Contractions: to align or not to align, that is the question

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

Fernando Batista
Instituto Universitário de Lisboa (ISCTE-IUL)
INESC-ID, Rua Alves Redol 9
1000-029 Lisboa, Portugal
fernando.batista@inesc-id.pt

Abstract

This paper performs a detailed analysis on the alignment of Portuguese contractions, based on a previously aligned bilingual corpus. The alignment task was performed manually in a subset of the English-Portuguese CLUE4Translation Alignment Collection. The initial parallel corpus was pre-processed and, a decision was made as to whether the contraction should be maintained or decomposed in the alignment. Decomposition was required in the cases in which the two words that have been concatenated, i.e., the preposition and the determiner or pronoun, go in two separate translation alignment pairs (e.g., [*no seio* ***de***] [***a*** *União Europeia*] | [*within*] [*the European Union*]). Most contractions required decomposition in contexts where they are positioned at the end of a multiword unit. On the other hand, contractions tend to be maintained when they occur in the beginning or in the middle of the multiword unit, i.e., in the frozen part of the multiword (e.g., [***no*** *que diz respeito a*] | [*with regard to*] or [*além* ***disso***] [*in addition*]. A correct alignment of multiwords and phrasal units containing contractions is instrumental for machine translation, paraphrasing, and variety adaptation.

1 Introduction

The past decade has seen a significant advance in the field of machine translation mainly due to the growth of publicly available corpora, from which an enormous amount of translation alignments have been extracted. Alignments of multiword units and other phrases represent the driving force in the development of translation systems and the success of systems like Google Translate and others has a great deal to do with huge lexical coverage available in the large amounts of corpora that they have access to (Barreiro et al., 2014b) and from which translation alignments are extracted. But the quality of these alignments is also very important. For example, several authors have pointed out that the integration of multiword units in translation models based on linguistic knowledge is considered as an impact factor in obtaining better quality translations (Chiang, 2005; Marcu et al., 2006; Zollmann and Venugopal, 2006). Expert participation extends to the gathering, enhancement and integration of language resources including non-contiguous multiword unit alignments (Barreiro and Batista, 2016). Above all, high quality machine translation depends on the quality of the alignments used in the processes of machine learning. Some systems use unsupervised learning, in which the machine itself decides which segments of a source-language phrase align with which target language phrase segments (Och and Ney, 2000), other systems use supervised learning based on previous alignments made manually by linguists (Blunsom and Cohn, 2006). In this paper, we focus on the **alignment of multiword units where contractions occur**, a challenge that has been overlooked in the existing literature and can be responsible for grammatical errors in translations.

A contraction is a word formed from two or more words of different parts-of-speech (most frequently) or the same part-of-speech (more seldom) that would otherwise appear next to each other in a sequence. For example, in English the most common contractions are those where the word *not* is added to an

auxiliary verb in negative sentences, with omission of internal letters (e.g., *is not* → *isn't*) or those consisting of combinations of pronouns with auxiliary verbs, in which a word or a syllable is substituted by an apostrophe (e.g., *it is* → *it's*). These contractions are mainly used in speech and informal writing, but not in formal writing as in the Romance languages, where contractions are non-optional. The most common contractions in the Romance languages are those where prepositions are contracted with articles or pronouns with addition, replacement, or omission of letters. For example, in Portuguese the contraction *nas* '*in*' | '*at the*' results from the concatenation of the preposition *em* with the feminine plural definite article *as*; in Italian, the contraction *degli* '*of*' | '*from*' results from the concatenation of the preposition *di* with the masculine plural definite article *gli*; in Spanish, the contaction *al* '*to*' | '*at the*' results from the concatenation of the preposition *a* with the masculine singular definite article *el*; in French, the contaction *aux* '*at*' | '*for*' | '*to the*' results from the concatenation of the preposition *à* with the masculine plural definite article *les*. However, contractions can also be composed of two words with the same part-of-speech, e.g., two determiners (*la une* → *l'une*) or two prepositions (*de après* → *d'après*), as in French.

We describe a linguistically motivated approach to the alignment of multiword units where contractions occurring in these multiword units are required to be decomposed, except in specific circumstances determined by the context, such as when they constitute a non-variable (non-inflecting) element of a frozen multiword unit. Decomposition allows the correct alignment of a multiword unit, such as the prepositional compound *apesar de* | *in spite of*, in the sense that it separates the preposition *de* (*of*) that is part of the multiword from a concatenated element, in this case, the feminine singular definite article *a* (*the*) that is not part of the multiword, but rather belongs to the phrase or expression that immediately follows it (e.g., *apesar* ***da*** → [*apesar* ***de***] [***a*** NP]. Similarly, the masculine plural definite article *os* (*the*) in the expression *à luz de* | *in the light of*, requires to be split from the preposition *de* (e.g., [***à*** *luz* ***dos***] → [***à*** *luz* ***de***] [***os*** NP). However, the contraction of the preposition *a* (*at*) with the feminine singular definite article *a* in this expression is not decomposed from its composed form *à*, because it represents a fixed element of the multiword unit, never changing its form. Failure to align and process correctly these multiword units involving contractions containing elements that are external to them leads to errors in the translated texts. Even if these errors do not affect the understanding of the translated text, they may compromise the quality of the translation leading to greater post-editing efforts.

In our experiment, a linguist has pre-processed manually a subset of the reference Europarl parallel corpus (Koehn, 2005) containing 400 Portuguese-English parallel sentences. From this subset corpus, the EN–PT CLUE4Translation Alignment Collection was achieved by adopting the methodology described in Section 3 for the alignment of Portuguese multiwords and other phrasal units involving contractions in the original corpus. This methodology was achieved during the development of the CLUE Alignment Guidelines, a set of linguistically-informed guidelines for the alignment translation or paraphrastic units in bitexts. In other words, the Guidelines were developed in two separate sets of documents containing statements by which to determine courses of action regarding the alignment of multiwords and other phrasal units, depending on whether these linguistic units are used in translation (CLUE4Translation Alignment Guidelines) or in paraphrasing (CLUE4Paraphrasing Alignment Guidelines). The approach reinforces the weight of multiwords as objects of representation in the alignment between the source and the target languages, independently of the source-target being two different languages, in the case of translation, the same language, in the case of paraphrases, or between language varieties, in the case of variety adaptation. The annotation of the subset corpus was performed with the CLUE-Aligner tool (Barreiro et al., 2016), a paraphrastic and translation unit aligner built to provide an efficient solution in the alignment of non-contiguous multiword units. CLUE-Aligner was developed within the eSPERTo project[1], whose objective is to develop a context-sensitive, linguistically enhanced paraphrase system that can be used in natural language processing applications, such as intelligent writing aids, summarization

[1] eSPERTo stands for **S**ystem of **P**araphrasing for **E**diting and **R**evision of **T**ext (in Portuguese, **S**istema de **P**arafraseamento para **E**dição e **R**evisão de **T**exto). eSPERTo's core linguistic resources were extracted from OpenLogos bilingual resources (Barreiro et al., 2014a), the free open source version of the Logos System (Scott, 2003), adapted and integrated into NooJ linguistic engine (Silberztein, 2016). eSPERTo is available at `https://esperto.l2f.inesc-id.pt/esperto/esperto/demo.pl`

tools, smart dialogue systems, language learning, among others. Our broader research aims to contribute to new machine translation systems that produce high quality translation for which linguistically-based alignments are extremely important.

2 Related Work

In NLP tasks, contractions are problematic for several reasons, among them: (i) two or more function words[2] mostly with different parts-of-speech overlap, which makes syntactic analysis and generation difficult; (ii) in cross-language analysis, the contrast between languages that have contractions and languages that do not have them, or do not have them in the same contexts, may present additional difficulties. Although, most parsers and part-of-speech taggers can process contractions successfully, the alignment of segments in a parallel pair of sentences, where one particular segment corresponds to a contraction in one language and to more than one segment (no contraction) in the other language has not been adequately addressed in alignment annotation guidelines or alignment research (cf. (Och and Ney, 2000; Lambert et al., 2005; Graça et al., 2008), or (Tiedemann, 2011), among others). For example, the Portuguese contraction of the preposition and the demonstrative pronoun *neste* corresponds to two words in English (*in this*) and in Spanish (*en esta*), as illustrated in Example (1). In addition, the freely available parallel corpora most used in alignment tasks (Koehn, 2005) have not been pre-processed in order to make possible the correct alignment of the pairs of multiword units involving contractions. These shortcomings and lack of adequate directives to guide annotators in alignment tasks are responsible for machine translation errors, but they also affect negatively other NLP tasks involving alignment resources, such as paraphrasing, among others. Our contraction pre-processing task aims to advance the state of the art alignment taking into consideration the correct alignment of multiword units where contractions existed in the original corpus.[3] The methodology used to decide whether contractions need to be decomposed for the alignment of their canonical forms or whether they are required to be maintained inside the multiword unit is presented in Section 3.

(1) $_{EN}$ - *to make further progress* **in this** *area*
$_{ES}$ - *a fin de avanzar* **en esta** *dirección*
$_{PT}$ - *com o intuito de conseguir um avanço* **neste** *(***em + este***) domínio*

The Romance languages have peculiar behaviour with regards to the use of contractions. Some languages require a particular contraction, other languages require another type of contraction. Our methodology is consistent with regards to decomposition of contractions when they refer to aligning canonical forms, i.e., separate words like a preposition and a determiner cannot align with a contraction or when they are part of a frozen compound or fixed expression. For example, the English lexical bundle ***in that*** *sense* requires the contraction in the Portuguese translation [**nesse** *sentido*] to be maintained. The equivalents in the remaining Romance languages do not contain contractions ([**en ese** *sentido*] in Spanish, and [**en ce** *sens*] in French).

3 Methodology

In our alignment task, the PT–EN CLUE4Translation parallel corpus was pre-processed for a framework decision regarding whether its contractions should be decomposed or maintained. Sections 3.1 and 3.2 discuss the alignment issues specific to each one of the decisions, with a set of real-world alignment examples, which aid in the understanding of the issues raised. Initially, the pre-processing task consisted of a semi-manual decomposition by a linguist of all contractions. Decomposition allowed for the correct alignment of multiword units where contracted forms required to be split so that those multiwords and the phrases that follow them could be mapped to the corresponding elements in the source language, as

[2]Function or structure words, such as prepositions, determiners, auxiliary verbs and pronouns, among others, have little lexical or ambiguous meaning, and are used to express grammatical (or structural) relationships with other words within a sentence. They are extensively described in grammars. Function words are in contrast with content or lexical words, which include nouns, verbs, adjectives, and most adverbs, normally containing very specific meanings listed in the dictionaries.

[3]This topic has been only superficially described in (Barreiro and Mota, 2017).

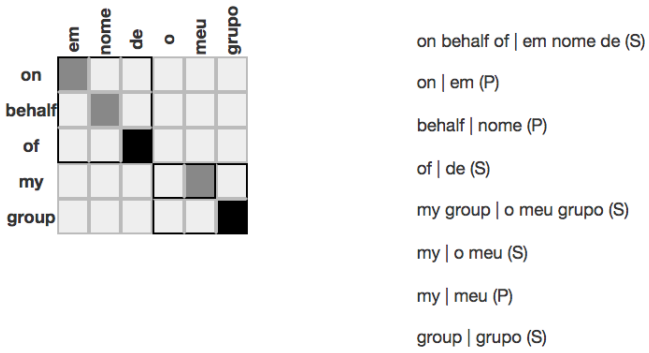

Figure 1: Alignment of the compound word *em nome de* | *on behalf of* and the noun phrase *o meu grupo* | *my group* with their internal elements (individual words)

illustrated in Section 3.1. Subsequently, all the decomposed forms were reviewed and the decomposed forms in multiwords and frozen expressions were changed back to contractions, as described in Section 3.2. This methodology prioritized decomposition for statistical reasons only. The number of contractions that need to be decomposed in the corpus is much greater than the number of contractions that require to be maintained.

3.1 Decomposed Contractions

The Portuguese word *do* (*of the*) occurring in the original corpus corresponds to the contraction of the preposition *de* with the masculine definite article *o* that agrees with its masculine noun modifier *grupo* in the phrase *em nome do meu grupo* | *on the behalf of my group*. This contraction was decomposed in two elements, the preposition *de* (*of*) and the masculine singular definite article *o* (*the*) (*de* + *o*) in order to align correctly both the canonical form (lemma) of the compound word *em nome de* | *on behalf of*, and the noun phrase *o meu grupo* | *my group*, where the preposition of the contraction goes with the compound and the definite article goes with the noun phrase, i.e., the decomposition is required to make possible that the two concatenated words go in two different alignment pairs, as illustrated in Figure 1. Similar decomposition has taken place in contractions such as those illustrated in examples (2)–(5).

(2) *EN* - **across** + [the Atlantic]
PT - **do outro lado de** (`do = [de+o]`) [o Atlântico]

(3) *EN* - **issues like** + [the NP]
PT - **questões como a de** (`dos = [de+os]`) [os NP]

(4) *EN* - **with respect to** + [the N]
PT - **quanto a** (`ao = [a+o]`) [o N]

(5) *EN* - **fully approves** [NP: the joint position of the council]
PT - **dá a sua total aprovação a** (`à = [a+a]`) [NP: *a posição comum do conselho*]

Decomposition of contractions also has implication in coordination. For example, the coordinated noun phrases *o parlamento* | *the parliament* and *o conselho* | *the council* illustrated in Figure 2 are direct complements of the Portuguese prepositional verb *realizado por* | *carried out by*. While in English the preposition *by* of the prepositional verb is not repeated before the second noun phrase, in Portuguese there is repetition of the preposition *por* in the coordination introduced by the prepositional verb *realizado* ***por*** [NP] *e* ***por*** [NP]. The CLUE-Aligner alignment tool allows the alignment of the non-contiguous coordinating structure, excluding the NP elements (gaps), which are the variable elements of the coordination, and making possible to align them separately. Alignment methodologies require these linguistic nuances captured in translation to be handled correctly.

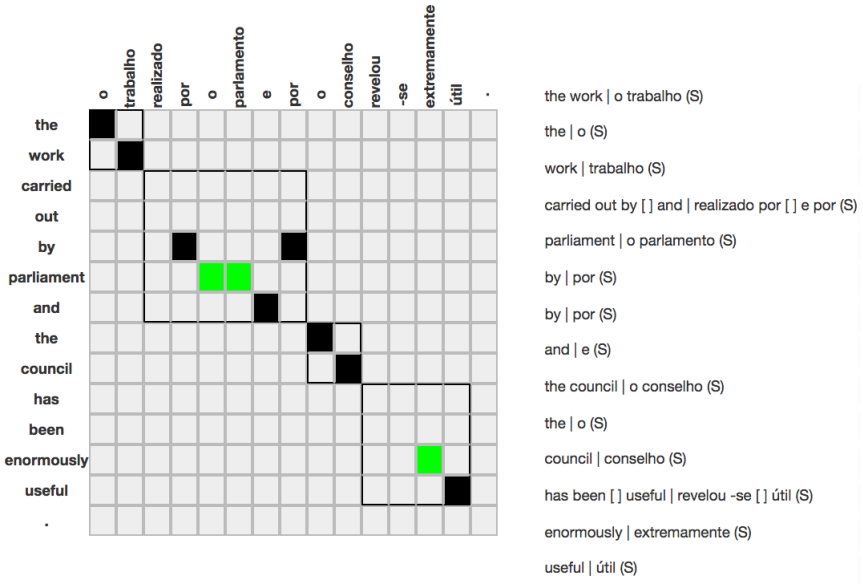

Figure 2: Alignment of the coordinated multiword *realizado* ***por*** [NP] *e* ***por*** [NP], implying the double decomposition of the contraction *pelo* into the preposition *por* of the prepositional verb and the masculine singular definite article *o* of the coordinated NP.

3.2 Non-decomposed Contractions

In a second pre-processing step, decomposed contractions *à = a + a* '*to + the*', *na = em + a* '*in + the*', and *do = de + o* '*of + the*' were restored in non-compositional multiword units, such as the fixed expressions *à luz de* | *in the light of* already mentioned in Section 1, *à data* | *at the time* illustrated in Figure 3, and *na ordem do dia* | *the next item* in the corpus (better translated as *in the agenda*), illustrated in Figure 4.

4 Analysis of Preliminary Results

Preliminary results confirmed that most contractions require decomposition in contexts where they are part of a multiword unit. For example, the most frequent contractions in the corpus, (*de + a*, *de + o*, *de + os*, and *em + a*), with more than 50 occurrences each, establish syntactic relationships between multiwords, such as compounds, prepositional nouns, etc., some of which are discontinuous (e.g. *centrar-se* [] *em* | *to deal with*). In these contractions, the preposition establishes the final border of the first phrase (i.e., the last word in the phrase), and the determiner establishes the initial border of the phrase immediately after (i.e., the first word in the phrase). The second noun phrase can be a named entity (e.g., *União Europeia* | *European Union*, *Ásia* | *Asia*), or a term (e.g., *capital de risco* | *risk capital*, *fundos de pensão* | *pension funds*), but, there are also occurrences of contractions that require decomposition in contexts where the preposition is part of a multiword unit (the last word of the multiword, e.g., *em relação a* | *with regard to* and the determiner is part of a regular noun phrase, e.g., *as observações* | *the comments*. Table 1 presents the frequency of contractions in contexts in which they require decomposition.

With regards to contractions that cannot be decomposed, most of them occur in the beginning or in the middle of the multiword unit, seldom in the end. For example, the contractions *no*, *neste*, *pelo*, and *às* in the multiwords *no que diz respeito a* | *with regard to*, *neste momento* | *at this time*, *pelo contrário* | *on the contrary*, and *às 12h30* | *at 12.30 p.m.* cannot be decomposed, because they are not positioned in the border with the next phrase. The same goes for the contraction *à* in the multiword unit *até à data* | *so far*, which occurs in a middle position. Exceptionally, the contraction *disso* in the multiword *além disso* | *in addition* also remains undecomposed, because it corresponds to a fixed adverbial expression. Table 2 presents the frequency of contractions in contexts in which they cannot be decomposed.

A few observations are worth noting with regards to undecomposeable contractions. One of them is that there are some semantico-syntactic patterns that function as linguistic constraints. For example, the

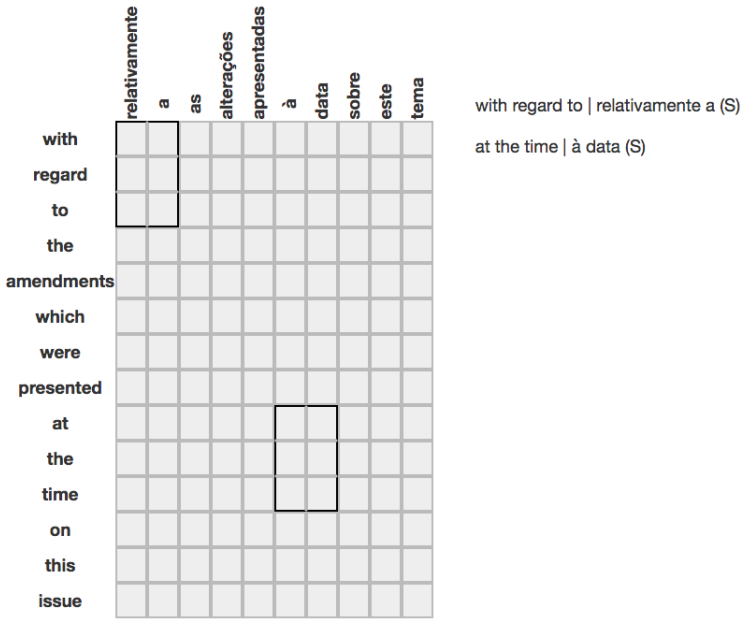

Figure 3: Alignment of the fixed expression *à data*

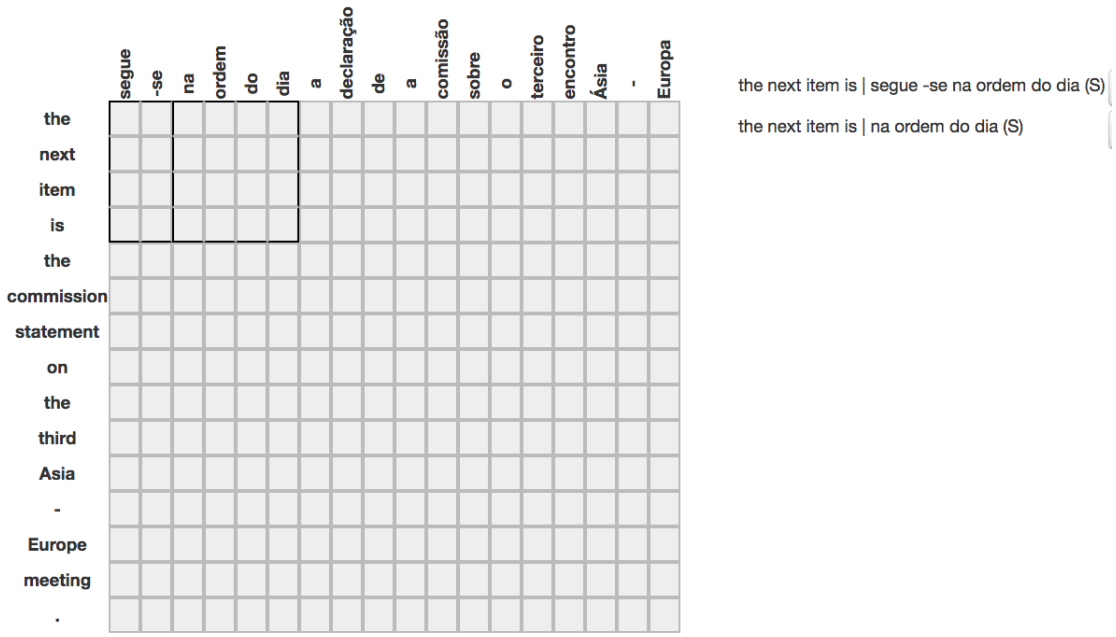

Figure 4: Alignment of the fixed expression *na ordem do dia*

Decomp.	Freq	PT example	EN translation
de a	113	[no seio **de**] [**a** União Europeia]	[within] [the European Union]
de o	93	[a promoção **de**] [**o** capital de risco]	[encouraging] [risk capital]
de os	68	[o favorecimento **de**] [**os** fundos de pensão]	[to favour] [pension funds]
em a	61	[integração **em**] [**a** Ásia]	[integration in] [Asia]
a a	44	[dar prioridade **a**] [**a** extensão de]	[focusing on] [the extension of]
a o	34	[prestar-se [] atenção **a**] [**o** trabalho infantil]	[attention must [] be paid to] [child labour]
de as	29	[o objectivo **de**] [**as** redes transeuropeias]	[the purpose of] [the trans-European networks]
em o	29	[fusões **em**] [**o** mercado de capitais]	[mergers on] [the capital market]
a os	20	[no concernente **a**] [**os** fundos de pensões]	[as for] [pension funds]
em as	16	[centrar-se [] **em**] [**as** questões comuns]	[to deal with] [questions which unite us]
a as	15	[em relação **a**] [**as** observações]	[with regard to] [the comments]
em os	12	[com base **em**] [**os** mesmos critérios]	[to use the same yardstick]
por o	10	[realizado **por**] [**o** parlamento]	[carried out by] [parliament]
por os	10	[angariados **por**] [**os** mercados de capital de risco]	[raised from] [venture capital]
por a	9	[influenciados [] **por**] [**a** instalação de]	[compromised [] by] [fitting]
em uma	7	[assenta **em**] [**uma** relação de igualdade]	[based on] [a relationship of equality]

Table 1: Frequency of contractions in contexts in which they require decomposition

contracted	freq	PT example	EN translation
no	43	**no** que diz respeito a	with regard to
do	34	inclusão [na ordem **do** dia]	added [to the agenda]
da	33	[**da** mesma forma que]	[in the same way that]
nos	17	**nos** dois sentidos	on both sides
dos	17	a carta **dos** direitos fundamentais	the charter of fundamental rights
na	17	**na** sua quase unanimidade	almost unanimously
neste	13	**neste** momento	at this time
à	13	até **à** data	so far
ao	13	**ao** dar prioridade a	by focusing on
disso	7	além **disso**	in addition
pelo	6	**pelo** contrário	on the contrary
das	5	redução [] **das** despesas	reducing [] expenditure
nesse	2	**nesse** sentido	to this effect
às	2	**às** 12h30	at 12.30 p.m.
desse	2	**desse** modo	hence
consigo	2	em paz **consigo** próprio	at peace with itself

Table 2: Frequency of contractions in contexts in which they cannot be decomposed

contractions *às*, *nos*, or *pelos* cannot be decomposed when used with time-related named entities, such as *às seis horas da tarde* '*at 6 p.m*', *às sextas-feiras* '*on fridays*', *nos anos sessenta* '*in the sixties*, or *pelos anos seguintes* '*for all years ahead*', among others. Another important observation is that, in normal circumstances, contractions of prepositions with pronouns, such as *consigo* in the expression *consigo próprio* '*with itself*' should not be decomposed.

The alignment task has given us cause to reflect on how certain linguistic units have been aligned in previous research work. As far as alignments involving the contraction phenomenon, have there been discussions on whether the contraction should be maintained or decomposed in cases such as *muitos dos presentes nesta assembleia* | *many in the house*, or *pelas mais variadas razões* | *for a variety of reasons*? What about other linguistic phenomena? Is there scientific ground to establish "strict" borders for aligning paraphrastic units or translation units or are alignment decisions sometimes arbitrary? While this is not the first attempt to establish guidelines for alignment tasks, we have made an attempt to treat contractions in a scientific way, either maintaining the contraction at the beginning and the middle of a multiword unit or decomposing the contraction at the end of the multiword unit. The resulting alignment data may still contain errors, but we tried to make decisions in more than an ad-hoc fashion.

5 Final Remarks

Language experts enrolment in machine translation is essential in pre-editing tasks to improve the quality of the text to be translated (input or source text), and in post-editing tasks to improve the translated text (output or target text). High quality machine translation is directly related to the human factor, namely to the intervention of specialists of the languages involved in translation and their role in the validation of correct translation alignments. When used in machine translation systems, alignments containing linguistic knowledge contribute to improved accuracy, reduced computational complexity and ambiguity, and improved translation quality, as it happens in the particular case of contractions described in this paper. Given that contractions can be a frequent phenomenon in a language, the results that can be obtained through their correct alignment in a system can be significantly better than those obtained in a purely statistic or ad-hoc manner. But, there are other linguistic phenomena that require further examination. Without a suitable linguistic approach to the alignment task, and limited to the capacity of the algorithms, systems will continue to be overloaded with poor quality alignments, which will create translation of limited quality, requiring a greater post-editing effort. However, there is still a shortage of manually annotated alignments that can be used in training and evaluation for many language pairs or language variants, especially those with scarce resources. In this paper, we have used a methodology to align multiword units involving contractions, which pose a challenge to their correct alignment. The proposed alignment methodology does not depend on the application, so the pairs of aligned multiwords and phrases can be used in translation, paraphrasing, variety adaptation and other NLP tasks. We also hope that the linguistic knowledge learned in our alignment task can help solve problems related to the alignment of multiword units, provide better solutions to process and align them, and ultimately serve to build a more sophisticated automatic alignment tool.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under project eSPERTo, with reference EXPL/MHC-LIN/2260/2013, and with reference UID/CEC/50021/2013. Anabela Barreiro was also funded by FCT through post-doctoral grant SFRH/BPD/91446 /2012.

References

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

Barreiro, A., Batista, F., Ribeiro, R., Moniz, H., and Trancoso, I. (2014a). OpenLogos Semantico-Syntactic Knowledge-Rich Bilingual Dictionaries. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC 2014, pages 3774–3781, Reykjavik, Iceland. European Language Resources Association.

Barreiro, A., Monti, J., Orliac, B., Preuss, S., Arrieta, K., Ling, W., Batista, F., and Trancoso, I. (2014b). Linguistic Evaluation of Support Verb Constructions by OpenLogos and Google Translate. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC 2014, pages 35–40, Reykjavik, Iceland. European Language Resources Association.

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

Barreiro, A., Raposo, F., and Luís, T. (2016). CLUE-Aligner: An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference*, LREC 2016, pages 7–13. European Language Resources Association.

Blunsom, P. and Cohn, T. (2006). Discriminative Word Alignment with Conditional Random Fields. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 65–72, Sydney, Australia. Association for Computational Linguistics.

Chiang, D. (2005). A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL 2005, pages 263–270, Ann Arbor, Michigan, USA. Association for Computational Linguistics.

Graça, J., Pardal, J. P., Coheur, L., and Caseiro, D. (2008). Building a Golden Collection of Parallel Multi-Language Word Alignment. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the 6th International Conference on Language Resources and Evaluation*, LREC 2008, pages 986–993, Marrakech, Morocco. European Language Resources Association.

Koehn, P. (2005). EuroParl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, Asia-Pacific Association for Machine Translation.

Lambert, P., De Gispert, A., Banchs, R., and Mariño, J. B. (2005). Guidelines for Word Alignment Evaluation and Manual Alignment. *Language Resources and Evaluation*, 39(4):267–285.

Marcu, D., Wang, W., Echihabi, A., and Knight, K. (2006). SPMT: Statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2006, pages 44–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL 2000, pages 440–447, Hong Kong. Association for Computational Linguistics.

Scott, B. (2003). The Logos Model: An Historical Perspective. *Machine Translation*, 18(1):1–72.

Silberztein, M. (2016). *Formalizing Natural Languages: the NooJ Approach*. Wiley Eds.

Tiedemann, J. (2011). *Bitext Alignment*. Synthesis Digital Library of Engineering and Computer Science. Morgan & Claypool.

Zollmann, A. and Venugopal, A. (2006). Syntax Augmented Machine Translation via Chart Parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, SMT 2006, pages 138–141, New York City. Association for Computational Linguistics.