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-

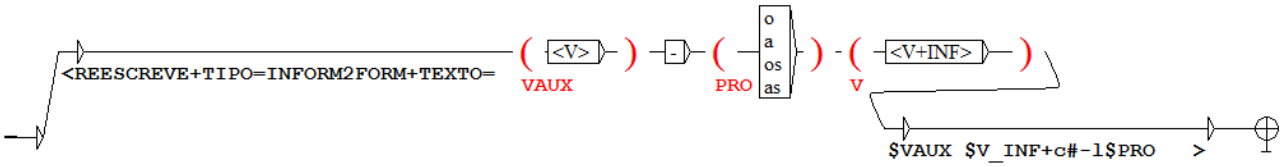

Figura 1: Gramática para normalizar linguagem informal em linguagem formal com o uso de clíticos.

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-

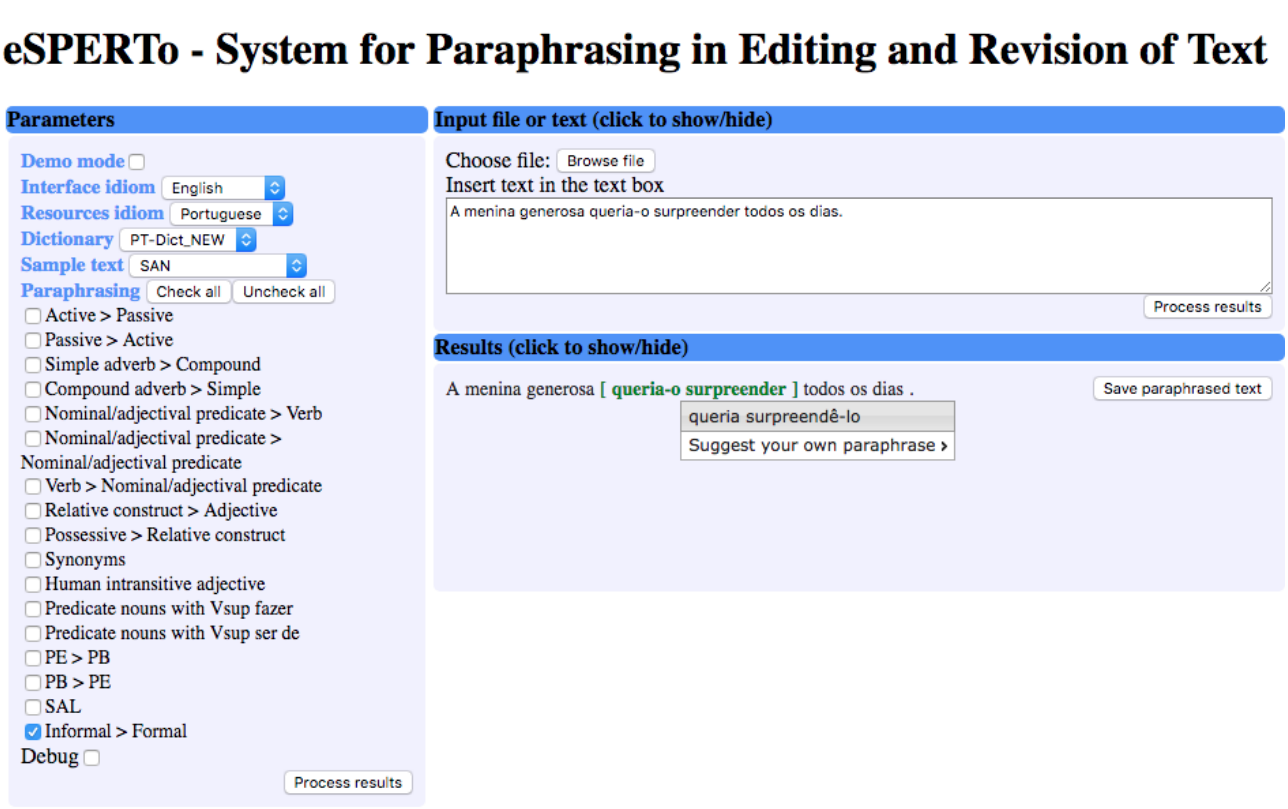

Figura 2: Conversão de um composto verbal informal com um pronome clítico num equivalente formal onde o clítico surge depois do verbo principal.

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.