10.17771/PUCRio.TradRev.30591

e-PACT: eSPERTo Paraphrase Aligned Corpus of EN-EP/BP Translations

Anabela Barreiro and Cristina Mota*

1. Introduction

Paraphrasing is an indispensable practice in translation. Translation can even be considered as a paraphrasing process, where the paraphrasing of a sentence or part of a sentence in a source language takes place in a target language. In translation, paraphrases can be more or less literal, depending on the translator's choice and the characteristics of the source text, domain, or genre. In this sense, translation is the richest source of paraphrases, where one can find different ways of expressing the same concept or idea. The translation of a text by the same author may contain several paraphrases for the same source expression. The translation of a text by different authors may contain even more paraphrases, because different translators use different linguistic strategies and stylistic preferences.

Our goal in this paper is to present a sampled parallel corpus of aligned paraphrases, e-PACT (acronym for **e**SPERTo **P**araphrase **A**ligned **C**orpus of EN-EP/BP **T**ranslations) briefly introduced in Barreiro & Batista (2016). The e-PACT corpus results from aligning the translations of two English books by the same author, David Lodge: *Therapy* and *Changing Places*. The goal in building the e-PACT corpus was twofold: (1) contribute to the development of the eSPERTo paraphrase acquisition system, and (2) contrast the European and Brazilian variants of Portuguese by exploring monolingual alignments, taking into account both similar and differing forms of expression. Contrasting European and Brazilian Portuguese allows to find the differences between them, but mostly the common vocabulary

[*] Anabela Barreiro and Cristina Mota are post-doctoral researchers at INESC-ID Lisboa.

Submetido em 02/3/2017 Aceito em 12/5/2017

10.17771/PUCRio.TradRev.30591

and ways of expression that can be used to minimize barriers between the variants and reach for an international variant of Portuguese, as proposed in Santos (2014) and Santos (2015).

The object of our research draws on literary texts because, being notoriously more idiosyncratic, literary language is richer in paraphrasing, i.e., in using different wording to express identical or similar ideas. Literary language is often figurative using devices like metaphor and other linguistic strategies to avoid repetition. Therefore, two literary translators are less likely to translate a particular sentence in a similar way than two technical translators, who use terminology consistently and do not have much scope for linguistic creativity. In this context, a wider variety of pairs of paraphrastic units can be achieved from this domain and be of good utility in writing aids or in tools to help students rephrase text or understand the mechanisms used to form sentences.

Another reason why we have chosen European and Brazilian Portuguese translations of the same book is related with our hypothesis that a paraphrase in European Portuguese is not always used or adequate in Brazilian Portuguese and vice-versa. With this in mind, our purpose is to allow for paraphrase adaptation to each variant of Portuguese, where the user can choose variant-specific paraphrases (e.g., *brincar de pegapega* versus *brincar à apanhada* (*play hide and seek*)), while making possible the selection of a more variant-independent paraphrase, if that choice is available.

In e-PACT, we annotated a set of sentences contrasting semantically identical multiwords, phrases, and expressions in European and Brazilian Portuguese following a set of alignment guidelines, the CLUE Paraphrase Alignment Guidelines, which resulted in the Gold CLUE4Paraphrasing annotated subcorpus. The annotation was performed with the CLUE-Aligner tool. The e-PACT corpus, the CLUE-Paraphrasing Guidelines, the CLUE-Aligner tool, and the Gold CLUE4Paraphrasing annotated subcorpus were developed within the eSPERTo project (Section 3), whose goal is to develop a linguistically enhanced paraphrase system that can be used in a large variety of natural language processing applications.

10.17771/PUCRio.TradRev.30591

2. Related Work

Barreiro et al. (1996) studied the lexical differences between European and Brazilian Portuguese in a contrastive dictionary and measured the degree of difference between the two variants of Portuguese on current language and technical vocabulary. They were the first authors to put forward a proposal for the structure of two contrastive computational dictionaries, which would later be named Lusolex and Brasilex (Wittmann et al., 2000). The research was performed during a preliminary phase of a machine translation project, whose aim was to develop a system to translate from English into both European and Brazilian Portuguese. That research was based essentially on existing computational lexicons for both variants of the Portuguese language. Our current work follows and extends that line of research into lexical and paraphrastic relations between multiword units used in both variants of Portuguese found in translation corpora. In Barreiro (2010), the concept of a dictionary of multiword units, DicTUM (Dicionário de Termos e Unidades Multipalavra), was put forward with the goal to include compounds, lexical bundles and other grammatical and content expressions, such as predicate compound nouns, support verb constructions, frozen and idiomatic expressions, among others. The research presented in this paper uses the e-PACT corpus to launch the process of gathering a larger and broader set of lexical and paraphrastic resources to enlarge DicTUM and make it available for the two variants of Portuguese, contemplating their differences and similarities, and contrasting them with the corresponding English equivalents.

3. The eSPERTo Project

eSPERTo stands for "System of Paraphrasing for Editing and Revision of Text" (in Portuguese, "Sistema de Parafraseamento para Edição e Revisão de Texto").[1] The main objective of the eSPERTo project is the development of a context-sensitive and linguistically enhanced paraphrase system that can be used as the lever for developing intelligent writing aids, summarization tools and smart dialogue systems, among other natural language processing

[1] eSPERTo is available at https://esperto.l2f.inesc-id.pt/esperto/esperto/demo.pl

10.17771/PUCRio.TradRev.30591

applications. The system comprises a paraphrase generator, a paraphrase acquisition module, and a web interactive application that was designed to help Portuguese language learners, translators and editors in revising their texts. eSPERTo recognizes semantico-syntactic, multiwords and other phrasal units, and transforms them into semantically equivalent phrases, expressions, or sentences. This semantically-driven paraphrasing system combines statistics and local grammars to acquire linguistic knowledge applied in the identification and generation of new and increasingly more complex paraphrases. The utility of eSPERTo's paraphrasing capabilities is now being explored in two other application scenarios: (i) in a question-answering system to increase the linguistic knowledge of an intelligent conversational virtual agent, and (ii) in a summarization tool to assist the paraphrasing task. Another application that we want to explore is the "adaptation" from one language variant into another and the translation between different languages.

4. Corpus of EN-EP/BP Translations

The alignment pairs in e-PACT were extracted from COMPARA[2] (Frankenberg-Garcia & Santos 2003), a bidirectional parallel corpus of English and Portuguese comprised of 32 English source texts and 40 Portuguese source texts, which are linked sentence by sentence to their corresponding translations in Portuguese and English, respectively. More specifically, the corpus was created by sampling the alignments between two texts by David Lodge, *Therapy* (EBDL1[3]) and *Changing Places* (EBDL3), and each of their two translations: one in European Portuguese (EBDL1T1 and EBDL3T1) and the other in Brazilian Portuguese (EBDL1T2 and EBDL3T2). These David Lodge books together with *Iracema* (PBJA1), by Brazilian writer José de Alencar, are the only books in COMPARA with more than one translation[4]. It is worth mentioning that the translations in Brazilian Portuguese of these two David Lodge books were done by the same

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

[3] In COMPARA, each bilingual parallel text is referred by an alphanumerical code.

[4] In the case of *Iracema*, one translation is in American English (PBJA1T1) and the other in British English (PBJA1T2).

10.17771/PUCRio.TradRev.30591

translator, and that one of the Portuguese translators also translated two other books in COMPARA: one also by David Lodge, *How Far Can You Go?* (EBDL4) and the other by Julian Barnes, *Talking it Over* (EBJB3). Having different books translated by the same Portuguese and Brazilian translators may help clarify whether differences arise from different styles or different Portuguese variants.

The texts were sampled by querying COMPARA for alignments containing, in the Portuguese translation of those texts, words that can potentially trigger the eSPERTo's paraphrase-generation grammars. These word forms are: 94,130 different human intransitive adjectives, 96,199 different words marked with the feature VSUP, which indicates the support verb of a nominal or adjectival construction, and 14,128 different predicative nouns. The sets are not exclusive, therefore, the total number of different words used to query COMPARA was 186,150. The list does not include the following forms that are frequent enough to make the query return a random set of 1,000 aligments[5] containing those words: auxiliary and support verbs, the forms *entre*, *deste*, *desse*, *destes*, and *desses*, and forms associated to the lemmas *casa*, *casar*, *dizer*, *ver*, *vir*, *suar*.[6] Each query included 100 forms that occured both in COMPARA and in our list, in a total of 169 queries.

This query process resulted in a list of 19,901 English/Portuguese alignment pairs: 11,890 for EBDL1 and 8,011 for EBDL3. These alignments included duplicates, because the same alignment may be found through different queries. From this list, we extracted pairs of European/Brazilian Portuguese that had the same source sentence, which resulted in e-PACT, a corpus of 1,628 and 1,041 unique European/Brazilian Portuguese sentence pairs in EBDL1 and EBDL3, respectively. In cases where we only obtained

[5] Due to copyright reasons, COMPARA does not return more than 1,000 alignments at a time. If a query returns more than 1,000 results, COMPARA will only show a random subset of 1,000. Since we want other researchers to be able to obtain the same alignments as we did, we queried COMPARA in a way that it returned less than 1,000 alignments per query.

[6] Most words in this list are ambiguous. The word *entre* is both a preposition (*between*) and the first and third persons singular of the verb *entrar* (*come in*) in the present subjective. The words *deste*, *desse*, *destes*, and *desses* are both the contraction of the preposition *de* (*of*) with the demonstrative pronoun; the masculine singular *este* (*this*) and *esse* (*that*), and the masculine plural *estes* (*these*) and *esses* (*those*) and several forms of the verb *dar* (*give*). The word *casa* can be the singular noun *house* and the third person singular form of the verb *casar* (*marry*) in the present tense. The words *dizer*, *ver*, *vir*, and *suar* are the infinitive forms of the verbs *say*, *see*, *come*, and *sweat*.

10.17771/PUCRio.TradRev.30591

the translation in one of the variants, we considered finding the corresponding alignment in the other variant by querying COMPARA again with a sequence of words from the corresponding English source sentence. However, in that way, we would have obtained almost all the alignments of the David Lodge texts, which could not be acceptable for copyright reasons.

We also created the Gold CLUE4Paraphrasing, a subcorpus of e-PACT corresponding to 30% of the aligned sentenced pairs in a total of 802 pairs: 489 pairs from EBDL1 and 313 pairs from EBDL3. In this corpus, we sub-aligned semantically identical and equivalent words, multiwords, phrases, and expressions (short paraphrases) between the two variants of Portuguese (cf. Sections 5.2 and 6).

Translation	PT Variant	# Alignment units	# Tokens	# Words	# Types
EBDL1T1	EP	2,150	45,184	39,005	7,920
EBDL1T2	BP	2,150	45,185	39,095	7,531
EBDL3T1	EP	1,481	29,271	24,296	6,595
EBDL3T2	BP	1,481	30,842	26,244	6,315

Table 1. David Lodge translations in COMPARA

5. Paraphrase Alignment Tool and Guidelines

The e-PACT alignment task required the development of an alignment tool, the CLUE-Aligner (Cross-Language Unit Elicitation Aligner)[7] (Barreiro et al. 2016) in which the European/Brazilian Portuguese pairs were identified, annotated, and collected. CLUE-Aligner was designed to annotate either paraphrastic or translation units representing contiguous and non-contiguous multiwords and other phrases or expressions found in monolingual or bilingual pairs of parallel sentences. Non-contiguous block alignments are necessary to express alignments between multiwords or phrases, which contain insertions, i.e., words that are not part of the multiword or phrase (Barreiro 2016). The tool allows the alignment of smaller individual or multiword units inside non-contiguous multiword units.

[7] https://esperto.l2f.inesc-id.pt/esperto/aligner/index.pl?

10.17771/PUCRio.TradRev.30591

Our annotation process consisted in using the CLUE-Aligner tool fed with 30% of the parallel sentences of the two variants of Portuguese in e-PACT where paraphrastic units were manually assigned based on linguistic information. We sub-aligned semantically identical and equivalent words, multiwords, and phrasal expressions (short paraphrases). The alignment task was performed by following alignment guidelines to guide paraphrastic unit annotation, the CLUE4Paraphrasing Alignment Guidelines, described in Section 5.1. The resulting annotated corpus is made available from our research and constitutes the Gold CLUE4Paraphrasing described in Section 5.2.

5.1. CLUE4Paraphrasing Alignment Guidelines

The CLUE4Paraphrasing Alignment Guidelines[8] is a *work in progress* set of guidelines that summarize the most important monolingual alignment recommendations collected through the European/Brazilian Portuguese paraphrastic unit alignment task in the e-PACT corpus. In addition to the conventional word alignments, the CLUE4Paraphrasing Alignment Guidelines are linguistically-informed and motivated and take into special account the annotation of multiwords, expressions and other phrasal and paraphrastic units. The monolingual alignment recommendations were collected and revised throughout the alignment task that lead to the creation of the Gold CLUE4Paraphrasing resources.

5.2. Gold CLUE4Paraphrasing

The Gold CLUE4Paraphrasing corpus corresponds to 30% of the e-PACT sentence pairs. It was created by sampling the first 3 sentences every 10 sentences. This sampling method was inspired by Gale and Church (1994), who sampled a corpus of bigrams by extracting every other bigram to a corpus for estimating probabilities and the other for testing purposes. In this way, they generated "as close to two samples of the same universe of discourse as possible'".

[8] https://esperto.l2f.inesc-id.pt/guidelines

10.17771/PUCRio.TradRev.30591

The corpus was initially pre-processed manually for decomposition of all contracted forms.[9] For example, the contracted form *nos* (*in the*) was decomposed in two elements, the preposition *em* (*in*) and the masculine plural definite article *os* (*the*) (*em+os*) in the phrases [*é típico em]* [*os nossos jogos*] ([*it is typical of*] [*our games*]), where the possessive pronoun is used with the definite article *os* in the European Portuguese corpus. In turn, the contracted form *das* (*of the*) was decomposed in two elements, the preposition *de* (*of*) and the feminine plural definite article *as* (*the*) (*de* + *as*) in the phrases [*é mais característico de*] [*as nossas partidas*] ([*it is more characteristic of*] [*our matches*]), where the possessive pronoun is used with the definite article *as* in the Brazilian Portuguese corpus. These two decompositions made possible the correct alignment of the European/Brazilian Portuguese phrases [*é típico em*]/[*é mais característico de*] ([*it is typical of*]/[*it is more characteristic of*]), and [*os nossos jogos*]/[*as nossas partidas*] ([*our games*]/[*our matches*]). Subsequently, all the decomposed forms were reviewed and the decomposed forms in multiwords and frozen expressions were changed back to contractions. For example, the decomposed contractions *a* + *a* (*to* + *the*) (*à*) and *de* + *o* (*of* + *the*) (*do*) were re-contracted in the non-compositional multiword unit *à luz do dia* (literally, *to the light of the day* = *during daylight*).

The annotation process resulted in a list of 26,101 European/Brazilian Portuguese paraphrastic unit pairs: 17,452 for the EBDL1 corpus and 8,649 for the EBDL3 corpus in a total of 13,075 unique pairs, which correspond to the union of 8,297 and 5,423 unique paraphrastic pairs, respectively. The distribution of alignment units according to the number of tokens in the pair is shown in Table 2. Inside parentheses is the number of pairs where the unit is identical (the same) in European and Brazilian Portuguese.

[9] In the future, we intend to pre-process the contracted forms automatically with STRING (Baptista et al. 2014).

10.17771/PUCRio.TradRev.30591

EP	EB	EBDL1	EBDL3
Single word	Single word	8,348 (6,099)	5,328 (3,841)
MWU	MWU	7,185 (2,083)	2,712 (809)
MWU	Single word	759	203
Single word	MWU	455	230
Non-cont	MWU	225	64
MWU	Non-cont	264	56
Non-cont	Non-cont	210 (44)	50 (12)
Single word	Non-cont	0	1
Non-cont	Single word	6	5

Table 2. Distribution of alignment units according to the number of tokens in pair.

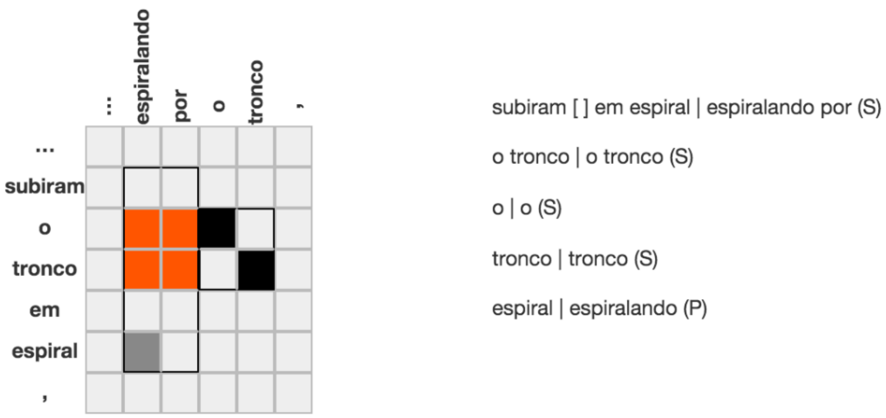
5. Discussion of Paraphrastic Units

Generally speaking, the term paraphrase refers to the relation between two or more morpho-syntactically and/or semantically related constructions. We consider that the level of equivalence in paraphrases goes from a multiword unit, such as a support verb construction (e.g., *ter febre* (*develop a fever* or *have a fever*) = *estar febril* (*be feverish*)) to a broader context, such as a sentence (e.g. *podemos, logo à primeira vista, detectar uma diferença* (*we can, even on first sight, detect a difference*) (active voice) = *uma dessas diferenças pode ser observada de relance* (*one of these differences can be observed with just a glance*) (passive voice)). In many cases, a paraphrasing relation is established between constructions corresponding to the same syntactic unit, but it can also be established between different syntactic units in a semantico-syntactic relation between two or more sentences and/or their constituents.

Figure 1 illustrates a paraphrase in the European and Brazilian Portuguese variants. This paraphrastic pair contains a non-contiguous multiword unit (*subiram* [INSERTION-NP] *em espiral* (*spiralled* [INSERTION-NP] *up*) in European Portuguese versus the verb *espiralando por* (*spiralling through*) in Brazilian Portuguese). The noun phrase insertion *o tronco* (*the tree*

10.17771/PUCRio.TradRev.30591

trunk) in the multiword unit *subiram o tronco* (*climbed the tree trunk*) in European Portuguese, can be now aligned with the verb *espiralar* (*spiral up*), which requires the direct object noun phrase to appear after the verb. The alignment pairs are represented in the list on the right side of this figure.



subiram [ ] em espiral | espiralando por (S)

o tronco | o tronco (S)

o | o (S)

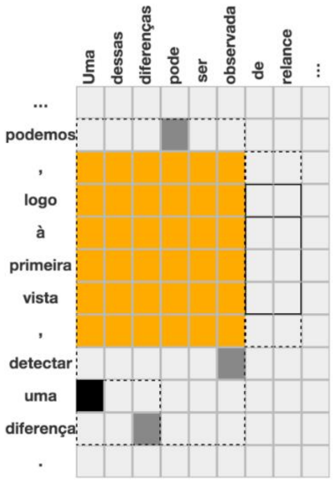
tronco | tronco (S)

espiral | espiralando (P)

Figure 1. Alignment of a non-contiguous multiword

Another interesting paraphrastic case can be seen in Figure 2, which contains an adverbial apposition inserted in an active construction in the European Portuguese sentence. This insertion, *logo à primeira vista* (*at first sight*), aligns with the adverb *de relance* (*at a glance*) in the Brazilian Portuguese sentence, which uses a passive construction, *pode ser observada* (*can be observed*), represented in the list of pairs of paraphrastic units.

10.17771/PUCRio.TradRev.30591



podemos [ ] detectar uma diferença | Uma dessas diferenças pode ser observada (P)

podemos | pode (P)

, logo à primeira vista , | de relance (P)

logo à primeira vista | de relance (S)

à primeira vista | de relance (S)

detectar | observada (P)

uma diferença | Uma dessas diferenças (P)

uma | Uma (S)

diferença | diferenças (P)

Figure 2. Non-contiguous multiword unit alignment pair contrasting active--passive constructions

Figure 3 illustrates different level paraphrases in the European and Brazilian Portuguese variants. In the sentence pair alignment, (i) identical individual segments (e.g., European and Brazilian Portuguese *janela* (*window*)); (ii) synonyms (European Portuguese *ramos* (*branches*) versus Brazilian Portuguese *galhos* (*twigs*), European Portuguese *escritório* (*office*) versus Brazilian Portuguese *estúdio* (*studio*), or European Portuguese *apanhada* (*catch*) and Brazilian Portuguese *pegapega* (*catch-catch*)); (iii) identical phrases (European Portuguese *estive a observar* (*I was watching*) versus Brazilian Portuguese *fiquei observando (I kept watching*)) or European Portuguese *da janela do meu escritório* (*from the window of my office*) versus Brazilian Portuguese *em frente à janela do meu estúdio* (*in front of the window of my studio*) with internal block alignments of *da janela de* (*from the window of*) with *em frente à janela de* (*in front of the window of*) and *o meu escritório* (*my office*) with *o meu estúdio* (my studio); (iv) paraphrases of contiguous multiword units (European Portuguese *brincarem à apanhada* (*playing catch*) versus Brazilian Portuguese *brincavam de pegapega* (*playing catch-catch*) inside larger syntactic blocks); (v) paraphrases of non-contiguous multiword units (EP *subiram* [INSERTION-NP] *em espiral* versus BP *espiralando por*), as seen above in Figure 1; and (vi) paraphrases of support verb construction with a single verb (European Portuguese *fazendo negaças* (*make feints*) versus Brazilian Portuguese

10.17771/PUCRio.TradRev.30591

ziguezagueando (*zigzagging*)).[10] Translation errors and literal translations that may sound unnatural can be useful, nevertheless, to identify whether a particular text is an original or a translation.

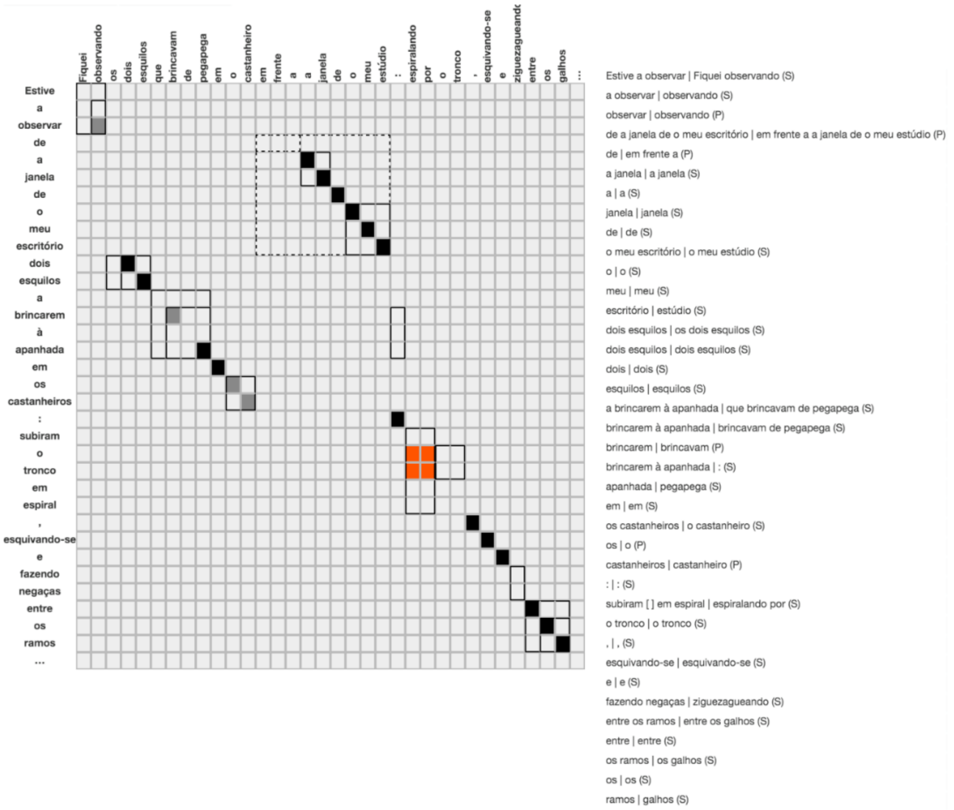


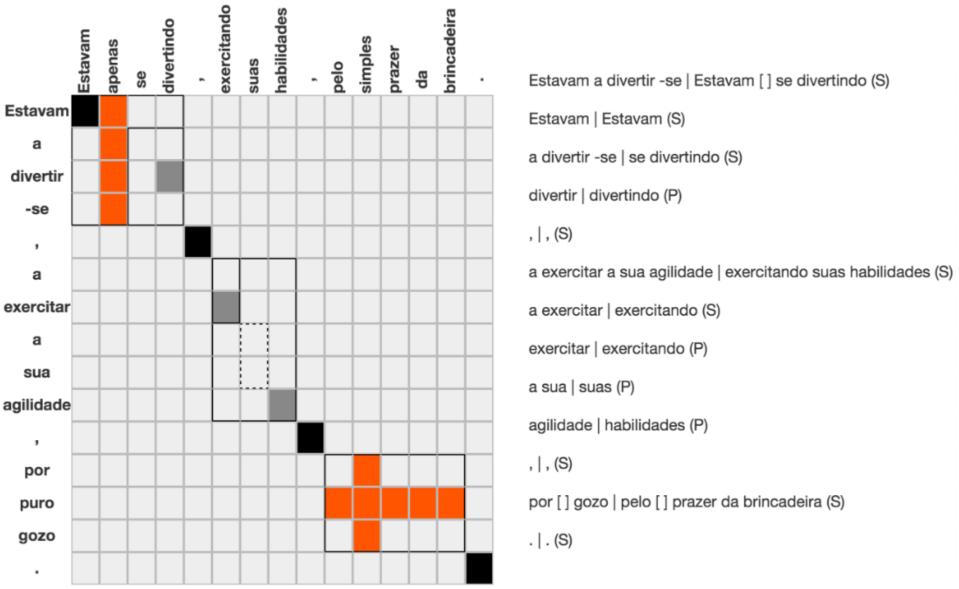
Figure 3. Alignment of EP/BP paraphrastic units.

In Figure 4 the verbal expression *estavam a divertir-se* (*were having fun*) in European Portuguese aligns with the expression *estavam se divertindo* in Brazilian Portuguese, which contains an adverbial insertion, *apenas* (*only*). This adverbial is not part of the expression and it only exists in the Brazilian Portuguese sentence. It is also worth to point out that the idiomatic expression *por puro gozo* (*for sheer pleasure*) in the European Portuguese sentence aligns with the expression *pelo simples prazer da brincadeira* (*for the simple pleasure of playing*) in the Brazilian Portuguese sentence in a less literal

[10] We are aware that some of the contrasts found in the corpus may not be valid due to translation errors or different ways of interpreting the English source expression. The collected resources require validation by a linguist/translator. Translation errors and literal translations that may sound unnatural can be useful, nevertheless, to identify whether a particular text is an original or a translation.

10.17771/PUCRio.TradRev.30591

translation. The resulting pairs of paraphrastic units are listed on the right side of the alignment grid.



Estavam a divertir -se | Estavam [ ] se divertindo (S)

Estavam | Estavam (S)

a divertir -se | se divertindo (S)

divertir | divertindo (P)

, | , (S)

a exercitar a sua agilidade | exercitando suas habilidades (S)

a exercitar | exercitando (S)

exercitar | exercitando (P)

a sua | suas (P)

agilidade | habilidades (P)

, | , (S)

por [ ] gozo | pelo [ ] prazer da brincadeira (S)

. | . (S)

Figure 4. Non-contiguous multiword unit alignment pair.

6. Conclusions

In this research work, we created a corpus of 2,669 European/Brazilian Portuguese aligned sentence pairs, e-PACT, and annotated 802 European/Brazilian Portuguese pairs with paraphrastic units using the CLUE-Aligner tool that resulted in the Gold CLUE4Paraphrasing. We have chosen European/Brazilian Portuguese translations of the same David Lodge's books to test our hypothesis that a paraphrase in European Portuguese is not always used and/or adequate in Brazilian Portuguese, and vice-versa. With this, our purpose is to make available in the eSPERTo paraphrasing system paraphrase adaptation to each variant of Portuguese, where the user can choose variant-specific paraphrases if it is desired, or opt for a more international Portuguese way of expression. If the goal of the user is to choose one of the variants, a required task will consist in ensuring whether each unit is used frequently in both variants of Portuguese or more frequently in only one of them.

In subsequent work, we wish to explore other variants of Portuguese, regional preferences, preferences among age groups, etc., to offer to the user

10.17771/PUCRio.TradRev.30591

of the eSPERTo paraphrasing system a greater variety of choices, according to preferences or purposes of the paraphrasing task. After exploring COMPARA, we wish to explore CHAVE--PT and BR journalistic texts from 1994-1995 (not annotated), as well as Portuguese movies and series subtitles from OpenSubtitles. A broader and larger set of texts will help validate and explore our initial results and extend the paraphrastic resources extracted from each text. Larger corpora will be designed to allow a closer comparative study of the different Portuguese variants targeting multiword units and fixed expressions. The systematic empirical methodology set forth herein constitutes the foundation for the development of lexicographic, educational, and writing applications for the Portuguese language.

Acknowledgements

This research work was supported by Fundação para a Ciência e Tecnologia (FCT), under project eSPERTo -- EXPL/MHC-LIN/2260/2013, project UID/CEC/50021/2013, and post-doctoral grant SFRH/BPD/91446/2012. The authors would like to thank Diana Santos and Ana Frankenberg-Garcia for answering questions about COMPARA.

References

BAPTISTA, Jorge; MAMEDE, Nuno; MARKOV, Ilia. “Integrating a lexicon-grammar of verbal idioms in a Portuguese NLP system”. PARSEME General Meeting, Athens, March 10-11, 2014 (poster session). Available at: http://www.inesc-id.pt/ficheiros/publicacoes/10099.pdf

BARREIRO, Anabela. “Linguistic Resources and Applications for Portuguese Processing and Machine Translation”. In: Judit Kuti, Max Silberztein, and Tamás Váradi (Eds.), *Applications of Finite-State Language Processing*. Cambridge Scholars Publishing, p. 41–51, 2010.

BARREIRO, Anabela. “Contributos para o Aumento de Qualidade na Língua Digital”. In: José Teixeira (ed.), *O Português como Língua num Mundo Global: problemas e potencialidades*, Centro de Estudos Lusíadas da Universidade do Minho, p. 31–47, 2016.

10.17771/PUCRio.TradRev.30591

BARREIRO, Anabela; BATISTA, Fernando. "Machine Translation of Non-Contiguous Multiword Units". In: *Proceedings of DiscoNLP 2016*. San Diego, California: Association for Computational Linguistics, p. 22 –30, 2016.

BARREIRO, Anabela; RAPOSO, Francisco; LUÍS, Tiago. "CLUE-Aligner: An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units". In: Nicoletta Calzolari et al. (eds.), *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*. Portorož, Slovenia, p. 7–13, 2016.

BARREIRO, Anabela¸ WITTMANN, Luzia; PEREIRA, Maria de Jesus. "Lexical Differences between European and Brazilian Portuguese". *The INESC journal of Research and Development* 5.2, p. 75–101, 1996.

FRANKENBERG-GARCIA, Ana; SANTOS, Diana. "Introducing COMPARA: the Portuguese-English Parallel Corpus". In: Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds.), *Corpora in Translator Education*. Manchester: St. Jerome, p. 71–87, 2003.

GALE, William; CHURCH, Kenneth. "Whats wrong with adding one". In: *Corpus-Based Research into Language: In honour of Jan Aarts*, 1994, p. 189–200.

SANTOS, Diana. "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*. 2014, University of Stockholm. Available at:
http://www.linguateca.pt/Diana/download/VariantesPIGSCP.pdf

SANTOS, Diana. "Português internacional: alguns argumentos". In: José Teixeira (ed.), *O Português como Língua num Mundo Global: problemas e potencialidades*, Centro de Estudos Lusíadas da Universidade do Minho, p. 49-66, 2016.

WITTMANN, Luzia; RIBEIRO, Ricardo; PÊGO, Tânia; BATISTA, Fernando. "Some Language Resources and Tools for Computational Processing of Portuguese at INESC". In: *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*. 2000, Athens – Greece.

10.17771/PUCRio.TradRev.30591

Resumo

Este artigo apresenta o e-PACT, um corpo de paráfrases alinhadas entre português europeu e português do Brasil. O corpo representa uma amostra dos alinhamentos à frase resultantes das traduções de duas obras literárias de David Lodge, disponíveis no COMPARA, um corpo bilingue paralelo. Uma parte do e-PACT foi anotada com unidades parafrásticas mais pequenas do que a frase correspondentes a multipalavras, elementos frásicos e expressões semanticamente equivalentes entre as duas variantes de português. A tarefa de anotação foi realizada através do uso da ferramenta de alinhamento CLUE-Aligner tendo em conta um conjunto de linhas diretrizes e recomendações, as CLUE4Paraphrasing Alignment Guidelines. O corpo resultante da anotação com pares de unidades parafrásticas constitui o subcorpo dourado Gold CLUE4Paraphrasing. Todos estes recursos: o corpo e-PACT, o subcorpo dourado Gold CLUE4Paraphrasing, as diretrizes de alinhamento CLUE4Paraphrasing Alignment Guidelines e a ferramenta de anotação de unidades parafrásticas CLUE-Aligner foram desenvolvidos no âmbito do projeto eSPERTo.

Abstract

This paper presents e-PACT, a corpus of paraphrase aligned European and Brazilian Portuguese sentences sampled from the translations of two literary English books by David Lodge available in the COMPARA bilingual corpora. We used a subcorpus of the e-PACT sentence-aligned corpus as a baseline to annotate paraphrastic units that correspond to semantically equivalent multiwords, phrases, and expressions between the two variants of Portuguese. The annotation task was performed through the use of an alignment tool called CLUE-Aligner and by following a set of guidelines, the CLUE4Paraphrasing Alignment Guidelines. The resulting corpus annotated with pairs of paraphrastic units constitutes the Gold CLUE4Paraphrasing. All the resources, the e-PACT corpus, the Gold CLUE4Paraphrasing subcorpus, the CLUE4Paraphrasing Alignment Guidelines, and the CLUE-Aligner tool were developed in the scope of the eSPERTo project.