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.

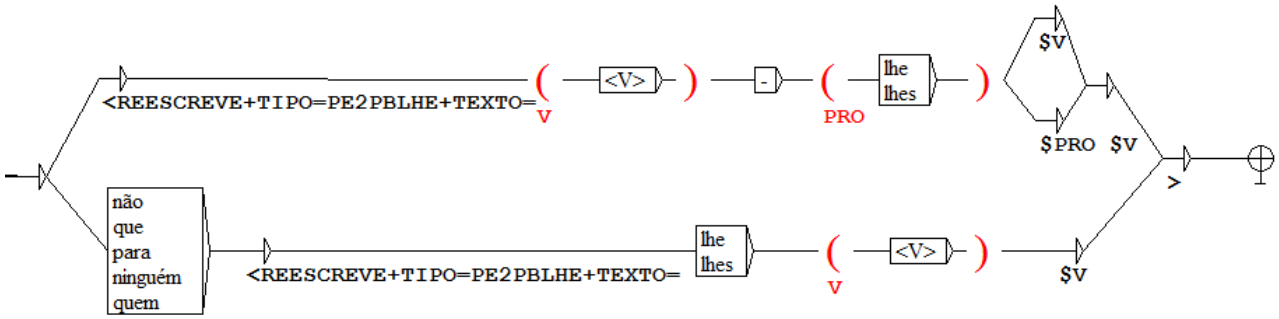

Figura 1: Gramática para formalizar a conversão de predicados verbais com *lhe* de PE em PB.

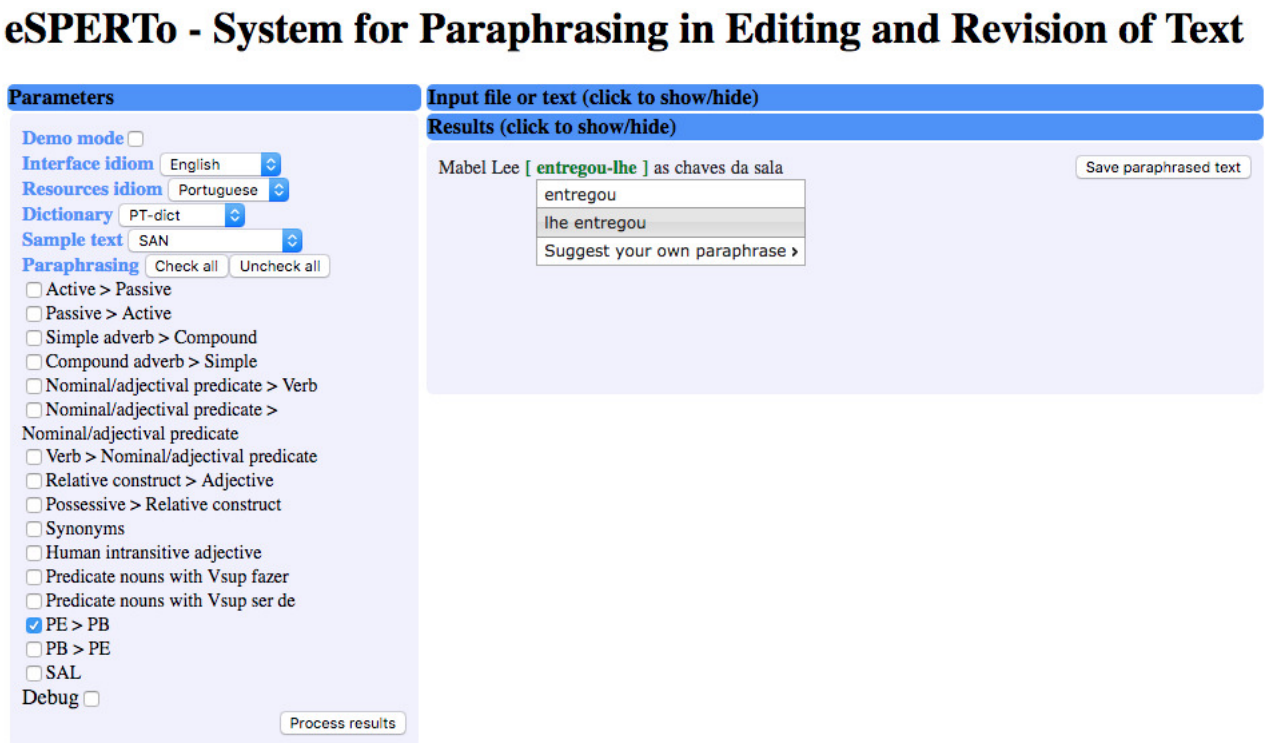

Figura 2: Conversão de um predicado verbal com o pronome clítico *lhe* de PE em PB.

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.