Paraphrastic Variance between European and Brazilian Portuguese

Anabela Barreiro
INESC-ID, Rua Alves Redol 9
1000-029 Lisboa, Portugal
anabela.barreiro@inesc-id.pt

Cristina Mota
INESC-ID, Rua Alves Redol 9
1000-029 Lisboa, Portugal
cristina.mota@inesc-id.pt

Abstract

This paper presents a methodology to extract a paraphrase database for the European and Brazilian varieties of Portuguese, and discusses a set of paraphrastic categories of multiwords and phrasal units, such as the compounds *toda a gente* vs *todo o mundo* "everybody" or the gerundive constructions [*estar a* + V-Inf] vs [*ficar* + V-Ger] (e.g., *estive a observar* vs *fiquei observando* "I was observing"), which are extremely relevant to high quality paraphrasing. The variants were manually aligned in the e-PACT corpus, using the CLUE-Aligner tool. The methodology, inspired in the Logos Model, focuses on a semantico-syntactic analysis of each paraphrastic unit and constitutes a subset of the Gold-CLUE-Paraphrases.[1] The construction of a larger dataset of paraphrastic contrasts among the distinct varieties of the Portuguese language is indispensable for variety adaptation, i.e., for dealing with the cultural, linguistic and stylistic differences between them, making it possible to convert texts (semi-)automatically from one variety into another, a key function in paraphrasing systems. This topic represents an interesting new line of research with valuable applications in language learning, language generation, question-answering, summarization, and machine translation, among others. The paraphrastic units are the first resource of its kind for Portuguese to become available to the scientific community for research purposes.

1 Introduction

Paraphrases are linguistic devices that allow to recognize and generate equivalent forms of expressing the same content, either oral or written, i.e., of saying and writing the same thing/idea using different wording or syntactic structure. Paraphrases are essential in human communication, both in language production and understanding. They can occur at various levels: multiword or phrasal unit, phrase, expression, sentence, paragraph, full text, etc.. Given the scale and nature of paraphrases, paraphrase research has become an activity of growing importance in natural language processing, and a vital and strategic area for future language technology industries, ranging from text production, language learning, dialogue systems and machine translation applications, among others. The work presented here lies within the scope of ongoing research activities of the eSPERTo project[2], which aims to develop an automated paraphraser to assist writers and language learners in text production and revision. eSPERTo has the challenging objectives of guaranteeing thorough knowledge of the context, fluency of language, appropriate style and consistent terminology. Within these objectives, eSPERTo is designed to enable the adaption of a text within the different varieties of the Portuguese language.

In order to enable variety adaptation, we have analyzed the contrastive pairs of paraphrastic units aligned and collected from the corpus e-PACT (eSPERTo Paraphrase Aligned Corpus of EN-EP/BP Translations), a parallel corpus of aligned paraphrases (Barreiro and Mota, 2017). One of the motivations behind the creation of this corpus was to contrast the European (EP) and Brazilian (BP) varieties

[1] An approach based on word-level alignment clues is often referred to as the "clue alignment approach" (Tiedemann, 2003) (Tiedemann, 2011)). In our approach, CLUE is an acronym that stands for "**C**ross-**L**anguage **U**nit **E**licitation" that is based on manual alignments of multiwords and other phrasal units, which can be monolingual or bilingual.

[2] https://esperto.l2f.inesc-id.pt/esperto/esperto/demo.pl

Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 111–121
Santa Fe, New Mexico, USA, August 20, 2018.

of Portuguese by exploring monolingual alignments taking into account both similar and differing forms of expression between them. This approach allows finding vocabulary and expressions common to both varieties, but also linguistic constructions that constitute lexico-syntactic and stylistic differences between EP and BP. Breaking away from ad-hoc and random alignment practices, our methodology centers around the Logos Model (cf. (Scott, 2003), (Barreiro et al., 2011), (Scott, 2018)) and its semantico-syntactic approach, which results from over 30 years of experience in successful commercial machine translation.

The paraphrastic units collected, which are common between the two language varieties, are useful to increase eSPERTo's paraphrasing capabilities, whereas the paraphrastic variants, i.e., multiwords, phrases or expressions not used in one of the varieties, are useful for variety adaptation. Variety adaptation allows, for example, the necessary amending proposals to ensure that eSPERTo's user text, ways of expression or style can add clarity to the text and improve its readability in the other Portuguese variety. Adaptation will also attempt to reduce communication barriers among the Portuguese varieties, and eventually, contribute to an international variety of Portuguese (cf. (Santos, 2014) and (Santos, 2015)).

The alignments were performed with the support of the CLUE-Aligner tool (Barreiro et al., 2016), developed to facilitate the alignment of both paraphrasing and translation units in monolingual and in bitexts, including the alignment of discontinuous multiwords and phrasal units, such as the support verb constructions *fazer* [] *caminhadas por = dar* [] *passeios por* "taking [] walks through", or *ficar contente* "be happy". Within this line of research, we have developed a set of guidelines – CLUE4Paraphrasing Alignment Guidelines – that use information about the syntactic and semantic properties of phrases to align paraphrastic correspondences in a monolingual EP–BP sentence pair. Our alignment research focuses mainly on lexical and semantico-syntactic phenomena that can be, to a greater or lesser extent, challenging to a paraphrasing system. As the paraphrastic database grows, our aim is to create an automated alignment model with pre-defined elements and concepts that can be used for future applications involving monolingual or bilingual alignment tasks.

2 Related Work on Alignments

Paraphrasing systems can be trained using similar methods to those used in machine translation systems[3], i.e., they can be trained with **paraphrastic alignments**[4], which are representations of semantically-equivalent words, phrases, expressions or sentences within the same language or language variety, such as EP and BP. The paraphrastic alignment process consists of identifying, analyzing and registering corresponding phrasal equivalents within pairs of parallel sentences, where the source and the target sentences correspond to the same language.

Paraphrastic alignments extracted from parallel corpora may be either of high quality or of questionable quality depending on the quality of those corpora or the quality of the work performed during the alignment task, respectively. For Portuguese, there is a lack of freely available parallel corpora that can be used to train and test paraphrasing systems. Linguistic knowledge-based alignments extracted from good quality corpora can contribute to increased precision and, subsequently, improve the quality of generated paraphrases. In particular, alignments of paraphrastic units can be extremely useful to collect data and obtain an adequate dimension of the work to be executed prior to linguistic validation and integration of good quality data into real-world systems.

Our alignment task consisted of identifying, aligning, and collecting paraphrastic equivalences, i.e., multiwords and phrasal units or expressions that represented semantic correspondences in the aligned sentence pairs of the EP–BP parallel text. The outcome of our alignment task contained a set of individual paraphrastic alignments between meaningful sequences of words, i.e., linguistically-motivated pairs of paraphrastic units.[5] From an applicational perspective for Portuguese, no research has been done at a

[3] In machine translation, several works have been published on alignment annotation guidelines or other aspects of alignment research (cf. (Och and Ney, 2000), (Lambert et al., 2005), (Graça et al., 2008), or (Tiedemann, 2011), among others)

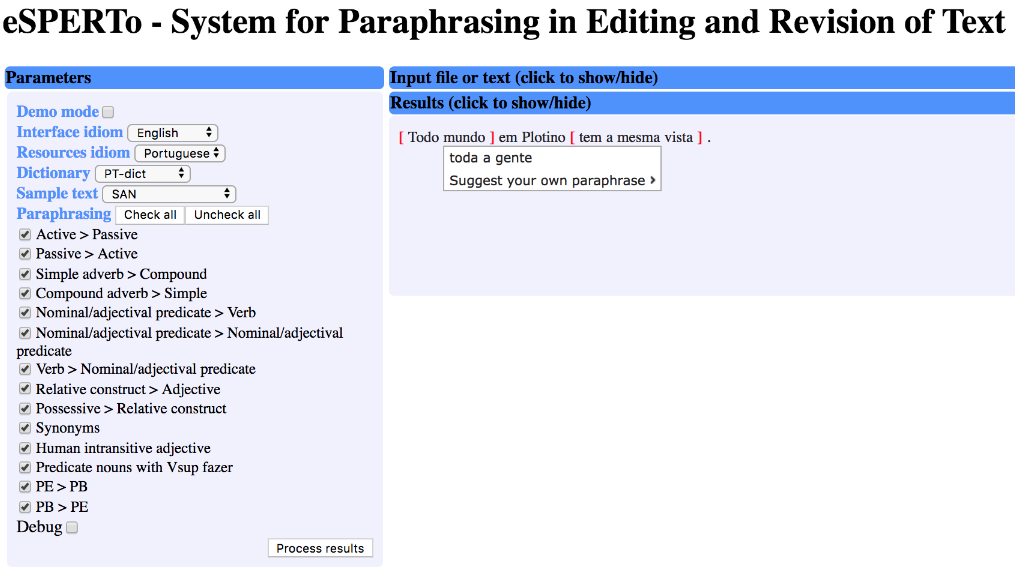
[4] In comparison to *translation alignments*, which are representations of semantically-equivalent words, phrases, expressions or sentences within the source and target sentences of a bilingual or multilingual parallel corpus (Brown et al., 1990).

[5] In statistics, a sequence of more than one n-gram is commonly called "phrase". Our alignments do not contain statistical phrases, but linguistic phrases or other linguistic units. Alignments based on random n-grams or statistical phrases do not

level beyond the lexicon. Early work on EP–BP standard and technical language lexical distinctions has been compiled in a contrastive lexicon (Barreiro et al., 1996) that led to INESC's Lusolex and Brasilex dictionaries (Wittmann et al., 2000), but no alignment methods have been used. Despite meagre initial resources, manually annotated paraphrastic alignments represent an important step in the development of paraphrasing systems.

3 The eSPERTo Project

Variety adaptation is an important feature of the eSPERTo project, whose main focus is the development of a paraphrasing system with capacity to produce semantically equivalent sentences and ways of expression, also when these are contrasting, as in the case of varieties of the same language. Figure 1 illustrates the usefulness of paraphrases in eSPERTo's variety adaptation capability, where for a sentence written in EP, the system offers suggestions to paraphrase and rewrite it in BP (and vice-versa). For example, for the BP sentence *Todo mundo em Plotino tem a mesma vista* "Everybody in Plotinus has the same view", eSPERTo presents *toda a gente* as the EP suggestion for the BP phrase *todo mundo* and the EP suggestion *tem a mesma vista* for the BP phrase *tem vista igual*. This adaptation is extremely useful when the user wants to reach an audience that speaks the variety that he/she is less familiar with.



eSPERTo - System for Paraphrasing in Editing and Revision of Text

Parameters

Demo mode □
Interface idiom: English
Resources idiom: Portuguese
Dictionary: PT-dict
Sample text: SAN
Paraphrasing

☑ Active > Passive
☑ Passive > Active
☑ Simple adverb > Compound
☑ Compound adverb > Simple
☑ Nominal/adjectival predicate > Verb
☑ Nominal/adjectival predicate > Nominal/adjectival predicate
☑ Verb > Nominal/adjectival predicate
☑ Relative construct > Adjective
☑ Possessive > Relative construct
☑ Synonyms
☑ Human intransitive adjective
☑ Predicate nouns with Vsup fazer
☑ PE > PB
☑ PB > PE
Debug □

Process results
Input file or text (click to show/hide)
Results (click to show/hide)

[Todo mundo] em Plotino [tem a mesma vista] .

toda a gente
Suggest your own paraphrase ›

Figure 1: EP–BP paraphrastic variants *toda a gente* | *todo mundo* and *tem a mesma vista* | *tem vista igual*

eSPERTo uses semantico-syntactic knowledge to identify multiwords and other phrasal units, and applies local grammars to transform them into semantically equivalent phrases, expressions, or sentences. The quantity and quality of the resources have been increasing considerably with the integration of tables developed within the lexicon-grammar theoretical and methodological framework (cf. (Gross, 1984) and (Gross, 1987)), based on the transformational operator grammar (cf. (Harris, 1952), (Harris, 1965), (Harris, 1991), among others). Lexicon-grammar tables contain distributional and transformational properties of nominal predicates that can be used in paraphrasing tasks with successful results. Several lexicon-grammar research works have been describing these predicates in great detail, establishing relations between different types of predicate, and defining properties in tables that can be adapted and converted into dictionary entries, becoming a useful resource for paraphrasing. Predicates are not necessarily verbal, they can be nominal too, and they are often used interchangeably without any significant difference in meaning. There are nominal predicates, both nouns and adjectives, which, like verbs,

have a linguistic motivation or contrastive analysis lying behind them. Even though they represent an efficient intermediate representation developed for engineering purposes in natural language processing and machine translation systems, they present shortcomings from a linguistic point-of-view. In "n-grams in search of theories", (Maia et al., 2008) raised the question of the need to create linguistically more robust n-gram tools, which imply a supporting theoretical or practical framework for the research on word alignment.

have argumental selection properties. For example, there are adjectives that require complements (e.g., *ele está desejoso de ir à praia* "he is eager to go to the beach"), being classified as transitive adjectives, and there are adjectives that do not require any complement (e.g., *ele está doente* "he is sick"), classified as intransitive adjectives. In these cases, it is not the verb, which in both sentences is the same auxiliary, *está* "is", that imposes these argument restrictions, it is the adjective instead. The same can be said with regard to predicate nouns.

In our research work, three lexicon-grammar tables formalized for EP have recently been added to expand eSPERTo's paraphrastic capabilities: (i) the lexicon-grammar of human intransitive adjectives (Mota et al., 2015), (ii) the lexicon-grammar of predicate nouns co-occurring with the support verb *fazer* "do" or "make" (Mota et al., 2017), and (iii) the lexicon-grammar of predicate nouns which co-occur in constructions with the support verb *ser de* "be of" (Mota et al., 2018). These resources allow the generation of paraphrases such as *de origem portuguesa* "of Portuguese origin/roots" = *portugueses* "Portuguese" = *de Portugal* "from Portugal"; *fez uma classificação de NP* "made a classification of NP" = *classificou NP* "classified NP"; *é de uma certa cortesia* "is of a certain courtesy" = *é cortês* "is courteous". So far, we have not managed to integrate any lexicon-grammar tables for BP, but we would only need to formalize those entries which are exclusive or differ from the ones in EP.

Even though eSPERTo has been explored in a question-answering system and in a summarization tool (Mota et al., 2016), the lexicon-grammar integrated resources have not been tested in these applications. We envisage to test the new paraphrastic resources in an e-learning environment to assist Portuguese language learners with the editing and revision of texts. But, precise paraphrases can also be helpful in professional translation, editing, and proofreading, among other tasks.

4 Description of the Paraphrastic Alignment Task

Our paraphrastic alignment task was facilitated by the use of the CLUE-Aligner, an alignment tool that permits the alignment and storage of both continuous and discontinuous multiwords and other phrasal units to be used in paraphrasing (and also in translation), i.e., in monolingual or bilingual parallel sentences. Based on the CLUE4Paraphrasing Alignment Guidelines[6], we built a gold collection of pairs of EP–BP paraphrastic variants for the e-PACT corpus. The CLUE4Paraphrasing Alignment Guidelines summarize the most important recommendations and decisions for the alignment of multiwords and phrasal units found in monolingual parallel sentences corresponding to the EP and BP translations of two books by David Lodge, *Therapy* and *Changing Places*.[7] The initial e-PACT corpus contained 30% of the two novels extracted from the COMPARA English and Portuguese bidirectional parallel corpus. To create the initial corpus (Barreiro and Mota, 2017), we extracted the first 3 parallel sentences of each group of 10 parallel sentences, the EBDL1 batch contains 489 sentence alignments and the EBDL3 batch contains 313 sentence alignments, in a total of 802 parallel sentences. For the current work, we enlarged the original e-PACT with 10% more of the total number of sentences for the 2 novels, which correspond to the first 4 parallel sentences in each group of 10. This 10% increase corresponds to 163 sentence alignments in a first batch and 312 sentence alignments in a second batch. Therefore, so far, we manually annotated 40% of the total number of aligned sentences for both novels, in total 1,277 parallel sentences. From the enlarged e-PACT corpus, we have collected a few thousands of paraphrastic alignments that still need to be revised by a Portuguese and a Brazilian linguist before making them publicly available. From this collection, not all the paraphrastic alignments correspond to contrasts between the EP and BP varieties of Portuguese. Our goal in exploring monolingual alignments of two varieties of the same language was not only to capture differing forms of expression between these varieties, but also take into account paraphrases that can apply to one variety or the other. These variety-free/independent paraphrases can contribute to the development of the eSPERTo paraphrase acquisition system.

[6] The CLUE4Paraphrasing Alignment Guidelines are a set of CLUE Alignment Guidelines.

[7] We have used EP and BP translations of the same English novels as alignment data, because this is a popular and straightforward approach to gather parallel paraphrastic data. However, one drawback of this kind of corpora is that there may be a syntactic and lexical bias carrying over from the English original. It is possible that non-translation corpora may be less biased, but more difficult to find or prepare.

5 Examples of EP–BP Contrastive Paraphrasing Phenomena

In this Section, we will illustrate several types of EP–BP paraphrases, which constitute real examples from the e-PACT corpus. We provide the English source sentence for each EP–BP paraphrase case. Lack of space precludes a detailed description of most of the paraphrasing phenomena found in the corpus. We have selected a few examples of paraphrastic alignments. All multiwords, phrases and expressions, including discontinuous ones, have been aligned as illustrated in Figure 2. Due to space limitation we present only this illustrative image, which represents the paraphrastic alignment pair highlighted in example (1). The alignment refers to the EP–BP variety contrast between the discontinuous support verb construction *subiram* [] *em espiral*, literally "climbed [] in a spiral" in EP and the BP prepositional verb *espiralando* "spiraling up". In EP the predicate noun *espiral* "spiral" is placed apart from the support verb *subir* "climb". The direct object noun phrase insertion in the support verb construction, *o tronco* "the branch", aligns independently (not highlighted in this image).

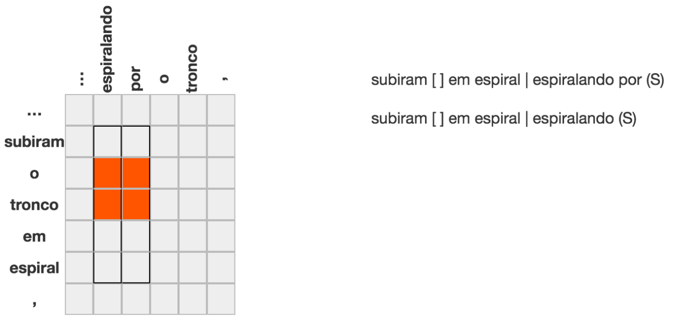

...
espiralando
por
o
tronco
,
...
subiram
o
tronco
em
espiral
,
subiram [ ] em espiral | espiralando por (S)
subiram [ ] em espiral | espiralando (S)

Figure 2: Paraphrastic S-alignment $_{EP}$ - *subiram* [ ] *em espiral* | $_{BP}$ - *espiralando* (*por*)

5.1 Verbal Constructions

5.1.1 Verbs and Support Verb Constructions

Verbs and support verb constructions are frequent in commutative conditions. Often paraphrastic variance between EP and BP results from the use of a support verb construction or of a verb, in one or another direction. There are cases where their respective use is simply related to an arbitrary decision by the translator. In other cases, the frontier between stylistic choice and variety adaptation is not straightforward. In example (1), EP and BP adopt different surface structures (i.e., syntax); EP uses a support verb construction, while BP uses a verb. These could be simply considered stylistic variants resulting from the fact that the BP translator translated less conventionally by using a new verb instead of the more conventional support verb construction, but in a sense there seems to be a more evident translation permissibly that is allowed or fostered in BP as far as new vocabulary is concerned.[8] A clear stylistic choice was the translation of the support verb into the past tense, *subiram*, by the EP translator who arbitrarily or voluntarily did not maintain the gerundive form used in the English source and in the BP translation.

(1) $_{EN}$ - *I watched two playing tag [...] just outside my study window:* ***spiralling up a trunk****...*
$_{EP}$ - *Estive a observar da janela do meu escritório dois esquilos a brincarem à apanhada [...]:* ***subiram o tronco em espiral****...*
$_{BP}$ - *Fiquei observando os dois esquilos que brincavam de pegapega [...] em frente à janela do meu estúdio:* ***espiralando pelo tronco****...*

5.1.2 EP [*estar a* + V–Inf] versus BP [*ficar* + V–Ger] Constructions

The use of progressive constructions when aligning EP–BP paraphrases is extremely frequent and there are many interesting cases that are worth analyzing. However, due to space limitations we focus on: (i)

[8] While this verb is rather far-fetched even for a Brazilian speaker, it is known that BP speakers are in general less conservative as far as the formation of new words is concerned. It appears less likely that the verb *espiralar* would be employed by a EP native speaker or translator. But, it is also possible that the English gerund *spiralling* induced a corresponding Portuguese form *espiralando* from a translator that would not otherwise have used it, because it is a lexeme less likely to cross someone's mind.

the gerundive infinitive, formed with the auxiliary verb *estar* "be" (or *ficar* "remain", or *ir* "go") plus the preposition *a* "at" plus the infinitive form of the main verb: [*estar a* + V–Inf] (e.g., *estar* + *a* + *trabalhar* = "to be + working"), used in written EP, and (ii) the gerundive construction made up of the auxiliary verb *estar* (or *ficar*, or *ir*) plus the present participle (gerundive) of the main verb, which ends with the suffix -*ndo* "-ing": [*ficar* + V–Ger] (e.g., *ficar/estar* + *trabalhando* = "to be + working), used in BP. Example (2) illustrates the contrast between the EP gerundive infinitive, which is formed with the verb *estar* in the past imperfect tense (*estive*) "(I) was", followed by the preposition *a*, and by the main verb in the infinitive form, *observar* "observe", and the BP gerundive form, which is formed with the auxiliary verb *ficar* in the preterit tense, *fiquei* ("(I) remained/stayed"), followed by the gerundive form of the main verb, *observando* "observing".

(2) $_{EN}$ - ***I watched** two playing tag [...] just outside my study window:*
$_{EP}$ - ***Estive a observar** da janela do meu escritório dois esquilos a brincarem à apanhada [...]:*
$_{BP}$ - ***Fiquei observando** os dois esquilos que brincavam de pegapega [...] em frente à janela do meu estúdio:*

5.2 Word Order – Placement of Clitic Pronouns: V–Pro versus Pro V

The placement of the clitic pronoun is different in EP and BP. In EP, the normal position for the clitic is after the verb and connected to it by an hyphen[9] ([(Prep)V]–Pro), while in BP, the normal position for the clitic is before the verb with no attaching hyphen (Pro [(Prep)V]). As illustrated in example (3), the prepositional verb[10] *puseram-me em*, literally "(they) put me in", in the EP sentence is a paraphrastic variant of the prepositional verb *me mandaram para* "(they) sent me to" in the BP sentence. While there is a stylistic difference with regards to the translator's choice of the prepositional verb, the word order difference is a clear case of EP–BP paraphrastic variance.

(3) $_{EN}$ - *My Mum and Dad **sent me to** Sunday school when I was a nipper...*
$_{EP}$ - *Os meus pais **puseram-me** na [**em** + a] catequese quando ainda era pequeno...*
$_{BP}$ - *Minha mãe e meu pai **me mandaram para** a Escola Dominical quando eu era pequeno...*

5.3 Lexical versus non-Lexical Realization in Nominal Constructions

Within noun phrases, it is common to find paraphrastic alignments where one element of the pair of paraphrases contains a lexically-realized determiner or a pronoun and the other element does not contain them. Section 5.3.1 discusses the alignment of phrases containing determiners with phrases containing what is known as *zero determiners*. Section 5.3.2 discusses the EP–BP variance cases involving subject pronouns, or lack of them, which is normally designated as *subject pronoun drop* or simply *pro-drop*.

5.3.1 Determiners and Zero Determiners

The presence of zero articles is common in BP, and less frequent in EP. Aligning a zero determiner with a lexically realized determiner implies association of the determiner to the noun. Determiners are aligned together with the noun (single or compound) when they do not appear in one of the varieties (mostly BP) of an alignment pair. When determiners appear in both variants of the alignment pair, they are also aligned individually. For example, the noun phrase *o Nizar* with the definite article *o* and the named entity *Nizar* in EP aligns with the single noun "Nizar" (no determiner) in BP, i.e., the alignment of $_{EP}$ - [DET N] | $_{BP}$ - [Ø-DET N]. The alignment of phrases with determiners with phrases with no determiners implies that the lexically realized determiner is associated to the phrase. In example (4), the noun phrase containing the definite article *os* in the noun phrase *os meus grupos* "my groups" in the EP sentence aligns with the noun phrase without a determiner *meus grupos* in the BP sentence.

(4) $_{EN}$ - *I gather from Busby that you'll probably be taking over **my** tutorial **groups**.*
$_{EP}$ - *Soube pelo Busby que vai ficar com **os meus grupos**.*
$_{BP}$ - *Pelo que Busby me contou, o senhor vai ser o orientador de **meus grupos** de estudos.*

[9] Although in both EP and BP there are differences between written and spoken language (some correspond to regional differences), and also between verb tenses or antecedents in the sentence, we will not enter into any of these details here.

[10] Prepositional verbs are preposition-governing transitive verbs, where the preposition is at the right-hand side of the verb.

EP	BP	EN
braço de um gira-discos	*agulha de um toca-discos*	*the stylus arm of a* [] *record deck*
comboio	*trem*	*train*
revisor	*cobrador*	*ticket-collector*
serviço de mesas	*serviço de garçom*	*table-service*
fio do berbequim	*fio da furadeira*	*lead on [] Black and Decker*
maçã-de-adão	*pomo-de-adão*	*Adam's apple*
desporto	*esporte*	*sport*
fato	*terno*	*suit*
blusão de cabedal	*jaqueta de couro*	*leather jacket*

Table 1: EP–BP lexical contrasts

5.3.2 Subject Pronoun Drop

The contrast of overt pronouns with omitted or null pronouns is a recurring phenomenon in the alignment task (e.g., ***Ø*** *sugeri* = ***Eu*** *disse* "I said"). If a personal pronoun is overt in one of the varieties and omitted in the other variety, alignment should be made on a one-by-two basis. In example (5), the subject pronoun *Ele* "He" together with the verb *era* "was" of the adjectival support verb construction *era um desportista* "(he) was a sportsman" in BP aligns with its equivalent without the pronoun *Ele* in BP.

(5) $_{EN}$ - ***He*** *was in fact a keen sportsman*
$_{EP}$ - ***Ø*** *Era de facto um desportista hábil*
$_{BP}$ - ***Ele*** *era de fato um esportista aplicado*

5.4 Forms of Address – 2nd versus 3rd Person

EP and BP have different forms of addressing people and different forms of courtesy. In EP, the pronoun *tu* "you" is used as an informal way of addressing friends and family in casual situations. In formal situations, it is used the pronoun *você* (sometimes omitted) with the verb conjugated in the third person singular. In BP, the most common form of address is *você* in both formal and informal contexts.[11] For example, *tens a certeza* is used in EP, while *tem certeza* "you're sure" is used in BP. Similarly, *não te importas* is used in EP and *não se importa* "you don't mind" is used in BP. The form of address is a very frequent source of EP–BP paraphrastic variance.

(6) $_{EN}$ - ***You're sure*** [Ø] ***you don't mind****?*
$_{EP}$ - ***Tens a certeza*** *de que* ***não te importas****?*
$_{BP}$ - ***Tem certeza*** *de que* ***não se importa****?*

6 Variety Differences

The most important issue to be considered, at this particular point in our research, is to distinguish between those paraphrastic alignments that represent stylistic differences but that are natural and fluent multiwords, expressions or phrases in both the EP and BP varieties, and those paraphrastic alignments that represent contrastive variance between EP and BP and they cannot be used in commutative conditions in both varieties, i.e., they are exclusively used either in EP or in BP. In the list of contrasts, we have also registered lexical contrasts, illustrated in Table 1, even though they are not the focus of our discussion. Table 2 illustrates contrasts of a syntactic nature, i.e., multiwords expressions or phrases that are used only in EP or only in BP. Future work should focus on the categorization of each one of these syntactic phenomena, and the creation of grammars that can use these phenomena in more generalized contexts. This research needs to be deepened and sustained with validation of the contrasts by expert linguists on both varieties. Table 4 illustrates stylistic contrasts that correspond to valid paraphrases for both varieties of Portuguese.

Many of the EP–BP contrasts that we have collected have insertions, i.e., elements that are external to the multiword or phrasal unit, either in the English source or in any of the Portuguese varieties, such

[11] However, in BP, *você* can be combined both with the second and third person singular personal pronouns to distinguish between a more or less familiar person.

EP	BP	EN
viu-me *vir a correr por* [NP]	***me viu*** *correndo por* [NP]	*saw me running down* [NP]
estaria a querer dizer	*estaria insinuando*	*What was he implying*
a brincarem ***à apanhada***	*brincavam* ***de pegapega***	*playing tag*
A Alexandra ***perguntou-me***	*Alexandra* ***me perguntou***	*Alexandra asked me*
há pessoas *que*	***tem gente*** *que*	*people*
pessoas *que só querem comprar selos*	***gente*** *que só quer comprar selo*	*people who just want to buy stamps*
se ***hei-de*** *acender* [NP]	*se* ***devo*** *acender* [NP]	*whether I should turn on* [NP]
Não gosto que	*não gosto* ***do jeito*** *que*	*I don't like* ***the way*** *that*
Talvez seja altura de ***acabar***	*Talvez devêssemos* ***dar um xeque-mate*** *em*	*Perhaps we should call it a day*
campo de girassóis ***de pernas para o ar***	*campo de girassóis* ***de ponta-cabeça***	*inverted field of sunflowers*
Dá ***que*** *pensar*	*Dá* ***o que*** *pensar*	*Makes you think*
em que era ***mergulhado*** *em*	*sobre ser* ***jogado*** *em*	*being dunked in*
queres *ficar sozinho com* [NP]	***você quer*** *ficar sozinho com* [NP]	*you want to be alone with* [NP]

Table 2: Examples of EP–BP contrasts of a syntactic nature

EP	BP	EN
Estamos a falar de	*Estamos falando de*	*We're talking* [NP] *here*
nunca mais me obrigaram a ir a [NP]	*não me fizeram ir mais a* [NP]	*didn't make me go to* [NP] *any more*
vou de [N(CO-clothes)]	*estou usando / estou com* [N(CO-clothes)]	*I'm wearing / I'm in* [N(CO-clothes)]
Já contei a [NP]	*Já botei* [NP] *a par de*	*I've put* [NP] *in the picture about*
ir ocupar [PRO-Poss] *lugar em*	*tomar* [PRO-Poss] *lugar em*	*go take* [PRO-Poss] *place in*
há alturas em que	*Tem hora* [ADV] *que*	*There are times*
as mulheres [ADV] *fazem coisas estranhas*	*mulher faz coisa estranha*	*women do funny things*

Table 3: Examples of EP–BP contrasts with insertions or SAL categories

as *We're talking [sitcom] here*, *didn't make me go to [Sunday School] any more*, *as mulheres [às vezes] fazem coisas estranhas*, or *Já botei [Hal] a par do problema*. According to the methodology described in (Barreiro and Batista, 2016) for translation, we have extracted all these insertions from the paraphrastic alignments and subsequently have assigned generic categories to these insertions, such as [NP] for noun phrase, [ADV] for adverb, [PRO-Poss] for possessive pronoun, and so on and so forth. In the derived generic grammars, extracted alignments are generalized by replacing dependents words with constituent variables such as NP and ADV, etc., or SAL categories. The reason for this, is that we want grammars to apply independently of the word (noun, adverb, pronoun, etc.) inserted. So, for example, instead of the proper name *Hal*, the grammar would still be able to transform the expression no matter which proper name would appear as an insertion. In other cases, we have defined semantico-syntactic (SAL) categories so that grammars apply to a certain group of words (See (Scott, 2003; Barreiro et al., 2011; Scott, 2018) for a description of SAL). For example, the English expressions *I'm wearing a suit* and *I'm in jeans and leather jacket* were found in the same sentence illustrated in example (7) in the e-PACT corpus with the EP translations *vou de fato* and *vou de jeans e blusão de cabedal*, and with the BP translations *estou usando um terno* and *estou com jeans e jaqueta de couro*. The use of the SAL category [N(CO-clothes)] ([**CO**ncrete noun + **clothes** that one can wear/dress]) for the noun *suit* and the coordinated nouns *jeans and leather jacket* and their corresponding translations in both EP and BP allows the grammar to apply the paraphrases with any noun or coordinated nouns that are classified with the same SAL category (e.g., *pants*, *dress*, *sweatshirt*, etc.).

(7) $_{EN}$ - ***I'm wearing*** *a suit myself today [...], but sometimes, when* ***I'm in*** *jeans and leather jacket*
$_{EP}$ - *Hoje também* ***vou de*** *fato [...] mas, às vezes, quando* ***vou de*** *jeans e blusão de cabedal*
$_{BP}$ - *Hoje* ***estou usando*** *um terno [...], mas às vezes,* ***estou com*** *jeans e jaqueta de couro*

7 Conclusions and Future Work

This paper describes the methodology to build a new linguistic resource of manual paraphrastic alignments representing multiwords and phrasal units in EP and BP collected from the e-PACT corpus. The

EP	BP	EN
todos os problemas ***que já tenho***	*todos os* ***meus*** *problemas*	*all my other problems*
era mais normal que *tivesse ido para*	***normalmente*** *teria ido para*	*I would normally have gone into*
Vou ***fazer*** *uma pequena* ***cirurgia***	*Vou* ***ser operado****. Uma operação simples*	*I'm having a minor operation*
ficar ***naquelas*** *filas intermináveis*	*ficar* ***numa dessas*** *longas filas*	*stand in one of those long [] queues*
fazerem o seu trabalho ***o melhor possível***	*produzirem* ***o melhor que puderem***	*make it as good as it possibly can be*
discotecas ***duvidosas***	*discotecas* ***de reputação duvidosa***	*dubious discos*
perto *de*	***nas imediações*** *de*	*near*
não me cruzei com *ninguém*	***não vi*** *ninguém*	*I haven't seen anybody*
coloquei *esta questão*	***levantei*** *essa questão*	*I raised this question*
tem disponibilidade *para*	***fica livre*** *para*	*free to*
fazer *teatro*	***atuar em peças de*** *teatro*	*do live theatre*
não me ***importo*** *de*	*eu não me* ***importaria*** *de*	*I wouldn't mind*
Parti do princípio de *que*	***Imaginei*** *que*	*I assumed*
que nem *um doido*	***como*** *um louco*	*like a drain*
a maior parte *deles*	***a maioria*** *deles*	*most of them*

Table 4: Examples of stylist variants = paraphrases both possible in EP and BP

paraphrastic alignment can provide a sort of contrastive dictionary function after validation. We have illustrated a few cases of paraphrasing phenomena, but many more could be brought for reflection. Our main goal was to show how short paraphrastic variants can contribute to the development of a paraphraser that handles variety adaptation. This is a fertile research field that still needs to mature in order to bear fruit to enrich technological applications for language learning, writing and editing, among others.

A first observation to be made concerns the validation of the words, multiwords, phrases, expressions, structures and sentences of each variety that are taken into consideration in the paraphrastic alignments. Some ways of expression may vary according to the translator, the translator's experience or professional performance and be less related to the variety itself. For example, the use of the word *estúdio* with the meaning of *office* is questionable (context-specific knowledge is important). Normally, it refers only to an artist's work place, not a regular office. While the kind of corpora used may be a rich source of paraphrases, not all paraphrases are reliable, and some of them are not indicative of language variance.

We have targeted several morpho-syntactic alignment problems that have not been consistently considered up to now, such as the alignment of articles together with the nouns with zero articles, a solution for a significant number of gender and number agreement problems between an article, and a noun, or the alignment of the preposition with a noun in noun adjunct cases. Due to the extent of the work at hand, a large amount of paraphrastic phenomena was left undiscussed. A detailed analysis of these phenomena is important for the improvement of alignment techniques and for the enhancement of the quality of paraphrasing. One of the phenomenon that we are currently revisiting is the alignment of multiwords when there are contracted forms involved (Barreiro and Batista, 2018). Another one is the alignment of verbal constructions involving clitic pronouns (Rebelo and Barreiro, 2018 forthcoming). As we move along the development process of a manually aligned dataset and definition of a typology of linguistic phenomena, we wish to attempt an automated alignment tool.

Finally, it is worth noting that the computational tools available for alignment also present shortcomings and limitations. The process of collecting paraphrastic alignments is a far-reaching work that is far from complete. In future work, we envisage to create grammars from many of these contrasts that will use semantico-syntactic knowledge and apply to a larger number of cases whenever that is possible.

Acknowledgements

This work was supported by Fundação para a Ciência e a Tecnologia under project eSPERTo, with references EXPL/MHC-LIN/2260/2013 and UID/CEC/50021/2013. Anabela Barreiro was also funded by FCT through post-doctoral grant SFRH/BPD/91446 /2012. We would like to thank Ida Rebelo-Arnold for her interest in our paper as well as for her helpful comments and suggestions that contributed to improving the quality of our study.

References

Anabela Barreiro and Fernando Batista. 2016. Machine Translation of Non-Contiguous Multiword Units. In Wolfgang Maier, Sandra Kübler, and Constantin Orasan, editors, *Proceedings of the Workshop on Discontinuous Structures in Natural Language Processing*, DiscoNLP 2016, pages 22–30, San Diego, California, June. Association for Computational Linguistics (ACL).

Anabela Barreiro and Fernando Batista. 2018. Contractions: to align or not to align, that is the question. In Anabela Barreiro, Kristina Kocijan, Peter Machonis, and Max Silberztein, editors, *Proceedings of the COLING-Workshop on Linguistic Resources for NLP, New Mexico, USA*,

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

Anabela Barreiro, Luzia Wittmann, and Maria Pereira. 1996. Lexical differences between European and Brazilian Portuguese. *INESC Journal of Research and Development*, 5(2):75–101.

Anabela Barreiro, Bernard Scott, Walter Kasper, and Bernd Kiefer. 2011. OpenLogos Rule-Based Machine Translation: Philosophy, Model, Resources and Customization. *Machine Translation*, 25(2):107–126.

Anabela Barreiro, Francisco Raposo, and Tiago Luís. 2016. CLUE-Aligner: An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference*, LREC 2016, pages 7–13. European Language Resources Association.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

João Graça, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro. 2008. Building a Golden Collection of Parallel Multi-Language Word Alignment. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the 6th International Conference on Language Resources and Evaluation*, LREC 2008, pages 986–993, Marrakech, Morocco, May. European Language Resources Association.

Maurice Gross. 1984. Lexicon-grammar and the syntactic analysis of French. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, COLING 1984, pages 275–282, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maurice Gross. 1987. The use of finite automata in the lexical representation of natural language. In Maurice Gross and Dominique Perrin, editors, *Electronic Dictionaries and Automata in Computational Linguistics*, volume 377 of *Lecture Notes in Computer Science*, pages 34–50. Springer.

Zellig Sabettai Harris. 1952. Discourse analysis. *Language*, 1(28):1–30.

Zellig Sabettai Harris. 1965. Transformational Theory. *Language*, 41(3):363–401.

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

Patrik Lambert, Adrià De Gispert, Rafael Banchs, and José B. Mariño. 2005. Guidelines for Word Alignment Evaluation and Manual Alignment. *Language Resources and Evaluation*, 39(4):267–285.

Belinda Maia, Rui Sousa Silva, Anabela Barreiro, and Cecília Fróis. 2008. N-grams in search of theories. In Barbara Lewandowska-Tomaszczyk, editor, *Corpus Linguistics, Computer Tools, and Applications - State of the Art*, volume 17, pages 71–84. Peter Lang.

Cristina Mota, Paula Carvalho, Francisco Raposo, and Anabela Barreiro. 2015. Generating Paraphrases of Human Intransitive Adjective Constructions with Port4NooJ. In Tatsiana Okrut, Yuras Hetsevich, Max Silberztein, and Hanna Stanislavenka, editors, *Automatic Processing of Natural Language Electronic Texts with NooJ - Selected Papers of the 9th International Conference*, Communications in Computer and Information Science, pages 107–122, Cham. Springer International Publishing.

Cristina Mota, Anabela Barreiro, Francisco Raposo, Ricardo Ribeiro, Sérgio Curto, and Luísa Coheur. 2016. eSPERTo's Paraphrastic Knowledge Applied to Question-Answering and Summarization. In Linda Barone, Mario Monteleone, and Max Silberztein, editors, *Automatic Processing of Natural Language Electronic Texts with NooJ - Selected Papers of the 10th International NooJ Conference*, Communications in Computer and Information Science, pages 208–220, Cham. Springer International Publishing.

Cristina Mota, Lucília Chacoto, and Anabela Barreiro. 2017. Integrating the lexicon-grammar of predicate nouns with support verb fazer into port4nooj. In Samir Mbarki, Mohammed Mourchid, and Max Silberztein, editors, *Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications - Selected Papers of the 11th International Conference*, Communications in Computer and Information Science, pages 29–39, Cham. Springer International Publishing.

Cristina Mota, Jorge Baptista, and Anabela Barreiro. 2018. The Lexicon-Grammar of Portuguese Predicate Nouns with *ser de* in Port4NooJ. In Anabela Barreiro, Kristina Kocijan, Peter Machonis, and Max Silberztein, editors, *Formalising Natural Languages With Nooj and Its Natural Language Processing Applications - Selected Papers of the 11th International Conference*, Communications in Computer and Information Science, pages –, Cham. Springer International Publishing.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL 2000, pages 440–447, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ida Rebelo and Anabela Barreiro. 2018 (forthcoming). EP–BP Paraphrastic Alignments of Verbal Constructions Involving the Clitic Pronoun *lhe*. In *International Conference on Computational Processing of Portuguese*, PROPOR 2018. Springer.

Diana Santos. 2014. Como estudar variantes do português e, ao mesmo tempo, construir um português internacional? Presentation at Contact, Variation and Change: corpora development and analysis of Iberoromance language varieties workhop.

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

Bernard Scott. 2003. The logos model: An historical perspective. *Machine Translation*, 18(1):1–72, March.

Bernard Scott. 2018. *Translation, Brains and the Computer: A Neurolinguistic Solution to Ambiguity and Complexity in Machine Translation*. Machine Translation: Technologies and Applications. Springer International Publishing.

Jörg Tiedemann. 2003. Combining Clues for Word Alignment. In Ann Copestake and Jan Hajic, editors, *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2003, pages 339–346, Budapest, Hungary.

Jörg Tiedemann. 2011. *Bitext Alignment*. Synthesis digital library of engineering and computer science. Morgan & Claypool.

Luzia Wittmann, Ricardo Ribeiro, Tânia Pêgo, and Fernando Batista. 2000. Some Language Resources and Tools for Computational Processing of Portuguese at INESC . In Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis, and Gregory Stainhauer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation*, LREC 2000, pages 347–350, Athens – Greece.