandra Aluísio, Universidade de São Paulo/ICMC/NILC (Brasil)
Jorge Baptista, Universidade do Algarve L2F/INESC-ID, Lisboa (Portugal)
Anabela Barreiro L2F/INESC-ID Lisboa (Portugal)
Lucília Chacoto, Universidade do Algarve (Portugal)
Luísa Coheur, IST/INESC-ID, Lisboa (Portugal)
Ariani Di Felippo, Universidade Federal de São Carlos (Brasil)
Claudia Freitas, PUC-Rio (Brasil)
Pablo Gamallo, Universidade de Santiago de Compostela (Espanha)
Hugo Gonçalo Oliveira, Universidade de Coimbra (Portugal)
Éric Laporte, Universidade de Paris-Est Marnela-Vallèe (França)
Belinda Maia, Universidade do Porto (Portugal)
Thiago Pardo, Universidade de São Paulo (Brasil)
Paulo Quaresma, Universidade de Évora (Portugal)
Ama manda Rassi, Lionbridge (Brasil)
Ricardo Ribeiro, INESC ID Lisboa/ISCTE-IUL (Portugal)
Paolo Rosso, Universidade Politécnica de Valência (Espanha)
Diana Santos, Universidade de Oslo (Noruega)
Max Silberztein, Universidade de Franche-Comtè (França)
Alberto Simões, 2Ai Lab - IPCA Braga (Portugal)
Oto Vale, Universidade Federal de São Carlos (Brasil)
Renata Vieira, Pontifícia Universidade Católica - Rio Grande do Sul, PA (Brasil)

Comissão Científica

Alberto Álvarez Lugrís,
Universidade de Vigo

Alberto Simões,
Universidade do Minho

Aline Villavicencio,
Universidade Federal do Rio Grande do Sul

Álvaro Iriarte Sanroman,
Universidade do Minho

Ana Frankenberg-Garcia,
University of Surrey

Anselmo Peñas,
Univers. Nac. de Educación a Distancia

Antón Santamarina,
Universidade de Santiago de Compostela

Antoni Oliver González,
Universitat Oberta de Catalunya,

Antonio Moreno Sandoval,
Universidad Autónoma de Madrid

António Teixeira,
Universidade de Aveiro

Arantza Díaz de Ilarraza,
Euskal Herriko Unibertsitatea

Arkaitz Zubiaga,
Dublin Institute of Technology

Belinda Maia,
Universidade do Porto

Carmen García Mateo,
Universidade de Vigo

Diana Santos,
Linguateca/Universidade de Oslo

Ferran Pla,
Universitat Politècnica de València

Gael Harry Dias,
Université de Caen Basse-Normandie

Gerardo Sierra,
Univers. Nacional Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Iñaki Alegria,
Euskal Herriko Unibertsitatea

Irene Castellón Masalles,
Universitat de Barcelona

Joaquim Llisterri,
Universitat Autònoma de Barcelona

José João Almeida,
Universidade do Minho

José Paulo Leal,
Universidade do Porto

Joseba Abaitua,
Universidad de Deusto

Juan-Manuel Torres-Moreno,
Lab. Informatique d'Avignon - UAPV

Kepa Sarasola,
Euskal Herriko Unibertsitatea

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Marcos Garcia,
Universidade da Corunha

María Inés Torres,
Euskal Herriko Unibertsitatea

Maria das Graças Volpe Nunes,
Universidade de São Paulo

Mercè Lorente Casafont,
Universitat Pompeu Fabra

Miguel Solla Portela,
Universidade de Vigo

Mikel Forcada,
Universitat d'Alacant

Pablo Gamallo Otero,
Universidade de Santiago de Compostela

Patrícia Cunha França,
Universidade do Minho

Rui Pedro Marques,
Universidade de Lisboa

Susana Afonso Cavadas,
University of Sheffield

Tony Berber Sardinha,
Pontifícia Univ. Católica de São Paulo

Xavier Gómez Guinovart,
Universidade de Vigo

POP: Por Outras Palavras

Proposta recebida em October 2018 e aceite para publicação em December 2018.

Alinhamentos Parafrásticos PE–PB de Construções de Predicados Verbais com o Pronome Clítico *lhe*

EP–BP Paraphrastic Alignments of Verbal Constructions Involving the Clitic Pronoun *lhe*

Ida Rebelo-Arnold
Universidad de Valladolid
imdamotar@funge.uva.es

Anabela Barreiro
INESC-ID
anabela.barreiro@inesc-id.pt

Paulo Quaresma
Universidade de Évora
pq@uevora.pt

Cristina Mota
INESC-ID
cmota@islt.utl.pt

Resumo

Este artigo apresenta o alinhamento de construções contendo predicados verbais com o clítico *lhe* nas variedades de Português Europeu (PE) e Português do Brasil (PB), como nas frases *Já lhe arrumaram a bagagem — Sua bagagem está seguramente guardada*, onde a próclise do dativo *lhe* em PE contrasta com o pronome possessivo *sua* em PB. Selecionámos vários pares contrastivos de paráfrases, tais como pronomes clíticos em próclise e ênclise, pronomes ocorrendo em presença de pronomes relativos e de advérbios de negação, entre outras construções a fim de ilustrar esse fenómeno linguístico. Algumas diferenças correspondem a contrastes reais entre as duas variedades de Português, enquanto que outras representam escolhas puramente estilísticas. As variantes contrastivas foram alinhadas manualmente a fim de estabelecer um conjunto padrão, e a tipologia estabelecida de forma a poder ser futuramente ampliada e disponibilizada ao público. Os alinhamentos dos pares de paráfrases foram executados no corpus e-PACT usando a ferramenta CLUE-Aligner. Esta pesquisa foi desenvolvida no âmbito do projeto eSPERTo.

Palavras chave

Paráfrases, parafraseamento automático, compostos verbais, pronomes clíticos, português europeu, português do Brasil, alinhamentos parafrásticos

Abstract

This paper presents the alignment of verbal predicate constructions with the clitic pronoun *lhe* in the European (EP) and Brazilian (BP) varieties of Portuguese, such as in the sentences *Já lhe arrumaram a bagagem — Sua bagagem está seguramente guardada* "His baggage is safely stowed away", where the EP dative proclisis *lhe* contrasts with the BP possessive pronoun *sua*. We have selected several different paraphrastic contrasts, such as proclisis and enclisis, clitic pronouns co-occurring with relative pronouns and negation-type adverbs, among other constructions to illustrate the linguistic phenomenon. Some differences correspond to real contrasts between the two Portuguese varieties, while others purely represent stylistic choices. The contrasting variants were manually aligned in order to constitute a gold standard dataset, and a typology has been established to be further enlarged and made publicly available. The paraphrastic alignments were performed in the e-PACT corpus using the CLUE-Aligner tool. The research work was developed in the framework of the eSPERTo project.

Keywords

Paraphrases, automated paraphrasing, verbal compounds, clitic pronouns, European Portuguese, Brazilian Portuguese, paraphrastic alignments

1 Introdução

Neste artigo propomo-nos abordar o uso do clítico *lhe* em Português Europeu (PE) e Português do Brasil (PB). Nossa metodologia consiste em aplicar conhecimento linguístico ao alinhamento de pares de paráfrases contrastivas entre as duas variedades, e nosso objetivo principal é discutir os diferentes comportamentos semântico-sintáticos do pronome clítico *lhe* nas construções em que ocorre, assim como definir uma tipologia para os diferentes usos.

Analisamos pares de unidades parafrásticas alinhadas e retiradas de um subcorpus do e-PACT[1], um corpus paralelo escrito de paráfrases

[1] e-PACT é um acrónimo para eSPERTo Paraphrase Alignment Corpus of Translations, em português, Corpus de Traduções de Paráfrases Alinhadas do eSPERTo.

DOI: 10.21814/lm.10.2.287

alinhadas—(Barreiro & Mota, 2017). O subcorpus compreende dois romances, o Romance 1 tem 1.628 frases e 35.495 palavras em PE e 35.572 em PB, enquanto que o Romance 2 tem 1.041 frases e 22.001 em EP e 24.113 palavras em BP, respetivamente. Os exemplos ilustrativos apresentados neste artigo representam traduções EN–PE e EN–PB das mesmas obras de ficção de David Lodge. Essas obras incluem alguns diálogos e exemplos de comunicação oral informal, com uma mistura de construções simples e complexas. Frases bem formadas, ou construções contendo um uso convencional dos clíticos em PE e em PB, incluindo *lhe*, o que envolve exemplos não convencionais. Depois de analisar um conjunto de ocorrências com *lhe* no corpus, estabelecemos uma tipologia que cobre os usos mais fequentes desse clítico. Além disso, exploramos também um tipo de anotação computacional na qual os alinhamentos parafrásticos podem ser usados para criar gramáticas locais genéricas, e que servirão de base para o processamento automático de paráfrases. Os alinhamentos foram realizados através do uso da ferramenta CLUE-Aligner[2] (Barreiro et al., 2016).

Os pares de paráfrases contrastivas que resultaram deste estudo serão integrados numa ferramenta de paráfrases. Esses pares contrastivos possibilitarão a conversão, de uma variedade para a outra, das construções com *lhe*, como na frase *A Philip só ocorria um nome — Apenas um nome lhe veio à cabeça*, onde o complemento *A Philip* em PE representa um contraste com a próclise do dativo *lhe* em PB. É importante salientar que a maioria das ocorrências encontradas nos textos e apresentadas aqui foram mencionadas por autores que reconhecem a existência de uma variação em curso tanto em PE como em PB (Kato & Martins, 2016; Castilho, 2011, 2010; Perini, 2002; Neves, 2000; Cunha & Cintra, 1985), entretanto, nenhum dos casos foi descrito ou categorizado da maneira como é feito neste estudo, i.e., sob uma perspetiva computacional para uso em um sistema gerador de paráfrases, empregando uma ferramenta de alinhamento, usando corpora dos quais os pares de paráfrases são extraídos, e analisando os dados levantados para definir uma tipologia de contrastes entre as variedades PE–PB.

A pesquisa apresentada aqui foi desenvolvida no âmbito do projeto eSPERTo[3], que visa construir um sistema automatizado de paráfrases inovador, sensível ao contexto e linguisticamente aperfeiçoado, com capacidade para produzir construções semanticamente equivalentes e formas de expressão que auxiliem escritores e estudantes da língua portuguesa, tanto como língua estrangeira, quanto como língua nativa, na produção de textos, revisão, ou adaptação. Futuros desenvolvimentos do eSPERTo visam possibilitar a adaptação de um texto nas diferentes variedades do Português, como PE e PB (Barreiro & Mota, 2018; Barreiro et al., 2018).

2 Revisão da Literatura

Os clíticos são pronomes usados para substituir objetos diretos e indiretos, e podem assumir as funções acusativa e dativa. Em português, um pronome clítico exerce uma função sintática ao nível da frase, e pode ocorrer antes, no meio, ou depois do verbo, conforme a variedade usada. Nesse sentido, há importantes contrastes nas preferências sintáticas entre as variedades de PE e de PB. Evidências empíricas mostram que as regras de colocação nem sempre são claras para os falantes de ambas as variedades, por isso, a relevância em fornecer paráfrases entre as duas variedades reside no fato de evitar mal-entendidos ou em solucioná-los.

Os dados revelam que cada variedade tende a pôr em evidência as suas próprias preferências em relação ao uso dos clíticos. São elementos que podem ocorrer depois do verbo (ênclise), no meio do verbo, i.e., entre o radical e o morfema de tempo/pessoa (mesóclise), ou antes do verbo (próclise). A Tabela 1 apresenta a frequência de ocorrência dos clíticos nas duas obras completas (romances) de David Lodge das quais foram extraídas 40% das frases que constituem o corpus e-PACT. Para obter esses valores foi usado o analisador FreeLing (Padró, 2011) de forma a identificar os clíticos em uso. A seguir, foi desenvolvido um programa para contar as diferentes ocorrências, próclise e ênclise, levando-se em conta a estrutura de cada frase analisada.

Em geral, em PE, o clítico ocorre, com mais frequência, junto ao verbo do qual depende e ligado a esse por um hífen (enclítico). Em PB, por sua vez, ocorre como um item autônomo precedendo o verbo (proclítico) em frases declarativas.

[2] http://www.esperto.l2f.inesc-id.pt/esperto/aligner/index.pl

[3] Os experimentos usaram o eSPERTo para enriquecer os recursos parafrásticos em um sistema de diálogo, por exemplo, para aumentar o conhecimento linguístico de um agente virtual inteligente, e para produzir reduções de texto "inteligentes" em uma ferramenta de sumarização. Experimentos recentes visam fornecer novos recursos parafrásticos em um ambiente de aprendizagem da língua, e gerar paráfrases precisas para serem usadas em tradução automática e em tradução profissional, produção, edição e revisão de textos. http://www.esperto.l2f.inesc-id.pt/esperto/esperto/demo.pl

David Lodge		**me**	**te**	**se**	**lhe**	**nos**	**vos**	**lhes**	**a**	**o**	**as**	**os**
Romance 1	PE	407	16	331	109	48	0	13	52	74	9	14
	PE-Enclítico	221	9	168	64	23	0	6	38	48	6	11
	PE-Proclítico	186	7	163	45	25	0	7	14	26	3	3
	PB	281	2	285	26	28	0	0	50	53	4	22
	PB-Enclítico	69	1	75	6	7	0	0	35	36	3	18
	PB-Proclítico	212	1	210	20	21	0	0	15	17	1	4
Romance 2	PE	29	7	296	127	7	0	10	20	80	3	20
	PE-Enclítico	18	4	146	67	6	0	4	17	52	2	14
	PE-Proclítico	11	3	150	60	1	0	6	3	28	1	6
	PB	22	0	291	41	1	0	0	20	56	5	18
	PB-Enclítico	7	0	98	12	1	0	0	17	31	4	15
	PB-Proclítico	15	0	193	29	0	0	0	3	25	1	3

Tabela 1: Pronomes clíticos nas traduções em PE e PB dos romances de David Lodge.

Encontram-se, ainda assim, muitas nuances na colocação do clítico, que serão ilustradas neste artigo com exemplos do corpus. Importa assinalar que o uso mesoclítico é bastante comum em PE, sendo encontrados vários casos no nosso corpus, nenhum deles, porém, representa o clítico *lhe*. Em consequência disso, esse tipo de ocorrência deverá ser tratado em futuros trabalhos, e não neste. Alguns contrastes aqui tratados foram apontados e, esporadicamente, contemplados em análises de vários autores com mais ou menos detalhe (Costa & Grolla, 2017; Kato & Martins, 2016; Castilho, 2011; Castro, 2011; da Costa Pacheco, 2008; Pereira, 2007; Bagno, 2001). Ao comparar nossa perspetiva com trabalhos anteriores, de natureza teórica ou prática, as propriedades sintáticas revelam-se insuficientes para dar conta do fenómeno dos clíticos de maneira eficiente. Além disso, como o uso dos clíticos em português constitui um campo de pesquisa muito amplo, neste estudo nos concentramos nas paráfrases PE–PB envolvendo o clítico de terceira pessoa com valor dativo, *lhe*. E na análise deste clítico nos cingimos a uma porção do corpus que constui as duas obras, mais concretamente a um total de 475 frases.

Sob uma perspetiva linguística, a primeira gramática explicita o afastamento das regras estabelecidas pela gramática tradicional em relação a diferentes representações validadas pelo uso efetivo nas variedades em questão (Cunha & Cintra, 1985). Resumimos, a seguir, as particularidades encontradas na literatura mencionada acima, tanto em PE como em PB. De um lado, o PE (i) tem preferência pela ênclise e, apenas em alguns casos, seleciona a próclise, aceita também a mesóclise, considerada, em PB, como um uso arcaico para os clíticos; (ii) admite a elisão do dativo e do acusativo num mesmo item lexical; (iii) rejeita o pronome pessoal de uso nominativo em posição de acusativo; e (iv) apresenta o uso generalizado dos clíticos dativos como possessivos. Por outro lado, o PB (i) tem preferência pela próclise e, apenas em alguns casos, seleciona a ênclise; (ii) a mesóclise é inexistente tanto na modalidade escrita padrão como na modalidade falada, ainda que possa ser encontrada em um corpus literário; (iii) não admite elisão do dativo e do acusativo; (iv) seleciona o pronome pessoal de uso nominativo em posição de acusativo.

Sob uma perspetiva computacional, tomamos um conjunto de paráfrases entre PE e PB, resultantes de uma tarefa de alinhamento prévia e descrita para utilização em um sistema gerador de paráfrases (Barreiro & Mota, 2018). Tal sistema, com a ambição de incluir um módulo que tenha em conta a adaptação entre variedades de modo a lidar com as diferenças culturais, linguísticas e estilísticas entre essas variedades, requer um conjunto alargado de pares de paráfrases entre as variedades do português. Tanto esse sistema, como o conjunto de paráfrases, são recursos que ainda não se encontram disponíveis para o português. Nossa tarefa mais ampla consiste em reunir pares de paráfrases, incluindo unidades lexicais multipalavra e outras unidades frásicas, tais como os compostos *toda a gente* versus *todo o mundo* ou construções gerundivas [estar a + V-Inf] versus [ficar + V-Ger] (e.g., estive a observar — fiquei observando), entre outras. Neste artigo, seguimos nessa linha de investigação (Barreiro & Mota, 2018), mas com o enfoque no alinhamento de construções que ocorrem com o pronome clítico *lhe*. A recolha desses contrastes em corpora é muito importante, pois a ocorrência de fenômenos linguísticos em textos é indispensável a uma cobertura ampla e eficiente do processo de adaptação entre variedades. A conversão (semi-)

automática de textos de uma variedade em outra representa uma importante função em sistemas geradores de paráfrases. Além disso, os recursos resultantes da tarefa de alinhamento adicionam valor a outras aplicações, entre as quais, ensino-aprendizagem de línguas, sumarização, resposta a perguntas, diálogo, deteção de plágio, autoria e revisão de textos e tradução automática.

3 O Uso dos Clíticos em Português

Alinhamentos semi-automáticos permitem-nos avaliar o grau de aceitabilidade das frases selecionadas, já que são feitos por linguistas que são, também, falantes nativos de PE ou PB. Ainda assim, alguns comentários se fizeram necessários nos casos que apresentam uma distância considerável entre as variedades, resultando em paráfrases aproximadas, com valor semântico passível de mais interpretações ou com diferentes graus de precisão. Essas características nos parecem relevantes, principalmente, ao considerarmos a aplicação almejada. As seleções encontradas nas paráfrases vão variar conforme o objetivo: ensinar Português como Língua Estrangeira (PLE), ferramenta de revisão e edição de textos ou para um motor de buscas com alternativas entre variedades. Este artigo se concentra nesses usos, visando a descrição de ocorrências e a criação de tabelas lexico-gramáticas que as sistematizem; deixamos para pesquisas futuras a distinção entre paráfrases possíveis em ambas as variedades e paráfrases predominantemente estilísticas, apropriadas a qualquer das variedades.

3.1 Clíticos após Advérbios e Pronome Relativo *que*

O PE segue a regra geral pela qual o pronome relativo atrai o clítico mantendo-o em posição proclítica. Esse fato parece sugerir que a regra do antecedente, pelo menos em PE, sobrepõe-se à tendência à ênclise dessa variedade. O exemplo (1) expressa essa tendência, em que, em PE temos o *lhe* proclítico a seguir ao pronome relativo, enquanto que na paráfrase em PB o clítico é omitido. Em PB, há uma possível ambiguidade em relação ao sujeito que será desfeita no contexto mais largo.

(1) $_{EN}$ - *listened to what he took, at the time, to be a very funny parody*
$_{PE}$ - *ouvira o que* ***lhe*** *pareceu ser uma paródia muito divertida* - [V-PRO$_{DAT}$(PROCL) / ANTEC-QUE]
$_{PB}$ - *ouviu o que parecia ser [ ] uma paródia muito engraçada* - [V-PRO$_{\theta}$]

3.2 Dativo versus Pronome Nominativo

O uso de pronome dativo versus o nominativo nessa mesma posiçao é comum no contraste PE–PB. O exemplo (2) ilustra o contraste entre o uso do clítico com valor dativo *lhe* na posição enclítica, ou seja, depois do verbo *vendo*, em contraste com o uso do predicado preposicionado formado pela preposição *para* seguida de um pronome sujeito, ou seja, com valor nominativo (NOM), *ele*, i.e., *para ele*. Esse fenômeno também pode ocorrer com outras preposições, tais como *em* ou *com* (e.g., *nunca aconteceu com ele*).

(2) $_{EN}$ - *I'll sell* ***him*** *my [plane] ticket*
$_{PE}$ - *vendo*-***lhe*** *o bilhete* – [V–PRO$_{DAT}$(ENCL)]
$_{BP}$ - *vou vender a passagem* ***para ele*** – [V–PREP PRO$_{NOM}$]

Por outro lado, uma das ocorrências mais interessantes do nosso estudo é o uso do dativo clítico em PE, que tem como paráfrase em PB uma expressão com pronome possessivo. Esse caso foi amplamente referenciado e analisado por Santos (2015) e cuja leitura vem, indubitavelmente, ampliar a compreensão desse fenômeno já assinalidado no passado por Cunha & Cintra (1985)[4].

4 Tipologia das Construções com *lhe* no Corpus

A Tabela 2 apresenta a tipologia das construções com *lhe* no sub-corpus e-PACT por nós selecionado, em que contrastamos as variedades PE e PB. Como não existe nenhum caso de mesóclise no corpus analisado, este tipo de construção não se encontra ilustrado na Tabela.

O **primeiro exemplo** na Tabela ilustra, em PE, o verbo *sair* seguido de dois complementos, cada um precedido da preposição *a*. O primeiro complemento é [+HUM] e o segundo [+VALOR/dinheiro]. O complemento [+HUM] está expresso pelo clítico dativo *lhe* e o complemento referente ao valor monetário está expresso pela preposição seguida da entidade nomeada *a 300 dólares*. O segundo complemento desta paráfrase em PB não é precedido de preposição pela natureza do verbo *custar*, que não seleciona preposição. Temos Identidade Semântica

[4] A tradução em objetos nulos em PB parece mais frequente do que em PE, mas essa afirmação requer suporte de dados quantitativos. Esse apoio poderia ser dado pela análise mais detalhada de parágrafos em nosso corpus ou pelo uso de mais dados do COMPARA ou de outros corpora paralelos inglês-português.

	Alinhamentos Parafrásticos	**Clíticos**			**Identidade da Paráfrase**		
		VComp	Posição	Ant	Lex	Sem	Syn
1	it would cost him 300 dollars						
PE	ia sair-**lhe** a 300 dólares	Dat	Encl	-	-	+	Parcial
PB	ia custar-**lhe** 300 dólares	Dat	Encl	-		+	
2	looking out of the window still gives him vertigo						
PE	olhar pela janela continua a dar-**lhe** vertigens	Dat	Encl	-	Parcial	+	Inversão
PB	sentía vertigens só de olhar pela janelinha	Ø				+	
3	with which she prepared his breakfasts						
PE	com que **lhe** preparava os pequenos-almoços	Dat	Procl	Rel	-	+	Parcial
PB	com que preparava o seu café da manhã	Poss				+	
4	it had never happened to him						
PE	nunca tal **lhe** acontecera	Dat	Procl	Neg	+	+	-
PB	isso nunca tinha acontecido com ele	Prep+Nom	Encl	Neg	+	+	-
5	bestowing upon them the title						
PE	**lhe** conferia o título	Dat	Procl	-	-	+	-
PB	agraciava-os com o título	Acus	Encl	-	-	+	-

Tabela 2: Tipologia das construções com *lhe* no sub-corpus e-PACT em PE e PB.

integral (Sem), mas Identidade Sintática parcial (Syn) pela não-correspondência de todos os termos que organizam a sequência. Essa não-correspondência, entretanto, parece ficar apenas na estrutura superficial, já que a sequência [V + PREP + N+HUM + N+VALUE/dinheiro] revela-se adequada para expressar ambas as paráfrases. Mesmo que, para concretizar integralmente a paráfrase, não haja as mesmas imposições no que se refere à seleção de preposição no segundo argumento, a Identidade Semântica mantém-se integralmente.

O **segundo exemplo** revela algumas paráfrases um pouco mais complexas, à primeira vista, devido à alternância do elemento topicalizado nas construções do verbo-suporte, *dar-lhe vertigens* em PE e *sentia vertigens* em PB. Esta alternância é uma conseqüência do verbo-suporte selecionado em cada variante, *dar* em PE e *sentir* em PB. Se traduzirmos ambas as paráfrases de forma esquemática, teríamos: [isso dá-me N], em PE e [sinto / tenho N quando isso acontece], em PB. O pronome demonstrativo é o tópico em PE, porque é o agente do verbo *dar*. No entanto, em PB, a paráfrase seleciona o verbo estático *sentir*. Na paráfrase em PB, porém, a construção do verbo-suporte *sentia vertigens* tem um significado resultante devido à existência de uma unidade lexical multipalavra idiomática que tem um significado incoativo com o significado de fazer com que algo aconteça (por exemplo, *sinto enjôo só de olhar para a comida / só de entrar no carro / só de ver a estrada*). Assim, no caso de um aplicativo NLP, precisaríamos criar uma fórmula que possa ter em conta não apenas o deslocamento do tópico, mas também o verbo aspetual em PB com o significado da construção do verbo-suporte *continua a dar-lhe vertigens*. A seleção de um verbo-suporte diferente provoca o desaparecimento do clítico Dativo em PB. A Identidade Lexical (Lex) continua existindo pelo menos parcialmente. Vale a pena ressaltar que, em PE, o aspecto original de continuidade do inglês resultante do uso do advérbio *ainda* não é preservado na frase do PB, onde a noção de continuidade foi eliminada. Consideramos isso mais uma opção estilística do que um contraste real entre PE e PB.

No **terceiro exemplo**, a variação que salta à vista é Lexical, com a alternância entre *pequeno-almoço / café da manhã*, exemplos bem conhecidos de contraste entre PE e PB. Ambos ocorrem com o pronome relativo *que* como antecedente que obriga à próclise do clítico (*com que lhe preparava NP* e a Identidade Sintática das paráfrases seria total se não fosse pelo fato de que PB seleciona o possessivo *o seu* posposto ao verbo *preparava*, em vez do proclítico *lhe* mais o verbo em PE. Essa alternância *lhe*/possessivo entre PE e PB é amplamente registrada e relatada em gramáticas (Cunha & Cintra, 1985) e também pode ser confirmada em corpora, como no COMPARA[5] (Frankenberg-Garcia & Santos, 2003). De fato, parece ser uma escolha estilística constante no PE, criar uma estrutura mais complexa com a participação do clítico com verbos que o permitem, onde um possessivo é perfeitamente aceitável.

No **quarto exemplo**, o par de unidades parafrásticas, como nas anteriores, manifesta a presença de um antecedente, um advérbio de negação *nunca*, que requer a próclise do pronome

[5] http://www.linguateca.pt/COMPARA/

(PROCL), *nunca // lhe acontecera*. Este clítico é necessário para completar o significado do verbo *acontecer* no contexto. No entanto, não é indispensável para o verbo em si, dado o seu valor intransitivo. Esta informação adicional, que em PE é transmitida pelo clítico, em PB é [PRO$_{NOM}$ *ele*] precedido pela preposição PREP *com*, ou seja, *com ele*, no corpus.

No **quinto exemplo**, em PE, o clítico *lhe* ocorre em uma posição pré-verbal sem a presença de um antecedente, apenas uma conjunção coordenada *e*, *e lhes conferia NP*. Em PB, o verbo *agraciar*, seleciona um complemento direto na paráfrase pelo clítico *os* em posição enclítica (ENCL). Novamente, a Identidade Sintática é desfeita pela ocorrência de elementos lexicais que preenchem a Identidade Semântica sem corresponder à mesma estrutura de complementos de predicado verbal. O proclítico Dativo ([PRO$_{DAT}$ *os*) em PE corresponde a um enclítico Acusativo ([PRO$_{ACC}$ *os*) em PB.

Algumas ressalvas devem ser feitas ao observar os dados da Tabela 2. A primeira ressalva é que parece que o PB busca formas de evitar o uso do pronome clítico *lhe*, sem necessariamente violar as regras da gramática. Isso se dá através da substituição do clítico por outros elementos, como em exemplos anteriores, ou selecionando outros itens lexicais que preenchem a Identidade Semântica (Sem). Essa seleção implica, com frequência, uma mudança total ou parcial na Identidade Sintática (Syn), como pode ser visto nas paráfrases selecionadas (exemplos 1, 2 e 5 da Tabela 2). A segunda ressalva diz respeito aos tempos verbais selecionados nas paráfrases, parece que são irrelevantes em termos de identidade semântica, pertencendo ao domínio do estilo e das escolhas de cada tradutor. Em qualquer caso, os tempos verbais selecionados não interferem no compartilhamento de informação entre paráfrases. Por fim, é importante lembrar que todos os contrastes apresentados neste artigo foram encontrados no contexto da tradução onde as escolhas de um indivíduo, o tradutor, determinam as construções da língua de destino encontradas no corpus analisado.

A respeito do montante de frases alinhadas, temos 475 paráfrases, compostas por 13.585 palavras em PE e por 14.126 palavras em PB. Nesse total de 475 paráfrases alinhadas encontramos 91 ocorrências do clítico *lhe*, sendo que uma dessas ocorrências constitui uma expressão idiomática com baixíssimo grau de composicionalidade — *tem que se lhe diga* — e não faz parte do escopo desta análise. Dentre as 90 ocorrências analisadas, temos uma maioria de casos em que o clítico ocorre em PE e, na respectiva paráfrase em PB, encontram-se soluções diversas que explicitamos a seguir.

Mais da metade (46) das ocorrências classificadas correspondem às categorias DAT/NONE 33% e DAT/POSS 17%, havendo, na paráfrase em PB, uma construção sem a presença do clítico *lhe*, no primeiro caso e, no segundo caso, a construção em PE corresponde a uma categoria recorrente de paráfrases entre as variedades onde a noção de pertença expressa em PE pelo clítico *lhe* é inexistente como tal em PB, sendo vertida por uma paráfrase com o possessivo.

Em outras duas categorias, DAT/ACC e DAT/PREP+NOMI PRON, encontram-se clíticos distintos de *lhe*, reforçando a percepção de evitamento da seleção desse clítico em PB. Por fim, temos a categoria em que tanto PE como PB selecionam o *lhe*. Essa categoria, representada por DAT/DAT é, porém, escassamente representada nas ocorrências analisadas no nosso corpus, constituindo 10% do total de ocorrências, ou seja, apenas 9 paráfrases alinhadas. Há, ainda, dois tipos de ocorrências excluídos da Tabela 1: NONE/DAT e DAT/Paráfrases aproximadas. Essas paráfrases não foram incluídas por representarem exemplos idiossincráticos no corpus e com baixa reprodutibilidade.

A Figura 1 ilustra uma gramática local que permite converter uma construção verbal predicativa em PE, onde aparece o enclítico após advérbios diferentes ou após os pronomes relativos *que* e *quem* (guardados numa variável $PRO), em uma construção equivalente em PB, onde o pronome clítico foi elidido, como em *não lhe disse que* → *Você não contou que*. A mesma gramática também permite a conversão mais comum entre ênclise e próclise, como por exemplo, em *entregou-lhe as chaves* e *lhe passou as chaves*. Em ambos os casos, o verbo (denotado por `<V>`) fica guardado na variável $V, mas em PE a frase será etiquetada com a anotação

`<REESCREVE+TIPO=PE2PBLHE+TEXTO=$V` e
`<REESCREVE+TIPO=PE2PBLHE+TEXTO=$PRO$V`,

enquanto que em PB será etiquetada com `REESCREVE+TIPO=PB2PELHE+TEXTO=$V`, onde $V e $PRO serão substituídos pelo respectivos verbo e pronome encontrados no texto. A gramática local foi desenvolvida no NooJ (Silberztein, 2016) e está disponível publicamente através do módulo do Port4NooJ v3.0 (Mota et al., 2016). Os resultados podem ser reproduzidos através do sistema de parafraseamento eSPERTo.

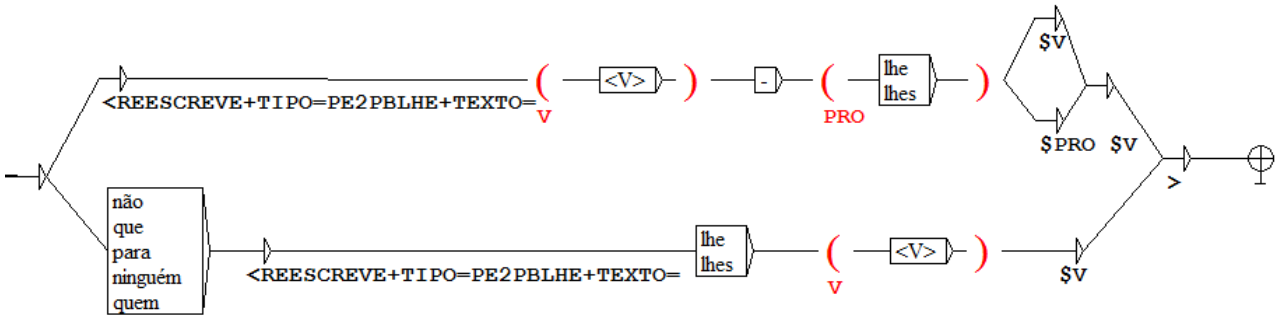


Figura 1: Gramática para formalizar a conversão de predicados verbais com *lhe* de PE em PB.

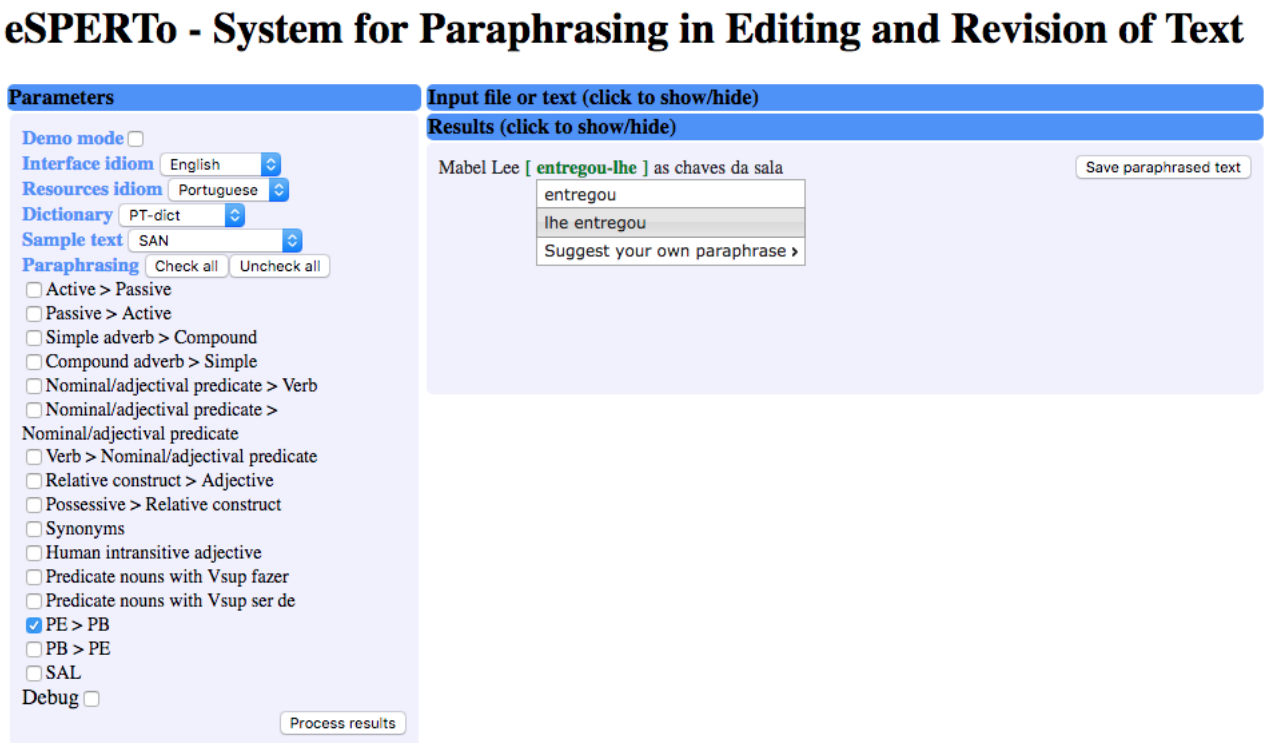


Figura 2: Conversão de um predicado verbal com o pronome clítico *lhe* de PE em PB.

A Figura 2 ilustra a capacidade de adaptação entre variedades dentro do eSPERTo, onde para uma frase escrita em PE, há sugestões para reescrevê-la em PB e vice-versa. Por exemplo, para a sentença do PE *Mabel Lee entregou-lhe as chaves da sala*, que no e-PACT corresponde à frase em PB *Mabel Lee lhe passou as chaves da sala*, o eSPERTo apresenta como opções de conversão para o predicado verbal com o enclítico *entregou-lhe* em EP, (i) o predicado verbal sem o clítico, *entregou*, e (ii) o predicado verbal com o proclítico *lhe entregou*. Se existe alguma das seguintes palavras: *não*, *que*, *para*, *ninguém* ou *quem* (a lista de palavras é muito maior), o pronome clítico migra para uma posição antes do verbo, como em *para lhe dizer*. A gramática permite a geração do PB *não digo* a partir do PE *não lhe digo* e a geração do PB *digo* e *lhe digo* a partir do PE *digo-lhe*. A capacidade de adaptação à variedade dentro do eSPERTo significa que, para uma frase escrita em PE, o sistema oferece sugestões para parafraseá-lo em PB. Em muitos casos, essa adaptação é extremamente útil quando o usuário deseja alcançar um público que fala a variedade com a qual ele está menos familiarizado.

5 Conclusões e Pesquisa Futura

A adaptação à variedade é uma característica importante do projeto eSPERTo, cujo foco principal é o desenvolvimento de um sistema de parafraseamento inovador, com capacidade de produzir frases e formas de expressão semanticamente equivalentes, mesmo quando contrastantes, como no caso de variedades da mesma língua. A colocação ou posicionamento de clíticos difere consideravelmente entre PE e PB, constituindo um desafio para a adaptação (semi-)automatizada entre

estas variedades. Um contraste claro é aquele que é exibido pelo pronome clítico *lhe*, para o qual mostramos as diferenças no comportamento sintático. Fizemos uma primeira tentativa de definir uma tipologia de contrastes parafrásticos e analisamos as diferentes formas de expressão. Alguns dos pares parafrásticos indicam um valor aproximado, que apesar de não assumirem uma correspondência semântica completa, são extremamente úteis e válidos para tarefas de parafraseamento, ou seja, na conversão entre variantes. No entanto, não fazemos (e não podemos, dado o tamanho e as características dos nossos dados) a distinção entre paráfrases que são possíveis de estabelecer, independentemente da variedade de português envolvido, e paráfrases contrastivas que são "obrigatórias", ou fortemente sugeridas, pelas diferenças entre as duas variedades.

Nossa tipologia e resultados iniciais foram alcançados pela análise de um subconjunto reduzido de ocorrências. No futuro próximo, pretendemos continuar o alinhamento das correspondências parafrásticas nos pares de frases PE-PB do corpus existente com relação à ampla variedade de pronomes clíticos. Planejamos alinhar a totalidade do corpus, pois ele pode fornecer uma fonte mais rica de paráfrases relacionadas com o fenômeno dos clíticos, que representa uma fonte relevante de contrastes entre as variedades PE-PB. Além do alinhamento completo do corpus, a fim de obter conclusões mais significativas sobre os contrastes de variedade envolvendo o pronome clítico, também é recomendável comparar esses resultados com dados maiores, a saber, comparar os pares contrastantes obtidos com os originais em PE e PB. Atualmente, a única ferramenta à nossa disposição é o CLUE-Aligner, que permite analisar duas línguas ou duas variedades da mesma língua simultaneamente. Podemos procurar a frase original em inglês, mas isso não está imediatamente disponível durante a tarefa de alinhamento. Para obter as frequências dos diferentes tipos de construções, mesmo que não estejam alinhadas, pode ser relevante obter uma imagem mais precisa do fenômeno, o que não incluímos aqui devido a restrições de espaço. No entanto, é importante criar corpora paralelos mais livremente disponíveis para o PE-PB para treinar e testar nossos resultados em sistemas de parafraseamento do mundo real, incluindo fenômenos que só podem ser encontrados em outros tipos de corpora paralelos, abrangendo não apenas textos genéricos, mas também ter em consideração paráfrases de diferentes gêneros textuais e domínios específicos ou especializados. Além disso, as legendas podem ser uma fonte interessante de corpora. O projeto Opus[6] contém subcorpora de OpenSubtitles, onde a língua portuguesa está incluída, apresentando grande quantidade de informação útil nesta área, oferecendo o alinhamento de legendas PE-PB em quantidade abrangente, apesar de apresentar "ruído".

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e Tecnologia através do projeto com a referência UID/CEC/50021/2013, do projeto exploratório eSPERTo com a referência EXPL/MHC-LIN/2260/2013, e através da bolsa de pós-doutoramento com a referência SFRH/BPD/91446/2012.

Referências

Bagno, Marcos. 2001. *Português ou Brasileiro: um convite à pesquisa*. Parábola.

Barreiro, Anabela & Cristina Mota. 2017. e-PACT: eSPERTo Paraphrase Aligned Corpus of EN-EP/BP Translations. *Tradução em Revista* 1(22). 87–102.

Barreiro, Anabela & Cristina Mota. 2018. Paraphrastic variance between European and Brazilian Portuguese. Em *5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 111–121.

Barreiro, Anabela, Francisco Raposo & Tiago Luís. 2016. CLUE-Aligner: An alignment tool to annotate pairs of paraphrastic and translation units. Em *10th Language Resources and Evaluation Conference (LREC)*, 7–13.

Barreiro, Anabela, Ida Rebelo-Arnold, Jorge Baptista, Cristina Mota & Isabel Garcez. 2018. Parafraseamento automático de registo informal em registo formal na língua portuguesa. *Linguamática* 10(2). 53–61.

Castilho, Ataliba. 2010. *Nova gramática do Português Brasileiro*. Contexto.

Castilho, Ataliba. 2011. O Português do Brasil. Em Rodolfo Ilari (ed.), *Linguística Românica*, 237–269. Ática.

Castro, Ivo. 2011. *Introduçao à história do Português*. Colibri.

Costa, João & Elaine Grolla. 2017. Pronomes, clíticos e objetos nulos: dados de produção e compreensão. Em *Aquisição de língua materna e não materna: questões gerais e dados do português*, 177–199. Language Science Press.

[6] http://opus.nlpl.eu

da Costa Pacheco, Juliana. 2008. *As construções médias do português do Brasil sob a perspectiva teórica da morfologia distribuída*: Universidade de São Paulo. Tese de Mestrado.

Cunha, Celso & Lindley Cintra. 1985. *Nova gramática do Português Contemporâneo*. Nova Fronteira.

Frankenberg-Garcia, Ana & Diana Santos. 2003. Introducing COMPARA: the Portuguese-English parallel corpus. Em Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds.), *Corpora in Translator Education*, 71–87. St. Jerome.

Kato, Mary & Ana Maria Martins. 2016. European Portuguese and Brazilian Portuguese: an overview on word order. Em *The Handbook of Portuguese Linguistics*, 15–40. Wiley-Blackwell.

Mota, Cristina, Paula Carvalho & Anabela Barreiro. 2016. Port4NooJ v3.0: Integrated linguistic resources for portuguese NLP. Em *10th Language Resources and Evaluation Conference (LREC)*, 1264–1269.

Neves, Maria Helena Moura. 2000. *Gramática de usos do Português*. UNESP.

Padró, Lluís. 2011. Analizadores multilingües en freeling. *Linguamatica* 3(2). 13–20.

Pereira, Shirley. 2007. *Estudio contrastivo del régimen verbal en el Portugués de Brasil y el Español Peninsular*: Universidade de Santiago de Compostela. Tese de Doutoramento.

Perini, Mário A. 2002. *Modern Portuguese: a reference grammar*. Yale University.

Santos, Diana. 2015. Os possessivos estão-me a complicar o ensino ;-) um estudo do dativo possessivo baseado em corpos. *Linguística: Revista de Estudos Linguísticos da Universidade do Porto* 10. 107–130.

Silberztein, Max. 2016. *Formalizing natural languages: the NooJ approach*. Wiley.

Proposta recebida em Novembro 2018 e aceite para publicação em Dezembro 2018.

Construções Conversas do Português do Brasil: Descrição e Classificação Iniciais

Converse constructions in Brazilian Portuguese: preliminary description and classification

Nathália Perussi Calcia
Universidade Federal de São Carlos
nathalia.perussi@gmail.com

Oto Araujo Vale
Universidade Federal de São Carlos
otovale@ufscar.br

Resumo

Os estudos que descrevem as construções com os verbos-suporte (*Vsup*) *dar*, *ter* e *fazer* apontam que grande parte dos substantivos predicativos (*Npred*) construídos com esses verbos aceitam a transformação denominada *Conversão*. A conversão é uma operação formal que estabelece uma relação não-orientada de equivalência sintática e semântica (parafrástica) entre duas frases elementares, tal como *dar um beijo/receber um beijo*. Nessa relação o nome predicativo é mantido e a posição dos argumentos é alterada, sem alterar os papeis semânticos. Nessas construções, a sentença de orientação ativa e o *Vsup* ativo são considerados *standard*; enquanto a sentença equivalente, de orientação passiva, é considerada conversa. Este trabalho apresenta os primeiros passos de uma descrição dessas construções no português brasileiro. O estudo baseia-se na metodologia de descrição do Léxico-Gramática, a partir de matrizes binárias nas quais as colunas representam as propriedades sintático-semânticas de cada construção. Os resultados do estudo de construções com verbo-suporte podem contribuir para análise de textos, identificando as informações e a forma da estrutura, e consequentemente, enriquecendo a descrição do Português Brasileiro. Além disso, a representação dos resultados em matrizes binárias prevê uma descrição formal, que poderá ser utilizada em aplicações no Processamento de Língua Natural.

Palavras chave

Conversão, Construção Conversa, Verbo-suporte, Léxico-Gramática

Abstract

Approaches to constructions with the support verbs *dar* (to give), *ter* (to have) and *fazer* (to make) in Brazilian Portuguese indicate that most of the predicative nouns combined with these verbs accept the transformation called Conversion. Conversion is a formal operation that establishes a non-oriented relation of syntactic and semantic (paraphrastic) equivalence between two elementary sentences, such as *Ana dá um beijo em Rui/ Rui recebe um beijo de Ana (Rui gives Ana a kiss/ Ana gets a kiss from Rui)*. In this relation the predicative noun is maintained and the argument position is changed without affecting their semantic roles. In these constructions, the active sentence and the active support verb are considered standard; while the equivalent passive sentence is considered a converse construction. This work presents the first steps of a description of these constructions in Brazilian Portuguese. The study is based on Lexicon-Grammar binary matrices, in which the columns represent the syntactic-semantic properties of each construction. This study results may contribute to the analysis of texts, identifying the information and form of the structure, and consequently, improving the description of Brazilian Portuguese. Also, the representation of the results in binary matrices provides a formal description that can be used in applications in Natural Language Processing.

Keywords

Conversion, Converse Construction, Support Verbs, Lexicon-Grammar

1 Introdução

O objetivo deste trabalho foi o de iniciar uma análise sobre um fenômeno da Conversão que ainda não tinha sido profundamente estudado em Português Brasileiro (PB), formalizar os dados obtidos por meio de critérios sintático-semânticos estudados no quadro do Léxico-Gramática de Maurice Gross (1981, 1975). A hipótese de base era que a construção com verbo-suporte resultante de uma Conversão transmite a mesma informação da construção com verbo-suporte *standard*, como nos exemplos[1] a seguir:

[1] A maioria dos exemplos apresentados foram retirados ou adaptados de ocorrências na Web encontradas por meio da ferramenta WebCorp (Renouf et al., 2007). Nos exemplos, a notação `<E>` significa elemento vazio, ou não ocorrência de nenhum elemento.

DOI: 10.21814/lm.10.2.290

(1) *Antes de vir, Zico me deu um conselho ⇔ Antes de vir, recebi um conselho do Zico.*

(2) *A seleção inglesa fez um convite a Felipão ⇔ Felipão recebeu convite da seleção inglesa.*

(3) *Os colegas têm respeito por Selton Mello pelo seu talento⇔ Selton Mello tem o respeito dos colegas pelo seu talento*

Os verbos *dar, fazer* e *ter* possuem grande produtividade na Língua Portuguesa, no caso das construções em que esses verbos ocorrem como verbo-suporte, como é visto nos exemplos, há a possibilidade de se formar outra construção por meio da Conversão. Para abordar as características dessa operação e seus principais objetivos de pesquisa, o trabalho será dividido em: i) introdução da definição de Conversão e suas principais características; ii) forma de obtenção dos dados e metodologia usada; iii) principais regularidades acerca dos resultados preliminares obtidos; e iv) conclusão e perspectivas futuras do estudo.

2 Conversão

Gaston Gross (1989) define a Conversão como uma transformação que estabelece uma relação não-orientada de equivalência sintática e semântica entre duas frases elementares. Em outros termos, é uma operação sintática em que há a permuta do argumento que está na posição de sujeito pelo argumento que está na posição de complemento preposicionado, sem que a informação de base da frase sofra alterações. Nessa relação, o nome predicativo é sempre mantido, pois ele é núcleo predicativo em uma frase com verbo-suporte, e seus argumentos (sujeito e complemento) trocam de ordem sem ocasionar alteração nos papeis semânticos. Algumas dessas construções podem apresentar uma relação parafrástica com outras. Essa operação é equivalente à passiva nas construções verbais, sendo, assim, considerada como um tipo de passiva nominal, segundo Gross (1993, 1989). A transformação de conversão foi estudada, entre outros, por Ranchhod (1990) e Baptista (1997, 2005) no Português Europeu (PE) e mais recentemente por Rassi et al. (2016) em um estudo comparativo dessas construções em PB e PE, e Calcia et al. (2017) acerca das construções conversas do *Vsup fazer*. Os seguintes exemplos apresentam uma frase *standard* e sua construção conversa equivalente:

(4) a. *A polícia deu instruções aos motoristas sobre o uso das balsas.*
b. *[Conversão] Os motoristas receberam instruções da polícia sobre o uso das balsas.*

No exemplo (4-a), *polícia* é, simultaneamente, o sujeito e agente da frase, enquanto *motoristas* é o complemento do nome predicativo *instruções*, com papel semântico de paciente. Já em (4-b), observa-se a troca dos argumentos em torno do núcleo predicativo, sem haver a alteração dos papéis semânticos e a substituição do *Vsup* elementar *dar* na frase standard pelo verbo *receber*, de orientação inversa (passiva), chamado de *Vsup* converso, por Gross (1989).

A construção com o verbo-suporte *dar* e o próprio verbo são designados de *standard*, enquanto a construção com o verbo-suporte *receber* e o próprio verbo, são denominados de conversos. Baptista (1997) complementa dizendo que o verbo-suporte *standard* condiz a uma frase de orientação ativa e o verbo-suporte converso a uma frase de orientação passiva. Como forma de mostrar as semelhanças existentes entre as construções conversas e as passivas verbais, Gross (1993) apresenta algumas propriedades comuns às duas construções, como:

i. Inversão dos argumentos:

(5) a. *Rui beijou Ana.*
b. *[Passiva] Ana foi beijada por Rui.*
c. *Rui deu um beijo em Ana.*
d. *[Conversão] Ana recebeu um beijo de Rui.*

ii. Apagamento do agente:

(6) a. *Os vizinhos ameaçavam frequentemente o músico.*
b. *Os vizinhos faziam ameaças frequentes ao músico.*
c. *O músico recebia ameaças frequentes (<E> + dos vizinhos + da parte dos vizinhos).*

iii. Bloqueio da passiva quando há elementos correferentes ao sujeito:

(7) a. *Ana deu uma ajuda a Maria, arrumando o quarto.*
b. **Maria recebeu uma ajuda de Ana, arrumando o quarto.*

Segundo Gross (1989, p. 9), a frase conversa deve possuir a mesma distribuição dos determinantes e o mesmo tipo e número de argumentos da frase *standard*. Outra característica das frases conversas é o fato de aceitarem a relativização,

porém, sem a redução do *Vsup* converso e, por consequência, sem a formação de grupo nominal (GN), como se nota em:

(8) a. *O jornal fez uma crítica (ao+do) projeto.*
b. *[Conversão] O projeto recebeu uma crítica do jornal.*
c. *[Relativização] A crítica que o projeto recebeu do jornal <foi admirável>.*
d. **A crítica (ao+do) projeto do jornal <foi admirável>.*
e. *[Redução do Vsup] = A crítica (a+do) projeto por parte do jornal <foi admirável>.*

O que causa a inaceitabilidade do GN como conversa em (8-d) é o fato de haver dois elementos introduzidos pela preposição *de*, fato que gera um problema de interpretabilidade, pois não se sabe qual deles é o sujeito, ou, no caso da preposição *a*, a interpretação da sequência *projeto do jornal* como um GN independente. Nota-se, porém, que a frase com a locução prepositiva *por parte de* torna-se aceitável, como em (8-e).

Nesse sentido, refere-se às construções conversas como passivas nominais e evidencia a importância do fenômeno do ponto de vista teórico. Nota-se que no exemplo (5-d) a mudança de orientação do sentido ativo para passivo numa construção verbal dá origem a uma construção passiva, já a mudança de orientação de ativo para passivo numa construção nominal dá origem a uma construção conversa. A principal diferença entre uma construção verbal e uma construção nominal, é que o núcleo predicativo de uma frase verbal é o próprio verbo, enquanto o núcleo predicativo de uma construção nominal é o nome predicativo.

3 Obtenção dos dados e metodologia do Léxico-Gramática

Os dados substanciais para a realização deste estudo foram os *Npred* associados aos *Vsup* obtidos por meio de três fontes: as descrições sobre as construções com o *Vsup dar* (Rassi, 2015); *fazer* (Barros, 2014); e *ter* (Santos, 2015). Após esse levantamento foram descritas e formalizadas as construções que apresentavam a transformação de conversão, segundo a metodologia do Léxico-Gramática (LG).

O LG propõe que seja feita uma investigação e descrição linguística formalizada em matrizes binárias, onde as linhas representam as entradas lexicais que não são simplesmente palavras, mas frases simples que correspondem a um predicado semântico. As colunas indicam as propriedades formais, distribucionais e transformacionais que as entradas lexicais podem apresentar. Na intersecção de cada linha e coluna é colocado um sinal (+ ou -) referente à entrada lexical apresentar ou não alguma propriedade, como mostra a Tabela 1.

Como se pode perceber, o valor de cada entrada lexical dá-se a partir da sua relação com as outras entradas, sendo assim, poucos itens apresentam a mesma distribuição que outros, dado que cada um deles tem comportamentos específicos. Com a publicação dessas matrizes, Laporte (2008) salienta que é possível observar se o julgamento e as precauções tomadas pelo linguista estão de acordo com os mesmos julgamentos dos demais falantes da língua. Desse modo, possíveis erros como a existência de colunas que correspondem a propriedades equivocadamente definidas, por exemplo, podem ser corrigidos pelo linguista. Além de sua publicação, as informações linguísticas formalizadas nas matrizes possuem interesse científico e técnico, pois podem ser facilmente adaptadas e implementadas em sistemas de Processamento de Linguagem Natural.

Os exemplos de frases representados na matriz não são frases encontradas em corpus, mas representam a constituição básica de cada construção. Porém os exemplos construídos foram atestados empiricamente por meio da ferramenta WebCorp (Renouf et al., 2007) que utiliza a Web como corpus.

4 Análise e classificação

Dentre as construções analisadas, cerca de 700 apresentaram a relação de conversão e foram dispostas em uma única matriz binária. Em um primeiro momento, optou-se em separá-las em grandes classes, classificando-as de acordo com os pares de *Vsup* que compõem a construção *standard* e a construção conversa, ambos elementares, devido à heterogeneidade dos nomes predicativos. Devido a isso, essa classificação não levou em consideração o conjunto de variantes estilísticas ou aspectuais dos verbos-suporte conversos, nem a homogeneidade sintática e semântica de certos nomes predicativos, que serão um dos objetivos de trabalhos futuros.

As construções conversas do PB, então, foram dispostas em quatro grandes classes: a classe DR (*dar- receber*), a classe DL (*dar-levar*), a classe FR (*fazer-receber*) e classe TT (*ter-ter*).

	classificação		*standard*				propriedades N1								propriedades N0			converso			nominalização				variante conv.							
Nome pred.	Classe PE (1997)	Classe PB	Vsup=:dar	Vsup=:fazer	Vsup=:ter	Argumentos	N1=:Nhum	N1=:N-hum	N1=:Red Npc	N1=:papel semântico	DET=:E	DET=:Def.	DET=:Indef.	Prep. conversa	N0=:Nhum	N0=:N-hum	N0=:papel semântico	Vsup=:receber	Vsup=:levar	Vsup=:ter	V-n	Vpleno corresp.	N-n	Nome corresp.	Vsup=:contar com	Vsup=:obter	Vsup=:ganhar	Vsup=:tomar	Vsup=:possuir	Vsup=:aceitar	Vsup=:sofrer	Exemplo
abertura	-	DR	+	-	-	3	+	-	-	patient	+	+	+	de, por parte de	+	-	experiencer-vol	+	-	+	+	abrir	-	-	-	+	-	-	-	-	-	João recebeu uma abertura da Ana [para falar].
abocanhada	-	DL	+	-	-	2	+	+	+	patient	-	-	+	de	+	+	agent-gen	+	+	-	+	abocanhar	-	-	-	-	-	+	-	-	-	João levou uma abocanhada do cão.
abordada	-	DL	+	-	-	3	+	-	-	object-gen	-	-	+	de, por parte de	+	-	agent-gen	+	+	+	+	abordar	-	-	-	-	-	-	-	-	-	João levou uma abordada do segurança [até a sáida].
abordagem	-	DL	+	+	-	2	+	-	-	object-gen	-	+	+	de, por parte de	+	-	agent-gen	+	+	+	+	abordar	-	-	-	-	-	+	-	-	+	João levou uma abordagem do policial.
abraço	-	DR	+	-	-	2	+	-	-	patient	-	-	+	de	+	-	agent-gen	+	-	-	+	abraçar	-	-	-	-	+	-	-	-	-	João recebeu um abraço da Ana.
abrigo	-	DR	+	-	-	2	+	-	-	patient	+	-	-	de	+	+	locative-place	+	-	+	+	abrigar	-	-	-	+	+	-	-	+	-	João recebeu abrigo da Ana.
absolvição	-	DR	+	-	-	2	+	-	-	patient	+	+	-	de, por parte de	+	-	agent-gen	+	-	+	+	absorver	-	-	-	+	+	-	-	-	-	O acusado recebeu absolvição da justiça.
acabamento	DL1	DR	+	+	-	2	-	+	-	object-gen	+	+	+	de	+	-	agent-cause	+	-	+	-	-	-	-	-	-	-	-	-	-	-	O quadro recebeu acabamento [de verniz] do artista.
ação	-	TT	-	-	+	2	-	+	-	object-gen	-	-	+	de			agent-cause	+	-	+	-	-	-	-	-	-	-	-	+	-	-	A empresa tem uma ação do empregado.
acariciada	-	DR	+	-	-	3	+	-	+	patient	-	-	+	de, por parte de	+	-	agent-gen	+	-	-	+	acariciar	+	carícia	-	-	+	-	-	-	-	João recebeu uma acariciada da Ana.
aceitação	-	DR	+	-	-	3	+	-	-	patient	+	+	-	de, por parte de	+	-	experiencer-vol	+	-	+	+	aceitar	-	-	+	+	+	-	-	-	-	João recebeu a aceitação da Ana.
acenada	-	DR	+	-	-	2	+	-	-	patient	-	-	+	de, por parte de	+	-	agent-gen	+	-	-	+	acenar	+	aceno	-	-	-	-	-	-	-	João recebeu uma acenada da Ana.
aceno	-	DR	+	+	-	2	+	-	-	patient	-	-	+	de, por parte de	+	-	agent-gen	+	-	-	+	acenar	-	-	-	-	-	-	-	-	-	João recebeu um aceno da Ana.
açoite	DL22	DL	+	-	-	2	+	-	-	patient	+	-	+	de	+	-	agent-gen	+	+	-	+	açoitar	-	-	-	-	-	+	-	-	-	João levou um açoite da Ana.
acolhimento	-	DR	+	+	-	2	+	-	-	patient	+	+	-	de, por parte de	+	-	locative-place	+	-	+	+	acolher	-	-	+	+	-	-	-	-	-	João recebeu acolhimento da Ana.
acomodação	-	DR	+	-	-	2	+	-	-	patient	+	+	-	de, por parte de	+	-	locative-place	+	-	+	+	acomodar	-	-	+	+	-	-	-	-	-	O hospede recebeu acomodação da Ana.
acompanhamento	-	DR	+	-	-	2	+	+	-	patient	+	+	+	de, por parte de	+	-	agent-gen	+	-	+	+	acompanhar	-	-	+	+	-	-	-	-	-	João recebeu acompanhamento do médico.
acondicionamento	-	DR	+	-	-	2	-	+	-	patient	+	-	+	de, por parte de	+	-	locative-place	+	-	+	+	acondicionar	-	-	-	-	-	-	-	-	-	A bebida recebeu acondicionamento [correto] do metre
aconselhamento	-	DR	+	+	-	3	+	-	-	addressee	+	-	+	de, por parte de	+	-	agent-speaker	+	-	+	+	aconselhar	+	conselho	+	+	-	-	-	+	-	João recebeu aconselhamento da Ana.
acordo	-	FR	-	+	+	3	+	-	-	object-gen	-	+	+	de, por parte de	+	-	agent-gen	+	-	+	+	acordar	-	-	-	+	-	-	+	-	-	João recebeu o acordo da Ana.
acusação	-	FR	-	+	-	3	+	-	-	patient	-	+	+	de	+	-	agent-cause	+	-	+	+	acusar	-	-	-	-	-	-	-	-	+	João recebeu uma acusação da polícia.
adaptação	-	DR	+	+	-	3	-	+	-	object-gen	+	-	+	de	+	-	agent-gen	+	-	+	+	adaptar	-	-	+	-	+	-	-	-	-	O anime recebeu uma adaptação da Ana.
adendo	-	FR	-	+	-	2	-	+	-	object-gen	-	-	+	de			agent-gen	+	-	-	-	-	-	-	-	+	-	-	-	-	-	O documento recebeu um adendo da Ana.
adesão	-	DR	+	+	-	2	+	+	-	object-gen	+	+	-	de, por parte de	+	-	agent-gen	+	-	+	+	aderir	-	-	+	-	+	-	-	-	-	O corretor recebeu adesão da administradora.
adestrada	-	DR	+	-	-	2	+	+	-	patient	+	-	+	de, por parte de	+	-	agent-gen	+	-	+	+	adestrar	-	-	-	-	-	-	-	-	-	O gato recebeu uma adestrada da Ana.
adeus	-	DR	+	-	-	2	+	+	-	patient	+	+	+	de, por parte de	+	-	agent-gen	+	-	-	-	-	-	-	-	-	-	-	-	-	-	O jogador recebeu o adeus da torcida.
adiada	-	DR	+	-	-	3	-	+	-	object-gen	-	-	+	por parte de	+	-	agent-gen	+	-	+	+	adiar	-	-	-	-	-	-	-	-	-	A adoção recebeu uma adiada da Ana.
adiantada	-	DR	+	-	-	3	-	+	-	object-gen	-	-	+	por parte de	+	-	agent-gen	+	-	+	+	adiantar	-	-	-	-	-	-	-	-	-	O trabalho recebeu uma adiantada do aluno.
admissão	-	DR	+	+	-	2	+	+	-	patient	+	+	-	de, por parte de	+	-	agent-gen	+	-	+	+	admitir	-	-	-	-	+	-	-	-	-	O aluno recebeu admissão da escola.
admoestação	-	DR	+	+	-	2	+	-	-	addressee	+	-	-	de, por parte de	+	-	agent-speaker	+	-	+	+	admoestar	-	-	-	+	-	-	-	-	-	O país recebeu admoestação do comandante.
adulação	-	FR	-	+	-	3	+	-	-	patient	+	-	+	de, por parte de	+	-	agent-cause	+	-	-	+	adular	-	-	-	-	+	-	-	-	-	O jogador recebeu uma adulação dos torcedores.
advertência	-	DL	+	+	-	3	+	-	-	addressee	+	+	+	de, por parte de	+	-	agent-speaker	+	+	+	+	advertir	-	-	-	-	-	+	-	-	-	O filho levou uma advertência da mãe.

Tabela 1: Léxico-Gramática referente às Construções Conversas (Calcia, 2016).

4.1 Construções conversas da classe DR

Além do *Vsup* receber, os nomes predicativos da classe DR, podem aceitar as variantes: *ter* (*Os bombeiros tiveram o apoio do Samu*); *contar com* (*O repórter contou com o auxílio da tradutora de sinais*); *obter* (*O cientista obteve suporte da Agência Espacial Brasileira*); *ganhar* (*O estudante ganhou uma ajuda de custo da secretaria*); *possuir* (*O estabelecimento possui o alvará de funcionamento*); e *aceitar* (*Derek Warwick aceitou uma carona de Gerhard Berger, piloto da Ferrari*). Dentre as variantes mencionadas, o *Vsup* ter é muito presente nas construções conversas da classe DR, mostrando sua forte tendência em ser tomado como um *Vsup* converso. Porém, esse fato não implica a criação de uma classe especifica para o par *dar-ter*, pois em todas as construções analisadas o *Vsup* receber também é aceito:

(9) a. *O Samu deu apoio aos bombeiros.*
b. *[Conversão] Os bombeiros (receberam + tiveram) o apoio do Samu.*

Na grande maioria das construções *dar-receber*, tanto sujeito como complemento são do tipo humano. Quando um argumento do tipo não-humano é encontrado, sempre ocupa a posição de sujeito da construção conversa e, consequentemente, a posição de complemento na construção standard (N_1 e N_0, respectivamente), como mostram os exemplos:

(10) a. *Acioli tratou a madeira com inseticida e deu duas demãos de verniz.*
b. *[Conversão] A madeira recebeu duas demãos de verniz de Acioli.*

Em relação às propriedades estruturais das construções conversas, alguns *Npred* não admitem a presença de um determinante:

(11) a. *O Tribunal deu (<E> + *a+ *uma) ciência à empresa da abertura do procedimento administrativo.*
b. *[Conversão] A própria empresa afirma que recebeu (<E> + *a + *uma) ciência da abertura de procedimento administrativo pelo Tribunal.*

Os nomes que designam atos de fala ou cumprimentos e que estão em sua forma plural foram destacados:

(12) a. *A migração internacional era maior em uma época que construiu, inclusive, uma Estátua da Liberdade para dar as boas-vindas aos imigrantes.*
b. *[Conversão] Os imigrantes recebiam as boas vindas da Estátua da Liberdade.*

Os demais nomes predicativos que pertencem à classe DR apresentam uma variação no que diz respeito aos determinantes, ou seja, apresentam determinantes definidos, indefinidos ou ambos, e isso ocorre pelo fato de os nomes predicativos dessa classe serem bastante heterogêneos.

4.2 Construções conversas da classe DL

Diferente da classe DR, que predominantemente não aceita nenhum nome construído com o *Vsup* converso *levar*, a maioria dos *Npred* da classe DL podem aceitar ambos os verbos, porém *receber* é muito menos representativo se comparado com *levar* nas construções desta classe, como é observado em:

(13) a. *No segundo tempo, André deu uma cotovelada em Jorge Andrade e foi expulso.*
b. *[Conversão] Jorge Andrade levou uma cotovelada de André.*

Grande parte dos nomes predicativos da classe DL aceita a variante *tomar* (*Rui tomou um tapa da Ana*) na construção conversa. Outras variantes aceitáveis são: *ter* (*Rui teve uma abordagem do policial*); *sofrer* (*O cantor sofreu um golpe da própria funcionária*); e em casos excepcionais podem aceitar a variante *ganhar* (*Giovanna Antonelli ganhou uma apalpada de Deborah Secco,*).

Quanto à estrutura sintática, foram encontradas regularidades mais precisas nas construções da classe DL. Enquanto na classe DR os determinantes e preposições alternam-se constantemente dependendo do *Npred*, na classe DL isso não ocorre com tanta frequência e os *Npred* parecem aceitar determinantes e preposições mais fixas. É possível notar a prevalência do determinante indefinido e da preposição de nas construções conversas desta classe, como mostram os exemplos:

(14) *Bruna levou (*<E> + *a + uma) agulhada (*por parte da + de) Ana.*

(15) *Richards levou (*<E> + *o + um) carrinho (*por parte de + de) Fagner.*

Outra diferença em comparação à classe DR está relacionada aos tipos de nomes predicativos, os

quais parecem ser muito mais homogêneos na classe DL, nos níveis sintático e semântico. Em outras palavras, o uso do *Vsup levar* nessas construções é mais frequente e comum em relação ao uso do *Vsup receber*, apesar de também ser aceitável em algumas construções da classe DL.

Pode-se dizer que a maioria dos nomes predicativos desta classe possui uma polaridade negativa, uma vez que se referem a um tipo de agressão (*bofetada, murro*), xingamento (*fodase*), punição (*castigo*), golpe (*machadada*), entre outros. Em geral, os *Npred* da classe DL aceitam que a posição sintática de segundo argumento da construção *standard* seja preenchida por um nome parte-do-corpo, como mostra o exemplo (16):

(16) a. *Edmundo, antes de ser substituído por Paulo Nunes, deu um soco no rosto de Cristaldo.*
b. *[Conversão] Cristaldo levou um soco de Edmundo.*
c. **Cristaldo levou um soco no rosto de Edmundo*

Na construção conversa há uma reestruturação do nome parte-do-corpo ao apagar o substantivo *rosto* da construção. Isso acontece devido à confusão que pode ocorrer caso esses nomes sejam mantidos, como em (16-c) . Além disso, há ainda, *Npred* que são derivados de nomes parte-do-corpo e de nomes de objetos, respectivamente:

(17) a. *Terry deu uma joelhada nas costas do atacante do Barça.*
b. *[Conversão] O atacante do Barça levou uma joelhada de Terry.*

Na classe DL há uma prevalência de nomes terminados em *-ada* e isso se dá pelo fato desses nomes serem nominalizações construídas a partir de um lema de um verbo (*apalpar* corresponde a *apalpada*); a partir de um lema de um substantivo concreto (*faca* corresponde a *facada*); ou por derivarem de nomes parte-do-corpo (*cotovelo* corresponde a *cotovelada*). A classe DL compreende ainda, uma parcela de *Npred* que fazem parte da terminologia do futebol (*carrinho, cartão amarelo/vermelho, cruzado, drible, finta, penalidade, expulsão*, entre outros). Há também, alguns nomes predicativos da classe DL que não se relacionam com nenhum dos tipos citados até então, mas que ainda assim, possuem propriedades sintáticas e semânticas comuns desta classe (como, por exemplo, *abordagem, autuação, cantada, esnobada, flagrante, olé*).

4.3 Construções conversas da classe FR

Dentre os *Npred* construídos com o *Vsup fazer*, há nomes que, predominantemente, também aceitam o *Vsup dar* na construção *standard* (*advertência, agradecimento, elogio*, entre outros). Na construção conversa, o *Vsup* elementar é *receber* (*Rui recebeu uma ameaça da Ana*), porém os *Npred* também podem aceitar as variantes *ter* (*Rui teve a companhia de Ana*), *sofrer* (*O vereador sofreu a cassação da Câmara*), *contar com* (*Rui contou com a caridade da Ana*), *possuir* (*O exemplar possui uma dedicatória do autor*), *ganhar* (*Neymar ganhou os elogios do técnico*), e *obter* (*O projeto obteve o fomento da instituição*). Durante a análise dos *Npred* da classe FR, foi constatado que os nomes que apresentam carga semântica negativa (por exemplo, *traição, cassação, conspiração*) aceitam muito bem a variante *sofrer* na construção conversa (*Rui sofreu uma traição por parte da Ana*).

Sobre as propriedades estruturais, as construções conversas da classe FR aceitam determinantes variados e as preposições *de* ou *por parte de*, como mostra o exemplo (18). Em alguns casos, o agente da construção pode ser apagado e isso ocasiona a exclusão da preposição do complemento da construção conversa, como é visto em (19):

(18) a. *A funcionária fez uma gentileza ao idoso.*
b. *[Conversão] O idoso recebeu (a + uma) gentileza (da + por parte da) funcionária.*

(19) a. *Ana fez uma injustiça com o Rui.*
b. *[Conversão] Rui sofreu uma injustiça .*

4.4 Construções conversas da classe TT.

Os *Npred* elencados nesta classe são os que apresentam o *Vsup ter*, como elementar, na construção standard e na construção conversa. Esses nomes também podem ser construídos com as variantes conversas *receber* (*A escola recebeu o investimento do governo*), *contar* com (*O corretor contou com a adesão da administradora*) e *ganhar* (*Miguel Trauco ganhou o afeto da fanática torcida rubro-negra*).

Sobre as propriedades distribucionais, pode-se destacar que a maioria dos *Npred* da classe TT possui sujeito e complemento do tipo humano, como é visto em (20). Em poucos casos, os argumentos podem ser do tipo não-humano, como mostra o exemplo (21). É importante lembrar

que o sujeito da construção conversa corresponde ao paciente da construção *standard* e o complemento da construção conversa ao agente da construção *standard*, pois os papeis semânticos não sofrem nenhum tipo de alteração.

(20) a. *Tite tem confiança em Neymar.*
b. *[Conversão] = Neymar tem a confiança de Tite.*

(21) a. *O clima tem uma grande influência sobre a agricultura.*
b. *[Conversão] = A agricultura tem grande influência do clima.*

A distribuição das preposições em algumas construções da classe TT é feita de maneira um pouco diferente das outras classes. Na construção *standard* o complemento preposicionado é introduzido pela preposição *sobre*, enquanto o complemento da construção conversa, assim como na maioria dos casos, é introduzido pela locução prepositiva *por parte de*:

(22) a. *O gestor tem controle sobre a alocação dos recursos.*
b. *[Conversão] A alocação de recursos tem controle por parte do gestor.*

5 Considerações finais e trabalho futuro

Em resumo, o que se observou foi que, a Conversão é uma propriedade transformacional que as construções nominais com os verbos-suporte *dar, fazer* e *ter* podem apresentar e que as construções com o verbo-suporte *dar* são as que mais produzem construções equivalentes com os verbos-suporte *receber* e *levar*, como mostra o quadro abaixo:

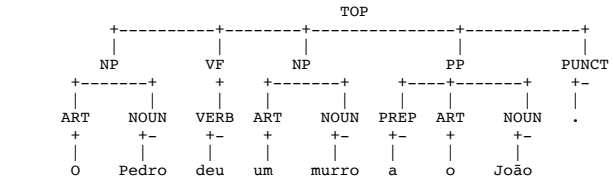
Classe	Estrutura	Exemplo	
DR	*N1(hum +nhum) receber Det N Prep N0 (hum)*	*alta, alvará, notícia, parecer, resposta, sinal, suporte*	406
DL	*N1(hum +nhum + RedNpc) levar Det N Prep N0(hum)*	*ataque, bronca, chute, facada, golpe, puxão, surra, susto*	204
FR	*N1 (hum + nhum) receber Det N Prep N0 (hum)*	*agressão, ameaça, falta, ofensa, solicitação, sugestão*	107
TT	*N1 (hum + nhum) ter Det N Prep N0 (hum).*	*amor, atenção, comando, cuidado, recorde, respeito*	16
		Total	733

Tabela 2: Classificação das construções conversas do PB (Calcia, 2016)

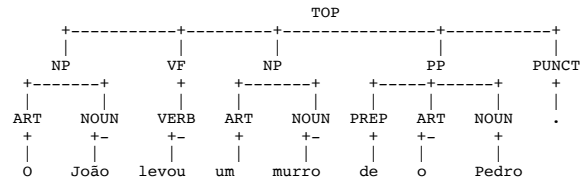
Este estudo resultou em um recurso linguístico que pode ser implementado em sistemas de Processamento de Linguagem Natural, referente a tarefas de identificação de paráfrases, por exemplo. Um exemplo de sistema é a *STRING* (Mamede et al., 2012)[2]. Trata-se de uma cadeia híbrida de processamento de língua natural que se baseia tanto métodos estatísticos quanto o processamento por regras. Utilizando para tanto tabelas do léxico-gramática como recurso para um *parser*. Naquele sistema já estão incorporados, para o português brasileiro, as tabelas do léxico-gramática de nomes predicativos de Barros (2014), Santos (2015) e Rassi (2015).

Na Figura 1 pode-se ver uma análise da STRING para o par de construções *standard* e conversa do nome predicativo *murro* no português europeu. Nosso próximo passo, portanto, será uma adaptação das tabelas para a inclusão naquele sistema.

```
"O Pedro deu um murro ao João. O João levou um murro do Pedro."
| xip/string.sh -t -tr -f -indent
```




```
MAIN(deu)
QUANTD_PRE(murro,um)
DETD(Pedro,O)
DETD(João,o)
VDOMAIN(deu,deu)
SUPPORT_VSUP-STANDARD(murro,deu)
MOD_POST_DAT(deu,João)
SUBJ_PRE(deu,Pedro)
CDIR_POST(deu,murro)
NE_PEOPLE_INDIVIDUAL(Pedro)
NE_PEOPLE_INDIVIDUAL(João)
EVENT_LEX(murro,outro)
EVENT_OTHER(murro)
0>TOP{NP{O Pedro} VF{deu} NP{um murro} PP{a o João} .}
```




```
MAIN(levou)
QUANTD_PRE(murro,um)
DETD(João,O)
DETD(Pedro,o)
VDOMAIN(levou,levou)
SUPPORT_VSUP-CONVERSE(murro,levou)
MOD_POST(murro,Pedro)
SUBJ_PRE(levou,João)
CDIR_POST(levou,murro)
NE_PEOPLE_INDIVIDUAL(João)
NE_PEOPLE_INDIVIDUAL(Pedro)
EVENT_LEX(murro,outro)
EVENT_OTHER(murro)
1>TOP{NP{O João} VF{levou} NP{um murro} PP{de o Pedro} .}
```

Figura 1: Exemplo de parsing de construções *standard* e conversas pelo sistema *STRING* (figura gentilmente fornecida por Jorge Baptista em comunicação pessoal).

Além disso, a partir da classificação geral feita por este trabalho, outras possibilidades de agru-

[2]C.f. https://string.l2f.inesc-id.pt

pamento poderão ser discutidas e realizadas, em trabalhos futuros. Por exemplo, haveria como estabelecer uma gradação da polaridade negativa dos nomes predicativos dessas construções a partir dos verbos suporte selecionados? Ou ainda, em termos de polaridade, qual seria o papel dos modificadores que aparecem junto aos nomes predicativos?

Pretende-se ainda estudar a abrangência e utilização dos *Vsup standards* e conversos em *corpora* de especialidades, aprofundar o estudo sobre a relação que existe entre as construções com os *Vsup fazer* e *dar* e, posteriormente, indexar os dados formalizados em bases de dados de predicados nominais, além da identificação de novas variantes dos verbo-suporte.

Agradecimentos

Este trabalho foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES – Código de Financiamento 001 e pela Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP (proc. 2016/24670-3).

Referências

Baptista, Jorge. 1997. Sermão, tareia e facada. uma classificação das construções conversas dar-levar. *Seminários de Linguistica* 1. 5–37.

Baptista, Jorge. 2005. *Sintaxe dos predicados nominais: com ser de*. Fundação Calouste Gulbenkian.

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

Calcia, Nathalia Perussi. 2016. *Descrição e classificação das construções conversas no Português do Brasil*: Universidade Federal de São Carlos. Tese de Mestrado.

Calcia, Nathalia Perussi, Cláudia Dias de Barros & Oto Araújo Vale. 2017. Sofrer uma ofensa, receber uma advertência: Verbos-suporte conversos de fazer no PB. Em *Symposium in Information and Human Language Technology*, 240–246.

Gross, Gaston. 1989. *Les constructions converses du français*. Genebra: Librairie Droz.

Gross, Gaston. 1993. Les passifs nominaux. *Langages* 109. 103–125.

Gross, Maurice. 1975. *Méthodes en syntaxe*. Hermann.

Gross, Maurice. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages* 63. 7–52.

Laporte, Eric. 2008. Exemplos atestados e exemplos construídos na prática do léxico-gramática. *(Con)textos Linguísticos* 2(2). 26–51.

Mamede, Nuno, Jorge Baptista, Cláudio Diniz & Vera Cabarrão. 2012. STRING: an hybrid statistical and rule-based natural language processing chain for Portuguesep. Em *10th International Conference on Computational Processing of Portuguese*, s. pp.

Ranchhod, Elisabete Marques. 1990. *Sintaxe dos predicados nominais com "estar"*. Instituto Nacional de Investigação Científica.

Rassi, Amanda Pontes. 2015. *Descrição, classificação e processamento automático das construções com o verbo 'dar' em português brasileiro*: Universidade Federal de São Carlos. Tese de Doutoramento.

Rassi, Amanda Pontes, Nathalia Perussi Calcia, Oto Araújo Vale & Jorge Baptista. 2016. Estudo contrastivo sobre as construções conversas em PB e PE. Em *Léxico e suas Interfaces: descrição, reflexão e ensino*, 199–218.

Renouf, Antoinette, Andrew Kehoe & Jay Banerjee. 2007. WebCorp: an integrated system for web text search. *Language and Computers* 59. 47–67.

Santos, Maria Cristina Andrade dos. 2015. *Descrição dos predicados nominais com o verbo-suporte ter no Português do Brasil*: Universidade Federal de São Carlos. Tese de Doutoramento.

Proposta recebida em Novembro 2018 e aceite para publicação em Dezembro 2018.

Paráfrase de advérbios terminados em *–mente* em Português

Paraphrasing Portuguese Adverbs ending in *–mente*

Jorge Baptista
U.Algarve-FCHS, Campus de Gambelas, P-8005-139 Faro, Portugal
INESC-ID Lisboa, L2F-Spoken Language Lab, R. Alves Redol 9, 1000-029 Lisboa, Portugal
jbaptis@ualg.pt

Resumo

Neste artigo, partimos da análise léxico-sintático-semântica das propriedades que foram usadas para classificar advérbios terminados em *–mente* mais frequentes em português e exploramos a geração de diferentes padrões de paráfrase, tanto estruturas regulares ou muito gerais, tais como os advérbios de modo e de ponto de vista, bem como outros, menos produtivos (e às vezes idiomáticas). O objetivo é fornecer um abrangente conjunto de estratégias de paráfrase, que podem ser usadas em várias aplicações de processamento de linguagem natural, como a simplificação de texto ou até mesmo tradução automática.

Palavras chave

advérbios, terminado em *–mente*, paráfrase, desambiguação de sentido, léxico-gramática, português

Abstract

In this paper, we depart from the lexical-syntactic-semantic properties that were used to classify the most frequent adverbs in Portuguese ending in *–mente* '–ly', and explore the generation of different paraphrasing patterns, both regular or very general structures, such as those for manner and view point adverbs, as well as other, less productive (and sometimes idiomatic) structures. The goal is to provide a comprehensive set of paraphrasing strategies, which can be used in several natural language applications, like text simplification or even machine translation.

Keywords

adverbs, ending in *–mente*, paraphrase, word sense disambiguation, lexicon-grammar, Portuguese

1 Introdução

As construções adverbiais são uma parte importante do conteúdo de qualquer texto e mostram uma sintaxe complexa, que pode ser vista como um desafio para muitas aplicações de Processamento de Linguagem Natural (PLN). As suas propriedades formais (sintáticas) incluem: (i) o escopo dos advérbios (um único constituinte ou uma frase inteira); (ii) posição (básica), (iii) quantificação, (iv) a parte do discurso ou categoria morfossintática (PoS) que modificam; (v) o tipo de paráfrase(s) que podem permitir. É este último tipo de propriedade que será o foco deste artigo.

Além disso, adotamos o conceito de *paráfrase* na perspetiva de Harris (1991) e do Léxico-Gramática (Gross, 1975, 1996b), isto é, sempre que há uma relação de equivalência transformacional entre frases, o que requer que o mesmo material lexical plenamente significativo (mesmo que de uma forma diferente) esteja envolvido, excluindo-se, portanto, situações de mera sinonímia. A paráfrase é, portanto, uma ferramenta teoricamente motivada para a descrição linguística, embora, até onde sabemos, as questões decorrentes da exploração sistemática de mecanismos parafrásticos envolvidos nas construções adverbiais portuguesas em textos reais (concatenados) não tenham sido descritas anteriormente.

Neste artigo exploratório, partimos das propriedades léxico-sintático-semânticas (Molinier & Levrier, 2000) que foram usadas para classificar os advérbios mais frequentes terminados em *–mente* em português (Fernandes, 2011). A nossa hipótese de partida é que a classificação léxico-sintática é a chave para produzir paráfrases adequadas para esses advérbios e testamos a geração de diferentes padrões de paráfrase em exemplos reais selecionados aleatoriamente a partir de um corpus.

Por outro lado, pretendemos obter um conjunto abrangente de estratégias de paráfrase, juntamente com algumas restrições à sua aplicação, que podem ser usadas como diretrizes para a descrição sistemática de construções adverbiais, e podem ser usadas em várias aplicações de PLN, como a simplificação de texto ou até mesmo tradução automática.

DOI: 10.21814/lm.10.2.289

A tarefa de desambiguação de sentido de palavra (ing. *word-sense disambiguation*, WSD) é considerada como um passo anterior para a tarefa de parafrasear estas expressões. Uma abordagem anterior baseada de aprendizagem de máquina para WSD dos advérbios derivados mais frequentes, terminados em *–mente*, do Português do Brasil (Fernandes, 2011) relatou uma precisão geral de 81%. Portanto, somente (ou principalmente) as paráfrases de advérbios não ambíguos (monossémicos) serão consideradas aqui.

Este artigo está organizado do seguinte modo: Começamos por uma sucinta revisão da literatura (secção 2) e apresentamos os métodos aqui utilizados (secção 3), para, logo de seguida, enumerar as diferentes paráfrases consideradas neste estudo (secção 4). Apresentamos, então, os resultados obtidos (secção 5), que comentamos em pormenor. O artigo termina (secção 6) com uma síntese das principais conclusões e apontando perspetivas de trabalho futuro.

2 Revisão da literatura

Embora muitos autores tenham produzido descrições perspicazes sobre a sintaxe e a semântica dos advérbios (Costa, 2008; Ernst, 2002; Real Academia Española, 2010; Kovacci, 2000), esses estudos consistem principalmente em observações esparsas e em alguns esclarecimentos quanto às suas propriedades semânticas e (mais raramente) sintáticas (ou seja, formais). Além disso, as taxonomias e esquemas de classificação, quando produzidos, frequentemente mostram critérios que se sobrepõem. É, portanto, seguro dizer que, até onde sabemos, nenhuma descrição sistemática e abrangente desta categoria morfossintática foi produzida para qualquer língua natural, exceto talvez para advérbios compostos franceses (Gross, 1996a) e particularmente para advérbios derivados que terminam em *–ment* (Molinier & Levrier, 2000).

Neste artigo, adotamos a abordagem do Léxico-Gramática (Gross, 1996b) para a descrição da língua e a classificação de advérbios originalmente proposta por (Gross, 1996a), e posteriormente adaptada à descrição dos advérbios portugueses, nomeadamente, aos advérbios que ocorrem com maior frequência e terminam em *–mente* do português brasileiro (Fernandes, 2011); e aos advérbios compostos do português europeu (Palma, 2009). Neste quadro teórico, as construções adverbiais são organizadas em 9 classes: 3 tipos principais de advérbios modificadores de frases (classes Px) e 6 tipos principais de advérbios modificadores internos de proposição (classes Mx). Por falta de espaço, uma descrição detalhada da sintaxe e semântica dessas construções não pode ser apresentada aqui, pelo que remetemos o leitor para os trabalhos acima referidos.

3 Métodos

A partir do corpus CETEMPúblico (Santos & Rocha, 2001), extraímos primeiro todos os lemas de todas as palavras analisadas como advérbios e terminados em *–mente* (4.384). Destes, um número considerável é constituído claramente por formas com erros ortográficos ou de digitação (e.g. *abolutamente* e *aboslutamente*, por *absolutamente*), incluindo nomes terminados em *-mento* com um erro na última vogal (e.g. *adiamente*, por *adiamento*, e *afastamente*, por *afastamento*), erros de acentuação (e.g. *simultâneamente*, por *simultaneamente*), etc.

Praticamente todas as formas válidas dos advérbios encontrados no corpus já estão presentes no léxico do sistema STRING (Mamede et al., 2012)[1]. Neste momento, este léxico contém mais de 6.800 advérbios terminados em *-mente*. Note-se que as variantes ortográficas devidas à presença de consoantes surdas etimológicas (e.g. *actualmente*) são consideradas como formas corretas e associadas ao respetivo lema (v.g. *atualmente*), o qual é estabelecido segundo o Acordo Ortográfico e para a variante europeia.

Naturalmente, nem todos estes advérbios dispõem ainda de uma descrição sintática no léxico do sistema. Das 974 construções adverbiais já descritas no léxico apenas 28 ($> 3\%$) não ocorreram no corpus.

Usando a informação codificada nesse léxico, observamos que 110 (11,3%) dos advérbios encontrados no corpus são ambíguos, com 2 ou mais sentidos (e distinto comportamento sintático-semântico correspondente). Por exemplo, *pontualmente* 'ocasionalmente/atempadamente' pode funcionar: (i) como um modificador interno de frase, relacionado com tempo/aspeto (frequência) (classe MT), como em *Pontualmente, o Pedro faz isso* (com aproximadamente o mesmo significado que *ocasionalmente*); ou (ii) como um advérbio modificador interno à frase, orientado ao sujeito (classe MS), como em *O Pedro chegou pontualmente à reunião* =*O Pedro foi muito pontual a chegar à reunião.*

Em seguida, para cada classe sintático-semântica, foram descritas as principais estratégias de paráfrase, utilizando transdutores de

[1] `https://string.l2f.inesc-id.pt` [31/12/2018].

estados finitos com a ferramenta de processamento de corpus Unitex 3.1 (Paumier, 2016). A descrição linguística baseou-se nas concordâncias dos advérbios de cada classe que foram obtidas a partir do corpus, excluindo primeiro aqueles que seriam utilizados na avaliação.

Um conjunto de transdutores de estados finito foi construído em duas etapas, usando as informações mencionadas na Secção 4 e codificadas no léxico para substituir os advérbios pelas paráfrases apropriadas. Estas informações correspondem às propriedades léxico-sintáticas das construções adverbiais e são representadas num formato tabular.

O grafo da Figura 1 ilustra a primeira etapa do processo e constitui um grafo de referência (aqui ligeiramente simplificado) em que essas propriedades são representadas por variáveis (`@X`), em que `X` representa o número da coluna correspondente dessa tabela. A variável `@B` representa o advérbio-alvo; `@D` e `@E` as variantes *PrepC* (Secção §4.1); `@F` a forma adjetival equivalente; as variáveis `@H` a `@H` são propriedades binárias, que só dão origem às transduções indicadas a seguir sse tiverem o valor '+'. Desse modo, apenas são produzidos nos transdutores finais os caminhos relevantes para cada construção adverbial.

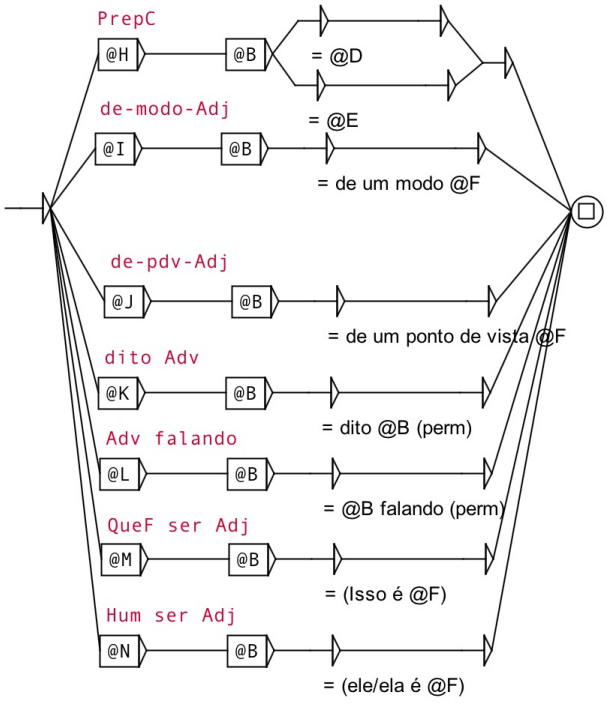


Figura 1: Grafo de referência

Os transdutores finais (aqui apresentados separadamente para cada propriedade, para uma maior clareza) são ilustrados pela Figura 2.

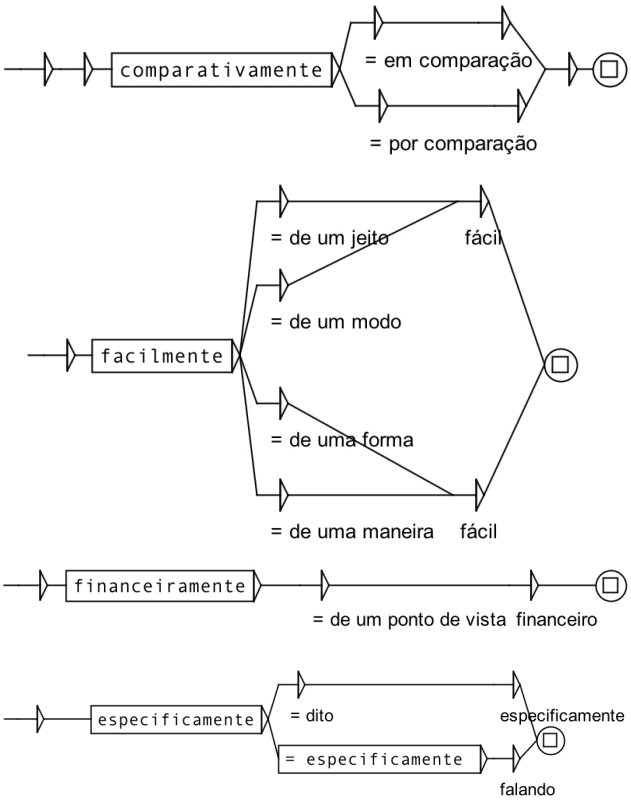


Figura 2: Exemplos de transdutores.

Estes transdutores limitam-se a inserir no texto, imediatamente a seguir ao advérbio, a paráfrase adequada.

Para a avaliação, selecionaram-se os 2 advérbios mais frequentes de cada classe sintático-semântica. No caso de algum desses advérbios ser ambíguo, isto é, apresentar mais do que uma construção sintático-semântica, descartámo-lo e escolhemos o advérbio seguinte, por ordem de frequência, dentro da mesma classe[2].

De seguida, foram aleatoriamente extraídas do corpus 10 concordâncias para cada um dos advérbios selecionados.

Finalmente, aplicaram-se os transdutores previamente construídos para obter as corrrespondentes paráfrases. A *qualidade* das frases parafraseadas, isto é, sua *equivalência semântica* à frase original extraída do corpus e a sua *aceitabilidade*, foi avaliada independentemente por dois anotadores, ambos falantes nativos de português europeu, e qualquer divergência discutida e resolvida. Como o objetivo do artigo é essencialmente exploratório, nenhuma avaliação do grau de concordância entre anotadores será fornecida aqui.

[2]Por razões práticas, que apresentaremos mais adiante, nem sempre foi possível fazê-lo de forma sistemática.

4 Paráfrases

Nesta secção, destacamos algumas das estruturas de paráfrase mais comuns encontradas para advérbios terminados em *–mente* em português europeu. Para maior clareza, eles serão organizados por classes, embora algumas destas estratégias possam dizer respeito a várias classes.

Devido à sua sintaxe mais complexa e por não serem propensos a se deixar parafrasear regularmente, os advérbios relacionados com o conceito de *tempo* (classe MT) (Hagège et al., 2009, 2010), os advérbios quantificadores (MQ) e os advérbios de foco (MF) (Baptista & Català, 2011), até porque constituem classes praticamente fechadas, não serão considerados neste artigo. São exemplos destas construções:

Atualmente/ correntemente/ antigamente/ diariamente, o Pedro faz/fazia isso [MT]
O Pedro ficou completamente/ totalmente esgotado [MQ]
O Pedro foi duplamente enganado [MQ]
O Pedro lê essencialmente/ basicamente este tipo de livros [MF]

4.1 PC (advérbio conjuntivo)

Uma das estratégias de paráfrase mais interessantes encontradas para advérbios que terminam em —*mente* é a possibilidade de produzir um grupo preposicional (notado *PrepC*) cuja cabeça é o substantivo morfologicamente associado ao adjetivo base de que o advérbio foi derivado, e.g. *consequentemente* → *em consequência*, *por consequência*.

Essas paráfrases nominais apresentam frequentemente uma certa fixidez quanto às combinações desse nome com a preposição e o determinante, ou apenas permitem uma variação muito limitada, lexicalmente determinada. Por isso, muitos deles já estão codificados no léxico da cadeia STRING como advérbios compostos (Mamede et al., 2012). Devido às idiossincrasias dessas combinações de palavras, essas paráfrases devem ser diretamente codificadas ou associadas à entrada lexical advérbio terminado em —*mente*.

Os diferentes valores de frequência dessas variantes podem sugerir uma estratégia de paráfrase. Por exemplo, além de *consequentemente* (204 ocorrências), ambas as paráfrases seguintes são encontradas: *em consequência* (767 ocorrências); e *por consequência* (354).

4.2 PS (advérbio disjuntivo de estilo)

Na perspetiva harrissiana (Harris, 1991, p. 91), advérbios disjuntivos de estilo (classe PS) operam como um advérbio de modo que modificam um operador metalinguístico performativo *Eu digo* subjacente a qualquer enunciado efetivamente produzido. Neste sentido, eles poderiam, em princípio, ser parafraseados reconstruindo tal operador (e transformando o enunciado de discurso direto em discurso indireto). Assim, para um enunciado como:

`par=ext803257-pol-95b-1:` *Especificamente, o principal motivo das consultas é a artrose do joelho, ...*

deve ser possível produzir uma frase como:

Eu digo especificamente que a principal razão para as consultas médicas é a artrite do joelho.

É possível observar variantes destas formas de base, nomeadamente uma oração reduzida de particípio (notada *dito Adv*), especialmente quando o advérbio é modificado por *mais*:

... dito mais especificamente, a principal razão para as consultas médicas ...

ou uma oração reduzida gerundiva (notada *Adv falando*), com *falar*:

Especificamente falando, o principal motivo das consultas ...

No entanto, como *especificamente* é ambíguo, sendo classificado tanto como um advérbio de modo (MV) como um advérbio de foco (MF), nos casos em que esta ambiguidade não foi adequadamente resolvida, são esperadas paráfrases incorretas.

4.3 PA (advérbios de atitude disjuntiva)

Esta classe compreende várias subclasses. Comecemos pelos advérbios modificadores de frase avaliativos (*PA:eval*):

`par=ext4944-eco-97b-2:` *Surpreendentemente, o juiz nem leu mais nada ...*

Nestas construções, o advérbio poderia ser considerado uma paráfrase de um verbo opinativo como *Eu acho/penso* tendo como argumento uma frase adjetival:

Eu acho (que é) surpreendente que o juiz nem tenha lido mais nada.

Consideramos ainda os advérbios modais (*PA:modal*), que atribuem uma modalidade específica à frase que modificam:

`par=ext436372-des-96a-2:` *Provavelmente,* [*ele*] *é mais inteligente do que Jean-Jacques.*

Independentemente da classe, estes tipos de construções adverbiais são muitas vezes equivalentes a uma construção adjetival com uma oração completiva sujeito (notado *QueF ser Adj*):

É provável que ele seja mais inteligente do que Jean-Jacques.

No entanto, esta é uma propriedade que sofre fortes restrições lexicais, exigindo descrição explícita, pois alguns advérbios impedem tal transformação:

Aparentemente, ele é mais inteligente do que Jean-Jacques;
cf. **É aparente que ele seja/é mais inteligente que Jean-Jacques.*

Devido à modalidade particular que estes advérbios introduzem, o tempo-modo da oração subordinada na construção adjetival equivalente tem de ser adequado em conformidade, o que representa um grau suplementar de complexidade na formulação da paráfrase:

Eu acho (*que é*) *surpreendente que o juiz nem tenha/ tivesse/ *tinha/ *tem lido mais nada.*
*É provável que ele seja/*é mais inteligente do que Jean-Jacques.*

Neste momento, ignoramos este tipo de alterações no tempo-modo das orações completivas das construções adjetivais equivalentes às construções adverbiais.

Além das subclasses mencionadas aqui, a classe PA também inclui duas outras subclasses: advérbios disjuntivos de hábito (*PA:habit*), e. g. *habitualmente*, e advérbios de frase orientados para o sujeito (*PA:subj-oriented*), por exemplo *inteligentemente*, e.g. *Inteligentemente, o Pedro não fez isso.* No entanto, como ainda existem poucos advérbios destas subclasses no léxico da STRING e muitas vezes são eles ambíguos com outros empregos, de outras classes, eles não serão abordados neste artigo.

4.4 MV (advérbios de modo)

Quantitativamente, esta é a mais importante classe léxico-sintática de advérbios. A típica estrutura de paráfrase envolve um grupo preposicional com os nomes-operadores de modo (Gross, 1996a) *modo*, *maneira*, *forma* e *jeito* (este último somente para o português do Brasil).

A aceitabilidade da paráfrase está estreitamente dependente da parte do discurso do elemento predicativo que estes advérbios modificam (adjetivo ou verbo), da sua posição relativa, e (muitas vezes) do grau de fixidez da combinação de palavras (colocação) (Vieira et al., 2012).

Considere-se, por exemplo, o advérbio *abertamente*, que pode combinar-se tanto com verbos (1.401), antes (1.256) ou depois (145) deles; como com adjetivos, mas apenas antes destes (131). Apesar do sentido um tanto idiomático do advérbio (não há relação transformacional sincrónica com o verbo *abrir* nem o adjectivo *aberto*), este pode quase sempre submeter-se à paráfrase característica com o nome-operador de modo (e suas variantes) quando combinado com um verbo:

`par=ext559653-eco-91b-2:` *Os Verdes, o partido que abertamente* [= *de um modo mais aberto*] *criticou a cimeira ...*
`par=ext21956-pol-95b-1:` *Powell criticou abertamente* [= *de um modo aberto*] *Robert McNamara ...*

No entanto, como um modificador à esquerda de um adjetivo, e.g.

`par=ext391158-opi-97a-2:` *Sou abertamente favorável à autonomia regional ...* ,

o advérbio parece não poder ser submetido a esta operação, nem antes nem depois do adjetivo:

**Sou de forma aberta favorável à autonomia regional.*
**Sou favorável de forma aberta à autonomia regional.*

Finalmente, deve notar-se que, independentemente da classe sintático-semântica, alguns advérbios que não pertencem à classe MV mas que terminam em —*mente* ainda retêm a possibilidade de serem parafraseados pelas construções (derivadas analiticamente?) com nomes-operadores de modo. Por exemplo, o advérbio *paradoxalmente* (classe PC):

`par=ext1450788-nd -91a-2:` *Por esta razão e de modo paradoxal* [= *paradoxalmente*] *... , nos anos 70, a população estrangeira aumentou em vez de diminuir ...*

Este não é frequentemente o caso de advérbios que não são da classe MV, como, por exemplo, o advérbio ambíguo *consequentemente*, que pode ser tanto um advérbio conjuntivo (PC) quanto um advérbio de modo (MV), e cujas construções

se podem distinguir consoante a respetiva posição na frase. No início de frase e destacado por vírgulas:

> *Consequentemente* [→ *De uma maneira consequente*], *o Pedro fez isso*,

apenas encontramos a construção PC, já que a paráfrase com nome-operador só seria interpretada como uma forma de advérbio MV; no final da frase, sem ser separado por vírgulas dos restantes elementos da frase, apenas a análise como MV, ainda que rara, é natural:

> *O Pedro fez isso consequentemente* [= *de uma maneira consequente*].

4.5 MS (advérbios de modo orientados para o sujeito)

Como parte da sua definição de duplo escopo, os advérbios de maneira orientados para o sujeito (MS), como *discretamente* (nos exemplos, nomes próprios foram abreviados):

> `par=ext203154-pol-92a-1:` *O dirigente socialista MS vai apoiar discretamente uma candidatura de NM à liderança*,

além de permitirem a paráfrase com o nome-operador de modo (e suas variantes):

> *O dirigente socialista MS vai apoiar de modo discreto a candidatura de NM à liderança*,

também permitem uma paráfrase com uma construção adjetival que capta a relação entre o advérbio e o sujeito:

> *O líder socialista MS será discreto ao apoiar a candidatura do NM à liderança.*

4.6 MP (advérbios no ponto de vista)

No início de uma frase, os advérbios de ponto de vista (MP) podem ser parafraseados geralmente por um grupo preposicional com o nome-operador composto *ponto de vista* (notado *pdv*):

> `par=ext1326334-eco-92a-1:` *Financeiramente* [= *de_o/um ponto de vista financeiro*], *o mercado de ações foi afetado*, ...

O enquadramento do conteúdo da oração principal pelo advérbio MP apela para o ponto de vista do locutor, o que pode parcialmente explicar a paráfrase com a oração reduzida gerundiva com *falar*:

> *Financeiramente falando, o mercado de ações foi afetado*,...

4.7 Variação posicional

Concluímos esta secção com uma observação sobre variação posicional, outra importante propriedade sintática para caracterizar (e distinguir) construções adverbiais.

Para os advérbios conjuntivos PC como *consequentemente*, verificamos que este se encontra separado por vírgula(s), no começo da frase, em 204 instâncias; 1.199 ocorrências no meio da frase; e, embora teoricamente possível, não se encontraram instâncias deste advérbio em final de frase. O advérbio homógrafo, da classe MV, é bastante raro (11 instâncias) e aparece apenas próximo de (após) um conjunto limitado de verbos, e.g. *agir*.

A variação posicional, embora não envolva exatamente a reformulação do advérbio noutra expressão equivalente, pode ser considerada um tipo especial de paráfrase, pois implica alterações (orto)gráficas no texto (ou seja, alterações de letras de maiúsculas para minúsculas e uso de pontuação). Dadas as dificuldades que essa variação levanta à geração de paráfrases, ela não foi considerada neste trabalho.

5 Resultados

As concordâncias dos 14 advérbios mais frequentes terminados em *–mente* (a maioria não ambíguos), dois de cada uma das classes selecionadas, foram recolhidas do corpus e 10 instâncias de cada advérbio foram selecionadas aleatoriamente para a avaliação.

A Tabela 1 mostra a informação codificada para cada advérbio no conjunto de teste e os resultados (precisão) obtidos pela avaliação tanto da identidade do significado quanto da aceitabilidade das paráfrases produzidas desta maneira. As lacunas lexicais estão marcadas com '–' e as propriedades não relevantes estão marcadas com 'x'.

Os valores de precisão indicados com '*' representam frases analisadas para as quais não se esperava à partida a possibilidade de estabelecer a paráfrase assinalada. A precisão total por tipo de paráfrase é assim indicada duas vezes: primeiro, considerando apenas os casos julgados relevantes para a classe sintático-semântica considerada (Total 1); e, depois (Total 2), todos os casos analisados, incluindo, portanto, os assinalados com '*'.

Globalmente, considerando apenas os casos relevantes (Total 1), obteve-se uma precisão de 0,79; quando se considera todos os casos analisados, esse valor desce para 0,77. Para maior clareza, os resultados serão comentados por tipo

de paráfrase e os exemplos são fornecidos da seguinte forma: primeiro o texto original, depois a paráfrase produzida automaticamente e, finalmente, a forma correta/desejada (sinalizada por '→').

5.1 *PrepC*

O estabelecimento de uma equivalência entre o advérbio e um grupo preposicional encontrou algumas dificuldades. Quando o advérbio é também modificado por um advérbio comparativo (v.g. *mais*, *tão*), o qual precede o advérbio, a ordem inicial das palavras é alterada ou a escolha do elemento comparativo deve sofrer mudanças, e.g.:

adaptar-se mais facilmente
**adaptar-se mais com facilidade*
→ *adaptar-se com mais facilidade*;

não se entende tão facilmente
**não se entende tão com facilidade*
→ *não se entende com tanta facilidade.*

Em construções passivas, a paráfrase pelo grupo preposicional é mais natural após o particípio passado, embora o advérbio também possa ocorrer antes dele:

A embaixada da Rússia foi também violentamente atacada pelos manifestantes
(cp. *A embaixada da Rússia foi atacada violentamente pelos manifestantes*)
**A embaixada da Rússia foi também com violência atacada pelos manifestantes*
→ *A embaixada da Rússia foi também atacada com violência pelos manifestantes.*

Certas restrições às combinações de palavras parecem estar relacionados com o estatuto de colocação dessas combinatórias lexicais (Vieira et al., 2012):

Cerqueira estava judicialmente impedido de exercer
**Cerqueira estava na/pela justiça impedido de exercer*
→ *Cerqueira estava impedido *na/?pela justiça de exercer.*

5.2 *de modo Adj*

Como esperado, a maioria dos exemplos de advérbios de modo (classe MV), e menos os advérbios orientados para o sujeito (classe MS), permitem a paráfrase com o nome-operador de modo (*forma*, *maneira* e *modo*; *jeito*, este último apenas em BP), e os erros são parcialmente devidos à já mencionada combinação com comparativos e a construção passiva. Devido à sua natural ambiguidade com os advérbios de modo (MV), os advérbios disjuntivos de estilo (classe PS) também permitem essa paráfrase. Ainda assim, como um PS, o advérbio *literalmente* parece ser mais usado para expressar uma modalidade (*realis*, ou modalidade real) do que propriamente o modo e a aceitabilidade das frases é geralmente duvidosa ou apenas mantém a interpretação da construção MV. Num uso claro, este valor de modalidade torna a frase inaceitável de uma forma mais evidente:

[*Eu*] *não sei literalmente* [= *absolutamente*] *nada*
*[*Eu*] *não sei de um modo literal nada.*

O resultado obtido com *especificamente* deveu-se ao seu uso predominante como advérbio de foco (MF). Surpreendentemente, o advérbio *curiosamente* (*PA:eval*) permite esta paráfrase. Como esperado, todos os advérbios de ponto de vista (MP) permitem esta paráfrase com o nome-operador ponto de vista (notado *pdv*).

As próximas propriedades são específicas de advérbios modificadores de frase.

5.3 *Dito Adv–mente* e *Adv–mente falando*

Embora apenas advérbios PS devessem aceitar estas paráfrases, nenhum dos 2 advérbios selecionados o faz, pelas razões já explicadas acima. Vale ressaltar que *comparativamente* (PC) permite a equivalência com a estrutura gerundiva:

`par=ext313886-des-94a-2:` *A corrida feminina produziu, comparativamente* [*falando*]*, melhores resultados [do] que a masculina.*

Surpreendentemente, porém, a maioria dos exemplos de advérbios de ponto de vista (MP) também podem ser parafraseados da mesma maneira.

A avaliação das propriedades seguintes, pela complexidade de que a geração das paráfrases correspondentes se reveste, consistiu basicamente em verificar se seria possível construir uma construção adjetival equivalente com oração completiva sujeito (Secção 5.4) ou com sujeito humano (Secção 5.5) e não propriamente uma avaliação da equivalência/aceitabilidade de frases concretas.

Class	**Adv-*mente***	**PrepC**	**Prep-C**	***de-modo*-Adj**	***de*-pdv-Adj**	***dito* Adv**	**Adv *falando***	**QueF *ser* Adj**	**N_0^{hum} *ser* Adj**
PC	*comparativamente*	*em/por comparação*	0,9	x	x	x	1,0*	x	x
PC	*finalmente*	*por fim*	1,0	x	x	x	x	x	x
PS	*especificamente*	-	x	0,3*	x	0,1	0,2	x	x
PS	*literalmente*	-	x	0,7*	x	0,0	0,7	x	x
PA:eval	*curiosamente*	*por curiosidade*	1,0	1,0*	x	x	x	1,0	x
PA:eval	*infelizmente*	-	0,9	x	x	x	x	-	x
PA:modal	*aparentemente*	-	0,9	x	x	x	x	-	x
PA:modal	*realmente*	*na realidade*	1,0	x	x	x	x	-	x
MV	*facilmente*	*com facilidade*	1,0	0,6	x	x	x	x	x
MV	*imediatamente*	*de/no imediato*	0,9	0,9	x	x	x	x	x
MS	*cuidadosamente*	*com cuidado*	1,0	0,5	x	x	x	x	1,0
MS	*violentamente*	*com violência*	1,0	0,6	x	x	x	x	0,8
MP	*financeiramente*	-	x	x	1,0	x	0,9	x	x
MP	*judicialmente*	*na justiça*	1,0	x	1,0	x	0,8	x	x
		Total 1 (casos relevantes)	0,96	0,65	1,00	0,05	0,58	1,00	0,90
		Total 2 (todos casos)	0,96	0,60	1,00	0,05	0,66	1,00	0,90

Tabela 1: Propriedades parafrásticas de advérbios selecionados e precisão alcançada.

5.4 (*Eu acho*) *QueF é Adj*

Esta propriedade aplica-se especificamente a advérbios avaliativos (subclasse *PA:eval*). É praticamente impossível produzir automaticamente paráfrases, a menos que a frase seja muito simples, ou a oração na qual o advérbio opera tenha sido corretamente extraída. Um exemplo possível é:

`par=ext1154964pol-98a-1:` [*Ele*] *não desmente, curiosamente, a citação ...*,

o que produziria:

→ (*Eu acho que*) *é curioso que* [*ele*] *não desminta a citação.*

Observe-se a mudança do modo-tempo do verbo na oração subordinada completiva:

desmente (indicativo-presente)
→ *desminta* (conjuntivo-presente)

e que aqui ignorámos, para efeitos de avaliação.

O advérbio *infelizmente* parece particularmente resiliente a esta transformação, mesmo em frases simples, e.g.:

`par=ext1198599-nd-91b-2:` *A guerra, infelizmente, criou situações que nos vão dar muito material para escrever*

o que não parece corresponder naturalmente a:

*[*Eu*] *acho que é infeliz que a guerra tenha criado situações que nos vão dar muito material para escrever.*

5.5 *Nhum ser Adj*

Esta propriedade corresponde à elicitação do duplo-escopo dos advérbios de maneira orientados para o sujeito (classe MS), portanto, pressupõe não só a correta análise sintática do sujeito do verbo, o que nem sempre é possível de fazer automaticamente, como a correta atribuição de um valor semântico de *humano*, dado o contexto em que esse nome ocorre. Os dois casos assinalados correspondem justamente a construções com sujeito não-humano (v.g. *camioneta* e *excesso de sódio*), em que ambos os advérbios surgem encaixados sob o verbo de uma oração relativa, pelo que seria necessário dispôr de um sistema de resolução de correferência (Marques, 2013).

`par=ext1551708-soc-93b-2:` ... *despiste de uma camioneta que embateu violentamente numa estação de correios ...*
`par=ext1376835-clt-soc-93b-1:` ... *não é só o excesso de sódio que ... ataca violentamente cada célula do nosso corpo*

6 Conclusão e trabalho futuro

Em suma e para concluir, parece ser possível produzir as alterações estritamente locais autorizadas pelos modificadores internos às proposições (classes Mx) usando as ferramentas aqui usadas, enquanto as propriedades específicas de advérbios modificadores de frases (classes Px) são

geralmente mais difíceis de formalizar. A maioria das propriedades parafrásicas é específica da classe ou mesmo lexicalmente dependente, e deve ser dada atenção a certos contextos sintáticos, nomeadamente a presença de outros modificadores adverbiais, a coordenação, integração em construções passivas e, em geral, o fato de que o advérbio está a modificar um verbo ou adjetivo. Está claro, a partir da amostra de advérbios aqui descrita, que a maioria das propriedades é lexicalmente dependente e que a tarefa de paráfrase não pode ser abordada apenas usando a classificação geral das construções adverbais, aqui resumidamente esboçada.

No futuro, está prevista atingirmos uma maior cobertura do léxico-gramática dos advérbios, com vista ao seu potencial para uso na descrição linguística de paráfrases em diferentes aplicações, como simplificação de texto, ferramentas tutoriais de aprendizagem de idiomas e em tradução. Em particular, ainda é necessário elaborar um catálogo de sentidos de advérbios de elevada granularidade, cobrindo uma proporção maior do léxico dos advérbios portugueses terminados em –mente, e associando-lhes as propriedades formais que permitem distinguir diferentes construções de um mesmo vocábulo. Isso só pode ser feito pela paciente descoberta das diferenças entre os usos de palavras, a fim de poder, então, anotar com segurança esses sentidos em textos. Tal permitirá usar técnicas de aprendizagem automática para a desambiguação de sentido destas palavras e, consequentemente, a sua utilização em tarefas que envolvam paráfrase.

Agradecimentos

A pesquisa para este trabalho foi parcialmente financiada pelo Governo Português, através da Fundação para a Ciência e a Tecnologia (ref. UID/CEC/50021/2019).

Referências

Baptista, Jorge & Dolors Català. 2011. Adverbes focalisateurs et analyse syntaxique automatique de groupes nominaux. Em *Passeurs de mots, passeurs d'espoir: lexicologie, terminologie et traduction face au défi de la diversité. Actes des 8èmes Journées scientifiques du LTT-AUF*, 97–110.

Costa, João. 2008. *O Advérbio em Português Europeu*. Colibri.

Ernst, Thomas. 2002. *The syntax of adjuncts*. Cambridge University Press.

Fernandes, Gaia. 2011. *Automatic Disambiguation of* -mente *ending Adverbs in Brazilian Portuguese*: Universidade do Algarve e Universitat Autònoma de Barcelona. Tese de Mestrado.

Gross, Maurice. 1975. *Méthodes en syntaxe*. Hermann.

Gross, Maurice. 1996a. *Grammaire transformationnelle du français: 3 - syntaxe de l'adverbe*. ASSTRIL.

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

Hagège, Caroline, Jorge Baptista & Nuno Mamede. 2009. Portuguese temporal expressions recognition: from TE characterization to an effective TER module implementation. Em *7th Brazilian Symposium in Information and Human Language Technology (STIL)*, 36–43.

Hagège, Caroline, Jorge Baptista & Nuno Mamede. 2010. Caracterização e processamento de expressões temporais em português. *Linguamática* 2(1). 63–76.

Harris, Zellig Sabettai. 1991. *A Theory of Language and Information. A Mathematical Approach*. Clarendon Press.

Kovacci, Ofelia. 2000. El adverbio. Em Ignacion Bosque & Violeta Demonte (eds.), *Gramática Descriptiva de la Lengua Española*, vol. 1, chap. 11, 705–786. Real Academia Española/Espasa.

Mamede, Nuno, Jorge Baptista, Cláudio Diniz & Vera Cabarrão. 2012. STRING - a hybrid statistical and rule-based natural language processing chain for Portuguese. Em *Computational Processing of the Portuguese Language (PROPOR)*, s/p.

Marques, João. 2013. *Anaphora resolution*: Instituto Superior Técnico - Universidade de Lisboa. Tese de Mestrado.

Molinier, Christian & Françoise Levrier. 2000. *Grammaire des adverbes: description des formes en* -ment. Droz.

Palma, Cristina. 2009. *Estudo contrastivo Português-Espanhol de expressões fixas adverbiais*: Universidade do Algarve. Tese de Mestrado.

Paumier, Sébastien. 2016. *Unitex 3.1 - user manual*. Université de Paris-Est/Marne-la-Vallée - Institut Gaspard Monge. http://igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf.

Real Academia Española. 2010. *Nueva gramática de la lengua Española - manual*. Associación de Academias de la Lengua Española.

Santos, Diana & Paulo Rocha. 2001. Evaluating CETEMPúblico: A free resource for Portuguese. Em *39th Annual Meeting of the Association for Computational Linguistics*, 442–449.

Vieira, Lucas Nunes, Cláudio Diniz, Nuno Mamede & Jorge Baptista. 2012. A lexicon of verb and *–mente* adverb collocations in Portuguese: Extraction from corpora and classification. Em *31st International Conference on Lexis and Grammar*, 155–162.

Proposta recebida em Novembro 2018 e aceite para publicação em Dezembro 2018.

Detecção de Paráfrases na Língua Portuguesa usando Sentence Embeddings

Detecting Paraphrases for Portuguese using Word and Sentence Embeddings

Marlo Souza
Universidade Federal da Bahia
msouza1@ufba.br

Leandro M. P. Sanches
Universidade Federal da Bahia
leandrompsanches@gmail.com

Resumo

A detecção (ou identificação) de paráfrases é a tarefa de determinar se duas ou mais sentenças de comprimento arbitrário possuem o mesmo significado. Os métodos para resolver esta tarefa com potenciais aplicações em sistemas de Processamento de Linguagem Natural. Este trabalho investiga a combinação de diferentes métodos de representação de sentenças em modelos de linguagem por espaços vetoriais e classificadores lineares para o problema de detecção de paráfrases para a língua portuguesa. Os resultados obtidos nesse trabalho estão aquém daqueles obtidos para a tarefa relacionada de detecção de implicação textual na avaliação ASSIN para a língua portuguesa, porém nesse trabalho investigamos a aplicação das representações vetoriais de sentenças para a detecção de paráfrases, outras características usualmente exploradas em sistemas desse tipo podem trivialmente ser incorporadas ao nosso método para melhorar a performance.

Palavras chave

Detecção de Paráfrases, Similaridade Semântica Textual, *Sentence Embeddings*

Abstract

Paraphrase detection/identification is the task of determining whether two or more sentences of arbitrary length possess the same meaning. Methods to solve this task have many potential applications in Natural Language Processing systems. This work investigates the combination of different methods of sentence representation in a vector space model of language and linear classifiers to the problem of paraphrase identification for the Portuguese language. The results obtained in this work are inferior to those obtained for the related task of recognizing textual entailment in the ASSIN evaluation for the Portuguese language, but we point out that in this work we investigate the application of sentence embeddings to the problem of paraphrase detection, as such other features usually explored in systems for this task may be trivially incorporated into our method to improve performance.

Palavras chave

Paraphrase Identification, Semantic Textual Similarity, Sentence Embeddings

1 Introdução

A identificação de paráfrase é a tarefa de determinar se duas ou mais sentenças de comprimento arbitrário possuem o mesmo significado. Para fins desse trabalho, consideraremos uma noção funcional (ou informacional) do que significa duas sentenças terem o mesmo significado. Seguindo as definições de Fonseca et al. (2016) na tarefa ASSIN – Avaliação de Similaridade Semântica e Inferência Textual – ocorrida em 2016, nós consideramos duas sentenças S_1 e S_2 como parafrásticas, se ao ler ambas, uma pessoa conclui que S_1 será verdade se, e somente se, S_2 também o for.

Métodos para detecção de paráfrases possuem aplicações para problemas como Sumarização Automática (Jing & McKeown, 2000), Recuperação de Informação, Sistemas de Resposta a Perguntas (Marsi & Krahmer, 2005), construção automatizada de ontologias (Suresh & Kumar, 2016), entre outros. Não é de se entranhar, portanto, que recentemente muito trabalho tenha sido produzido investigando métodos de identificação de paráfrases e da tarefa relacionada de similaridade semântica textual (*Semantic Textual Similarity* em inglês) (Fonseca et al., 2016; Socher et al., 2011; Yang et al., 2018). Dentre os métodos propostos na literatura, podemos distinguir abordagens baseadas em medir a similaridade lexical, contextual e semântica de sentenças.

Entre os que seguem a última abordagem, recentemente, dada a popularidade da aplicação de representações vetoriais de palavras (*word embeddings*) a várias tarefas de Processamento de Linguagem Natural (PLN), muito trabalho se concentrou no uso de modelos de representação semântica de sentenças através de veto-

DOI: 10.21814/lm.10.2.286

res —comumente chamados de *sentence embeddings*— para detecção de similaridade semântica textual. Uma representação vetorial de sentenças é um modelo de representação que transforma uma sentença de uma determinada linguagem em um vetor em um dado espaço vetorial de alta dimensão. Semelhante a modelos de representações vetoriais de palavras, supõe-se que a geometria do espaço vetorial usado para representar as sentenças codifica aspectos importantes de seu significado.

Representações vetoriais de sentenças foram aplicadas a muitos problemas no Processamento de Linguagem Natural, como Tradução Automática (Bahdanau et al., 2014), Análise de Sentimentos (Kiros et al., 2015), Geração Automática de Diálogos (Yang et al., 2018), etc. Particularmente, para a língua inglesa, os *benchmarks* criados para availar sistemas que medem a similaridade semântica entre sentenças tornaram-se recursos populares para avaliar a qualidade de modelos de representação vetorial de palavras (*word embeddings*) e de sentenças (*sentence embeddings*). Um exemplo de tais *benchmarks* é o conjunto SICK (Marelli et al., 2014) para similaridade semântica entre textos.

Este trabalho consiste de uma versão estendida do trabalho "Detecting Paraphrases for Portuguese using Word and Sentence Embeddings" apresentado no POP 2018 —*1st Workshop on Linguistic Tools and Resources for Paraphrasing in Portuguese* ocorrido conjuntamente com o PROPOR 2018 na cidade de Canela no Brasil. Nele, apresentamos investigações iniciais da aplicação de métodos de representações vetoriais de sentenças para identificação de paráfrase para a língua portuguesa. Em relação à publicação original, nós apresentamos aqui algumas avaliações posteriores realizadas para responder a questionamentos surgidos nas discussões do evento, como testar o efeito da calibração de parâmetros experimentais assim como do uso de diferentes modelos de representações vetoriais de sentenças nos resultados obtidos.

O presente trabalho está organizado da seguinte forma. Na Secção 2, discutimos alguns dos métodos de representação vetorial de sentenças utilizados nesse trabalho; na Secção 3, apresentamos os trabalhos relacionados ao nosso, i.e. trabalhos que tratam sobre deteção automática de paráfrases, com um foco na apresentação daqueles que utilizam o mesmo corpus que o nosso em seus experimentos; na Secção 4, descrevemos nosso trabalho experimental e discutimos os resultados obtidos; finalmente, na Secção 5, apresentamos algumas considerações finais.

2 Representações Geométricas de Palavras e Sentenças

Modelos de representação vetorial de palavras são modelos de linguagem que exploram a similaridade distribucional entre palavras em um grande corpus para aprender representações de palavras de uma linguagem como vetores em um dado espaço vetorial de alta dimensão. Da mesma forma, modelos de representação vetorial de sentenças visam codificar sentenças como vetores em um determinado espaço vetorial de forma a representar, na geometria do espaço vetorial, o significado original da sentença.

Dado um modelo de representação vetorial de palavras, podemos construir modelos simples para representação de uma sentença através da *agregação* das representações das palavras que a compõem. Tal método, apesar de simplório em primeira análise, pode ser justificado através do princípio da composicionalidade de significados, que afirma que o significado de uma sentença é obtido por alguma transformação no significado de seus constituintes. Assim, métodos seguindo essa abordagem (Mihalcea et al., 2006; Conneau et al., 2018) visam estabelecer alguma transformação que realiza tal *agregação*, i.e. de forma a codificar o significado de uma sentença a partir do significado das palavras que a constituem. De uma forma geral, métodos baseados em agregação buscam representar como o significado de palavras individuais contribuem para o significado da sentença.

Uma maneira trivial de fazê-lo é tomar o centroide da representação vetorial de todas as palavras (ou pelo menos das palavras lexicais) que constituem uma sentença como sua representação. Isso corresponde à ideia de que cada palavra contribui igualmente para determinar o significado da sentença. Essa representação vetorial pode ser obtida tomando a média dos vetores representando todas as palavras que compõem a sentença.

Não está claro, entretanto, que cada palavra contribui igualmente para o significado da sentença. De fato, algumas palavras podem atuar como marcadores gramaticais na sentença e seu significado individual pode não contribuir para o significado da frase, e.g. o caso da palavra *pas* que foi gramaticalizada na negação verbal "*ne ... pas*" ("não") em francês. Para explicar a diferença na importância de cada palavra para o significado da sentença, a representação da frase pode ser tomada como a *agregação ponderada* do vetor de cada palavra.

Muitas estratégias diferentes de ponderação podem ser estabelecidas de forma a levar em consideração a estrutura da frase ou as propriedades distributivas das palavras. Uma abordagem comum, semelhante à abordagem de Mihalcea et al. (2006) para calcular a similaridade de sentenças, é tomar o Inverso da Frequência nos Documentos (IDF, do inglês *Inverso Document Frequency*) de cada palavra em um dado corpus representativo como uma medida de importância para a palavra. A ideia fundamental de tal abordagem é que as palavras menos comuns da língua contribuam mais —ou tenham alguma *saliência*— no significado da sentença.

Observe que a maioria dos modelos de representação vetorial de palavras visa capturar padrões de co-ocorrência de palavras presentes no corpus de treinamento. No entanto, a presença de palavras fora de contexto pode causar ruído no modelo treinado (Arora et al., 2017). Assim, o método de agregação de representações de palavras para calcular a representação da sentença pode resultar num acúmulo de ruído e, portanto, degradar o significado da sentença em sua representação vetorial. Para superar esse problema, Arora e colegas propõem o uso de métodos de fatoração de matriz para identificar o componente principal dos vetores de palavra, que é interpretado como o *ruído acumulado* na representação da sentença por agregação. Tal ruído é, então, eliminado da representação da sentença. Essa técnica é conhecida como *Smooth Inverse Frequency* (SIF, ou Frequência Inversa Suave).

Trabalhos como o de Kiros et al. (2015), por outro lado, visam aplicar métodos de aprendizagem de máquina para aprender a representação de uma sentença a partir de seus padrões de distribuição em um grande corpus, similar aos métodos para aprendizagem de representações vetoriais de palavras. Esses métodos geralmente se baseiam nas representações das palavras que constituem uma sentença e tentam aprender com o corpus a melhor maneira de agregar tais representações para calcular a representação das sentenças. Recentemente, muitos métodos diferentes para aprender representações vetoriais de sentenças foram propostos na literatura, geralmente empregando redes neurais profundas e recorrentes para aprender tais representações (Conneau et al., 2017; Kiros et al., 2015; Le & Mikolov, 2014; Patro et al., 2018; Socher et al., 2011). Esses métodos foram aplicados com sucesso em diversas tarefas de Processamento de Linguagens Naturais (Cer et al., 2018; Howard & Ruder, 2018; Logeswaran & Lee, 2018).

Particularmente interessante para nós, é o método de Skip-Thought proposto por Kiros et al. (2015). Skip-Thought é um método não supervisionado de aprendizagem de representações vetoriais de sentenças usando uma arquitetura codificador-decodificador de redes neurais para prever a vizinhança de uma certa sentença, similar ao método *Skip-Gram* para aprender representações vetoriais de palavras. O impacto de tal trabalho reside no fato do mesmo descrever um método não-supervisionado para aprendizagem de tal representação e, portanto, não requerem dados anotados para seu treinamento. Métodos supervisionados para representações de sentenças, como InferSent (Conneau et al., 2017), por outro lado, provaram ser bem-sucedidos para aplicações específicas, mas vêm com o preço de depender de dados anotados – que podem não estar disponíveis para todos os idiomas.

Recentemente, Cer et al. (2018) propuseram o codificador universal de sentenças (*Universal Sentence Encoder*, em inglês), um método para aprender representações vetoriais de sentenças baseado em redes neurais que, de acordo com os autores, gera representações de uso geral para serem aplicadas em diversas tarefas de PLN. Os autores aplicam seus codificadores universais para cinco tarefas distintas: análise de sentimentos, detecção de subjetividade, classificação de questões, similaridade semântica textual e testes de associação implícita. Os autores apresentam resultados positivos para o uso de representações vetoriais de sentenças em tarefas de PLN, especialmente para os casos de baixa disponibilidade de dados.

Neste trabalho, focaremos na aplicação de diferentes modelos de representação de sentenças, e nas medidas de similaridade semântica relacionadas a esses modelos, ao problema de identificação de paráfrase para a língua português. Em certo sentido, nosso trabalho é semelhante ao de Feitosa & Pinheiro (2017), ou de Fialho et al. (2016), que avaliam o uso de algumas medidas de similaridade para o problema da similaridade semântica textual.

3 Identificação Automática de Paráfrases

Trabalhos sobre identificação de paráfrase podem ser divididos em três grandes categorias. Primeiro, há os trabalhos baseados em heurísticas, como medidas de semelhança semântica e tesauros ricamente anotados, como os trabalhos de Cordeiro et al. (2007) e Fernando & Stevenson (2008). Outros trabalhos, como o de Shinyama

et al. (2002), computam semelhanças contextuais entre palavras, como co-ocorrência em uma frase ou sintagmas, e exploram tais semelhanças para detectar semelhanças de significado entre duas sentenças – geralmente aplicando algoritmos de aprendizado de máquina para identificar as paráfrases. Finalmente, a terceira categoria de métodos se baseia em princípios de semântica distribucional, como a hipótese distribucional[1]

O trabalho de Cordeiro et al. (2007) propõe uma métrica para calcular a semelhança de significados entre duas sentenças, baseada na sobreposição de unidades lexicais. Observe que trabalhos que usam medidas de variação lexical/estrutural para identificar paráfrases, como o de Dolan et al. (2004) que usa a distância de Levenshtein entre duas sentenças, são capazes de identificar apenas aqueles exemplos em que as sentenças têm estrutura quase idêntica. Enquanto o trabalho de Cordeiro et al. (2007) evita muitas dessas armadilhas, uma vez que não se baseia na estrutura da sentença, como a sobreposição lexical é uma condição bastante restritiva para identificar paráfrases, sua abordagem é limitada no sentido de que não pode detectar paráfrases nas quais há variação significativa nas descrições de entidades e ações nas sentenças, como o uso de nomes diferentes e descrições definidas[2]. Considere, por exemplo, as sentenças "*Lula foi libertado nesta madrugada*" e "*O presidente Luís Inácio da Silva foi solto no início desta manhã.*" As frases claramente possuem significado semelhante, porém apresentam baixa sobreposição lexical.

Trabalhos como o de Mihalcea et al. (2006) e de Fernando & Stevenson (2008), por outro lado, propõem a exploração de medidas de similaridade entre sentenças para identificar paráfrases na língua inglesa, baseadas não somente na sobreposição lexical e semelhança estrutural, mas também em similaridades semânticas, contextuais e distributivas. Esses trabalhos exploram informações ricas com base em dicionários de sinônimos anotados, como a WordNet (Miller, 1995), e grandes corpora, como explorados por Turney & Littman (2002). Eles são flexíveis no sentido de que podem ser empregados usando diferentes medidas de similaridade, que exploram recursos linguísticos ricamente anotados ou grandes corpora não-anotados disponíveis para uma determinada linguagem.

Socher et al. (2011) empregam autocodificadores recursivos (RAE, *Recursive AutoEncoders*), um tipo de rede neural profunda não-supervisionada seguindo o modelo codificador-decodificador, para codificar a estrutura sintática das sentenças. Essas representações são, então, aplicadas para medir a semelhança de duas sentenças a nível de palavras e de seus constituintes sintáticos (sintagmas), que são então usadas para treinar um classificador de paráfrase.

Da mesma forma, Yin & Schütze (2015) propõem o uso de redes neurais convolucionais profundas para resolver o problema da detecção de paráfrase. Eles propõem uma nova arquitetura de rede neural que, segundo eles, permite codificar múltiplos níveis de granularidade do significado das sentenças. Essas representações são então usadas para treinar um classificador logístico para identificar paráfrases.

Esses trabalhos mais recentes são semelhantes ao de Mihalcea et al. (2006) e de Fernando & Stevenson (2008) por também explorarem a similaridade da distribuição de palavras e sintagmas em grandes corpora não anotados para calcular similaridade semântica entre sentenças. A diferença entre esses trabalho reside no fato das abordagens mais recentes considerarem a estrutura sintática da sentença para calcular a semelhança semântica entre elas, enquanto aqueles anteriores não levam essa informação em consideração.

O trabalho de Kiros et al. (2015) descreve um modelo de aprendizado não supervisionado de um codificador de sentença genérico, que pode ser aplicado a diferentes tarefas subsequentes de PLN. Semelhante ao que é feito para modelos de representação de palavras, os autores treinam uma arquitetura codificador-decodificador que tenta reconstruir as sentenças circundantes de uma passagem codificada. Os autores avaliam os modelos gerados pelo seu método em 8 tarefas: semelhança semântica, detecção de paráfrase, ranking de sentenças e imagens, classificação de tipo de pergunta e análise de sentimento e subjetividade.

Nosso trabalho decorre dos mais recentes que aplicam redes neurais e modelos de espaço vetorial para representar a informação semântica expressa em uma sentença. Nosso objetivo é avaliar sua utilidade para o problema de detecção de paráfrase para a língua portuguesa.

[1] A hipótese distribucional afirma que itens linguísticos com distribuições estatísticas semelhantes em grandes corpora têm significados semelhantes (Sahlgren, 2008).

[2] Note que essa crítica se justifica para a noção de paráfrase adotada nesse trabalho. Devemos salientar, entretanto, que noutra perspectiva de paráfrase, como a adotada pelo trabalho de Baptista nesse volume, tal abordagem está bem justificada, uma vez que duas sentenças são ditas parafrásticas quando "há uma equivalência *transformacional* entres sentenças que requer que o mesmo material lexical significativo esteja envolvido" (Baptista, 2018, tradução nossa).

Outros trabalhos sobre a detecção de paráfrase para o português foram realizados, especialmente no contexto da tarefa de avaliação conjunta ASSIN – Avaliação de Similaridade Semântica e Inferência Textual (Fonseca et al., 2016). Embora alguns desses trabalhos empreguem características obtidas com modelos de representação vetorial de palavras, mais notavelmente o trabalho de Hartmann (2016) para o problema de similaridade semântica, até onde sabemos, nenhum deles avaliou o uso diferentes métodos de representação de sentenças para esse problema.

O trabalho de Fialho et al. (2016) apresenta o método INESC-ID que utiliza informações diversas métricas de similaridade e sobreposição textual típicas da área de tradução automática para o problema de Inferência em Linguagem Natural ou Reconhecimento de Implicação Textual (RTE do inglês*Recognizing Textual Entailment*). Essa é uma tarefa mais geral que a detecção de paráfrase e, de fato, a engloba. Os autores relatam uma performance de 0.64 e 0.66 de medida F1 na tarefa RTE sobre o corpus ASSIN para as variantes europeia e brasileira da língua, respectivamente. Note entretanto que, como esses autores não avaliaram separadamente a performance do seu sistema na detecção de paráfrases, não podemos comparar nossos resultados com o deles.

Barbosa et al. (2016) apresentam o sistema Blue Man Group para o problema de RTE utilizando classificadores treinados sobre vetores de características obtidos usando a cosntrução de redes semânticas entre as palavras das duas sentenças e atributos de nível textual, baseados no trabalho de Kenter & De Rijke (2015). Os autores relatam uma medida F1 de 0.58 para o problema de RTE, mas não avaliam a performance de seu sistema sobre detecção de paráfrases separadamente.

O trabalho de Feitosa & Pinheiro (2017) avalia a aplicação de diversas medidas de similaridade textual - baseadas em aspectos sintáticos e semânticos do texto - à tarefa de RTE utilizando o corpus do ASSIN. Os autores relatam uma medida F1 de 0.71 em seus experimentos para a tarefa RTE, porém não avaliam seus resultados no reconhecimento de paráfrases somente e, portanto, seus resultados não podem ser comparados aos nossos.

O trabalho de Rocha & Lopes Cardoso (2018) utilizam classificadores multi-classe considerando características lexicais, sintáticas e semânticas de textos para a tarefa de RTE sobre o corpus ASSIN. Esses autores apresentam um resultado de 0.71 de (Macro) F1 para a tarefa de RTE. Enquanto os autores não apresentam os resultados de seu método para detecção de paráfrase sobre o conjunto de dados de teste, eles apresentam uma avaliação sobre o conjunto de treino usando validação cruzado, para o qual obtém uma medida F1 de 0.6 para a variante europeia do corpus e 0.52 para o corpus de treino com ambas as variantes combinadas.

O trabalho de Fonseca & Aluísio (2018) explora o uso de diferentes informações sintáticas para a tarefa de RTE também utilizando o corpus do ASSIN e alcança resultados 0.72 de medida F1. Esses autores também não apresentam seus resultados discriminando a performance no seu método para detecção de paráfrases, de modo que não podemos comparar nossos resultados com os deles.

Note que alguns trabalhos recentes tratando do problema de determinação de similaridade semântica entre textos na língua portuguesa também utilizam o corpus do ASSIN. É importante pontuar, entretanto, que enquanto os problemas de similaridade semântica e detecção de paráfrases são certamente relacionados, não é claro que um possa ser reduzido ao outro. Alguns desses trabalhos recentes, por exemplo (Silva et al., 2017; Pinheiro et al., 2017; Gonçalo Oliveira et al., 2017; de Barcelos Silva & Rigo, 2018; Alves et al., 2018), fornecem evidências de quais tipos de características linguísticas de um texto, como informação sintática, sobreposição lexical, etc. podem ser utilizadas também por sistemas de detecção de paráfrase.

Nesse trabalho, entretanto, nos concentraremos no uso das representações vetoriais de sentença – e as informações que podemos derivar com as mesmas como medidas de similaridade entre essas representações – para a detecção de paráfrases, sem considerar características de outras naturezas. A razão para tal escolha se recai no fato que estamos interessados em investigar quanta informação sobre o conteúdo da sentença pode ser codificado em sua representação vetorial.

4 Usando Representações de Sentenças para Detecção de Paráfrases

Nesta secção, descrevemos a implementação de classificadores de paráfrase que recebem duas sentenças e decidem se estas são parafrásticas. Investigamos diferentes classificadores lineares treinados em dados de representação das sentenças e similaridades obtidos com o uso de quatro diferentes formas de representação de sen-

tença. Abaixo, descrevemos os dados que usamos em nossos experimentos, bem como os resultados obtidos em nossa investigação.

4.1 Dados

Neste trabalho, usamos três fontes de dados principais: um modelo de representação vetorial de palavras para o português, um corpus não anotado de textos em português para treinar o modelo Skip-Thought e o corpus ASSIN (Fonseca et al., 2016) para treinar e avaliar nossa classificadores.

Para o modelo de representação de palavras usado em nossos experimentos, optamos por usar o modelo FastText (Bojanowski et al., 2016) pré-treinado para a língua portuguesa da Facebook Research[3], que foi treinado no corpus de artigos da Wikipédia escritos em português. Estamos cientes da existência de outros modelos de *word embeddings* para a língua portuguesa que estão disponíveis para uso, particularmente aqueles no Repositório de *Word Embeddings* do NILC[4] analisados no trabalho de Hartmann et al. (2017). Escolhemos o modelo FastText da Facebook Research, entretando, por dois motivos simples: primeiramente, o FastText se tornou um dos modelos de melhor desempenho de representação de palavras na literatura, veja por exemplo os experimentos de Hartmann et al. (2017); segundo, os tamanhos dos modelos NILC de maior dimensionalidade são simplesmente muito grandes para os recursos computacionais disponíveis para nós, enquanto o modelo da Facebook tem uma dimensionalidade competitiva, embora ainda tenha um tamanho gerenciável que nos permite realizar nossos experimentos. De qualquer forma, na Subsecção 4.4, nós testamos o impacto do modelo utilizado nesses experimentos.

O corpus utilizado para treinar o método Skip-Thought é composto por 10.354.228 sentenças e 308.261.905 *tokens*. O corpus foi compilado tomando os artigos escritos em português da Wikipédia, um extrato de cerca de 1000 documentos do corpus de textos jornalísticos PLN-BR Full (Bruckschen et al., 2008) e cerca de 700 resenhas de filmes dos sites *CinePlayers* [5] e *Cinema com Rapadura*[6]. Escolhemos complementar o corpus da Wikipédia com novos documentos com a principal finalidade de aumentar a robustez do modelo treinado, dado o treinamento de um modelo neural como o Skip-Thought requer um grande quantidade de dados de treino. A escolha dos textos utilizados para compor esse corpus se deu pela imediata disponibilidade dos mesmos para que pudéssemos utilizá-los, assim como para garantir uma diversidade de estilos representados no corpus – textos enciclopédicos, jornalísticos e opinativos.

Para computar as representações por agregação ponderada, bem como a representação SIF, também foi utilizado um dicionário de valores IDF para palavras na língua portuguesa – tanto as variantes do português brasileiro quanto do português europeu – composto por 873.329 unidades lexicais. Este dicionário foi obtido processando uma fração do corpus usado para treinar o modelo Skip-Thought aleatoriamente selecionada. Por limitação de tempo e recursos computacionais, não pudemos realizar o cálculo de IDF para todas as palavras do corpus. Assim, preferimos selecionar um extrato do corpus e calcular esses valores.

Para treinar os classificadores, utilizou-se o fragmento de treino do corpus ASSIN (Fonseca et al., 2016) de semelhança textual e paráfrases. Tal corpus é composto por 5000 pares de frases anotadas com similaridade entre sentenças e relações de inferência textual, dentre os quais 295 são exemplos anotados de pares de sentenças parafrásticas. Os classificadores foram avaliados no fragmento de teste do mesmo corpus, contendo 4000 pares de sentenças, dos quais 239 são exemplos positivos de paráfrases. Ficou reservado o fragmento de desenvolvimento, composto por 1000 pares de sentenças, das quais 70 são exemplos positivos de paráfrase, para avaliação de parâmetros experimentais, que são apresentados na Subsecção 4.4.

4.2 Projeto Experimental

Para avaliar a aplicação de modelos de representação vetorial de sentenças ao problema da detecção de paráfrase em português, treinamos um modelo Skip-Thoughts para a língua portuguesa e aplicamos esse modelo, juntamente com o modelo FastText do Facebook. Neste experimento, utilizamos a implementação do método FastText da biblioteca Gensim 3.5[7] e empregamos a representação centroide (média dos vetores de palavras), a agregação ponderada baseada na medida IDF, a representação SIF e a representação Skip-Thought de sentenças.

Note que, na literatura relacionada, existem duas formas principais de representar vetores de pares de sentenças para determinar se são pa-

[3] https://research.fb.com/fasttext/
[4] http://nilc.icmc.usp.br/embeddings
[5] http://www.cineplayers.com
[6] http://cinemacomrapadura.com.br
[7] https://radimrehurek.com/gensim/

rafrásticas ou não. Trabalhos como o de Socher et al. (2011) e Yin & Schütze (2015) utilizam diretamente as representações vetoriais $\overrightarrow{u}$ e $\overrightarrow{v}$ para as sentenças s_1 e s_2 como entrada para os classificadores, enquanto trabalhos como o de Kiros et al. (2015) usam combinações desses vetores pelo produto componente-a-componente $\overrightarrow{u} \cdot \overrightarrow{v}$ e a diferença entre os vetores $\overrightarrow{u} - \overrightarrow{v}$ indicando a semelhança e diferença semântica entre ambos, respectivamente. Nesse trabalho exploraremos ambas as formas de representar o conteúdo da sentença.

Dessa forma, processamos os dados e obtivemos um conjunto de dados diferente para cada método de representação de sentença contendo as seguintes características (*features*):

1. a representação vetorial $\overrightarrow{u}$ da primeira sentença do par;
2. a representação vetorial $\overrightarrow{v}$ da segunda sentença do par;
3. o produto componente a componente entre os vetores $\overrightarrow{u}$ and $\overrightarrow{v}$, i.e. o vetor $\overrightarrow{u} \cdot \overrightarrow{v}$;
4. a norma do vetor $\overrightarrow{u} \cdot \overrightarrow{v}$;
5. a diferença vetorial entre os vetores $\overrightarrow{u}$ e $\overrightarrow{v}$, i.e. $\overrightarrow{u} - \overrightarrow{v}$;
6. a norma do vetor $\overrightarrow{u} - \overrightarrow{v}$;
7. o cosseno entre as representações vetoriais das duas sentenças;

Note que, normalmente, considera-se que o cosseno entre dois vetores que representam sentenças codifica alguma forma de similaridade semântica entre elas. Assim, criamos também um conjunto de dados diferente que contendo somente os valores de similaridade para cada par de sentenças no corpus usando todos os diferentes métodos de representação de sentença investigados neste trabalho. Queremos com tal conjunto de dados avaliar se a similaridade entre as sentenças pode ser usada como um indicador de paráfrase. Também agregamos todas as informações em um único conjunto de dados, no qual cada ponto é composto de todas as informações obtidas para cada método de representação. Queremos avaliar com este conjunto de dados se diferentes representações podem codificar diferentes aspectos do significado das sentenças e se esses diferentes aspectos podem ser compostos para identificar paráfrases.

Avaliamos os classificadores obtidos usando as métricas bem estabelecidas de: Precisão (Prec), computada como a porcentagem de exemplos corretos de paráfrases dentro daqueles que foram identificados pelo sistema como exemplos parafrásticos; Cobertura (Rec, do inglês *Recall*), computada como a porcentagem dos exemplos corretamente identificados como parafráses pelo sistema dentro de todos os exemplos parafrásticos no corpus de treino; e F1, média harmônica entre a Precisão e a Cobertura (Alpaydin, 2009).

4.3 Resultados

Treinamos diferentes classificadores usando dados obtidos com cada representação de sentença. Na Tabela 1, apresentamos os resultados obtidos para cada classificador explorado neste trabalho, ou seja, Máquinas de Vetor de Suporte (SVM, do inglês *Support Vector Machines*), Naïve Bayes (NB), e Árvores de Decisão usando o algoritmo J48 (AD). Tais classificadores foram treinados em dados obtidos por cada método de representação de sentenças, ou seja, o centroide dos vetores das palavras, i.e. sua média (Avg, do inglês *Average*), a agregação ponderada de vetores de palavras (Agg), a representação SIF (SIF) e a representação Skip-Thought (ST). Treinamos ainda os classificadores em um conjunto de dados contendo apenas os valores de similaridade obtidos (Sim) e um outro contendo todas as informações combinadas (Total).

Nestes experimentos, utilizamos a biblioteca SciKit-Learn [8] para linguagem Python para a implementação dos classificadores utilizados nesse trabalho e das técnicas de balanceamento de dados discutidas na Subsecção 4.4, assim para o cálculo das métricas de Precisão, Cobertura e F1.

Como os dados são severamente desbalanceados entre as classes, nós também avaliamos o impacto do balanceamento dos dados no desempenho dos classificadores. Para balanceamento dos dados, nós utilizamos a técnica de *oversampling* por amostragem aleatória nos dados. Os resultados dos classificadores treinados sobre esse conjunto balanceado de dados é exibido na Tabela 2.

Nos dados não balanceados, o classificador Naïve Bayes parece ter um comportamento marginalmente melhor (e mais estável) que os outros, entretanto não é possível afirmar que existem diferenças significativas na performance. Os métodos de representação com melhor desempenho foram os método de centroide (Avg), de similaridades (Sim) e o método com informações combinadas (Total). Para os dados balanceados, o classificador baseado em Máquinas de Vetor de Suporte (SVM) tem uma performance ligeiramente superior, porém ainda similar aos outros. Sobre esses dados, novamente as representações por centroide (Avg), similaridade (Sim) e in-

[8] https://scikit-learn.org/

formações combinadas (Total) alcançaram os melhores resultados.

Método	Classificador	Métricas		
		Prec	Rec	F1
Avg	SVM	0.38	0.19	0.25
	NB	0.20	0.72	**0.31**
	AD	0.21	0.24	0.22
Agg	SVM	0.13	0.04	0.05
	NB	0.10	0.63	0.17
	AD	0.11	0.14	0.12
SIF	SVM	0	0	0
	NB	0.09	0.71	0.15
	AD	0.10	0.10	0.10
Skip	SVM	0.20	0.06	0.09
	NB	0.06	0.94	0.12
	AD	0.11	0.12	0.11
Sim	SVM	0.50	0.08	0.13
	NB	0.13	0.90	0.22
	AD	0.25	0.26	0.25
Total	SVM	0.29	0.31	0.30
	NB	0.09	0.81	0.15
	AD	0.29	0.32	0.30

Tabela 1: Resultados da avaliação dos classificadores treinados para identificação de paráfrases

Método	Classificador	Métricas		
		Prec	Rec	F1
Avg	SVM	0.21	0.44	0.29
	NB	0.19	0.72	0.30
	AD	0.24	0.23	0.23
Agg	SVM	0.10	0.31	0.15
	NB	0.09	0.66	0.17
	AD	0.09	0.10	0.09
SIF	SVM	0.09	0.43	0.16
	NB	0.08	0.71	0.15
	AD	0.10	0.13	0.11
Skip	SVM	0.09	0.30	0.14
	NB	0.06	0.94	0.12
	AD	0.13	0.14	0.13
Sim	SVM	0.21	0.77	**0.33**
	NB	0.09	0.94	0.17
	AD	0.28	0.28	0.28
Total	SVM	0.27	0.33	0.30
	NB	0.09	0.81	0.16
	AD	0.30	0.31	0.31

Tabela 2: Resultados da avaliação de classificadores treinados sobre o conjunto de dados balanceados

Note que, em comparação com os resultados originais (Souza & Sanches, 2018), percebemos claramente uma melhora na performance dos classificadores SVM e AD, assim como da representação por informações globais. Atribuímos esse fato ao aumento na quantidade de dados de treino ao utilizar o corpus ASSIN completo, não somente a variante do Português Brasileiro como naquele trabalho[9]. Isso indica que, pelo fato da representação com informações combinadas gerar um espaço de representação com grande dimensionalidade, os resultados originais para tal representação podem ter sofrido por esparsidade de dados.

É importante salientar que a melhoria dos resultados ao utilizar ambas as variantes é relativamente surpreendente pois, enquanto ambas as variantes se comportam de forma similar para a tarefa de similaridade semântica, os participantes da ASSIN verificaram sistematicamente diferenças entre o comportamento dos sistemas nas duas variantes para a tarefa de RTE. Particularmente, Rocha & Lopes Cardoso (2018) argumenta que as características de implicação e paráfrase parecem ser diferentes em ambos conjuntos de dados.

4.4 Avaliação de Parâmetros Experimentais

É importante notar que o desempenho das técnicas investigadas neste trabalho está claramente abaixo do desempenho relatado para o idioma inglês (c.f. (Kiros et al., 2015), por exemplo) ou para a inferência textual relatada pelos concorrentes no desafio ASSIN (c.f. (Barbosa et al., 2016) ou (Fialho et al., 2016)). As razões para esse baixo desempenho podem surgir de inúmeros parâmetros experimentais utilizados, como o modelo de *word embeddings* adotado ou a técnica de balanceamento de dados utilizada nos experimentos. Para avaliar o efeito desses parâmetros experimentais, realizamos novos experimentos variando-os e comparando a performance do modelo. É importante salientar que nos experimentos discutidos nessa subsessão, usamos como corpus de treino o fragmento de treino do corpus ASSIN (Fonseca et al., 2016), como nos experimentos anteriores, e para teste, utilizamos o fragmento de desenvolvimento (*dev*) do mesmo corpus.

[9]Nesse ponto, agradecemos ao revisor por sua contribuição ao pontuar que a utilização dos dados do ASSIN para a variante europeia da língua poderia melhorar os resultados, como de fato foi observado.

Note que nos experimentos discutidos anteriormente foi utilizada uma técnica ingênua de balanceamento de dados por amostragem aleatória. Apesar do balanceamento dos dados ter apresentado um efeito positivo na performance de alguns classificadores (compare as Tabelas 1 e 2), o uso de tal técnica pode ter resultado num sobreajuste (*overfitting*) dos classificadores nos dados de treino, o que explicaria os baixos valores de Precisão obtidos. Existem na literatura, entretanto, técnicas mais avançadas de balanceamento de dados por sintetização de exemplos, como o SMOTE (Chawla et al., 2002) e o ADASYN (He et al., 2008). Nós decidimos, então, avaliar se o uso de tais técnicas pode melhorar a performance dos classificadores. Para avaliar tal o impacto, usamos o classificado Naïve Bayes e a representação de centroide (Avg), que obtiveram os melhores resultados nos experimentos apresentados anteriormente. Os resultados são apresentados na Tabela 3.

Técnica	Métricas		
	Prec	Rec	F1
Amostragem	0.23	0.79	0.36
SMOTE	0.37	0.54	**0.44**
ADASYN	0.34	0.54	0.42

Tabela 3: Resultados da avaliação do impacto de uso de técnicas de balanceamento de dados

Enquanto em valores absolutos, os resultados apresentados na Tabela 3 indicam que os métodos mais sofisticados de *oversampling* ocasionaram em classificadores com maior Precisão que a simples amostragem aleatória, a diferença entre a performance dos mesmos não parece ser estatisticamente significante. O que é possível observar, entretanto, é que os métodos SMOTE e ADASYM obtém maior precisão, provavelmente devido a uma melhor generalização sobre os dados de treino.

Um importante ponto a se considerar nesse resultado, entretanto, é o fato que a diferença de performance entre os métodos é mais pronunciada se considerarmos somente os dados da variante do Protuguês Brasileiro do corpus ASSIN (F1 de 0.24 para Amostragem Aletória, contra F1 de 0.37 para SMOTE), provavelmente pela menor quantidade e variedade de dados. Isso indica que o uso da técnica de *oversampling* por amostragem aleatória nos experimentos originais, publicados em (Souza & Sanches, 2018), pode ter um importante impacto nos resultados obtidos - devido a potencial sobreajuste dos classificadores.

Outra possível razão para o baixo desempenho dos classificadores testados pode ser a falta de robustez do modelo de *word embeddings* adotado. Tal modelo foi treinado no corpus de artigos da Wikipedia – um corpus pequeno para aprendizado não supervisionado deste tipo de modelos. Para avaliar o impacto do modelo de *word embeddings* usado, realizamos novos experimentos com os modelos treinados por Hartmann et al. (2017) usando os métodos FastText (Bojanowski et al., 2016), Word2Vec (Mikolov et al., 2013) e GloVe (Pennington et al., 2014) com 300 dimensões. Apesar da dimensionalidade dos modelos testados ser igual a do modelo usado originalmente, os modelos de Hartmann et al. (2017) foram treinados sobre um conjunto de dados muito maior que aquele da Facebook Research (FB Fasttext). Os resultados são apresentados na Tabela 4, que descreve as métricas obtivas pelo classificador Naïve Bayes utilizando o método de representação pelo centroide (Avg) com balanceamento usando a técnica SMOTE. Assim como nos experimentos anteriores, utilizamos as implementações da biblioteca Gensim 3.5[10] para os métodos Word2Vec, FastText e GloVe.

Modelo	Métricas		
	Prec	Rec	F1
FB FastText	0.37	0.54	0.44
FastText	0.37	0.57	**0.45**
Word2Vec	0.29	0.60	0.39
GloVe	0.30	0.41	0.35

Tabela 4: Resultados da avaliação do uso de diferentes modelos de *word embeddings*

Podemos perceber que, apesar dos modelos de Hartmann et al. (2017) serem treinados sobre um conjunto de dados maior que o do modelo da Facebook Research usados em nosso experimentos, os resultados apresentados na Tabela 4 não fornecem evidências de que o uso de modelos diferentes impactem significativamente a performance do classificador.

Por fim, perceba na Tabela 2 os valores de similaridade entre as sentenças demonstraram-se como um poderoso indicador de paráfrase.

É importante perceber, entretanto, que enquanto os classificadores treinados sobre os dados de similaridade apresentaram um valor de Cobertura (*Recall*) bastante elevado, os valores de Precisão obtidos foram bastante baixos. Uma possível explicação para tal fenômeno é que no corpus ASSIN, algumas sentenças possuem alto valor de similaridade entre si, porém não constituem exemplo de paráfrase. Tais sentenças podem se constituir em ruído para os classificado-

[10] https://radimrehurek.com/gensim/

res e impactar a performance. Para avaliar a influência de tais sentenças ruidosas, nós avaliamos o impacto de descartar do conjunto de treino todos os exemplos de sentenças que possuem um valor de similaridade acima de um determinado limiar e que não sejam para parafrásticas. Para identificar o impacto desse valor de limiar nos resultados, testamos os limiares no intervalo entre 1.5 e 5.0 em intervalos de 0.5. Os resultados podem ser observados na Tabela 5. Nesses experimentos, utilizamos o modelo FastText da Facebook Research, assim como o método SMOTE para balanceamento dos dados.

Limiar	Métricas		
	Prec	Rec	F1
1.5	0.09	0.95	0.17
2.0	0.17	0.89	0.29
2.5	0.25	0.63	0.35
3.0	0.31	0.57	0.40
3.5	0.31	0.54	0.40
4.0	0.35	0.54	0.42
4.5	0.36	0.54	0.43
5.0	0.37	0.54	**0.44**

Tabela 5: Resultados da avaliação da remoção de exemplos ruidosos

Dos resultados apresentados na Tabela 5, concluímos que a presença de pares de sentenças ruidosas parece possuir um efeito positivo no treinamento dos classificadores. Um explicação para esse fenômeno pode ser o fato desses exemplos servirem para informar os classificadores que enquanto a similaridade semântica textual parece ser uma importante evidência de paráfrase, o fenômeno de paráfrase não se limita o fenômeno de similaridade. Assim, tais sentenças informam ao classificador a existência de pares de sentença com alto grau de similaridade, mas que não são sentenças parafrásticas.

Por fim, calibrados os parâmetros experimentais, reavaliamos o nosso método sobre o corpus de teste, dessa vez usando o modelo FastText da Facebook Research, a representação pelo centroide, o classificador Naïve Bayes com balanceamento de dados usando a técnica SMOTE e sem exclusão de dados possivelmente ruidosos do conjunto de treino. Obtivemos então os seguintes resultados apresentados na Tabela 6.

Métricas		
Prec	Rec	F1
0.25	0.38	**0.30**

Tabela 6: Resultados da avaliação do método após calibração dos parâmetros experimentais

5 Considerações Finais

Este trabalho investigou a aplicação de diferentes métodos de representação de sentenças em um modelo de espaço vetorial da linguagem para o problema de identificação de paráfrase na língua portuguesa. Embora os resultados obtidos para a classificação de paráfrase tenham sido insatisfatórios, em comparação com os resultados relatados na literatura, acreditamos que nossos resultados indicam interessantes caminhos de investigação para detecção de paráfrases para a língua portuguesa. Particularmente, métodos simples de representação de sentenças e classificação, nomeadamente, a representação pelo centroide ou por semelhanças semânticas e um classificador Naïve Bayes, obtiveram os melhores resultados, indicando que uma grande quantidade de informações semânticas das sentenças são codificadas na geometria dos modelos de representação de palavras.

É importante notar que o desempenho das técnicas investigadas neste trabalho está claramente abaixo do desempenho relatado para o idioma inglês (c.f. (Kiros et al., 2015), por exemplo) ou para a inferência textual relatada pelos concorrentes no desafio ASSIN (c.f. (Barbosa et al., 2016) ou (Fialho et al., 2016)). Enquanto diversos parâmetros experimentais foram testados por nós, inúmeros aspectos do nosso projeto experimental devem ser considerados.

Primeiramente, por limitação dos recursos disponíveis para realização de experimentos, nós não pudemos avaliar a performance dos nossos prótipos quando treinados com os modelos com maior dimensionalidade treinados por Hartmann et al. (2017) – que obtiveram melhores resultados na avaliação por analogias. Enquanto nossa avaliação de parâmetros experimentais concluiu que o modelo de *word embedding* utilizado não foi fator determinante para os resultados, não pudemos avaliar o impacto da dimensionalidade de tais modelos – e portanto o poder de representação dos mesmos – nos resultados.

Note que o uso de técnicas de balanceamento de dados mais avançadas que a amostragem aleatória, utilizada nos experimentos originais, parecem ter efeito positivo na performance dos classificadores. Nesse sentido, resta realizar uma avaliação mais sistemática do efeito de técnicas de pré-processamento do conjunto de dados sobre a performance dos classificadores. Um importante ponto nesse sentido é a escolha de carac-

terísticas (*features*) para a descrição dos exemplos. A escolha feita por nós de utilizar duas formas de representação da relação semântica entre as sentenças presentes na literatura, tanto usando os vetores $\overrightarrow{u}$ e $\overrightarrow{v}$, quanto os vetores $\overrightarrow{u} \cdot \overrightarrow{v}$ e $\overrightarrow{u} - \overrightarrow{v}$, pode ter ocasionado um crescimento do espaço de representação, o que pode ter prejudicado o aprendizado dos classificadores.

Com relação ao desempenho dos métodos de representação pela agregação ponderada e SIF, notamos que apenas cerca de 393046 *tokens* no vocabulário do modelo FastText (composto por 592108 *tokens*) estão no dicionário IDF. Isso significa que cerca de 199062 *tokens* no modelo têm valor IDF de 0 e, portanto, não têm efeito na representação da sentença. Isso destaca que as diferentes estratégias de tokenização adotadas em nosso trabalho e a criação do modelo de *word embeddings* podem ter impactado nas representações que alcançamos e, portanto, nos resultados obtidos. É também digno de nota que o desempenho do método Skip-Thought pode ter sofrido com o fato de o corpus de treinamento ser relativamente pequeno em comparação com o utilizado para o idioma inglês (composto de 74.004.228 sentenças e 984.846.357 *tokens*).

É interessante observar que os métodos de melhor desempenho em nossos experimentos foram baseados na representação pelo centroide e pelas medidas de semelhança semântica entre as sentenças codificadas. Isso significa que a estrutura algébrica do espaço vetorial pode, na verdade, codificar uma grande quantidade de informações sobre a semântica composicional de sentenças e que um modelo simples de representação de sentenças pode ser adequado para muitas aplicações posteriores. Essas conexões teóricas e empíricas de *word embeddings* e semântica composicional, bem como as limitações do modelo codificador-decodificador, foram discutidas anteriormente na literatura, notadamente por Arora et al. (2018a,b); Dasgupta et al. (2018).

Note também que nossos experimentos utilizaram unicamente dados das representações vetoriais das sentenças e das similaridades obtidas através delas para treinar os classificadores. Outras características importantes e comumente utilizadas em sistemas de detecção de paráfrase e implicação textual, como medidas de sobreposição lexical, similaridades sintáticas, etc., poderiam ser trivialmente incorporados nos nossos modelos para melhorar a performance dos classificadores. Escolhemos, entretanto, não incorporar tais características em nosso modelo para avaliar quanta informação sobre o conteúdo semântico da sentença pode ser codificada na representação vetorial dessas sentenças.

Por fim, é importante salientar que o corpus ASSIN é constituído de exemplos difíceis, como evidenciado pelos resultados, assim como aqueles obtidos no trabalho de Gamallo e Pereira-Fariña (nesse volume) que utilizam o mesmo corpus para o problema de identificação de similaridade textual. Para avaliar o impacto da estrutura do ASSIN nos nossos resultados, pretendemos, no futuro, testar nossos métodos sobre o conjunto SICK-BR (Real et al., 2018) de inferência textual.

Referências

Alpaydin, Ethem. 2009. *Introduction to machine learning*. MIT Press.

Alves, Ana, Hugo Gonçalo Oliveira, Ricardo Rodrigues & Rui Encarnação. 2018. ASAPP 2.0: Advancing the state-of-the-art of semantic textual similarity for portuguese. Em *7th Symposium on Languages, Applications and Technologies (SLATE)*, 12:1–12:17.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma & Andrej Risteski. 2018a. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association of Computational Linguistics* 6. 483–495.

Arora, Sanjeev, Yingyu Liang & Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. Em *5th International Conference on Learning Representations*, s.pp.

Arora, Sanjeev, Andrej Risteski & Yi Zhang. 2018b. Do GANs learn the distribution? some theory and empirics. Em *6th International Conference on Learning Representations*, s.pp.

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computing Research Repository* http://arxiv.org/abs/1409.0473.

Baptista, Jorge. 2018. Paraphrasing portuguese adverbs ending in –mente. Apresentado no POP - Por Outras Palavras. 1st Workshop on Linguistic Tools and Resources for Paraphrasing in Portuguese.

Barbosa, Luciano, Paulo Cavalin, Victor Guimaraes & Matthias Kormaksson. 2016. Blue man group no ASSIN: Usando representações distribuídas para similaridade semântica e inferência textual. *Linguamática* 8(2). 15–22.

de Barcelos Silva, Allan & Sandro José Rigo. 2018. Enhancing brazilian portuguese textual entailment recognition with a hybrid approach. *Journal of Computer Science* 14(7). 945–956. doi:10.3844/jcssp.2018.945.956.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Bruckschen, Mírian, Fernando Muniz, José Guilherme C. de Souza, Juliana Thiesen Fuchs, Kleber Infante, Marcelo Muniz, Patrícia Nunes Gonçalves, Renata Vieira & Sandra Aluísio. 2008. Anotação lingüística em XML do corpus PLN-BR. Relatório técnico. Universidade de São Paulo.

Cer, Daniel, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope & Ray Kurzweil. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* .

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall & W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16. 321–357.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault & Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. Em *Empirical Methods in Natural Language Processing*, 670–680.

Conneau, Alexis, Germán Kruszewski, Guillaume Lample, Loïc Barrault & Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *Computing Research Repository* http://arxiv.org/abs/1805.01070.

Cordeiro, João, Gaël Dias & Pavel Brázdil. 2007. A metric for paraphrase detection. Em *International Multi-Conference on Computing in the Global Information Technology (ICCGI)*, 35–40.

Dasgupta, Ishita, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman & Noah D. Goodman. 2018. Evaluating compositionality in sentence embeddings. *Computing Research Repository* http://arxiv.org/abs/1802.04302.

Dolan, Bill, Chris Quirk & Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. Em *5th International Conference on Intelligent Text Processing and Computational Linguistics*, s.pp.

Feitosa, David & Vladia Pinheiro. 2017. Análise de medidas de similaridade semântica na tarefa de reconhecimento de implicação textual. Em *11th Brazilian Symposium in Information and Human Language Technology*, 161–170.

Fernando, Samuel & Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. Em *11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, 45–52.

Fialho, Pedro, Ricardo Marques, Bruno Martins, Luísa Coheur & Paulo Quaresma. 2016. INESC-ID@ ASSIN: Medição de similaridade semântica e reconhecimento de inferência textual. *Linguamática* 8(2). 33–42.

Fonseca, Erick & Sandra M. Aluísio. 2018. Syntactic knowledge for natural language inference in portuguese. Em *International Conference on Computational Processing of the Portuguese Language*, 242–252. Springer.

Fonseca, Erick Rocha, Leandro Borges dos Santos, Marcelo Criscuolo & Sandra Maria Aluísio. 2016. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática* 8(2). 3–13.

Gonçalo Oliveira, Hugo, Ana Oliveira Alves & Ricardo Rodrigues. 2017. Gradually improving the computation of semantic textual similarity in portuguese. Em *Progress in Artificial Intelligence*, 841–854.

Hartmann, Nathan, Erick R. Fonseca, Christopher Shulby, Marcos Vinícius Treviso, Jessica Rodrigues & Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *Computing Research Repository* http://arxiv.org/abs/1708.06025.

Hartmann, Nathan Siegle. 2016. Solo queue at ASSIN: Combinando abordagens tradicionais e emergentes. *Linguamática* 8(2). 59–64.

He, Haibo, Yang Bai, Edwardo A Garcia & Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Em *International Joint Conference on Neural Networks*, 1322–1328.

Howard, Jeremy & Sebastian Ruder. 2018. Fine-tuned language models for text classification. *Computing Research Repository* http://arxiv.org/abs/1801.06146.

Jing, Hongyan & Kathleen R. McKeown. 2000. Cut and paste based text summarization. Em *1st Annual Conference of the North American Chapter of the ACL*, 178–185.

Kenter, Tom & Maarten De Rijke. 2015. Short text similarity with word embeddings. Em *24th ACM International Conference on Information and Knowledge Management*, 1411–1420.

Kiros, Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun & Sanja Fidler. 2015. Skip-thought vectors. *Computing Research Repository* http://arxiv.org/abs/1506.06726.

Le, Quoc & Tomas Mikolov. 2014. Distributed representations of sentences and documents. Em *31st International Conference on Machine Learning*, 1188–1196.

Logeswaran, Lajanugen & Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893* s.pp.

Marelli, Marco, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini & Roberto Zamparelli. 2014. Semeval task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. Em *8th International Workshop on Semantic Evaluation*, 1–8.

Marsi, Erwin & Emiel Krahmer. 2005. Explorations in sentence fusion. Em *Tenth European Workshop on Natural Language Generation (ENLG)*, 109–117.

Mihalcea, Rada, Courtney Corley & Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. Em *21st National Conference on Artificial Intelligence*, 775–780.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Advances in Neural Information Processing Systems*, 3111–3119.

Miller, George A. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11). 39–41.

Patro, Badri N., Vinod K. Kurmi, Sandeep Kumar & Vinay P. Namboodiri. 2018. Learning semantic sentence embeddings using pair-wise discriminator. *arXiv preprint arXiv:1806.00807* s.pp.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: Global vectors for word representation. Em *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pinheiro, Anderson, Rafael Ferreira, Máverick Dionísio, Vitor Rolim & oão Tenório. 2017. Statistical and semantic features to measure sentence similarity in portuguese. Em *Brazilian Conference on Intelligent Systems (BRACIS)*, 342–347. doi:10.1109/BRACIS.2017.40.

Real, Livy, Ana Rodrigues, Andressa Vieira e Silva, Beatriz Albiero, Bruna Thalenberg, Bruno Guide, Cindy Silva, Guilherme de Oliveira Lima, Igor Câmara, Miloš Stanojević et al. 2018. SICK-BR: a Portuguese corpus for inference. Em *13th International Conference on Computational Processing of the Portuguese Language*, 303–312.

Rocha, Gil & Henrique Lopes Cardoso. 2018. Recognizing textual entailment: Challenges in the Portuguese language. *Information* 9(4). 76.

Sahlgren, Magnus. 2008. The distributional hypothesis. *Italian Journal of Disability Studies* 20. 33–53.

Shinyama, Yusuke, Satoshi Sekine & Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles. Em *2nd International Conference on Human Language Technology Research*, 313–318.

Silva, Allan, Sandro Rigo, Isa Mara Alves & Jorge Barbosa. 2017. Avaliando a similaridade semântica entre frases curtas através de uma abordagem híbrida. Em *11th Brazilian Symposium in Information and Human Language Technology*, 93–102.

Socher, Richard, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng & Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. Em *Advances in Neural Information Processing Systems*, 801–809.

Souza, Marlo & Leandro M. P. Sanches. 2018. Detecting paraphrases for Portuguese using word and sentence embeddings. Apresentado no POP - Por Outras Palavras. 1st Workshop on Linguistic Tools and Resources for Paraphrasing in Portuguese.

Suresh, Subhashree & P. Sreenivasa Kumar. 2016. Enriching linked datasets with new object properties. *Computing Research Repository* http://arxiv.org/abs/1606.07572.

Turney, Peter D. & Michael L. Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *Computing Research Repository* http://arxiv.org/abs/cs.LG/0212012.

Yang, Yinfei, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope & Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. *Computing Research Repository* http://arxiv.org/abs/1804.07754.

Yin, Wenpeng & Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. Em *Annual Conference of the North American Chapter of the ACL: Human Language Technologies*, 901–911.

Proposta recebida em Novembro 2018 e aceite para publicação em Dezembro 2018.

Análise da capacidade de identificação de paráfrase em ferramentas de resolução de correferência

Analyzing paraphrase identification capabilities in coreference resolution tools

Bernardo Scapini Consoli
Pontifícia Universidade Católica do Rio Grande do Sul
bernardo.consoli@acad.pucrs.br

Joaquim Francisco dos Santos Neto
Pontifícia Universidade Católica do Rio Grande do Sul
joaquim.santos@acad.pucrs.br

Sandra Collovini de Abreu
Pontifícia Universidade Católica do Rio Grande do Sul
sandra.abreu@acad.pucrs.br

Renata Vieira
Pontifícia Universidade Católica do Rio Grande do Sul
renata.vieira@pucrs.br

Resumo

Os fenômenos linguísticos de correferência e paráfrase compartilham certos aspectos. É comum, por exemplo, referir-se a uma mesma entidade de maneiras diferentes em um mesmo contexto, assim, a resolução de correferências pode auxiliar o processo de identificação de paráfrases. Este artigo apresenta uma análise das capacidades da ferramenta de resolução de correferência CORP, para Português, no contexto de identificação de paráfrases nos níveis de sentença e de sintagma.

Palavras chave

resolução de correferência, identificação de paráfrase

Abstract

The linguistic phenomena known as coreference and paraphrasing share certain aspects among themselves. It is common, for example, to refer to an entity in different ways within the same context, and as such the resolution of such coreferent mentions may be of aid to the process of identifying paraphrases. This paper presents an analysis of the capabilities of the coreference resolution tool CORP, created for use with the Portuguese language, within the context of paraphrase identification in the sentence and noun phrase levels.

Keywords

coreference resolution, paraphrase identification

1 Introdução

Aplicações de PLN que lidam com paráfrase e correferência tem o potencial de melhorar o entendimento e a geração de sistemas. Extração de paráfrases e resolução de correferência podem ser aplicadas nas tarefas de perguntas e respostas, extração de informação, tradução automática, entre outras (Recasens & Vila, 2010).

Paráfrase é definida como a relação entre duas expressões que possuem o mesmo sentido, enquanto correferência é definida como a relação entre duas expressões que possuem o mesmo referente no mesmo contexto (Jurafsky & Martin, 2009). Pares parafrásicos podem ser correferentes e vice-versa (Recasens & Vila, 2010). A relação entre correferência e paráfrase implica uma possibilidade da utilização de resolução de correferência para facilitar a identificação de paráfrases.

Contudo, correferência é um fenômeno linguístico muito mais dependente em contexto do que a paráfrase (Recasens & Vila, 2010). No exemplo: "Ana foi ao cardiologista entregar alguns exames. O médico pediu a ela para retornar na próxima semana". Podemos ver que [cardiologista] e [médico] são sintagmas correferentes e parafrásicos, enquanto o nome próprio [Ana] e o pronome pessoal [ela] são somente correferentes, pois possuem uma referência mas não um significado intrínseco (Recasens & Vila, 2010). Deste modo, é necessário distinguir quais menções de uma cadeia de correferências são parafrásicas.

DOI: 10.21814/lm.10.2.288

Este trabalho visa analisar a capacidade da ferramenta de resolução de correferência para o Português CORP descrita em (Fonseca et al., 2017) no contexto da área de identificação de paráfrases. Duas análises foram realizadas: a primeira estuda os tipos de sintagmas parafrásicos encontrados pelo CORP, comparando-os com resultados obtidos pela função de resolução de correferência para o Inglês da ferramenta Stanford CoreNLP apresentada em (Manning et al., 2014). Essa análise foi feita sobre um subconjunto de textos da revista Pesquisa FAPESP[1]; a segunda é uma análise da performance do CORP sobre um corpus de sentenças parafrásicas descrito em (Fonseca et al., 2016), visando identificar sintagmas nominais correferentes que possam servir como âncoras para um subsequente processo de identificação parafrásica.

O restante deste trabalho está organizado nas seguintes seções: a Seção 2 descreve o referencial teórico juntamente com trabalhos relacionados; a Seção 3 apresenta os recursos utilizados; a Seção 4 descreve a avaliação realizada; e por fim, na Seção 5 as considerações finais são apresentadas.

2 Correferência e Paráfrase

As ferramentas computacionais para resolução de correferência lidam com a tarefa de identificar as expressões textuais associadas a entidades ou eventos do mundo real. Recasens & Vila (2010) destacam que paráfrase e correferência são geralmente definidos como relações de similaridade, ou seja, dadas duas expressões que têm o mesmo significado, estas são parafrásicas; e dadas duas expressões referentes à mesma entidade em um discurso, estas são correferentes. Partindo desse princípio, quando usamos uma ferramenta de resolução de correferência e dela obtemos suas respectivas cadeias, podemos obter paráfrases nas cadeias identificadas.

Para um melhor entendimento das relações entre paráfrase, correferência e sistemas de resolução de correferência, considere o seguinte fragmento de texto[2], processado pelo sistema de resolução de correferência (CORP) (Fonseca et al., 2017):

"[...] Para **[Luiz_Eugênio_Mello [258]]**, vice-presidente de a Associação_Nacional_de_ Pesquisa_e_Desenvolvimento_das_Empresas_Inovadoras (Anpei) [...] Para **[Luiz_Mello [258]]**, há uma baixa intensidade de P&D mesmo entre empresas líderes [...]. Já as 10 mais em a Espanha, que foram HP, Airbus, Ericsson, CSIC, Fractus, Gamesa, Vodafone, Laboratórios_Dr._Esteve, Intel e Telefonica, depositaram 739 patentes em os Estados_Unidos, 88 % a mais, diz **[Mello [258]]**. [...] **[Luiz_Eugênio_Mello [258]]** também critica a dificuldade de trabalhar com prioridades. [...]"

Os termos destacados em negrito foram identificados como correferentes e categorizados com a categoria Pessoa, como ilustra a Tabela 1.

Cadeia Pessoa
[Luiz Eugênio Mello](2)
[Luiz Mello](1)
[Mello](1)

Tabela 1: Cadeia Pessoa do fragmento textual.

2.1 Trabalhos Relacionados

O estudo de identificação de paráfrase de Shinyama & Sekine (2003) utiliza Reconhecimento de Entidades Nomeadas para encontrar o que chamam de "âncoras", expressões que provavelmente não mudariam em frases parafrásicas (nomes de pessoas, datas, entre outros). De acordo com a definição de paráfrase dada previamente, sentenças parafrásicas devem possuir sintagmas nominais que se referem às mesmas entidades. Se identificadas as relações de correferência entre os sintagmas das sentenças, é possível removê-los do processo de identificação de paráfrase, substituindo-os por uma âncora. Um exemplo é a entidade nomeada [Presidente Temer] sendo mencionado em outras frases como [Michel Temer] ou até mesmo através de um pronome, no caso [ele].

O estudo de Regneri & Wang (2012) utilizou informações de estrutura e de contexto dos documentos analisados para auxiliar na coleta de sentenças parafrásicas. A resolução de correferência foi utilizada para adicionar mais informações de contexto à análise. Especificamente, a correferência incluiu a informação de quais partes das sentenças possuem o mesmo sentido no contexto analisado.

Nos estudos apresentados, por mais que os sintagmas em si não sejam parafrásicos, como no caso do pronome que por natureza não pode ser paráfrase de um nome, a relação de correferência entre um nome e um pronome ajuda a identificar sentenças parafrásicas.

Quanto a avaliação realizada no presente trabalho, a primeira análise é uma busca por sintagmas nominais parafrásicos encontrados pelo sistema CORP, com o objetivo de analisar os pa-

[1] http://revistapesquisa.fapesp.br/

[2] http://revistapesquisa.fapesp.br/2017/06/19/financiamento-em-crise/

drões em que estes sintagmas se encaixam, enquanto a segunda é um estudo de como o CORP poderia ser utilizado como um auxiliador para um sistema de identificação de paráfrases.

3 Recursos

Nesta Seção, nós descrevemos as ferramentas de correferência CORP e Stanford CoreNLP, que tratam as línguas Portuguesa e Inglesa, respectivamente, bem como o corpus ASSIN 2016 de sentenças parafrásicas.

CORP.
CORP é um recurso para a resolução de correferência em Português descrito por Fonseca et al. (2017). O CORP utiliza um conjunto de regras sintáticas e semânticas, propostas por Fonseca (2018), para decidir se dois sintagmas nominais (nomes próprios ou comuns) são correferentes, bem como as informações de *Part-of-Speech* (PoS) e sintáticas provindas da ferramenta Cogroo (Silva, 2013). Para um melhor entendimento, na Tabela 2 temos exemplos de cadeias de correferência de um texto da revista Pesquisa FAPESP[3]. A primeira coluna indica a categoria da cadeia (Sarmento et al., 2006) e na coluna seguinte as cadeias de correferência com a respectiva frequência de cada menção. Podemos notar que o primeiro exemplo indica a categoria Pessoa em que temos as diferentes menções que designam [Carlos Américo Pacheco]. Já para Organização/Local temos a cadeia referindo-se ao [Brasil], e por fim temos a cadeia com diferentes menções da [Universidade Estadual de Campinas] por meio do acrônimo [Unicamp].

Stanford CoreNLP.
Stanford CoreNLP (Manning et al., 2014) provê um conjunto de ferramentas de tecnologia de linguagem humana, incluindo Reconhecimento de Entidades Nomeadas, identificação de dependências sintáticas e resolução de correferências. Para este trabalho, utilizamos suas capacidades de resolução de correferência na versão determinística, que é baseada em regras sintáticas. O sistema determinístico foi utilizado para melhor comparar com o CORP, que também é um sistema determinístico. Para exemplificar, a Tabela 3 ilustra exemplos de cadeias de correferência de um texto da FAPESP[4]. A primeira coluna ilustra as categorias tratadas pelo Stanford CoreNLP, nas colunas seguintes temos as cadeias de correferência com a respectiva frequência de cada menção. Destaca-se as diferentes menções na cadeia de [Carlos Américo Pacheco] (Pessoa), a qual inclui o pronome pessoal [he].

Corpus ASSIN 2016.
O corpus ASSIN 2016[5] descrito por Fonseca et al. (2016) foi construído para as tarefas ASSIN da conferência PROPOR 2016. Este corpus possui 10.000 pares de sentenças parafrásicas anotados para grau de similaridade semântica e inferência textual. A Figura 1 demonstra um par do corpus ASSIN 2016, na linguagem XML. *Pair similarity* é a média da medida de similaridade semântica textual selecionada por 4 anotadores; o *ID* é o número identificador do par; *entailment* é a classe de inferência textual dada por anotadores; t e h são as sentenças e indicam qual delas é o texto e a hipótese para objetivos de inferência textual.

4 Avaliação

Duas análises foram realizadas sobre a capacidade do CORP de auxiliar na tarefa de identificação de paráfrase: a primeira trata de estudar a relação entre o fenômeno da correferência e o fenômeno da paráfrase no contexto de sintagmas nominais, subsequentemente realizando uma análise para discernir os padrões em que se encaixam os sintagmas nominais parafrásicos encontrados pelo CORP; a segunda trata de sua capacidade de encontrar relações de correferência que auxiliem em uma potencial tarefa de identificação de sentenças parafrásicas.

Análise 1.
Para a primeira análise foram utilizados 10 textos paralelos em Inglês e Português retirados da revista Pesquisa FAPESP. O método de identificação de padrões parafrásicos foi dividido em 2 passos:

1. O CORP e o Stanford CoreNLP são utilizados para extrair automaticamente as cadeias de correferência nos 10 textos paralelos em Inglês e Português da revista Pesquisa FAPESP.
2. Os sintagmas extraídos são manualmente analisados e classificados em padrões de acordo com características identificáveis.

Considerando que o Stanford CoreNLP é uma das melhores ferramentas disponível para a resolução de correferências, decidimos utilizar os

[3] http://revistapesquisa.fapesp.br/2017/06/19/financiamento-em-crise/
[4] http://revistapesquisa.fapesp.br/en/2017/12/10/funding-in-crisis/
[5] http://nilc.icmc.usp.br/assin/

Categorias	Menções (Frequência Individual)
Pessoa	[Carlos Américo Pacheco, professor de o Instituto de Economia de a Universidade Estadual de Campinas] (1) [Carlos Américo Pacheco] (2)
Organização/Local	[o Brasil] (11) [O Brasil] (3) [a Universidade Estadual de Campinas] (1) Unicamp (1)

Tabela 2: Cadeias de Correferência - CORP

Categorias	Menções (Frequência Individual)
Pessoa	[Carlos Américo Pacheco, a professor at the Institute of Economics of the University of Campinas (Unicamp)] (1) [Carlos Américo Pacheco] (4) [a professor at the Institute of Economics of the University of Campinas (Unicamp)] (1) [Pacheco, Chief Executive of the FAPESP Executive Board] (1) [Pacheco] (1) [he] (2)
País	[Brazil] (17) [Brazil,that filed the most patent applications in the U.S.] (1) [Brazil, which has become increasingly more complex in recent decades] (1)

Tabela 3: Cadeias de Correferência - Stanford

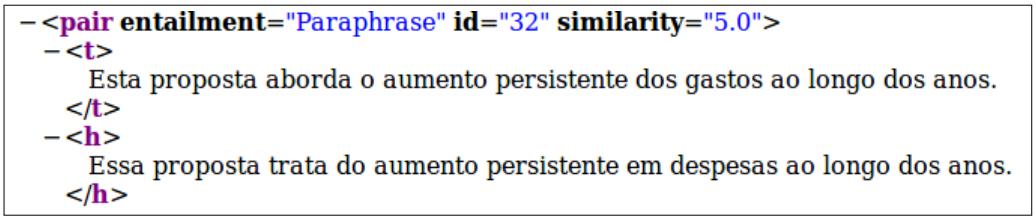


Figura 1: Amostra do corpus ASSIN 2016

resultados do toolkit como uma referência à qual podemos comparar os resultados encontrados pelo CORP. Cabe salientar que as ferramentas utilizadas possuem diferenças na identificação dos sintagmas nominais extraídos (menções) e no conjunto de categorias tratadas.

Como resultado da avaliação temos 3 padrões: Acrônimos, Nomes Próprios e Nomes, os quais são ilustrados nas Tabelas 4 e 5.

A análise dos sintagmas nominais parafrásicos identificados pelo Stanford CoreNLP mostrou que alguns padrões são melhor identificados do que outros. Exemplos de cadeias de correferência contendo menções parafrásicas de cada padrão são apresentados na Tabela 4. Podemos observar que a ferramenta conseguiu tratar poucos casos de sintagmas nominais parafrásicos envolvendo acrônimos, como por exemplo, [United States] - [US]. Em geral, o padrão de Nomes Próprios (ou Entidades Nomeadas) conseguiu identificar eficientemente as paráfrases a partir de cadeias bem completas, como por exemplo, na cadeia de Pessoa temos: [biochemist María Elena López of the Institute of Biological Sciences] - [López] - [María Elena López]. Podemos notar que na cadeia [O Rio de Janeiro] - [O Rio] a ferramenta não conseguiu identificar a categoria e classificou como Outro sendo a correta Local. O padrão de Nomes destaca-se por cadeias extensas contendo vários sintagmas parafrásicos, como no exemplo referindo-se a [gravitational wave] (Outro).

A avaliação do CORP com base nos 3 padrões identificados mostra que, em geral, o CORP consegue identificar eficientemente as paráfrases envolvendo acrônimos por meio das cadeias de correferência com diferentes menções para as categorias tratadas.

Padrão	Categoria	Menções (Frequência Individual)
Acrônimos	País	[United States] (1) [US] (1)
Nomes Próprios	Pessoa	[biochemist María Elena López of the Institute of Biological Sciences] (1) [López] (5) [María Elena López] (1)
	Outro	[o Rio de Janeiro] (2) [o Rio] (1)
Nomes	Outro	[another gravitational wave] (1) [this gravitational wave an identical instrument in Livingston, Louisiana] (1) [this gravitational wave] (1) [an identical instrument in Livingston , Louisiana] (1)

Tabela 4: Padrões de Paráfrases identificadas pelo Stanford CoreNLP.

Padrão	Categoria	Menções (Frequência Individual)
Acrônimos	Organização/Local	[o Instituto Brasileiro de Geografia e Estatística] (1) [IBGE] (1)
Nomes Próprios	Pessoa	[Gustavo Gomes] (1) [Gomes] (1)
	Organização/Local	[o Rio de Janeiro] (2) [o Rio] (1)
Nomes	Outro	[o debate] (1) [o desafio] (1)
	Comunicação	[um sinal sazonal , cujo pico máximo] (1) [O sinal] (1)

Tabela 5: Padrões de Paráfrases identificadas pelo CORP

Na Tabela 5 são ilustrados exemplos, como na cadeia [Instituto Brasileiro de Geografia e Estatística] - [IBGE] (Organização/Local). Um outro padrão são os nomes próprios (de Pessoas, Organização/Local, entre outros), como por exemplo, a cadeia [Gustavo Gomes] -[Gomes] (Pessoa) e a cadeia [o Rio de Janeiro] - [o Rio] (Organização/Local). Por fim, o padrão referente a Nomes refere-se aos sintagmas nominais que possuem o mesmo significado, como por exemplo, a cadeia de [o debate] - [o desafio] (Outro). O outro exemplo desse padrão tratou a categoria Comunicação em que as menções [um sinal sazonal, cujo pico máximo] e [O sinal] são parafrásicos. Nota-se que o primeiro sintagma nominal teve problemas na sua identificação por parte do Cogroo.

Análise 2.
Para a segunda análise foram utilizados 116 pares de sentenças parafrásicas do corpus ASSIN 2016. O método proposto de identificação dos sintagmas nominais para auxiliar na identificação de paráfrase entre sentenças foi dividido em 3 passos:

1. Pares de sentenças parafrásicas são anotados manualmente para correferência.

2. O CORP é utilizado para anotar automaticamente os pares de sentenças.

3. A anotação automática é comparada à manual.

Dada a falta de um corpus paralelo entre o Português e o Inglês anotado para correferência, o Stanford CoreNLP não foi utilizado para comparação, e julgamos que uma tradução automática geraria ruído demais nos dados. Desta forma, esta análise verifica o desempenho do CORP na identificação dos sintagmas nominais correferentes que possam servir como âncoras para um subsequente processo de identificação de paráfrases. Para isso, os sintagmas nominais parafrásicos contidos nos 116 pares de sentenças parafrásicas do corpus ASSIN 2016 foram anotados manualmente. Um total de 339 sintagmas foram anotados e serviram de referência para a avaliação da performance do CORP na tarefa proposta.

Sentenças	Menções - Referência	Menções - CORP
O tremor também deixou quase 100 mortos em a Índia e China.	[O tremor] - [O terremoto] [quase 100 mortos] - [cerca de 100 vítimas fatais]	[O tremor] - [O terremoto] —
O terremoto fez também cerca de 100 vítimas fatais em a Índia e em a China.	[Índia] - [Índia] [China] - [China]	[Índia] - [China] - [China] —
Esta proposta lida com o persistente aumento em os gastos a o longo de os anos. Esta proposta enfrenta o persistente aumento de despesas por anos.	[Esta proposta] - [Esta proposta] [os gastos] - [despesas] [os anos] - [anos]	— [os gastos] - [despesas] [os anos] - [anos]
Cerca de 5 mil pessoas trabalham usando a plataforma Uber hoje em o Brasil. Atualmente , cerca de 5 mil profissionais atuam usando a plataforma Uber em o país.	[Cerca de 5 mil pessoas] - [Cerca de 5 mil profissionais] [a plataforma Uber] - [a plataforma Uber] [o Brasil] - [o país]	— [a plataforma Uber] - [a plataforma Uber] [o Brasil] - [o país]

Tabela 6: Sintagmas correferentes identificados pelo CORP.

Como resultado temos 152 acertos; uma taxa de Precisão de 67%; Abrangência de 43% e F-measure de 53%. Na Tabela 6 são ilustrados exemplos de sintagmas correferentes identificados pelo CORP.

Na primeira coluna temos pares de sentenças parafrásicas, na segunda temos a anotação manual dos sintagmas nominais, e na última, os sintagmas nominais extraídos pelo CORP. Podemos observar que o CORP conseguiu identificar corretamente sintagmas parafrásicos, como por exemplo, nas cadeias [os gastos] - [as despesas]; [o tremor] - [o terremoto]; [o Brasil] - [o País]. Entretanto, como o CORP desconsidera sintagmas com dados numéricos, não foram identificados sintagmas contendo quantidades, como por exemplo, nas cadeias [quase 100 mortos] - [cerca de 100 vítimas fatais]; [cerca de 5000 pessoas] - [cerca de 5000 profissionais]. Além disso, ocorreram casos em que o CORP agrupou menções de cadeias diferentes em uma mesma cadeia, como no exemplo: [Índia] - [China] - [China].

5 Considerações Finais

Apresentamos neste trabalho as relações entre paráfrase e correferência, bem como uma avaliação da capacidade do CORP na identificação de paráfrases por meio de duas análises. A primeira análise resultou na identificação de três padrões de sintagmas correferentes que podem auxiliar na identificação de sintagmas parafrásicos. O CORP se mostrou capaz de identificar sintagmas correferentes e parafrásicos em textos do Português, destacando-se para os casos com acrônimos, como por exemplo, [o Estatuto de a Criança e de o Adolescente] - [ECA]. A segunda análise mostrou que o CORP auxilia na identificação de sintagmas nominais correferentes e parafrásicos em pares de sentenças parafrásicas. Uma das dificuldades da avaliação manual foi a delimitação dos sintagmas nominais, como por exemplo, no fragmento da sentença parafrásica "presidente Blatter não vai mais responder perguntas" o CORP identificou dois sintagmas: [presidente Blatter] e [Blatter]. Cabe ressaltar que a etapa de identificação dos sintagmas do CORP é provida pela ferramenta Cogroo.

Como trabalhos futuros, pretendemos tipificar/classificar os padrões para sintagmas parafrásicos propostos com auxilio de linguistas, e disponibilizar recursos para o Português anotados para correferência e paráfrases. Além disso, planejamos utilizar as cadeias de correferência extraídas pelo CORP para enriquecer uma análise de similaridade semântica deste mesmo corpus.

Agradecimentos

Agradecemos à PUCRS, CNPQ, CAPES e FAPERGS pelo seu apoio financeiro.

Referências

Fonseca, Erick, Leandro Borges dos Santos, Marcelo Criscuolo & Sandra Aluísio. 2016. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática* 8(2). 3–13.

Fonseca, Evandro. 2018. *Resolução de correferência nominal usando semântica em língua portuguesa*. PUCRS: Programa de Pós-Graduação em Ciência da Computação, PUCRS. Tese de Doutoramento.

Fonseca, Evandro, Vinicius Sesti, André Antonitsch, Aline Vanin & Renata Vieira. 2017.

CORP: uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências. *Linguamática* 9(1). 3–18.

Jurafsky, Daniel & James H. Martin. 2009. *Speech and language processing*. Prentice-Hall.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard & David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. Em *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Recasens, Marta & Marta Vila. 2010. On paraphrase and coreference. *Computational Linguistics* 36(4). 639–647.

Regneri, Michaela & Rui Wang. 2012. Using discourse information for paraphrase extraction. Em *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 916–927.

Sarmento, Luís, Ana Sofia Pinto & Luís Cabral. 2006. REPENTINO - a wide-scope gazetteer for entity recognition in Portuguese. Em *Computational Processing of the Portuguese Language*, 31–40.

Shinyama, Yusuke & Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. Em *Second International Workshop on Paraphrasing*, 65–71.

Silva, William Daniel Colen. 2013. *Aprimorando o corretor gramatical CoGrOO*: Universidade de São Paulo. Tese de Mestrado.

Proposta recebida em Outubro 2018 e aceite para publicação em Dezembro 2018.

Parafraseamento Automático de Registo Informal em Registo Formal na Língua Portuguesa

Automated Paraphrasing of Portuguese Informal into Formal Language

Anabela Barreiro
INESC-ID
anabela.barreiro@inesc-id.pt

Ida Rebelo-Arnold
Universidad de Valladolid
imdamotar@funge.uva.es

Jorge Baptista
Universidade do Algarve
jbaptis@ualg.pt

Cristina Mota
INESC-ID
cmota@islt.utl.pt

Isabel Garcez
Universidade de Lisboa
isabelgarcez@campus.ul.pt

Resumo

Este artigo apresenta o processo de automatização de parafraseamento em português e conversão de construções típicas do registo informal ou da linguagem falada em construções de registo formal usadas na linguagem escrita. Ilustraremos o processo de automatização com exemplos extraídos do corpus e-PACT, que envolvem a colocação normalizada de pronomes clíticos quando co-ocorrem com compostos verbais. A tarefa consiste em parafrasear e normalizar, entre outras, construções como *vou-lhe/posso-lhe fazer uma surpresa* em *vou/posso fazer-lhe uma surpresa*, em que o pronome clítico *lhe* migra de uma posição enclítica imediatamente a seguir ao primeiro verbo do composto verbal para uma posição enclítica a seguir ao verbo principal, que é o verbo responsável pela seleção do argumento pronominal. O primeiro verbo é um verbo auxiliar ou um verbo volitivo, e.g., *querer*. Este é um procedimento padronizado no processo de revisão em português europeu. Casos como este representam fenómenos linguísticos em que os estudantes de língua portuguesa e falantes em geral se confundem ou onde "tropeçam". O artigo enfatiza a língua padrão em que os fenómenos observados ocorrem, descreve exemplos de interesse encontrados no corpus e apresenta uma solução automática, baseada na aplicação de gramáticas transformacionais genéricas, que facilitam a normalização de inadequações ou falhas sintáticas (registos informais) encontradas nas construções pesquisadas em construções padronizadas típicas da escrita formal ou escrita profissional.

Palavras chave

paráfrases, parafraseamento automático, registo formal e informal, compostos verbais, pronomes clíticos, ordem das palavras, português europeu, português do Brasil, aprendizagem da língua, escrita profissional

Abstract

This paper presents the automation process of paraphrasing and converting Portuguese constructions typical of informal or spoken language into a formal written language. We illustrate this automation process with examples extracted from the e-PACT corpus that involve the placement of clitic pronouns in verbal compound contexts. Our task consists in paraphrasing and normalizing, among others, constructions such as *vou-lhe/posso-lhe fazer uma surpresa* into *vou/posso fazer-lhe uma surpresa* "lit: I will/can_to him/her make a surprise / I will/can make_to him/her a surprise; I will/can make him/her a surprise", where the clitic pronoun *lhe* migrates from an enclitic position immediately after the first verb of the verbal compound to an enclitic position after the main verb, which is the verb responsible for the selection of that pronominal argument. The first verb is either an auxiliary verb or a volitive verb, e.g., *querer* "want". This is a standard revision procedure in European Portuguese. Cases like this represent linguistic phenomena where language students and language users in general get confused or "stumble". The paper focuses on general language where the phenomena being observed occur, describes examples of interest found in the corpus, and presents an automatic solution for the normalization of informal syntactic inadequacies found in the researched structures into standard structures typical of formal or professional writing through the application of very generic transformational grammars.

Keywords

paraphrases, automated paraphrasing, formal and informal language, verbal compounds, clitic pronouns, word order, European Portuguese, Brazilian Portuguese, language learning, professional writing

DOI: 10.21814/lm.10.2.282

1 Introdução

A automatização da revisão de conteúdos é uma das funções mais desejadas para um revisor ou editor profissional, especialmente para aquelas tarefas enfadonhas que envolvem "lacunas" no tipo de registo formal, que consomem tempo e representam um entrave a uma revisão eficaz e rápida de textos de autoria. Aqui, o termo "lacuna" não significa necessariamente um erro gramatical, mas o uso de construções informais que são típicas do discurso oral, que são corrigidos pelos revisores na produção escrita de escritores profissionais. Além das vantagens ao nível da produção de escrita, um parafraseador com funções automáticas de normalização e/ou revisão poderá ser usado como uma aplicação de aprendizagem para estudantes, em particular, estudantes de línguas, entre outras aplicações. Neste artigo, apresentamos o processo de conversão de formas de expressão informais ou "menos polidas" em expressões formais utilizadas em textos escritos, dado que desejamos criar uma forma padronizada como as que existem em guias de autoria e estilo, por exemplo, ou em guias técnicos usados para obter uma publicação de qualidade.

Ilustramos este processo automatizado com construções de predicados verbais compostos (doravante, *compostos verbais*) envolvendo sequências de dois (algumas vezes mais) verbos e um pronome clítico, onde o clítico é um argumento do segundo verbo. O clítico pode ser colocado imediatamente a seguir ao verbo de que depende, e.g. *queria ver-te*. Esta é a construção que os livros e as gramáticas de estilo geralmente recomendam como "uso correto" no discurso formal; ou ser movido para junto do primeiro verbo, e.g. *queria-te ver* em português europeu (PE), *te queria ver* em português do Brasil (PB), que é muitas vezes considerado como menos formal ou até mesmo um uso "relaxado". Enquanto o segundo verbo do composto verbal é um verbo pleno, também conhecido como verbo *distribucional* (i.e., um item lexical que seleciona argumentos e com um significado lexical definido intencionalmente), o primeiro verbo pode ser um verbo *auxiliar*, no sentido definido por Cunha & Lindley-Cintra (1986, 393–396), muitas vezes designados como *perífrases verbais* ou *locuções verbais*[1], e.g. *estou a ver-te* versus *estou-te a ver* (PE), *te estou a ver* (PB), ou um verbo completo, incluindo os verbos volitivos, como *querer*, *desejar* e outras construções verbais. Em todos esses casos, a normalização exige que o pronome clítico migre para uma posição enclítica e seja anexado ao segundo verbo do composto verbal, por exemplo, *eu quero-o ver* → *eu quero vê-lo*. No exemplo normalizado, o verbo infinitivo sofre uma mudança de *ver* para *vê-* e o pronome clítico sofre uma mudança de *o* para *lo*, uma regra ortográfica motivada por razões fonéticas.

Em Processamento de Linguagem Natural (PLN), a maioria dos analisadores sintáticos (parsers) processa os verbos auxiliares portugueses da mesma maneira que qualquer outro verbo, isto é, como um verbo pleno e completo; veja-se, por exemplo, as árvores de análise produzidas pelo PALAVRAS (Bick, 2000)[2] e o LxParser (Silva et al., 2010)[3]. Uma proposta diferente é apresentada por Baptista et al. (2010), que processa construções auxiliares verbais de maneira diferente, distinguindo o auxiliar do verbo principal, tomando em conta as diferentes opções de posicionamento/colocação dos pronomes clíticos. De facto, os verbos auxiliares requerem uma proposta adequada de sistematização que considere não apenas as propriedades lexicais, mas também as propriedades semântico-sintáticas desses verbos. A descrição dos verbos em PE realizada no âmbito da Léxico-Gramática (Baptista, 2012, 2013; Baptista & Mamede, 2018) fornecem uma lista de mais de 100 construções verbais auxiliares (entre mais de 330 construções verbais auxiliares). Desta forma, será possível criar listas de ocorrências e construir gramáticas locais que podem ser usadas tanto por utilizadores humanos quanto por máquinas. É importante destacar que todos os verbos ilustrados e analisados neste artigo formam uma locução com outro verbo (o verbo principal). Em muitas co-ocorrências, o significado do verbo principal geralmente recebe um valor aspectual. Há também verbos cujos significados são construídos com a co-ocorrência de uma preposição seguida de outro verbo.

Como o tópico da nossa investigação é tão amplo em escopo e o nosso corpus inclui uma variedade tão vasta de casos de categorização e tratamento computacional difícil, decidimos focar-nos apenas nos casos de compostos verbais que co-ocorrem com clíticos. Os exemplos ilustrados no artigo foram extraídos do corpus e-PACT (Barreiro & Mota, 2017), que é composto por dois romances da autoria de David Lodge. Os alinha-

[1] Uma visão geral mais abrangente sobre o tópico pode encontrar-se em (Pontes, 1973; Gonçalves, 1999; Paiva Raposo, 2013). Também vale a pena mencionar as propostas de (Gross, 1998) para o sistema de verbos auxiliares em francês.

[2] http://www.visl.sdu.dk/visl/pt/parsing/automatic/dependency.php

[3] http://www.lxcenter.di.fc.ul.pt/services/pt/LXParserPT.html

mentos parafrásticos foram realizados por meio do uso da ferramenta de alinhamento CLUE-Aligner (Barreiro et al., 2016), já utilizada em outros trabalhos de investigação sobre alinhamentos de paráfrases.[4] O corpus contém exemplos simples e não padronizados, incluindo frases típicas de diálogos ou trechos de comunicação informal, que caracterizam o tipo de textos literários que constituem o corpus. Analisámos uma pequena quantidade de ocorrências no corpus e criámos uma tipologia de categorias de compostos verbais. Em seguida, usámos essas categorias para criar gramáticas locais genéricas que serviram de base para o processamento automatizado de paráfrases, nomeadamente geração e identificação em texto. Os pares não padronizados/padronizados de contrastes parafrásticos resultantes deste estudo serão validados para a sua integração na ferramenta de parafraseamento eSPERTo, que, entre outras aplicações, visa permitir a adaptação e revisão de textos. Atualmente, o eSPERTo está integrado numa aplicação online que fornece sugestões parafrásticas para ajudar alunos de língua portuguesa. À medida em que esta ferramenta for evoluindo, prevê-se que os seus recursos sejam utilizados na produção e revisão de textos.[5] Outra aplicação experimental envolve a construção de um conjunto de dados de contrastes parafrásticos entre as variedades europeia e brasileira da língua portuguesa, um recurso indispensável para a conversão e adaptação entre todas as variedades do português (Barreiro & Mota, 2018; Rebelo-Arnold et al., 2018). Esses esforços estão alinhados com a proposta de criar um padrão internacional de português (Santos, 2015). Finalmente, como uma abordagem inicial, começamos a explorar o tópico de ensinar aos alunos a distinção entre linguagem formal e informal através do uso de agentes conversacionais representando o papel de professores.

É relevante mencionar que, embora o corpus e-PACT não seja o ideal, é o melhor recurso publicamente disponível que serve os nossos propósitos, porque contém frases paralelas alinhadas que são traduções dos mesmos textos literários, e essas frases frequentemente contêm linguagem informal. A falta de corpora paralelos de paráfrases em geral, mas especialmente para o português, é uma necessidade que não foi tratada com a importância que merece. Outro fator instrumental é que as frases paralelas no e-PACT correspondem a duas variedades diferentes da língua portuguesa, a europeia e a brasileira, que temos contrastado em trabalhos recentes (Barreiro & Mota, 2018). Essas características-chave são essenciais para a adaptação e revisão das variedades. Neste artigo, concentramo-nos na revisão de texto, mas o artigo serve os dois propósitos, conversão de PE/PB informal em PE/PB formal e adaptação da variedade PB na variedade de PE e vice-versa. O artigo apresenta uma contribuição pequena mas positiva para a melhoria dos padrões de edição e revisão, bem como para a automatização de transformações específicas do discurso informal para o formal.

2 Trabalho Relacionado

Os compostos verbais, que são objeto do nosso estudo, têm a particularidade de incluir um pronome clítico tanto nas frases em PE como nas frases em PB ou ter esse clítico implicado numa paráfrase das construções dos compostos verbais numa ou noutra variedade da língua portuguesa (cf. exemplo (2)). Em português, um pronome clítico desempenha um papel sintático ao nível da frase e segue diferentes regras de colocação ou ordenação, dependendo da variedade da língua (PE ou PB), do número e da semântica dos predicados, co-ocorrência com uma preposição, entre outros fatores.

Existem estudos que se centram na aquisição de pronomes clíticos em PE, dos quais os trabalhos de Silva (2008) e Costa & Grolla (2017) são apenas exemplos entre muitos, que foram referenciados em trabalhos realizados recentemente (Rebelo-Arnold et al., 2018). Esses estudos estão relacionados principalmente com dificuldades no desempenho quando se trata do uso de clíticos em fases iniciais de aquisição da linguagem. As dificuldades de aquisição dos clíticos são materializadas, em particular, por escolhas fora da norma para a sua colocação em frases. Quando olhamos para os nossos dados, verificamos que as hesitações e dificuldades se estendem até à idade adulta, e há padrões de variação na seleção e posição dos clíticos em qualquer corpus de registo oral ou simplesmente de transcrição escrita da oralidade, onde a informalidade é recorrente na escrita moderna, incluindo meios de comunicação social (redes socais), mas também em canais de comunicação mais "sérios", como jornais, artigos de opinião ou escrita literária cuja revisão não é contemplada com a devida importância.

[4] Com o objetivo de economizar espaço neste artigo, apresentamos os exemplos no modo convencional, marcados a negrito em exemplos enumerados.

[5] A utilidade das capacidades parafrásticas do eSPERTo foi explorada em duas outras aplicações descritas por Mota et al. (2016a): (i) num sistema de perguntas e respostas para aumentar o conhecimento linguístico de um agente conversacional inteligente e (ii) numa ferramenta de sumarização para auxiliar a tarefa de parafraseamento.

Em PB, por sua vez, vários estudos enfocam a observação das construções espontâneas de falantes mais ou menos escolarizados envolvendo o uso de clíticos (Neves, 1999, 2000; Castilho, 2001; Naro & Scherre, 2007, entre outros). Essa observação revela uma distância entre as duas variedades em relação à aplicação das regras de seleção e colocação de clíticos em português. Tudo isso tem impacto tanto no trabalho dos revisores e tradutores quanto na aprendizagem de línguas, quer para o português como língua materna (PLM) quer para o português como língua estrangeira (PLE). O eSPERTo pode ser usado num ambiente de aprendizagem de língua(s), onde os estudantes de PLM e PLE podem aprender a produzir e aplicar paráfrases de grande precisão (ou seja, frases semanticamente equivalentes). Portanto, os recursos aqui criados podem ajudar a auxiliar escritores e revisores na produção, revisão ou adaptação de textos, mas também podem ser valiosos num ambiente de sala de aula. Neste artigo, continuamos uma linha de investigação anterior (Barreiro & Mota, 2018), onde foi apresentada uma primeira introdução geral a uma tarefa mais ampla de encontrar variantes parafrásticas PE-PB, seguida por uma abordagem mais restrita da questão das paráfrases entre PE e PB envolvendo o clítico de terceira pessoa com valor dativo, *lhe* (Rebelo-Arnold et al., 2018). Neste estudo, concentramo-nos no alinhamento das construções de compostos verbais, quando essas construções envolvem pronomes clíticos. A nossa pequena experiência mostra que a metodologia e a abordagem são viáveis num projeto autónomo maior, desde que haja uma quantidade suficiente de corpora adequados para fornecer uma cobertura suficientemente abrangente para um processo de normalização eficaz, como o que é exigido no desenvolvimento de um sistema de parafraseamento de larga escala. Esses dados também constituirão os pilares basilares para a criação de gramáticas aplicáveis a vários casos, não apenas para a língua portuguesa, mas para outras línguas.

3 Colocação dos Clíticos em Compostos Verbais

Os clíticos em português podem deslocar-se para a esquerda ou para a direita, quer do verbo auxiliar, quer do verbo principal. Algumas das nuances da colocação do clítico em compostos verbais serão ilustradas neste artigo com exemplos do corpus e-PACT. Parte das dificuldades em estabelecer categorias parafrásticas está relacionada com o valor aproximado de construções aparentemente "equivalentes". Os exemplos ilustram que, em cada par parafrástico PE–PB, uma frase contém um composto verbal com um clítico e a outra frase contém uma paráfrase da primeira. Às vezes, a paráfrase apresenta uma estrutura do composto verbal bastante diferente, que pode nem sequer incluir o pronome clítico que ocorre na frase equivalente.

3.1 PROCLDAT ou ACC VAUX-*ter* VPARTPASS

Os exemplos (1)–(3) representam contrastes importantes com a regra evidentemente produtiva de posição enclítica em PE. Esses contrastes ocorrem na presença do auxiliar *ter* (VAUX-*ter*) e são provavelmente o modelo que gera a incorreção na construção *lhes voltava a telefonar*. Este é o caso de uma falsa analogia porque, de facto, a regra de colocação de enclíticos deveria ter sido aplicada neste caso, e.g., *voltava a telefonar-lhes*. Na paráfrase em PB, o pronome clítico desaparece através da utilização de uma transformação mais "livre". Existe uma tendência notável em PB para evitar o uso pronomes clíticos em construções deste tipo e noutras.

(1) $_{EN}$ - *It was rumoured that he collected the phone numbers of likely-sounding girls and* ***called them back*** *after the programme to make dates.*
$_{PE}$ - *Dizia-se que colecionava os números das raparigas que mais lhe agradavam e* ***lhes voltava a telefonar*** *depois, a marcar encontros.*
$_{PB}$ - *Diziam até que ele colecionava números de telefone de garotas com voz macia* ***para ligar mais tarde*** *e marcar encontros.*

No exemplo (2), a paráfrase em PB, [N VAUX-*ter* NP[*boa viagem*]] (simplificada '[Y *ter boa* X]') apresenta uma inversão do tópico de modo a evitar o uso do clítico na 3ª pessoa exigido pelo verbo *agradar* como uma paráfrase do PE [SN[*a viagem*] VPRINC *agradou* PREP*a* N] (simplificado '[X agradar a Y]'). Em PB, a seleção lexical diferente explica a ausência de ENCLITDAT. Na frase em PE, a presença do pronome clítico *lhes* é suprimida em PB pela inversão do tópico. O verbo *agradar* em português exige o uso da preposição *a* (PREP*a*), que não é exigida pelo verbo *please* em inglês. A paráfrase em PE é mais formal enquanto que a paráfrase em PB é mais neutra. O pronome *lhe* nunca pode estar ligado a um particípio passado em construções auxiliares [VAUX-*ter* + VPP].

(2) $_{EN}$ - *he hopes they have enjoyed the flight*
$_{PE}$ - *diz esperar que a viagem* ***lhes tenha agradado****.*
$_{PB}$ - *ele desejava que* ***tivessem tido uma boa*** *viagem*

No exemplo (3), o PE também apresenta uma paráfrase mais formal (mais próxima da construção / forma de expressão original em inglês) do que em PB. A variação de uma paráfrase noutra presume uma escolha do tradutor. Em detalhe, a paráfrase em PB seleciona o mesmo item lexical em PE, *mudar*, que ocorre com o pronome reflexivo *se*, mas com um infinitivo pessoal composto e PROCLIT do clítico ao verbo principal (VPRINC). No entanto, o verbo *mudar-se* (*de X para Y*) é ambíguo, i.e., o reflexivo (*-se*) é opcional (a frase estaria, ainda assim, correta se o pronome reflexivo estivesse omitido como em *tivessem mudado para...*). Esta ocorrência (menos formal em PB) é atestada, contudo, na gramática do PB que rejeita o uso dos clíticos antes de VAUX. A variedade determina a ordem do clítico. Numa oração subordinada em PE o pronome reflexivo *se* aparece antes de VAUX.

(3) $_{EN}$ - *though they* ***moved*** *in due course* ***to*** *better insulated accommodation*
$_{PE}$ - *embora mais tarde* ***se tivessem mudado para*** *uma habitação bem isolada*
$_{PB}$ - *mesmo depois de* ***terem se mudado para*** *acomodações mais isoladas*

3.2 VAUX PREP VINF+ENCLITDAT-lhe versus VAUX2 *lhe* VGER NP

No exemplo (4), o composto verbal em PB é normalizado, mas a sua paráfrase em PE é muito mais próxima da estrutura usada na frase original do texto fonte em inglês, o que faz com que pareça um pouco estranha. Não existe evidência se isto está relacionado com uma fidelidade intencional à frase original, ou uma tentativa mal sucedida para usar linguagem controlada. A paráfrase em PE consiste na construção perifrástica [*continuar a* + VINF ENCLDAT]. Em PB, a paráfrase relativamente complexa envolve o auxiliar modal *dever* seguido de um advérbio, *ainda*, seguido da construção [VAUX-*estar* PROCL-*lhe* VGER *causando* NP]. Toda a sequência de elementos em PB tem como eixo semântico a noção aspetual de ação em progresso, idêntica à da paráfrase em PE, que é expressa numa construção muito mais simples e mais concisa. Este exemplo ilustra a necessidade, já mencionada neste artigo, de construir gramáticas para o fim específico de gerar paráfrases que são adequadas e úteis a revisores, editores e estudantes de português como língua estrangeira (PLE). Não podemos afirmar categoricamente que a versão em PB se deve ao uso recorrente da construção nesta variedade ou se se trata simplesmente de uma má interpretação por parte do tradutor. Além disso, pode incluir não apenas os pronomes com valor dativo DAT *lhe*, mas também os de valor acusativo ACC, quando o verbo principal está na forma infinitiva, VINF. Esta regra aplica-se até na presença do advérbio de negação *não* que precede o verbo na posição VAUX no composto verbal. O verbo *continuar* é um VAUX (*ter*, *ser*, etc.) típico de uma perífrase verbal, pelo que atribui um significado aspetual ao verbo principal *doer*, ocupando a posição de um auxiliar atípico, tal como em *não conseguiram dominá-la.*

(4) $_{EN}$ - *There's no bally reason why [ ]* ***should be giving you any more pain****.*
$_{PE}$ - *Não há a mínima razão para [ ]* ***continuar a doer-lhe***
$_{PB}$ - *Não há um pingo de razão por que [ ]* ***deva ainda estar lhe causando essa dor***

3.3 PREP-a VINF+REFLPRO-se → PROCLITse VGER

No exemplo (5), o PE determina o uso enclítico enquanto que o PB determina o uso proclítico. É interessante notar que ambas as variedades mantêm a noção aspetual de progressão. Esta noção é duplamente representada, tanto pela seleção de PREP-*a* VINF em PE e um gerundivo VGER em PB com a elipse do auxiliar *estar* em ambas as construções, e pela seleção lexical, pela qual ambos os verbos reflexivos *formar-se* e *preparar-se* expressam a noção de uma ação em curso. Estes não correspondem a paráfrases no sentido transformacional definido por Gross (1975, 1981), contudo, a tarefa de alinhamento parafrástico fornece candidatos que podem ser perfeitamente adicionados a um sistema de parafraseamento como pares parafrásticos. Esta é uma formalização importante e necessária que propõe sistematizar as paráfrases entre PE e PB, mesmo que a sua implementação seja, à partida, complexa. A importância deste exemplo reside no facto de a oposição PREP-*a* VINF → VGER ser uma marca distintiva entre as duas variedades do português. Assim, torna-se necessário oferecer listas exaustivas de possibilidades parafrásticas sempre com o maior cuidado para que o significado das paráfrases seja de boa qualidade, independentemente de o nosso objetivo ser estabe-

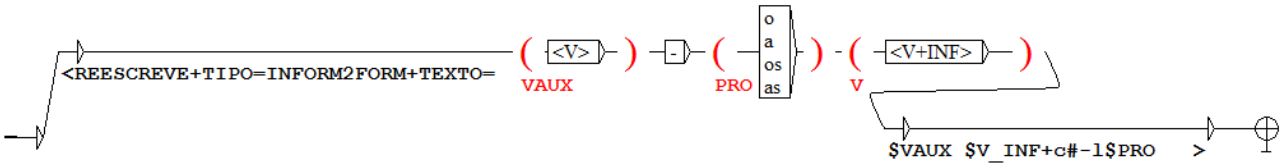


Figura 1: Gramática para normalizar linguagem informal em linguagem formal com o uso de clíticos.

lecer uma versão controlada do português, para dar assistência à tarefa da revisão, para apoiar a edição de texto ou o ensino de PLE.

(5) *EN - I sense a storm of depression flickering on the horizon, and a tidal wave of despair* ***gathering itself to swamp me****.*
PE - Sinto uma tempestade de depressão avolumar-se no horizonte e uma maré de desespero ***a formar-se para me engolir****.*
PB - Pressinto a chegada de uma tempestade de depressão se formando no horizonte e uma onda de desespero ***se preparando para me engolir****.*

4 Normalização de Linguagem Informal em Linguagem Formal

Baseados nas principais características apontadas na Secção 3 relativamente à colocação dos clíticos em compostos verbais em vários contextos: (i) co-ocorrência com modais (VMOD) em orações relativas; (ii) vários casos do uso de proclíticos ou enclíticos em contextos formais e informais (3.1); (iii) co-ocorrência com verbos aspetuais (VASP) em construções perifrásticas (3.2); ou (iv) co-ocorrência com verbos aspetuais com significado progressivo (3.3), propomos aqui a criação de uma gramática local que permite a normalização de uma construção verbal composta informal, onde o pronome enclítico aparece depois de um verbo (V). Este verbo pode ser um auxiliar (VAUX) ou qualquer outra forma verbal (VASP, VMOD, etc.). Esta construção verbal informal está normalizada numa construção formal equivalente através de uma gramática local ilustrada na Figura 1. O clítico, que na construção informal se encontra ligado ao verbo auxiliar (guardado na variável `$VAUX`), que por sua vez será guardado na variável `$PRO`, transita para uma posição a seguir ao verbo principal (que está na forma infinitiva `<V_INF>` e que será guardado na variável `$V`). Essa transição corresponde a delimitar a construção informal com a etiqueta

```
<REESCREVE+TIPO=INFORM2FORM+TEXTO=$VAUX$V_INF-#1$PRO>
```

atribuindo a `TEXTO` a concatenação dos valores de `$VAUX`, da forma infinitiva (`$V_INF`) do verbo principal modificada quando está na presença de um clítico `+c`, seguida do clítico antecedido por `-l` (`-l$PRO`) em que `#` é usado para garantir que `+c` e `-l` não são lidas como um todo, i.e., apenas como uma sequência `+c-l`, mas sim como duas sequências). Esta gramática foi desenvolvida no NooJ (Silberztein, 2016) e está disponível publicamente através do módulo do Port4NooJ v3.0 (Mota et al., 2016b).

Baseados na gramática proposta, centenas de procedimentos de normalização/parafraseamento ocorrem. Estas paráfrases normalizadas podem integrar o sistema de parafraseamento eSPERTo depois de validação por um linguista e os resultados podem ser reproduzidos através deste sistema. A Figura 2 ilustra a capacidade de revisão dentro do eSPERTo, onde uma frase escrita numa linguagem mais ou menos informal ou menos cuidada, pode ser revista com sugestões que são mais polidas, ou correspondem a uma norma da linguagem escrita. Por exemplo, para a frase *A menina generosa queria-o surpreender todos os dias*, o eSPERTo apresenta, como opção de conversão para o composto verbal informal com clítico *queria-o surpreender*, o seu equivalente formal *queria surpreendê-lo*. O sistema parafrástico oferece esta sugestão de parafraseamento ao utilizador, onde o clítico migra de uma posição enclítica ligada ao verbo *querer* para uma posição enclítica ligada ao verbo principal. Esta transformação faz com que a forma infinitiva do verbo principal, *surpreender*, mude para *surpreendê-* antes dos pronomes enclíticos com valor acusativo ACC *-lo*, *-la*, *-los*, *-las*, uma regra ortográfica motivada por razões fonéticas, como nos exemplos anteriores (cf. Secção 1).

5 Conclusões e Trabalho Futuro

A revisão estilística representa uma funcionalidade importante do projeto eSPERTo, cujo enfoque principal é o desenvolvimento de um sistema de parafraseamento inovador com capacidade para produzir frases semanticamente equivalentes e formas de expressão, sempre visando a melhoria da qualidade de cada texto. Neste artigo, tentámos estabelecer algumas categorias definidas com base na estrutura sintática das cons-

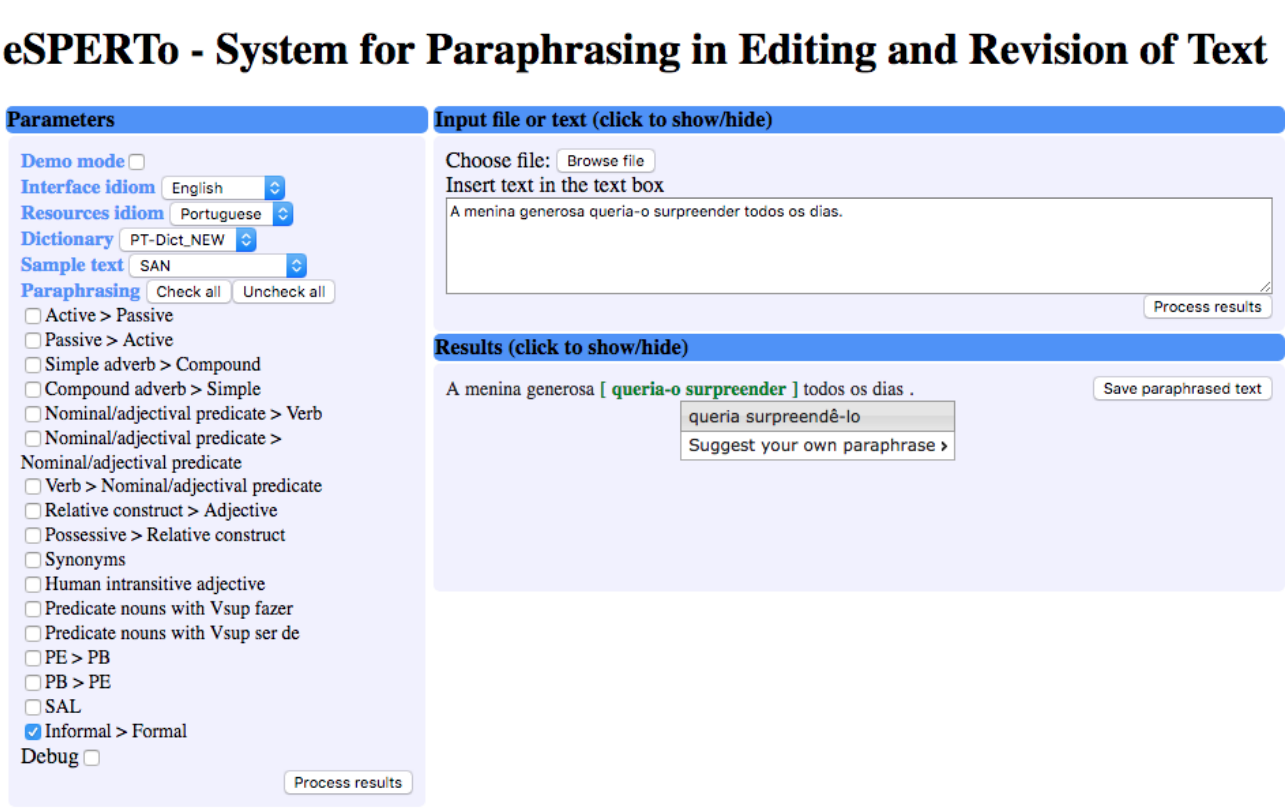


Figura 2: Conversão de um composto verbal informal com um pronome clítico num equivalente formal onde o clítico surge depois do verbo principal.

truções de compostos verbais envolvendo clíticos. Fizemos este estudo com base em pares de construções parafrásticas extraídas de frases de dois romances de David Lodge traduzidas para PE e PB. É importante notar que, especialmente em textos literários, os tradutores frequentemente usam uma tradução livre, que (idealmente) preserva o significado do texto original, mas envolve a reestruturação da sintaxe, às vezes com um uso flexível do léxico ou expressões para oferecer uma articulação natural das palavras na língua de destino. Daí resulta que o texto traduzido possa parecer "mais leve e flexível" ou mais ou menos idiomático relativamente ao texto original. Nesse processo, até mesmo os tradutores humanos profissionais podem introduzir erros, tornando uma parte específica de uma tradução infiel ao original. Em suma, a tradução pode ser vista como um processo de parafraseamento usando palavras noutro idioma, onde a introdução de diferentes palavras e estruturas pode criar uma certa distância entre as línguas de origem e de destino. Neste sentido, no nosso estudo, as paráfrases assumem uma equivalência semântica completa competindo com paráfrases que retêm uma equivalência conceptual aproximada (Barzilay & McKeown, 2001). As primeiras são indispensáveis para obter precisão, mas não podemos dispensar as segundas porque elas também desempenham um papel importante nas tarefas de parafraseamento, nomeadamente na revisão ou mudança estilística, ou quasi-parafraseamento (Barreiro, 2009).

Os dados extraídos dos corpora, embora sejam úteis e contenham significância estatística, requerem análise linguística e categorização de padrões e estruturas que comportam equivalências semânticas. Esperamos que a nossa tentativa de definir uma tipologia e usar conhecimento linguístico para normalizar construções informais tenha continuidade, porque revela uma tarefa crucial no desenvolvimento de uma ferramenta de revisão ou melhoria da língua. Este artigo esclarece a necessidade de incluir um recurso que distingue os registos formal/informal em várias aplicações para edição e revisão de texto, inclusivamente para ser usado num ambiente de aprendizagem de línguas, no qual os estudantes precisam de compreender as formas formais e informais de comunicação e de saber quando utilizar umas e outras. Num futuro próximo, discutiremos o tópico da utilização de agentes conversacionais que interagem com os alunos e lhes ensinam as diferenças entre a linguagem formal e a informal, com base na escrita do próprio aluno. Para textos escritos numa linguagem muito formal, os agentes conversacionais podem sugerir frases mais informais, ou vice-

versa, de acordo com o contexto comunicativo. Este tópico será explorado no âmbito de trabalhos colaborativos da Ação COST enetCollect, onde os agentes conversacionais terão um papel de professores numa aplicação de aprendizagem de línguas.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e Tecnologia através do projeto com a referência UID/CEC/50021/2013, do projeto exploratório eSPERTo com a referência EXPL/MHC-LIN/2260/2013, e através da bolsa de pós-doutoramento com a referência SFRH/BPD/91446/2012.

Referências

Baptista, Jorge. 2012. ViPEr: A lexicon-grammar of European Portuguese verbs. Em *31st International Conference on Lexis and Grammar*, 10–16.

Baptista, Jorge. 2013. ViPEr: uma base de dados de construções léxico-sintáticas de verbos do Português Europeu. Em *Actas do XXVIII Encontro da APL - Textos Selecionados*, 111–129.

Baptista, Jorge & Nuno Mamede. 2018. *Dicionário gramatical de verbos do português europeu*. Universidade de Aveiro.

Baptista, Jorge, Nuno Mamede & Fernando Gomes. 2010. Auxiliary verbs and verbal chains in European Portuguese. Em *Computational Processing of the Portuguese Language (PROPOR)*, 110–119.

Barreiro, Anabela. 2009. *Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation*: Universidade do Porto. Tese de Doutoramento.

Barreiro, Anabela & Cristina Mota. 2017. e-PACT: eSPERTo Paraphrase Aligned Corpus of EN-EP/BP Translations. *Tradução em Revista* 1(22). 87–102.

Barreiro, Anabela & Cristina Mota. 2018. Paraphrastic variance between European and Brazilian Portuguese. Em *5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 111–121.

Barreiro, Anabela, Francisco Raposo & Tiago Luís. 2016. CLUE-Aligner: An alignment tool to annotate pairs of paraphrastic and translation units. Em *10th Language Resources and Evaluation Conference (LREC)*, 7–13.

Barzilay, Regina & Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. Em *39th Annual Meeting on Association for Computational Linguistics*, 50–57.

Bick, Eckard. 2000. *The parsing system "palavras". automatic grammatical analysis of portuguese in a constraint grammar framework*. Arhus University Press.

Castilho, Ataliba. 2001. O português do Brasil. Em *Linguística Românica*, 237–269. Ática.

Costa, João & Elaine Grolla. 2017. Pronomes, clíticos e objetos nulos: dados de produção e compreensão. Em *Aquisição de língua materna e não materna: questões gerais e dados do português*, 177–199. Language Science Press.

Cunha, Celso & Luís Lindley-Cintra. 1986. *Nova gramática do português contemporâneo*. João Sá da Costa.

Gonçalves, Anabela. 1999. *Predicados complexos verbais em contexto de infinitivo não-preposicionado do português europeu*: Universidade de Lisboa. Tese de Doutoramento.

Gross, Maurice. 1975. *Méthodes en syntaxe: régime des constructions complétives* Actualités scientifiques et industrielles. Hermann.

Gross, Maurice. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages* 15(63). 7–52.

Gross, Maurice. 1998. La fonction sémantique des verbes supports. *Travaux de Linguistique: Revue Internationale de Linguistique Française* 37(1). 25–46.

Mota, Cristina, Anabela Barreiro, Francisco Raposo, Ricardo Ribeiro, Sérgio Curto & Luísa Coheur. 2016a. eSPERTo's paraphrastic knowledge applied to question-answering and summarization. Em *Automatic Processing of Natural Language Electronic Texts with NooJ*, 208–220.

Mota, Cristina, Paula Carvalho & Anabela Barreiro. 2016b. Port4NooJ v3.0: Integrated linguistic resources for Portuguese NLP. Em *10th Language Resources and Evaluation Conference (LREC)*, 1264–1269.

Naro, Anthony Julius & Maria Marta Pereira Scherre. 2007. *Origens do português brasileiro*. Parábola.

Neves, Maria Helena Moura. 1999. *Gramática do português falado*. UNICAMP.

Neves, Maria Helena Moura. 2000. *Gramática de usos do português*. UNESP.

Paiva Raposo, Eduardo. 2013. Verbos auxiliares. Em *Gramática do Português*, vol. 2, 1221–1281. Fundação Calouste Gulbenkian.

Pontes, Eunice. 1973. *Verbos auxiliares em português* Perspectivas Linguísticas. Vozes.

Rebelo-Arnold, Ida, Anabela Barreiro, Paulo Quaresma & Cristina Mota. 2018. Alinhamentos parafrásticos PE–PB de construções de predicados verbais com o pronome clítico *lhe*. *Linguamática* 10(2). 3–11.

Santos, Diana. 2015. Portuguese language identity in the world: adventures and misadventures of an international language. Em *Language - Nation - Identity: The questione della lingua in an Italian and non-Italian context*, 31–54. Cambridge Scholars Publishing.

Silberztein, Max. 2016. *Formalizing Natural Languages: the NooJ Approach*. Wiley Eds.

Silva, Carolina G. A. G. 2008. *Assimetrias na Aquisição de Clíticos Diferenciados em Português Europeu*: Universidade Nova de Lisboa. Tese de Mestrado.

Silva, João, António Branco, Sérgio Castro & Ruben Reis. 2010. Out-of-the-box robust parsing of Portuguese. Em *9th Conference on the Computational Processing of Portuguese (PROPOR)*, 75–85.

Proposta recibida en outubro 2018 e aceptada para publicación en decembro 2018.

Explorando Métodos Non-Supervisados para Calcular a Similitude Semántica Textual

Exploring Unsupervised Methods to Sematic Textual Similarity

Pablo Gamallo
CiTIUS
Univ. de Santiago de Compostela
pablo.gamallo@usc.es

Martín Pereira-Fariña
Departamento de Filosofía e Antropoloxía
Universidade de Santiago de Compostela
martin.pereira@incipit.es

Resumo

Neste traballo preséntanse varios métodos non-supervisados para a detección da similitude semántica textual, os cales están baseados en modelos distribucionais e no parseado de dependencias. Os sistemas son avaliados mediante datasets empregados na ASSIN Shared Task, celebrada conxuntamente co PROPOR 2016. Os métodos máis básicos ofrecen un mellor comportamento que aqueles, mais complexos, que inclúen información sintáctico-semántica na análise das oracións. Por último, o uso de modelos distribucionais construidos automaticamente a partir de corpus ofrece resultados comparábeis ás estratexias que utilizan recursos léxicos externos construídos manualmente.

Palabras chave

similitude textual, análise de dependencias, extracción de información aberta

Abstract

This paper presents some unsupervised methods for detecting semantic textual similarity, which are based on distributional models and dependency parsing. The systems are evaluated using the dataset realased by the ASSIN Shared Task co-located with PROPOR 2016. The more basic methods offer better behavior than the more complex ones, which include syntactic-semantic information in sentence analysis. Finally, the use of distributional models built automatically from corpora provides results comparable to strategies that use external lexical resources built manually.

Keywords

textual similarity, dependency analysis, open information extraction

1 Introdución

As paráfrases defínense como pares de oracións que conteñen a mesma ou case a mesma información (Androutsopoulos & Malakasiotis, 2010). Polo tanto, o recoñecemento de paráfrases consiste no recoñecemento de oracións (ou pequenos fragmentos de texto) que teñen aproximadamente o mesmo significado nun contexto dado. Unha tarefa similar a á identificación de paráfrases é a Similitude Semántica Textual (SST), a cal busca determinar o grao de equivalencia semántica entre dous fragmentos de texto.

SST pode empregarse en moitas das tarefas do Procesamento da Linguaxe Natural (PLN), dende recuperación de información ata a detección automática de plaxio. Existen varios métodos de SST na bibliografía, que van dende métodos non-supervisados e con recursos lixeiros ata métodos supervisados e con recursos intensos.

O principal obxectivo deste traballo é describir e avaliar métodos non-supervisados de SST baseados en modelos distribucionais e aplicados ao portugués. Máis concretamente, compararemos estratexias non-supervisadas de recursos lixeiros con outras estratexias, tamén non-supervisadas, mais que utilizan recursos máis intensos, como tesaurus, redes de coñecemento, ou mesmo información sintáctica. Todos os experimentos son levados a cabo usando o datasets proporcionado por ASSIN Shared Task (*Avaliação de Similaridade Semântica e Inferência Textual*), celebrado conxuntamente con PROPOR 2016 (Fonseca et al., 2016).

Na seguinte sección (2), describimos os modelos de SST para o portugués. A continuación, na Sección 3 presentamos tres métodos non-supervisados diferentes. Na Sección 4 expoñemos e discutimos os resultados dos nosos experimentos; por último, na Sección 5, resumimos as nosas principais conclusións e propoñemos algunas ideas para o traballo futuro.

DOI: 10.21814/lm.10.2.275

2 Similitude semántica textual para o Portugués

SST é unha das dúas tarefas avaliadas no ASSIN (Fonseca et al., 2016). A outra subtarefa, inferencia textual, está fóra do ámbito deste traballo. A tarefa SST consiste en asignar un valor numérico (entre 1 e 5) a pares de oracións segundo o grao de similitude semántica entre elas: canto maior sexa o valor numérico, maior é o grao de similitude entre elas. Esta tarefa está inspirada pola *SemEval Task 2* sobre similitude semántica textual (Agirre et al., 2015, 2016). Na tarefa compartida sobre SST no *SemEval 2016*, enviáronse 119 sistemas diferentes, o que denota o enorme interese deste campo.

A inmensa maioría (todos menos un) dos sistemas presentados en ASSIN estaban baseados no uso de métodos supervisados. A mellor equipa (Hartmann, 2016) aplicou regresión lineal para adestrar un clasificador cuxas características son os valores da medida do coseno que representan o grao de similitude de cada para de oracións. Estas modélanse de dous xeitos diferentes: a adición de valores TF-IDF (cada palabra da oración é un valor TF-IDF) e a adición de vectores de valores distribucionais, onde cada palabra se representa como un vector contextual a aprendido mediante redes neuronais (Mikolov et al., 2013). As similitudes do coseno entre estes tipos de representacións son valores de entrada do clasificador básico.

O segundo mellor sistema (e o mellor para o dataset do portugués europeo, de Fialho et al. (2016), adestrou un clasificador baseado en modelos de regresión (*Kernel Ridge Regression*) cun número maior de rasgos que os outros sistemas, inlcuíndo distancias de edición entre cadeas de caracteres, o tamaño da maior subcadea común de caracteres, distintas métricas de similitude dependentes dos valores de ocorrencia de TF-IDF. En total, o sistema usou máis de 90 características.

A única estratexia non suspervisada no ASSIN é a chamada *Reciclagem* e foi proposta por Alves et al. (2016). Este sistema usa medidas de similitude baseadas nas relacións semánticas extraídas desde tesauros externos e recursos léxicos. O preprocesamento é realizado co etiquetador morfosintático de OpenNLP (Apache) e o lematizador LemPort (Rodrigues et al., 2014). Entre os recursos léxicos utilizados, destaca PAPEL (Oliveira et al., 2010), que consiste en relacións extraídas do dicionario *Porto Editora da Língua Portuguesa*, mediante a elaboración de regras baseadas en regularidades atopadas nas definicións do dicionario. Alén deste recurso, os experimentos realizados con Reciclagem inclúen outras redes de coñecimento con maior cobertura, nomeadamente, *CARTÃO* (Oliveira et al., 2011), que a súa vez consta doutros recursos como PAPEL e as relacións extraídas do *Dicionário Aberto* (Simões et al., 2012), así como diferentes variantes do WordNet portugués: *OpenWordNet.PT* (de Paiva et al., 2012) e *PULO* (Simões & Guinovart, 2014).

Neste traballo, avaliaremos varias estratexias non-supervisadas baseadas fundamentalmente en modelos distribucionais sobre o mesmo dataset empregado en ASSIN.

3 Similitude semántica textual non-supervisada

Nesta sección, definimos tres estratexias non-supervisadas: a máis básica baséase na semántica distribucional e na etiquetaxe morfo-sintáctica (*PoS tagging*), mentres que os outros métodos dependen da análise sintáctica alén de técnicas de extracción de información aberta (*Open Information Extraction).*

3.1 Similitude distribucional

Unha das estratexias máis simples e básicas para calcular a similitude entre dúas oracións consiste en sumar os valores de semellanza entre cada par de palabras que aparecen nas dúas oracións comparadas. Neste caso, só tomamos en conta palabras léxicas, é dicir, nomes, verbos e adxectivos. O valor de similitude calcúlase concretamente mediante a medida de coseno entre vectores que conforman matrices de palabras encapsuladas (*word embeddings*) pre-adestradas. O algoritmo é o seguinte: escollemos a oración máis curta e seleccionamos a primeira palabra léxica de dita oración. A seguir, calculamos a similitude do coseno entre a palabra escollida e todas as palabras léxicas que conforman a oración máis longa, sumando todos os valores de semellanza de maneira a obtermos a relevancia léxica da primeira palabra seleccionada con respeito á oración máis longa. Despois, realizamos a mesma operación para o resto de palabras da oración máis curta e calculamos a media dividindo a suma final polo número total de palabras léxicas que conforman a oración curta. Máis formalmente, dado o vector da palabra $\mathbf{p_s}$ pertencente a U_s, onde U_s é o conxunto de vectores de palabras léxicas da oración curta, a *relevancia léxica*, LR, de $\mathbf{p_s}$ dada a oración máis longa, calcúlase do seguinte xeito:

$$LR(\mathbf{p_s}, U_l) = \sum_{\mathbf{p_i} \in U_l}^{L} Cosine(\mathbf{p_s}, \mathbf{p_i}) \quad (1)$$

onde U_l é o conxunto dos vectores de palabras léxicas da oración máis longa e L é o número de palabras léxicas nesa mesma oración. Por conseguinte, o valor final de similitude (DSim) para o par U_s e U_l é a media de LR:

$$\text{DSim}(U_s, U_l) = \frac{\sum_{\mathbf{p_i} \in U_s}^{S} LR(\mathbf{p_s}, U_l)}{S} \quad (2)$$

onde S é o número de palabras léxicas na oración máis curta. Convén salientar que esta estratexia non codifica a información sobre a orde dos elementos da oración.

3.2 Extracción de proposicións básicas

DSim só toma en conta relacións semánticas ao nivel da palabra sen considerar fenómenos máis complexos como a orde das palabras ou as dependencias sintácticas entre elas. Para podermos tomar en conta estes fenómenos, desenvolvemos unha nova metodoloxía na que a estratexia de similitude definida previamente (DSim) se aplica a *proposicións básicas* extraídas das oracións, en vez de aplicarse directamente ás oracións. As proposicións básicas (ou tripletas) son relacións suxeito-verbo-obxecto identificadas e extraídas mediante técnicas de Extracción de Información Aberta (OIE) (Etzioni et al., 2011; Gamallo & García, 2015). Unha oración pode conter varias proposicións básicas, como por exemplo, a oración seguinte:

> *En maio de 2010, os partidos da oposición boicotaron as eleccións despois de acusacións de fraude electoral.*

Esta oración, despois dunha análise sintáctica en dependencias, pode dividirse en, polo menos, tres tripletas ou proposicións básicas, tal e como se mostra no cadro 1.

O método de cómputo de similitude baseada en proposicións, que chamamos BPROP, só toma en conta as palabras léxicas contidas nas proposicións extraídas. Deste xeito, pódese computar a similitude DSim comparando as tres proposicións do cadro 1 (extraídas da oración do noso exemplo) coa seguinte proposición (extraída da oración máis curta: *os partidos boicotaron as eleccións*):

subject	**relation**	**object**
partido	boicotar	eleccións

Os dous conxuntos de vectores de palabras léxicas elabóranse directamente das proposicións extraídas. A partir do exemplo citado, U_s (o conxunto de vectores de lemas da oración máis curta) é constituido mediante a selección dos lemas léxicos seguintes:

> {*partido, boicotar, elección*}

Pola outra banda, U_l (os vectores de lemas da oración máis longa), consta de:

> {*partido, oposición, boicotar, elección, maio, 2010, acusación, fraude, electoral*}

Estes conxuntos de vectores de lemas serven para calcular tanto a relevancia léxica (ecuación 1) como a similitude DSim (2). Polo tanto, a estratexia BPROP só determina que lemas están nos conxuntos comparados, mais non modifica o método de cómputo da similitude en si mesmo.

3.3 Estrutura de argumentos

A terceira estratexia que imos utilizar é moi similar a BPROP, mais en vez de extraer todas as posíbeis relacións suxeito-verbo-obxecto, o obxectivo da mesma é seleccionar a estrutura argumental principal de cada oración. Definimos a estrutura argumental principal dunha oración como aquela formada pola raíz (verbo principal) e os núcleos dos seus constituíntes directos. Deste xeito, a similitude baseada na estrutura argumental, que chamamos ARGSTR, calcúlase a partir das palabras léxicas que se encontran dentro do esqueleto estrutural extraído das oracións comparadas.

Como no caso de BPROP, a estratexia ARGSTR só modifica as listas de lemas que se van utilizar para computar a similitude DSim. No caso das dúas oracións comparadas no exemplo anterior, a lista correspondente á oración máis curta, U_s, sería a mesma que no caso anterior:

> {*partido, boicotar, elección*}

Pois tanto *partido* como *elección* son os núcleos dos constituíntes directos do verbo raíz: *boicotar*. No entanto, a lista da oración máis longa, U_l, é máis restritiva que na estratexia BPROP:

> {*partido, boicotar, elección, maio, acusación*}

O resto de lemas léxicos: *oposición*, *2010*, *fraude* e *electoral* non son constituíntes directos do verbo raíz senon doutros constituíntes da cláusula.

subject	relation	object
partido de oposición	boicotar	elección
partido de oposición	boicotar elección en	maio de 2010
partido de oposición	boicotar elección despois de	acusación de fraude electoral

Cadro 1: Tres proposicións básicas extraídas de: *En maio de 2010, os partidos da oposición boicotaron as eleccións despois de acusacións de fraude electoral*

4 Expermimentos

Para avaliar a calidade das estratexias definidas na sección previa no seu uso na captura de similitude semántica textual (SST), probámolas nos datasets fornecidos pola tarefa partillada ASSIN (Fonseca et al., 2016). Os experimentos foron realizados utilizando varios modelos semánticos pre-adestrados e disponíbeis publicamente, nomeadamente os modelos distribucionais transparantes e baseados en sintaxe descritos en Gamallo (2017).

O texto das oracións do test foi procesado con diferentes módulos de LinguaKit, unha suite lingüística mulilingüe e de código aberto (Gamallo & Garcia, 2017).[1] Máis concretamente, para podermos implementar as tres estratexias introducidas previamente, utilizamos o módulo de etiquetaxe morfo-sintáctica (Garcia & Gamallo, 2015), o parser de dependencias incluído en LinguaKit (Gamallo & Garcia, 2018), que son necesarios para desenvolver a estratexia ARGSTR, así como o módulo OIE de extracción de tripletas, (Gamallo & García, 2015), requirido por BPROP.

O cadro 2 mostra os valores, en termos de correlación de Pearson, obtidos polas tres estratexias non-supervisadas obxecto de estudo (DSim, ARGSTR, and BPROP), en base a tres listas de pares de oracións: portugués europeo, portugués brasileiro e a unión dos dous (Total). A cada par de oracións se lle asigna un valor entre 1 e 5, de xeito que canto maior é o valor maior é a semellanza entre as dúas oracións comparadas. Cada sistema é avaliado mediante a medición da correlación entre os valores anotados por persoas e os valores devolvidos polo sistema. O cadro tamén mostra na última fila os resultados atinxidos polo único sistema non-supervisado, *Reciclagem*, que participou na tarefa partillada ASSIN.

Como se pode comprobar, as estratexias máis básicas (DSIM e Reciclagem) son as que conseguen os mellores resultados. Ambas abordaxes utilizan información lingüística moi básica: lematización e recursos semánticos externos (modelos distribucionais pre-adestrados, no caso de DSim, e tesaurus externos, no casos de Reciclagem). Polo contrario, as dúas estratexias baseadas en información lingüística mais elaborada, nomeadamente análise sintáctica e extracción de información aberta (ARGSTR e BPROP) devolven valores moito máis baixos. Mesmo se vai ser preciso realizar unha análise de erros en profundidade, unha análise superficial dos mesmos lévanos a afirmar que os erros sintácticos provocados polo analizador son determinantes nos resultados finais destas dúas estratexias.

É preciso tamén salientar que hai unha importante diferenza entre DSim e Reciclagem. O primeiro utiliza modelos distribucionais automaticamente construídos a partir de corpus, mentres que o segundo utiliza relacións semánticas extraídas de recursos elaborados manualmente. En consecuencia, a estratexia inherente a DSim é completamente non-supervisada, mentres que o método utilizado por Reciclagem require unha supervisión distante pois, en última instancia, depende de tesaurus manuais.

5 Conclusións

Neste traballo probamos e avaliamos diferentes estratexias non-supervisadas para medir a similitude semántica textual. O estándar de referencia, cimentado unicamente no cálculo das palabras compartidas e similares, claramente mellora os resultados de métodos máis complexos enriquecidos con análise sintáctica e extracción básica de proposicións. Os resultados tamén mostran como o uso de modelos distribucionais dentro de estratexias non-supervisadas conseguen valores comparábeis aos que usan recursos léxicos externos construídos manualmente.

Para o traballo futuro, analizaremos con detalle os tipos de erros xerados nas técnicas baseadas na análise sintáctica co obxectivo de propor novas estratexias non-supervisadas para SST. Tamén avaliaremos estas técnicas con datasets orientados a outras tarefas máis alá da SST, tales como a identificación de paráfrases, as cales poderían ser máis sensíbeis á información sintáctica.

[1] https://github.com/citiususc/Linguakit

Sistemas	PT Europeo	PT Brasileiro	Total
DSim	**0.54**	0.56	0.53
ARGSTR	0.27	0.22	0.24
BPROP	0.29	0.24	0.26
Reciclagem	0.53	**0.59**	**0.54**

Cadro 2: Valores (correlación Pearson) devoltos polos nosos tres sistemas e máis pola estratexia non-supervisada (*Reciclagem*), que participou na tarefa compartida ASSIN.

Agradecementos

Este traballo foi financiado polo proxecto TelePares (MINECO, ref:FFI2014-51978-C2-1-R), e a Consellería de Cultura, Educación e Ordenación Universitaria (acreditación 2016-2019, ED431G/08 e Programa de Formación Posdoutoral da Xunta de Galicia 2016) e European Regional Development Fund (ERDF).

Referencias

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria & Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic textual similarity, english, spanish and pilot on interpretability. En *9th International Workshop on Semantic Evaluation (SemEval)*, 252–263.

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau & Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. En *10th International Workshop on Semantic Evaluation (SemEval)*, 497–511.

Alves, Ana Oliveira, Ricardo Rodrigues & Hugo Gonçalo Oliveira. 2016. ASAPP: alinhamento semântico automático de palavras aplicado ao português. *Linguamática* 8(2). 43–58.

Androutsopoulos, Ion & Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research* 38. 135 – 187.

Apache. 2014. *Apache OpenNLP*. The Apache Software Foundation. http://opennlp.apache.org.

Etzioni, Oren, Anthony Fader, Janara Christensen, Stephen Soderland & Mausam. 2011. Open Information Extraction: the Second Generation. En *International Joint Conference on Artificial Intelligence*, 3–10.

Fialho, Pedro, Ricardo Marques, Bruno Martins, Luísa Coheur & Paulo Quaresma. 2016. INESC-ID@ASSIN: Medição de similaridade semântica e reconhecimento de inferência textual. *Linguamática* 8(2). 33–42.

Fonseca, Erick Rocha, Leandro Borges dos Santos, Marcelo Criscuolo & Sandra Maria Aluísio. 2016. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática* 8(2). 3–13.

Gamallo, Pablo. 2017. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation* 51(3). 727–743.

Gamallo, Pablo & Marcos García. 2015. Multilingual open information extraction. En *17th Portuguese Conference on Artificial Intelligence (EPIA)*, 711–722.

Gamallo, Pablo & Marcos Garcia. 2017. LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática* 9(1).

Gamallo, Pablo & Marcos Garcia. 2018. Dependency parsing with finite state transducers and compression rules. *Information Processing & Management* 54(6). 1244–1261.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. En *Languages, Applications and Technologies (CCIS)*, vol. 563, 65–75.

Hartmann, Nathan Siegle. 2016. Solo queue at ASSIN: combinando abordagens tradicionais e emergentes. *Linguamática* 8(2). 59–64.

Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. En *Conference of the North American Chapter of the ACL: Human Language Technologies*, 746–751.

Oliveira, Hugo Gonçalo, Leticia Antón Pérez, Hernani Pereira Costa & Paulo Gomes. 2011. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir

de dicionários electrónicos. *Linguamática* 3(2). 23–38.

Oliveira, Hugo Gonçalo, Diana Santos & Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e a sua avaliação. *Linguamática* 2(1). 77–93.

de Paiva, Valeria, Alexandre Rademaker & Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian wordnet for reasoning. En *International Conference on Computational Linguistics (COLING)*, 353–360.

Rodrigues, Ricardo, Hugo Gonçalo Oliveira & Paulo Gomes. 2014. LemPORT: a high-accuracy cross-platform lemmatizer for Portuguese. En *3rd Symposium on Languages, Applications and Technologies (SLATE)*, 267–274.

Simões, Alberto & Xavier Gómez Guinovart. 2014. Bootstrapping a Portuguese WordNet from Galician,Spanish and English wordnets. En *Second International Conference on Advances in Speech and Language Technologies for Iberian Languages (IberSPEECH)*, 239–248.

Simões, Alberto, Álvaro Iriarte Sanromán & José João Almeida. 2012. Dicionário-aberto: A source of resources for the portuguese language processing. En *Computational Processing of the Portuguese Language (PROPOR)*, 121–127.

http://www.linguamatica.com/

linguaMATICA

POP: Por Outras Palavras

Alinhamentos Parafrásticos PE–PB de Construções de Predicados Verbais com o Pronome Clítico *lhe*
Ida Rebelo-Arnold, Anabela Barreiro, Paulo Quaresma & Cristina Mota

Construções Conversas do Português do Brasil: Descrição e Classificação Iniciais
Nathália Perussi Calcia & Oto Araujo Vale

Paráfrase de Advérbios terminados em *–mente* em Português
Jorge Baptista

Detecção de Paráfrases na Língua Portuguesa usando Sentence Embeddings
Marlo Souza & Leandro M. P. Sanches

Análise da Capacidade de Identificação de Paráfrase em Ferramentas de Resolução de Correferência
Bernardo S. Consoli, Joaquim F. dos Santos Neto, Sandra C. de Abreu & Renata Vieira

Parafraseamento Automático de Registo Informal em Registo Formal na Língua Portuguesa
Anabela Barreiro Ida Rebelo-Arnold, Jorge Baptista, Cristina Mota & Isabel Garcez

Explorando Métodos Non-Supervisados para Calcular a Similitude Semántica Textual
Pablo Gamallo & Martín Pereira-Fariña