A Multilingual Paraphrasary of Multiwords

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 introduces the novel concept of a Multilingual Paraphrasary addressing its need for paraphrasing and translation. The multilingual paraphrasary is an ongoing work carried out in compliance with the CLUE-Alignments, a set of linguistically informed multilingual alignments, comprising several categories of multiword units. The CLUE-Alignments set has all possible combinations between English, French, Portuguese, and Spanish parallel texts of the common test set of the Europarl corpus. The gold collection of the manually annotated CLUE-Alignments is a refined Gold-CLUE. The paper also presents the CLUE-Aligner tool[1], developed to facilitate the alignment of the meaning and translation units in the bitexts, including the alignment of non-contiguous units. Our approach benefits from the Logos Model for machine translation, namely the semantico-syntactic abstraction language SAL and the semantic table SemTab. Finally, the paper illustrates how the collected paraphrases are used in the paraphrase generation tool eSPERTo[2], developed for Portuguese, as part of a larger multilingual generation project involving paraphrasing and translation.

[1] https://esperto.hlt.inesc-id.pt/esperto/aligner/index.pl

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

1 Introduction

Paraphrase generation is crucial in natural language processing (NLP) and quality machine translation (MT) cannot be achieved without comparable quality paraphrase knowledge because paraphrases are vital to deploying semantic knowledge to guarantee high fidelity translation. An important common issue in human translation and MT is to define equivalence and to define and establish paraphrasing capabilities. Therefore, one of the first tasks involved in the construction of a paraphrasing or MT system should be to collect pairs of alignments that correspond to semantically identical or similar units of meaning expressed with different vocabulary and/or syntactic structure. Some paraphrase extraction techniques may simply imply semi-automatic procedures, while others may consist of supervised alignment trained on manual alignments, which can be used for monolingual or bilingual term extraction.

We used the common test version of the European Parliament Proceedings taken from Q4/2000 portion of the data, 2000-10 to 2000-12 (Koehn, 2005). The bilingual texts are available on the European Parliament Proceedings Parallel Corpus website.[3] The reference sub-corpus is aligned at the sentence level, ranging from sentence number 101 to sentence number 500. Our work represents an extension of the work on multilingual alignments by (Graça et al., 2008). We manually annotated translation alignments for 400 sentences in 6 sets of the multilingual test corpus, representing 2,400 aligned sentences.

Our research led to the identification of four main classes of challenges to the alignment of

[3] http://www.statmt.org/europarl/archives.html#v1

units: (i) lexical and semantico-syntactic, (ii) morphological, (iii) morpho-syntactic, and (iv) semantico-discursive). Our focus is on the lexical and semantico-syntactic phenomena that MT systems, in general, do not translate well, namely the alignment of multilingual/cross-lingual expressions, multiwords, and other phrasal units as representation objects in the alignment between the source and target languages. In order to simplify the wording, we will use the designation of multiwords for the three types of semantico-syntactic translation units aforementioned.[4] The alignment task resulted in a paraphrase collection to be used in NLP applications including MT. We analysed the collection and created a novel linguistic computational object/concept, which we coined 'Paraphrasary', as a complex equivalent to a dictionary at a level larger than the word. A paraphrasary is to semantico-syntactic units' equivalences as a dictionary is to synonyms.

The structure of the remainder of the paper is as follows: in Section 2, we revisit the research on alignments, revise the concept of alignment, and justify our need for linguistic precision in the alignment task. In Section 3, we discuss the complexity of the alignment of multiwords. In Section 4, we explain how the Logos Model approach to the processing of units of meaning larger than the word (multiwords) helped configure our alignment model. In Section 5, we present the **C**ross-**L**ingual **U**nit **E**licitation (CLUE) approach, summarise the CLUE-Aligner tool and the gold collection Gold-CLUE. In Section 6, we describe how we choose what goes into the multilingual Paraphrasary. In Section 7, we illustrate how the collected paraphrases are used in the eSPERTo paraphrase generation system. Finally, in Section 8, we present some conclusions and future work.

2 Alignments Revisited

Word alignments were defined as representations of semantically equivalent words, phrases, or expressions within the source and target sentences of a parallel corpus (Brown et al., 1990), and the task of word alignment consists of identifying the translational equivalences that contain semantic correspondences in the aligned sentence pairs of a parallel text (Hearne and Way, 2011). As the outcome of the alignment task, a set of individual alignments or links, as some authors call them (Lambert et al., 2005), can be established between words or sequences of words, designated as n-grams. A sequence of more than one n-gram is usually called 'phrase'. Alignments based on random n-grams do not have a linguistic motivation or contrastive analysis lying behind them. However, MT systems built upon linguistic knowledge-based alignments extracted from high-quality translation corpora can contribute to increased precision, with the subsequent improvement of translation quality. Additional benefits can be gained for any natural language generation (NLG) task because "word alignments" is not a concept restricted to MT. They are used in a wide variety of applications, representing a highly valuable resource for evaluation and enhancement of word alignment algorithms, supervised word alignment, alignment evaluation, MT evaluation, automatic bilingual lexica, term extraction, and paraphrasing.[5]

Shortcomings in alignment tasks and alignment guidelines show that linguistic expertise and cross-lingual contrastive analysis are required to reduce the complexity and ambiguity in the alignment process, especially with regard to multiwords because linguistic principles can support alignment decisions independently of the annotator or the annotator's perception of what a translational equivalence should be. The paper "n-grams in search of theories" (Maia et al., 2008) claimed the need to create linguistically robust alignment tools for research based on a supporting theoretical and practical framework. As a follow up, the development of CLUE-Aligner[6] (Barreiro et al., 2016) appeared as a response to the demand for the alignment of not only contiguous multiwords, such as the support verb construction *to draw a distinction between* but also non-contiguous multiwords, i.e., units with insertions, such as the support verb construction *to bring* [INSERTION] *to a conclusion* (Barreiro and Batista, 2016). Our alignment task led to the development of a set of guidelines – CLUE-Alignments –

[4] Alignments are an efficient (and convenient) intermediate representation developed for engineering purposes in NLP and MT systems that present shortcomings from a linguistic point-of-view. We are trying to reduce the number of shortcomings in alignment tasks by adding scientific precision.

[5] Some basic annotation guidelines had been proposed, e.g. `http://www.cs.jhu.edu/~ccb/publications/paraphrase_guidelines.pdf`

[6] CLUE-Aligner is an alignment tool based on Linear-B (Callison-Burch and Bannard, 2004), enhanced in order to permit the alignment and storage of both contiguous and non-contiguous multiwords and other phrasal units to be used in paraphrasing and translation.

based on the fundamental principles of the Logos Machine Translation Model (henceforth, the Logos Model) (Scott, 2003) (Barreiro et al., 2011)[7], which relies on deep semantico-syntactic analysis to generate translation of multiwords, such as in the English-Portuguese (EN-PT) examples: (i) *give in without struggle — ceder sem resistência*; (ii) NHum/PRO *be settling down to* PRO *new job* — NHum/PRO *ir-se habituando ao novo emprego*; or (iii) *arrive first/second/last — chegar em primeiro/segundo/último lugar*. Quality texts and quality alignments based on the "SemTab" function of the Logos Model (Section 4) were key ingredients to build an efficient multilingual paraphrasary, which represents a step forward into meaningful quality translation, and a valuable resource for NLG. Section 3 discusses the challenges presented by multiwords to MT and the reasons why their correct and non-ambiguous alignment is important.

3 Multiwords

Multiwords, most commonly known as multiword expressions[8], have been defined by (Baldwin and Kim, 2010) as "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic, and/or statistical idiomaticity". The specification of several classes of multiwords reflects some progress in their classification. Literature draws attention to different types of multiwords: phrasal verbs, light or support verb constructions, noun compounds, proper names, and non-compositional idioms, among others. Nevertheless, the struggles of MT with multiwords are known and have been reported in several research works (Barreiro et al., 2010), (Barreiro et al., 2013), (Kordoni and Simova, 2014), (Barreiro et al., 2014), and (Barreiro, 2015), among others.

Multiwords are a source of mistranslations not only by MT systems, but also by professional translators, in part because they can be non-contiguous and the remote syntactic dependency may get lost or misunderstood, but also because they are a source of various **contextual nuances**, such as the prepositional verb *break into* in the EN-PT alignment pairs: (i) *break into* NPlace — *assaltar* NPlace as in *beak into a house — assaltar uma casa*; (ii) *break into a laugh — desatar a rir*; (iii) *break into a run — pôr-se em fuga*, *pôr-se a andar*; but (iv) *break into pieces — quebrar em bocados*, *estrilhaçar*.

The most important consideration with respect to multiwords is that they should never be processed on a word-for-word basis because they represent atomic semantico-syntactic and translation units and cannot be broken down into constituent parts in any alignment process.

Therefore, linguistic knowledge "elicited" in the alignment process and the use of a more refined alignment tool can solve some of the problems related to multiword alignment when it is so relevant that these alignments mirror the unity of the expression, a challenge that was addressed successfully in the Logos Model, as demonstrated next.

4 The Logos Model Approach

The Logos Model has been described with a great degree of detail in (Scott, 2003), (Scott and Barreiro, 2009), and (Scott, 2018), among others. We highlight in this paper only the SAL language and the SemTab function for the sake of illustrating how relevant they are for our approach to the processing and generation of multiwords and the establishment of bilingual and multilingual paraphrasaries.

4.1 Semantico-Syntactic Abstraction Language (SAL)

In the Logos Model, natural language is represented as a refined Semantico-Syntactic Abstraction Language (SAL), also designated as a hierarchical ontology, with categories for all parts of speech. When processing the sentence, word strings are converted into SAL patterns. SAL has four levels of abstraction: (i) a syntactic level (word class) and three levels referred to as (ii) superset, (iii) set, and (iv) subset. Figure 1 illustrates the hierarchical structure for the SAL Superset Animate-type nouns, where the Sets are in red and the Subsets are in blue. It is possible to apply the same techniques to the data, which in the Logos Model are not literal words, but SAL entities or SAL patterns. This is the reason why it makes sense to train machine learning (ML) systems to learn new SAL patterns based on alignments, instead of on the conventionally-used MT patterns.

.

[7]The Logos Model underlies both the commercial system and its open source version OpenLogos.

[8]This term has also been designated *inter alia* as "multiword lexical items", "phraseological units" and "fixed expressions", with slight variations in scope and meaning.

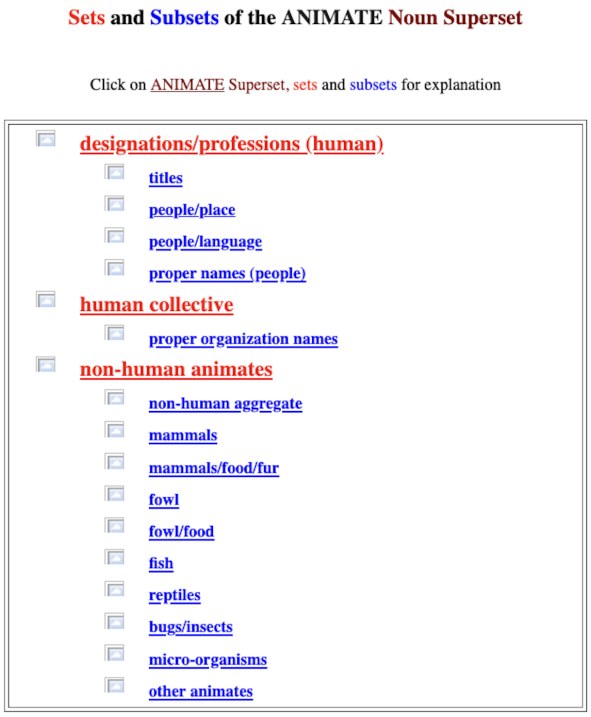

Figure 1: SAL Superset Animate-type Nouns

The objects of alignment provide a multilingual dictionary-type function, which we call **Paraphrasary**, that clarifies, simplifies, and adds precision to text, and to its translation (cf. Section 6).

4.2 The Sematic Table (SemTab)

In the Logos Model, all multiwords are represented as rules in a separate database, the Semantic Table or "SemTab", as described in (Orliac and Dillinger, 2003). In our alignment research work, we propose a methodological framework for the alignment task that relies on the use of multiwords as representation objects of alignment. The meaning is derived from the semantic processing in the SemTab function, where multiwords can be linguistically processed and translation fidelity can be improved. For example, SemTab allows distinguishing between the multiword (i) *be acquainted with* N(AN-Hum)/PRO — *conhecer* N/PRO *pessoalmente*, where the translation of the verb depends on the type of noun (human-type) and the multiword (ii) *be acquainted with* N-Abs — *estar ao corrente de* N-Abs, where the noun N is abstract (Abs) of the type "Information", e.g., a piece of news, a gossip, situation, etc. On the other hand, from the sentence (iii) *he was driving the car at full speed*, the noun *car* can be replaced by any type of concrete, vehicle: *drive* N(CO-Vehic) *at full speed* — *guiar/conduzir* N(CO-Vehic) *a toda a velocidade*. In the Logos Model, SemTab rules deploy SAL patterns or entities, such as the aforementioned N(AN-Hum), N(Abs-info-type), or N(CO-Vehic).

In the Europarl corpus, not all translations are optimal and often translational equivalents are approximate rather than exact. In example (1), the English prepositional verb *to deal with* is translated in the Romance languages as *dedicarse a* (*engage in*) in Spanish, the reflexive *s'attacher à* (*focus on/stick to*) in French, and *centrar-se em* (*concentrate/center/focus on*) in Portuguese.

(1) $_{EN}$ - our Asian partners prefer **to deal with** questions which unite us

$_{ES}$ - nuestros socios asiáticos prefieren **dedicarse a** las questiones que nos unen

$_{FR}$ - nos partenaires asiatiques préfèrent **s'attacher à** ([**a**+a]) ce qui nous unit

$_{PT}$ - os nossos parceiros asiáticos preferem **centrar-se** unicamente **n**as ([**em**+as]) questões comuns

The Logos Model allows for the application of a SemTab contextual rule, such as the one in Example 2, which is a deep structure pattern that matches on/applies to a great variety of surface structures.

(2) DEAL(VI) WITH N(questions) = S'OCCUPER DE N[9]

The differences in the translations of *deal* are related to the idiomatic ways that predicate nouns select their support verbs in different languages: *take a vow* in English, but "*make a vow*" in the Romance languages (*hacer* in Spanish, *faire* in French, and *fazer* in Portuguese). Verbal expressions such as the English prepositional verb *to deal with* take different senses (and translations) depending on contexts, typically their object or prepositional phrase complement. If the context of the verb is *to deal with questions*, as in (1), then the French translation should be *s'occuper de* (*to be busy with*). On the other hand, if the context is *he proved unable to deal with the problem*, then the translation should be the translation of its paraphrase *handle the problem*. However, if the context is *he refused to deal with the problem*, then the translation would be a translation of the paraphrase *analyse and try to solve the problem*. These different nuances are related to the ambiguity and

[9]Here we only display the comment line of the SemTab rule, not the rule itself or what it does in terms of the Logos language. The rule notation is arcane due to its numeric representation and it would take a larger effort to explain the use and meaning of the distinct codes in the Logos Model.

weakness of the verb *deal* and the different meanings of the predicate-like nouns *questions* (*issues*, *topics*, *interrogations*, etc.) or *problem* (*difficulty*, *exercise*, etc.). It is the meaning of these nouns that triggers the different translations of *deal*, just like the verb *take* will have different translations depending on the predicate noun it supports (*walk*, *responsibility*, *comfort*, etc.). Therefore, the two slightly different meanings for *problem* in the last two examples explain the distinct paraphrase: *handle*, in one case, and *analyse and try to solve*, in the other case. In the exemplified context, the SemTab rule states that when followed by the direct object noun *questions* or a noun with the same semantico-syntactic information, the verb is translated as *s'occuper de*, overriding the default dictionary translation of this verb. The power and advantage of the rule in the Logos Model with regard to non-contiguous multiwords is the ability of the MT system to recognise, analyse, and relate constituents that are apart (even far apart) in the sentence.

Former word alignment techniques, even when they contemplated multiword alignments, were unable to present a consistent and efficient solution to process non-contiguous expressions. In other words, SemTab is an effective way of analysing and translating words in context, especially when the context is remote.

In addition to the long-distance dependency capability, SemTab also allows generalising between alternative forms of the same multiword. For example, it presents the possibility of generalising translations of *take a walk* to translations of *walk*, if one of these two is found in the training corpus. Similarly, closed class items or highly frequent multiwords might be learned quickly and be translated correctly by state-of-the-art MT systems, but open class items or less frequent multiwords might present more challenging problems that can be observed in MT output, but also in non-native speakerisms, such as the choice of a support verb for a particular support verb construction (e.g., *make a visit to* N or *pay a visit to* N, which can be robustly corrected by the use of SemTab.

5 CLUE – Cross-Lingual Unit Elicitation

Under the umbrella of CLUE, we developed a set of alignment guidelines, an alignment tool, and a gold collection. For the alignment task, we used the bilingual corpora from the Europarl corpus.

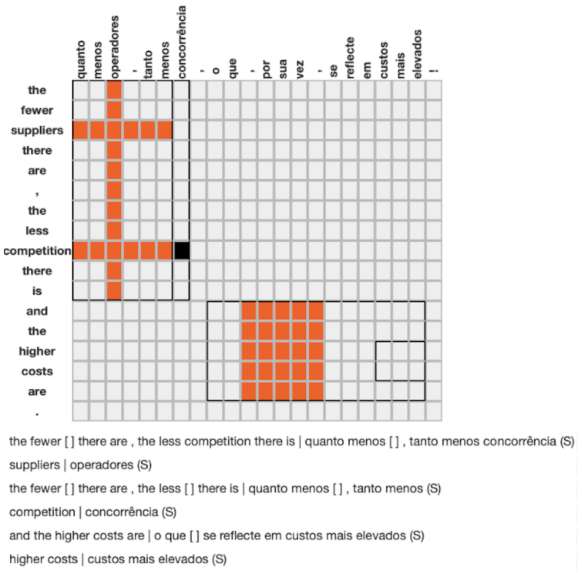

Figure 2: Alignment of comparison and metonymy *the fewer* [] *there are*, [] *the less* [] *there is* — *quanto menos* [], *tanto menos*

The Gold-CLUE was facilitated by the use of the CLUE-Aligner alignment tool. Both the Guidelines and the Gold-CLUE require revision and refinement. So far, the only revised Gold-CLUE was for the EN-PT language pair.

5.1 CLUE-Aligner

CLUE-Aligner is an alignment tool developed to annotate paraphrasing or translation units representing multiwords found in monolingual or bilingual pairs of parallel sentences (Barreiro et al., 2016). CLUE-Aligner is based on another alignment tool, Linear-B (Callison-Burch and Bannard, 2004), but it was extended in order to allow the alignment of contiguous and non-contiguous multiwords, addressing the long-distance dependency that characterises the majority of semantico-syntactic patterns.

CLUE-Aligner allows the loading of previously generated alignments (segments) for the corpora parallel sentences. During the annotation task, the annotator manually corrects any inaccurate alignments (either gathered manually or automatically), and defines the new alignments for multiwords, which represent translation (or paraphrasing) units.

Figure 2 illustrates the alignment of the non-contiguous comparison/metonymy *the fewer* [] *there are*, [] *the less* [] *there is* — *quanto menos* [], *tanto menos*, and *the higher* [] *are* — *o que se reflecte em* [] *mais elevados*. In this figure of speech,

the insertions were excluded and aligned independently: *suppliers* was aligned with *operadores*, *competition* was aligned with *concorrência*, and *costs* — was aligned with *custos*. On the CLUE-Aligner interface, in Figure 2, the linguistic annotator can immediately see the list of alignments in text format and correct any error that might have been done in the alignment task.

5.2 Gold-CLUE

The Gold-CLUE is the gold collection made of aligned multiwords resultant from our alignment task. The Gold-CLUE contemplates a set of linguistic phenomena that can be classified into four main classes: (1) lexical and semantico-syntactic challenges include multiwords, such as support verb constructions, compound/modal verbs, and prepositional predicates; (2) morphological challenges include contracted forms, lexical versus non-lexical realisation, that is, lexical items that are present in one language but not the other, such as determiners (articles and zero/missing articles), and pro-drop phenomena including subject pronoun dropping, and empty relative pronouns; (3) morpho-syntactic challenges include free noun adjuncts (noun-noun compounds); and (4) semantico-discursive challenges include emphatic linguistic constructions, such as pleonasm and tautology, repetition, and focus constructions. For lack of space in this paper to exemplify and discuss the most problematic alignment problems, and justify the annotation decisions for all the classes identified, we restrict our exemplification to class (1), specifically with support verb constructions' phenomena.

5.2.1 Support Verb Constructions

A support verb construction is a multiword or complex predicate consisting of a semantically weak verb (the support verb), and a predicate noun, a predicate adjective, or, much less frequently, a predicate adverb (*make a presentation*, *make it simple* or *go fast*) (Barreiro, 2009).[10] In the Europarl corpus, support verb constructions are either aligned with semantically equivalent single verbs (many-to-one correspondence) or with other semantically equivalent support verb constructions (many-to-many correspondence). For example, the English, French, and Portuguese prepositional transitive support verb constructions *draw a distinction* (*between*), *faire une distinction* (*entre*), and *estabelecer uma diferença* (*entre*), align with the Spanish prepositional transitive verb *distinguir* (*entre*) (*distinguish* (*between*)), as illustrated in Example (3). English and Portuguese use non-elementary support verbs *draw* and *estabelecer* (*establish*), while French uses an elementary support verb *faire* (*make*). Smaller alignments can be established between the intransitive support verb constructions *draw a distinction*, *faire une distinction*, and *estabelecer uma diferença* and the Spanish verb *distinguir*.These alignments would be necessary to translate the support verb construction when it is used intransitively.

(3) $_{EN}$ - we need **to draw a distinction between** north and south

$_{ES}$ - debemos **distinguir entre** norte y sur

$_{FR}$ - nous devons **faire une distinction entre** le nord et sud

$_{PT}$ - temos de **estabelecer uma diferença entre** norte e sul

5.2.2 Alignment Decisions

The Europarl corpus subset that we used contains several instances of non-contiguous support verb constructions. In translation, a non-contiguous expression in a source language can be maintained in the target language or replaced by an equivalent but contiguous expression that conveys the same meaning. It can also be transformed into a simpler contiguous syntactic structure, such as a single word.[11] In example (4), the non-contiguous English support verb construction *set in motion*, corresponding to the Portuguese equivalent single verb *empreender* (*undertake*), is used instead of maintaining the English structure, with a support verb construction to express a similar meaning. Both Spanish and French maintain the support verb construction (*llevar a cabo* and *mettre en chantier*), with the difference that in these languages the support verb constructions are contiguous and have no insertions. The existence of a non-contiguous expression in one of the sentences of the language pair causes additional complexity to the alignment task, which we are able to solve with the Logos approach.

[10] For a broader definition of support verb and support verb construction, see also (Jespersen, 1965), (Erbach and Krenn, 1993), and (Butt, 2010), among others.

[11] In some cases, the verbal expression is always expressed in the form of a support verb construction (cf. non-elementary support verb construction *play* [INSERTION] *role*) because there is no suitable corresponding single verb, which is semantically equivalent to the support verb construction (Barreiro, 2009).

(4) $_{EN}$ - many member states thus have the major task of **setting** structural reform **in motion**

$_{ES}$ - he aquí por lo tanto una tarea de gran importancia para que numerosos estados miembros **lleven a cabo** reformas estructurales

$_{FR}$ - il y a donc là une táche considérable pour beaucoup d'états membres, celle de **mettre en chantier** des réformes structurelles

$_{PT}$ - há, portanto, uma tarefa importante para muitos estados-membros em **empreender** reformas estruturais

To sum up, non-contiguous support verb constructions processing, recognition, and translation is a challenging problem when using alignment techniques and some previous methodologies violate the intrinsic property of the unit as an atomic group of elements when aligning them individually or when not respecting the correct boundaries of the unit. However, inspired by the Logos Model, we came up with a way of aligning them successfully in CLUE-Aligner. CLUE-Guidelines propose that individual word alignments should not be annotated inside the support verb construction block. There is no linguistic motivation to align the canonical form of the support verb and do a separate alignment for the optional and variable parts of the construction. However, when inserted constituents are equivalent in the source and target languages, these constituents are aligned as separate elements, outside the multiword unit.

Among several other somehow arbitrary decisions, we have not addressed whether the alignment of non-contiguous support verb constructions with pronominal insertions should be aligned. Would it be pragmatically justified the alignment of, for example, the expression *setting* PRON-*it in motion*? Probably, yes. Although, from a practical point of view, the alignment of this phrase can be justified, it needs to be tested what is pragmatically more adequate for a particular application, the inclusion of insertions or no inclusion of insertions of each grammatical category. For example, the alignment of pronominal elements, where there is a pronoun in the source language and a lexical element in the target language (or vice-versa), may be correct from a point of view of a text that needs to be analysed contrastively, but this does not teach correctly an MT system, and therefore, should be left out of the training data. On the other hand, the alignments where both source and target languages contain equivalent pronominal alignments, represent good training data. With regard to adverbs, they are free modifiers and normally less polysemous and less ambiguous than nouns, verbs, and adjectives, which makes the task easier for humans and machines. The alignment of insertions in a non-contiguous multiword unit needs to be further discussed for each particular application, due to considerations related to the word order of the insertions for each language, among others.

5.2.3 Methodology

In order to achieve a provisional first round of results, a polyglot linguist, with knowledge of the four languages covered in this study, annotated manually the total of 2,400 sentence alignments (400 x 6 language pairs) and built the CLUE-Guidelines based on linguistic knowledge as processed in the Logos Model, paying special attention to multiwords and other translation units. From the dataset of 400 sentences of the corpus, for the EN-PT language pair, a total of 3,700 multiword alignments were collected. They all represent candidates for entries in our Paraphrasary. Table 1 shows some examples.

Sentence Pair #	English– Portuguese
4	have [ ] margin for discretion ter [ ] margem de discricionalidade
181	between [ ] and [ ] million people entre [ ] e [ ] milhões de pessoas
207	have not [ ] been in favour of não se mostraram favoráveis a
237	would [ ] mainly focus on visa
279	cross - border services serviços além fronteiras
307	before [ ] even antes ainda de
308	what must underpin que deve subjazer a
316	avenues which could be explored pistas a seguir

Table 1: EN-PT Alignments

6 Multilingual Paraphrasary

Our research on paraphrasing applications shows that both monolingual and multilingual paraphrase generation require the development of paraphrasaries. Paraphrasary is a new concept of organising linguistic data in a repository (or several repositories), which can grow into a large body of paraphrastic knowledge. It is a database of multiword entries listed alphabetically validated by a linguist after these multiwords have been aligned

during the alignment task. For example, the alignment 181 in Table 1, *between* [ ] *and* [ ] *million people — entre* [ ] *e* [ ] *milhões de pessoas*, can enter the Paraphrasary via a SemTab-type rule that allows generating a large number of instances. Example (5) shows how the alignment can become much broader by using some constraints, such as [Num], a numeric expression.

(5) between [NUM] and [NUM] N = entre [NUM] e [NUM] de N

Via the power of generalisation that SAL categories allow, an alignment pair gathered from the corpus can be used in the generation of thousands of multilingual paraphrases. The development of paraphrasaries is, therefore, the kick-start of a paraphrasing tool.

7 eSPERTo Paraphrasing System

The eSPERTo paraphrasing system is an online platform[12] that allows rewriting different kinds of expressions using the NooJ linguistic engine (Silberztein, 2015; Silberztein, 2003). (Barreiro et al., 2022) present an overview of the system and lexicon-grammar resources that allow for the easy paraphrasing of constructions involving human intransitive adjectives, and also predicate nouns with support verbs *fazer* (do) and *ser de* (be of).

In Figure 3, we illustrate a simple example of using the multilingual paraphrasary to translate multiwords of a sentence in Portuguese into different paraphrases in English. eSPERTo uses grammars that identify multiwords in a source language, such as *constitui uma provocação* (literally, *it constitutes a provocation*) in Portuguese. When clicking on this multiword, the text changes to green and the translations of the multiword appear in a drop-down list. For the Portuguese multiword, eSPERTo shows 3 paraphrases in English: *is provoking*; *it is a public outrage*; and *is provocative*. The suggested translations were paraphrasing pairs in Gold-CLUE and entries in the (EN-PT) Paraphrasary where the same multiword in Portuguese were translations of different multiwords in English. Therefore, as illustrated in Figure 4, the multiword in Portuguese is represented as input of the graph by its constituents: `<constituir,V>` will match any form of the ver *constituir* in the text, and `<N+EN>` will match *provocação*, which will be stored in the variable `$Npred`. Then, the top path of the graph will output *it is a public outrage* whereas the bottom path will output the translation of *provocação* stored in the variable `$Npred` - `$Npred$EN` - preceded by *is*.

[12]eSPERTo stands for 'System for Paraphrasing in Editing and Revision of Text'. The system can be tested at: https://esperto.hlt.inesc-id.pt/esperto/esperto/demo.pl

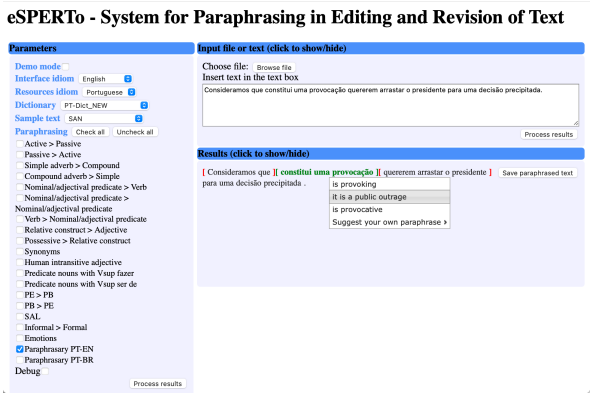

Figure 3: Translating Portuguese expression to English paraphrases in eSPERTo

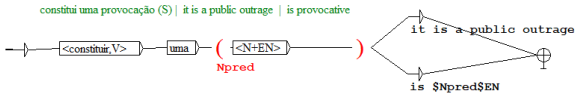

Figure 4: Paraphrasary grammar: *constituir uma provocação*

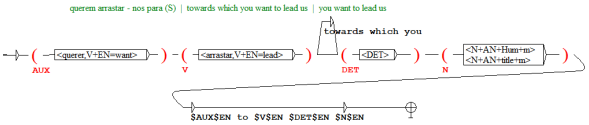

Figure 5: Paraphrasary grammar: *arrastar* [N+AN+Hum] *para*

In Figure 5, we illustrate the simplified paraphrasary grammar that allowed for the generation of the distinct translations into English of the multiword [QUERER] `arrastar` [NP+AN+HUM]. Each multiword constituent will be stored in different variables (`$AUX`, `$V`, `$DET`, and `$N`) in order to use them to translate them (respectively, `$AUX$EN`, `$V$EN`, `$DET$EN`, and `$N$EN`)). This grammar uses the SAL codes `+AN`, `+hum`, and `+title` to restrict the noun in the noun phrase to be human-type.

These grammars take advantage of the multilingual nature of NooJ and other properties included in the dictionary entries, but the full integration of the paraphrasary into the eSPERTo system is still under progress as it is not yet clear what is the best way of tackling this integration.

8 Conclusions and Future Work

An MT program that offers correct translation of multiwords, either via direct phrasal translation or via paraphrases demonstrates how applied linguistic knowledge helps improve output quality. In this paper, we reassessed the concept of alignment and justified our need for linguistic precision in the alignment task via the analysis of the complexity of multiwords, crucial in obtaining high-quality MT. We, then, described the Logos Model approach to the processing of multiwords and showed how the SemTab function can complement our alignment proxy. We presented the Cross-Lingual Unit Elicitation (CLUE) approach, which is based on the CLUE-Guidelines. These guidelines cover important linguistic phenomena that were left undiscussed in previously presented guidelines. With a special focus on multiwords, we added an extra level to the alignment process, with the hypothesis that this contributes to a deeper scientific process of alignments' annotation. The CLUE-Guidelines led to the gold data set Gold-CLUE, which includes efficiently-aligned non-contiguous multiwords. The linguistic analysis undertaken to establish the Gold-CLUE has allowed some advance in the establishment of a standard for the recognition, processing, translation, and evaluation of multiwords. Some limitations of previous alignment tools (and tasks) motivated the development of the CLUE-Aligner. All alignments were made by using this alignment tool, but only the EN-PT data set was reviewed. We are still in the process of reviewing the remaining language pairs. From the EN-PT Gold-CLUE, we selected which entries would go into the multilingual Paraphrasary, either as simple entries or comment lines for rules. The collection of multilingual paraphrasaries is used in the eSPERTo paraphrase generation system, as exemplified in the paper.

It is important to develop a more robust resource, with a joint discussion of the most challenging linguistic phenomena of the CLUE-Guidelines to improve areas that are known to be non-consensual, a more refined methodology, which supports linguistic phenomena in the four classes identified in this work. All data should be multi-annotated by more than two annotators so that no multiword is left unidentified and the coverage of multiword alignments in the data is complete and there are no disagreements between annotators.

Finally, due to the extent of the work at hand, most linguistic phenomena were left undiscussed. A detailed analysis of these phenomena is important for the improvement of the alignment techniques and for the enhancement of the quality of MT. Our goal is the development of an MT model that integrates linguistic knowledge where all sorts of multiwords are included at the alignment level and feed the paraphrasaries that set into motion and enrich the translation engine.

Acknowledgements

This article was also supported by Fundação para a Ciência e a Tecnologia (FCT), under projects PEst-OE/EEI/LA0021/2011, and also by the EC CIP METANET4U project #270893. We thank the anonymous reviewers for their careful reading of the manuscript, and their insightful and relevant comments and suggestions that helped improve this paper.

References

Baldwin, Timothy and Su Nam Kim. 2010. Multiword Expressions. In Indurkhya, Nitin and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Barreiro, Anabela and Fernando Batista. 2016. Machine Translation of Non-Contiguous Multiword Units . In *Proceedings of DiscoNLP 2016* , pages 22 – 30, San Diego, California, June. Association for Computational Linguistics.

Barreiro, Anabela, Annibale Elia, Johanna Monti, and Mario Monteleone. 2010. Mixed up with machine translation: Multi-word units disambiguation challenge. In *Proceedings of the ASLIB Conference: Translating and the Computer*, London, United Kingdom, november.

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

Barreiro, Anabela, Johanna Monti, Brigitte Orliac, and Fernando Batista. 2013. When Multiwords Go Bad in Machine Translation. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Machine Translation Summit XIV*.

Barreiro, Anabela, Johanna Monti, Brigitte Orliac, Susanne Preuss, Kutz Arrieta, Wang Ling, Fernando

Batista, and Isabel Trancoso. 2014. Linguistic Evaluation of Support Verb Constructions by OpenLogos and Google Translate. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 35–40. ELRA.

Barreiro, Anabela, Tiago Luís, and Francisco Raposo. 2016. CLUE-Aligner: An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units . In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 7–13, Portorož, Slovenia, May. ELRA.

Barreiro, Anabela, Cristina Mota, Jorge Baptista, Lucília Chacoto, and Paula Carvalho. 2022. Linguistic Resources for Paraphrase Generation in Portuguese: a Lexicon-grammar Approach. *Language Resources and Evaluation*, 56(1):1–35, March.

Barreiro, Anabela. 2009. *Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation: Universidade do Porto*. Ph.D. thesis, Tese de Doutoramento.

Barreiro, Anabela. 2015. Tradução automática, ma non troppo. *Oslo Studies in Language*, 7(1):207–222.

Brown, Peter F., 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.

Butt, Miriam. 2010. The Light Verb Jungle: Still Hacking Away. *Complex Predicates: Cross-Linguistic Perspectives on Event Structure*, 04.

Callison-Burch, Chris and Colin Bannard. 2004. Improving statistical translation through editing. European Association for Machine Translation (EAMT-04) Workshop. In *European Association for Machine Translation*.

Erbach, Gregor and Brigitte Krenn. 1993. Idioms and support-verb constructions in HPSG. Technical Report Report Nr. 28, Computerlinguistik an der Universität des Saarlandes (CLAUS), location=Saarbrücken: Universität des Saarlandes,.

Graça, João, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro. 2008. Building a Golden Collection of Parallel Multi-Language Word Alignment. In Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. ELRA.

Hearne, Mary and Andy Way. 2011. Statistical Machine Translation: A Guide for Linguists and Translators. *Language and Linguistics Compass*, 5(5):205–226.

Jespersen, Otto. 1965. *A modern English grammar on historical principles*. George Allen and Unwin Ltd.

Koehn, Philipp. 2005. EuroParl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT.

Kordoni, Valia and Iliana Simova. 2014. Multiword Expressions in Machine Translation. In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Lambert, Patrik, 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.

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

Orliac, Brigitte and Mike Dillinger. 2003. Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA, September 23-27.

Scott, Bernard and Anabela Barreiro. 2009. OpenLogos MT and the SAL Representation Language. In Pérez-Ortiz, Juan Antonio, Felipe Sánchez-Martínez, and Francis M. Tyers, editors, *Proceedings of the First International Workshop on Free-Open-Source Rule-Based Machine Translation*, pages 19–26, Alicante, Spain.

Scott, Bernard E. 2003. The Logos Model: An Historical Perspective. *Machine Translation*, 18:1–72.

Scott, Bernard. 2018. *Translation, Brains and the Computer: A Neurolinguistic Solution to Ambiguity and Complexity in Machine Translation*. Springer Publishing Company, Incorporated, 1st edition.

Silberztein, Max. 2003. *NooJ manual*.

Silberztein, Max. 2015. *La formalisation des langues: l'approche NooJ*. Collection science cognitive et management des connaissances. ISTE éd.