Building a golden collection of parallel Multi-Language Word Alignments

João Graça, Joana Paulo Pardal, Luísa Coheur and Diamantino Caseiro

L^{2}F – INESC-ID Lisboa/IST
{javg,joana,lcoheur,dcaseiro}@l2f.inesc-id.pt

Abstract

This paper reports an experience on producing manual word alignments over six different language pairs (all combinations between Portuguese, English, French and Spanish) (Graça et al., 2008). Word alignment of each language pair is made over the first 100 sentences of the common test set from the Europarl corpora (Koehn, 2005), corresponding to 600 new annotated sentences. This collection is publicly available at `http://www.l2f.inesc-id.pt/resources/translation/`. It contains, to our knowledge, the first word alignment gold set for the Portuguese language, with three other languages. Besides, it is to our knowledge, the first multi-language manual word aligned parallel corpus, where the same sentences are annotated for each language pair. We started by using the guidelines presented at (Lambert et al., 2005) and performed several refinements. Some due to under-specifications on the original guidelines, others because of disagreement on some choices. This lead to the development of an extensive new set of guidelines for multi-lingual word alignment annotation that, we believe, makes the alignment process less ambiguous. We evaluate the inter-annotator agreement obtaining an average of 91.6% agreement between the different language pairs.

1. Introduction

The concept of word alignment, introduced in (Brown et al., 1990) for statistical machine translation, consists in an object representing which words in a source language correspond to translations of other words in a foreign language, between two parallel sentences. A word alignment can be seen as a matrix of $n * m$ entries, where n is a position on the source sentence, and m is a position on the target sentence. An entry in that matrix $a_{n,m}$ specifies if the word at position n is part of a translation of the word at position m on the target language.

Figure 1 shows a possible word alignment between the English sentence $_{EN}$"*i did receive the request you sent me.*" and the Portuguese sentence $_{PT}$"*recebi de facto o pedido que me dirigiu*". A word alignment may contain a single link between two words, normally referred as a 1-1 link, meaning that the words are translated of each other, or n-m block, meaning that an expression is the translation of another expression. These block may be discontinuous as the order on which words appear in both languages may be significantly different. Furthermore, each alignment point may be marked as *sure*, meaning that both words are a translation of each other in any context, or *possible*, meaning that the words are a translation of each other in some contexts. The full semantics of the alignment points will be detailed further ahead.

These characteristics of word alignment makes it a very difficult task, and many works have addressed this issue in recent years.

Although the main use of word alignments is statistical machine translation, directly on a translation system as originally proposed in (Brown et al., 1990), as a primary resource for phrase base machine translation (Och and Ney, 2004) or syntax base machine translation (Galley et al., 2004), other applications of word alignments have been suggest in recent literature such as annotations' projections or extraction of bilingual lexica.

In fact, in the last years, the increase of freely available digitalized parallel texts led to a huge development in statistical machine translation systems. Many workshops and evaluation tasks were dedicated to multi-language word alignment[1], as well as some projects. For example, the Blinker project[2] aimed at aligning words between French and English texts. Also, many word alignments guidelines (Melamed, 1998; Och and Ney, 2000; Lambert et al., 2005; Kruijff-Korbayová et al., 2006) have been suggested.

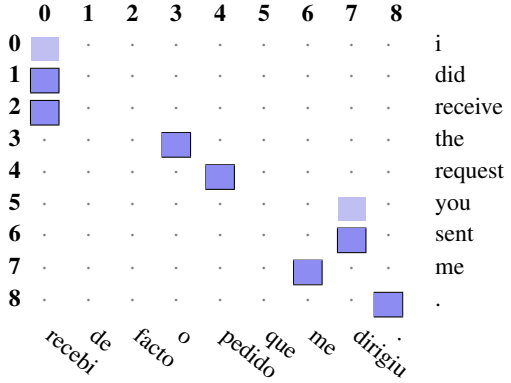

Figure 1: Word alignment between a Portuguese and an English sentence. Full dark blue box indicates a sure alignment point. Empty light blue box represents a possible alignment point.

Nevertheless, despite the growing number of available multi-language sentence aligned parallel corpora and word alignment tools, the number of publicly available manual word alignments is restricted to a few language pairs. Manual word alignments are a much desired resource, since they allow the evaluation of word alignment algorithms, training of supervised and semi-supervised algorithms, and tuning of parameters for all kinds of models. For instance, using posterior decoding instead of the usual Viterbi decoding has been showed to increase the quality of word alignment al-

[1] For instance, `http://www.cse.unt.edu/~rada/wpt/`, `http://www.statmt.org/wpt05` or `http://www.lpl.univ-aix.fr/projects/arcade`.

[2] `http://nlp.cs.nyu.edu/blinker/`.

gorithms. However, this decoding type requires the tuning of a threshold, requiring some amount, even if small, of annotated data.

This work provides six gold alignments sets, freely available at `https://www.l2f.inesc-id.pt/resources/translation/`. To the best of our knowledge, four of them are the first freely available for their language pairs (PT-EN, PT-ES, PT-FR, ES-FR), one for an existing language but a different domain (EN-FR), the Europarl corpus, since the existing and freely available ones are based on the Hansard corpus (Och and Ney, 2000) and on the Bible (Melamed, 1998); and a new set for the Europarl corpus (EN-ES). Alignments results can be directly comparable since they are performed over the same sentences and using the same alignment guidelines, making it easier to compare methods across language pairs.

Besides the gold alignment sets, this work contributes by providing guidelines for multi-language manual word alignments (also available at the same site). These guidelines were evaluated twice for inter-annotator agreement and the last evaluation resulted in an average result of 91.6% of agreement. The guidelines were further improved after the last evaluation.

We also develop a specific metric for word alignments where one can mark an alignment as possible and sure, that allows to rank different types of errors performed by the annotators.

The paper is organized as follows: in section 2. we described the used corpus; in section 3. we describe the alignment process and present a brief overview of some of the guidelines; in section 4. we describe the evaluation process and in section 5. we presents some statistics of the produced set. Finally, in 6. we conclude and provide direction for future work.

2. Corpus

We used the publicly available Europarl Corpus (Koehn, 2005) that contains proceedings of the European parliament in the different official languages.

The golden collection is built over the first 100 sentences of the common test set defined in (Koehn et al., 2003), which is taken from Q4/2000 portion of the data (2000-10 to 2000-12). The common test set can be download from Europarl archives[3]. The common test set is already tokenized and lowercased.

Table 1 presents some general statistics about the gold standard corpus.

3. Building the golden collection

Our starting point were the guidelines developed in (Lambert et al., 2005) for Spanish/English: general alignment rules were defined and then refined according to particular situations. The main goal was to leave the manual alignment process as unambiguous as possible. During this process a detailed manual alignment guideline was produced.

Annotations were performed by using the annotation tool described in (Callison-Burch et al., 2004). The tool is very intuitive and allows the annotation of possible and sure alignments as required. A very useful feature consists in associating a comment with each word alignment. Figure 2 shows a screenshot of the tool.

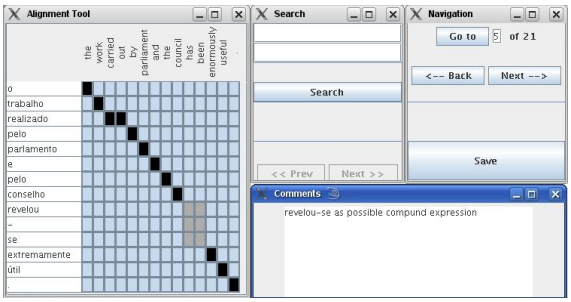

Figure 2: Screenshot of the tool used to produce the alignments. On the right: alignment for a given sentence. Black points correspond to a sure alignment point. Grey points correspond to a possible alignment point. On top the search interface, allows to find a sentence containing a given word. Also on the top the navigation toolbar that allows to navigate over different sentences. Left bottom, the comment window, allows to add comments associated with each sentence.

In what follows, all the examples that can be illustrated with English as one of the languages are preferred against the other possible pairs to ease the reading of the paper.

3.1. The alignment process

The process begun by annotating the first 20 sentences of each language pair using the existing guidelines (Lambert et al., 2005) by two annotators (h_1, h_2). During this first phase a new guideline was created containing many refinements to the existing ones, adding several examples, and changing some decisions.

	h_1	h_2
h_3	EN-PT	EN-FR PT-FR
h_4	EN-ES PT-ES	ES-FR

Table 2: Language pairs given to each annotator.

The second step included the four annotators and consisted in using the produced guidelines to annotate the next 20 sentences of the test set. In this step, each language pair was annotated twice (by two different annotators as shown in Table 2). The resulting alignments were compared and the differences discussed. The results of this process are described in the evaluation section. The feedback of the previous step was incorporated into the guidelines.

In the next step, each annotator was given 3 sets of 20 sentences (40-60) in different languages to be annotated using the improved guidelines. Again, each set was annotated twice. These alignments were the ones used to report a 91.6 % inter annotator agreement. The differences were corrected and the guidelines were again improved.

The last step was to annotate the remaining 40 sentences.

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

Number of sentences	100			
Language	English	Portuguese	French	Spanish
Words	1072	1131	1227	1106
Types	466	513	474	472
Aveg. Sent. size	10.72	11.31	12.27	11.06

Table 1: Test Corpus information

Each annotator was given three sets of 20 sentences. Table 3 resumes the annotation procedure.
At the end one of the annotators reviewed all 100 sentences sets for all language pairs to correct existing differences due to guidelines changes or specializations.

	1-20	20-40	40-60	60-80	80-100
EN-PT	h_1	h_1&h_3	h_1&h_3	h_1	h_3
EN-ES	h_1	h_1&h_4	h_1&h_4	h_1	h_4
EN-FR	h_2	h_2&h_3	h_2&h_3	h_2	h_3
PT-ES	h_1	h_1&h_4	h_1&h_4	h_1	h_4
PT-FR	h_2	h_2&h_3	h_2&h_3	h_2	h_3
ES-FR	h_2	h_2&h_4	h_2&h_4	h_2	h_4

Table 3: Annotations performed by each annotator. Annotations from sentence 20 to 60 were done twice for evaluation purposes. The guidelines were improved after each step.

We have to mention that in the early beginning of the alignment process, we found that aligning the same sentence across language pairs at the same time simplified the task, as it allowed to easier decide which were the minimal annotation units, since typically these were shared between different language pairs. These was the default annotation procedure that all annotators used. On the final correction of the 100 alignments, each sentence was done in turn for all language pairs to increase consistency.
When creating the final version of the golden collection, an interesting situation occurred that illustrates the differences in the writing style used by different translators: some words have a S-alignment in two languages but on a third language they only align as a Possible. For example, $_{EN}$"*another*" can be translated into $_{ES}$"*otra*" and $_{PT}$"*mais uma*", but $_{ES}$"*otra*" cannot be translated into $_{PT}$"*mais uma*" in all contexts (see Figure 3).

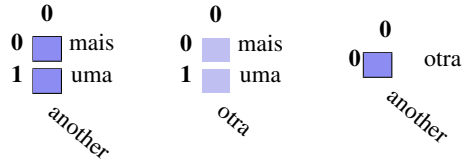

Figure 3: Alignments strength is not regular across languages.

3.2. Some guidelines

In the following we present some of the guidelines that drive the annotation process.

3.2.1. S-alignment and P-alignment semantics

Regarding the use of S-alignments and P-alignments we decided to give them the following semantics: an S-alignments is used when a translation is possible in every context, such as compound expressions that are always interchangeable (see again figure 1). On the other hand, we considered P-alignments when a translation was possible in certain contexts or in the presence of functional words that might be absent in one of the languages of a language pair. Notice that we do not use P-alignments for annotators disagreement. As we want guidelines to be as unambiguous as possible, if annotators disagree, they need to come up with an annotation solution in order to provide a precise guideline under that disagreement topic, as explained previously.

3.2.2. Aligning incorrect or incomplete translations

The first option we took was not to align the incorrect/incomplete parts of the translations. That is, if a sequence of words only appears in one language and has no correspondence in the others, we do not align it. For instance, consider the following sequences: $_{PT}$"*já foi distribuída*" and $_{EN}$"*have been distributed*". In this example, the word $_{PT}$"*já*" ($_{EN}$"*already*") has no correspondence in the English sentence and a correct translation would be something like $_{EN}$"*have already been distributed*"; as so, it is not aligned. For another example, consider again Figure 1: the Portuguese words $_{PT}$"*de facto*" (which can be translated as $_{EN}$"*indeed*") was left unaligned because it has no counterpart in the English sentence.
The difficulty related with these decisions is shown here too: it is arguable if the $_{EN}$"*did*" is used to highlight the fact that the report was distributed, as the particle could be omitted. In that case, $_{PT}$"*de facto*" could be aligned as Possible with it.
The same happens with words which are incorrectly translated: they are left non aligned. Although they do not change the semantic of the sentence, they would put weight on an incorrect translation pair. An example from the golden collection are the following chunks: $_{PT}$"*este semestre*", $_{ES}$"*este otoño*", $_{EN}$"*this autumn*". A correct translation of the Portuguese $_{PT}$"*este semestre*" would be $_{EN}$"*this semester*". As shown on Figure 4, we did not aligned Portuguese with English nor Portuguese with Spanish. However it was aligned between English and Spanish as they both use the same expression: $_{ES}$"*otoño*" and $_{EN}$"*autumn*" that, can be aligned in all contexts.

3.2.3. Aligning compound expressions

Compound expressions are aligned in different ways:

- If the compound expression is an exact translation in the sense that it can be used as a translation in

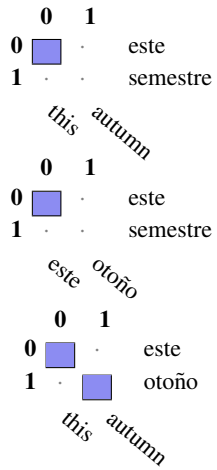

Figure 4: Incorrect word translation left unaligned.

many other contexts, we translate it as a block of S-alignments. For example $_{PT}$"*está carregada*" which literally translates to $_{EN}$"*is loaded*" but is better translated by the used expression $_{EN}$"*bears the weight*". Note that a compound expression may not behave as such in different language pairs. For instance, the translation of $_{PT}$"*está carregada*" translates in a one to one monotonic mapping to $_{ES}$"*está cargada*" (see Figure 5).

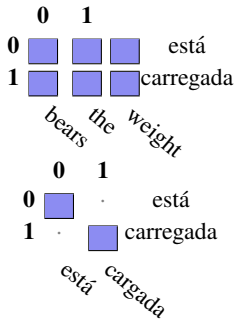

Figure 5: Compound expressions: different behavior in different pairs.

- If compound expressions are a translation of each other, but only in some contexts they are aligned as a block of possible alignments. Notice that within a block of possible alignments, it is possible to have some sure alignments. Figure 6 shows an example where the expression $_{ES}$"*orden del dia*" is aligned as a P-block with $_{PT}$"*ordem dos trabalhos*" and the two first words are aligned as Sure

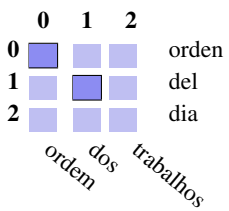

Figure 6: Possible compound term with sure links.

- In the case that a compound term has also a word to word correspondence between each language, then we align it as a block of S-alignments. For instance, consider the compound $_{PT}$"*médio oriente*". It translates as $_{EN}$"*middle east*", $_{ES}$"*oriente medio*", $_{FR}$"*moyen-orient*", but there is also a one-to-one word alignment. We take it as S-alignments where there are no hyphen, and as a block when there is (check Figure 7).

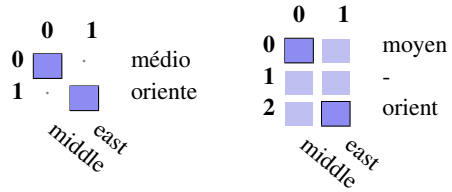

Figure 7: Compound expressions: when there are hyphens a block is needed.

Notice that non contiguous compound expressions can also occur and previous guidelines are followed. For example, consider the following sequences: $_{ES}$"*no ayuda$_x$ en absoluto*", $_{EN}$"*in no way helps$_x$*", $_{PT}$"*em nada contribui$_x$*" and $_{FR}$"*ne favorise$_x$ en rien*". In these situations, expressions are capture as possible expressions, but words appearing in $_x$ are not part of the alignment.

3.2.4. Linguistic details

Here we describe some decisions regarding certain morpho-syntactic categories:

- **Prepositions and Pronouns**: sometimes different writing styles lead to the occurrence of preposition or pronouns in only one language. In these situations they are taken as P-alignments with the corresponding noun/verb. Figure 8 shows the case where $_{EN}$"*presidency*" is translated into $_{PT}$"*da presidência*" and into $_{ES}$"*de la presidencia*"

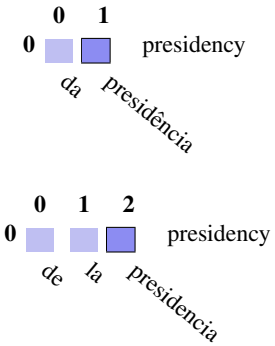

Figure 8: Compound expressions: different behavior in different pairs.

- **Contractions**: contractions are aligned as sure alignments to the not contracted parts in the other language. For instance, considering the preposition contraction $_{PT}$"*da*", which corresponds to $_{PT}$"*de a*" (and $_{ES}$"*de la*"), it is aligned as an S-alignment – one to one – to $_{EN}$"*of the*" (see Figure 9). In the case that a part of the contraction is missing in the other language we still consider an S-alignment.

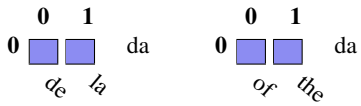

Figure 9: Contractions are aligned as S-blocks.

- **Date and Time** follow the same rules as fixed expressions. Figure 10 shows an example of a constant difference between dates description in Portuguese and English where Portuguese determiners are used. This should be aligned with a possible link. Also when translating times, we have Possible links connecting the elements that are omitted across the languages: English uses "PM", while in Portuguese the "h" between the number of hours and minutes refers to the omitted word "horas" (in English "hours"), while the Spanish includes the word "horas" (in English "hours").

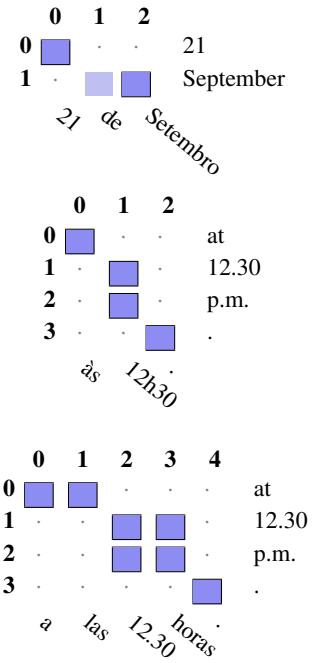

Figure 10: Alignment of dates and times.

- **Acknowledgments** are also treated as individual words whenever it is possible to keep the alignment blocks minimal. Figure 11 shows the alignments taken from the corpus between $_{PT}$"*muito obrigado*" and the equivalent expression in English: $_{EN}$"*thank you very much*".

Figure 11: Acknowledgment expressions.

- **Wrong punctuation** is aligned with possible when a different symbol is used but means the same in the current context as shown on Figure 12.

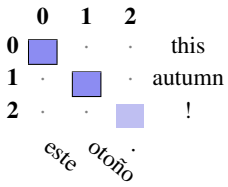

Figure 12: Wrong Punctuation.

3.2.5. Language-specific phenomena

- **Special constructions**: Certain types of ambiguity appear whenever we are aligning some constructions from some language with all the other languages. Examples of this are the English possessives like in $_{EN}$"*John's car*" (see Figure 13); or the French constructions like in $_{FR}$"*il n'y a pas*" (see Figure 14).

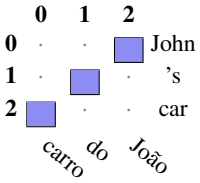

Figure 13: Contraction linked with head element example: common construction in English.

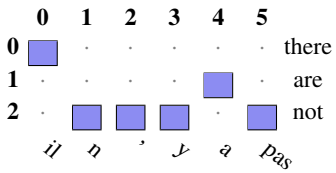

Figure 14: Negation particles: common construction in French.

- **Gender and Number variation**: In the vicinity of words that translate into words with different gender or number, articles, pronouns and adjectives also differ. For example, the noun phrase $_{ES}$"*el único obstáculo*" translates into $_{PT}$"*o único obstáculo*" which is aligned word by word, but can also be translated into $_{PT}$"*a única barreira*". In the later case, that means that $_{ES}$"*el*" that, as word, translates into $_{PT}$"*o*" is aligned with $_{PT}$"*a*" and the same happens with $_{ES}$"*único*" that is translated into $_{PT}$"*única*" (see Figure 15). The same occurs with number variations, specially in fixed expressions. We decided to ignore this kind of variations and align the words as Sure. To cope with this phenomena, another two different approaches could be used: align as Possible or create a new type of alignment ("Sure-Gender", for example) to represent this situations.
- **Spanish punctuation** questions and exclamations are marked by adding the inverted symbol as a mark at the

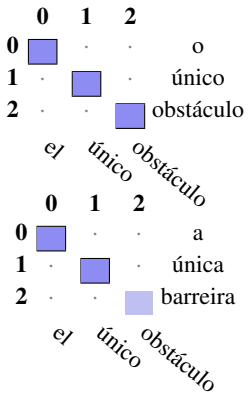

Figure 15: Gender variation is ignored.

beginning of the sentence. These pairs are translated as sure and as an indivisible token (Figure 16).

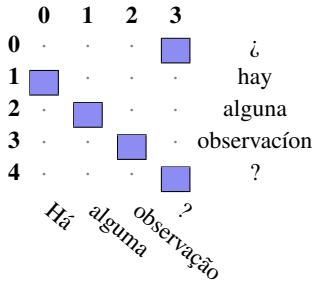

Figure 16: Spanish question mark as indivisible symbol.

4. Evaluation

While building a gold collection for word alignment each annotator has to decide for each two words in a sentence if they should be aligned (sure or possible) or unaligned. Most of the word pairs are unaligned, since in most cases each word aligns to another word in the corresponding language. Since this is a structured problem one cannot measure inter-annotator agreement by counting how many times the annotators agree for each possible decision link, since this value would be highly optimistic and overwhelmed by agreement on unaligned links. So we only consider alignments links where at least one of the annotators decided to place an alignment.

In order to measure agreement we studied two approaches. For illustration proposes, take the example from Figure 17 and ignore by now that there are sure and possible alignment points.

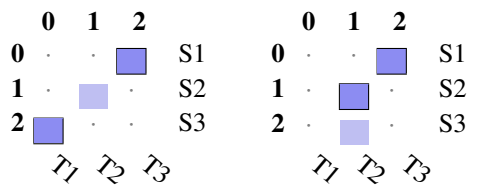

Figure 17: Two example annotations perform by different annotators.

One could calculate the disagreement by dividing the number of points in which the annotators agree by the total numbers of points that both annotators aligned. So agreement would be calculate by $\frac{|I_{1-2}|}{|A1|+|A2|}$, where $|I_{1-2}|$ is the number of alignment points in which the annotators agree and $|A_x|$ is the total number of aligned points from annotator X. Using this metric the inter-annotator agreement for this example would be of 50%. This metric is pessimistic in the sense that in assumes that the errors are independent of each other. Another approach used previously in other annotation projects (Melamed, 1998; Kruijff-Korbayová et al., 2006) is to count twice the points in which the annotators agree. This metric is calculated by $\frac{2*|I_{1-2}|}{|A1|+|A2|}$. Using this metric the inter-annotator agreement for this example would be of 66%. In our opinion this a more realistic metric, since it double counts the correct points as well as the incorrect points, and is the one we adopt in our work. However this metric does not take into account the types of alignment points. We think that different alignment errors have different weights. So it is worst to have a Sure point marked by one annotator and not aligned by the other, than to have a Sure point aligned by an annotator and aligned as Possible by the other.

To cope with this differences we defined the following agreement/disagreement types:

- Strong agreement (SA), which means that both annotators had the same alignment marked as sure or possible;
- Weak agreement (WA) meaning that one annotator had marked the alignment as sure and the other as possible;
- Weak disagreement (WD) meaning that one annotator had marked an alignment as possible and the other did not mark it;
- Strong disagreement (SD) where one annotator had marked the alignment with sure and the other did not mark it.

The percentages for each of the groups are calculated as follows:

- *Strong Agreement* - $\frac{2*(|I_{s-s}|+|I_{p-p}|)}{|A_1|+|A_2|}$
- *Weak Agreement* - $\frac{2*(|I_{p-s}|+|I_{s-p}|)}{|A_1|+|A_2|}$
- *Weak Disagreement* - $\frac{2*(|I_{p-0}|+|I_{0-p}|)}{|A_1|+|A_2|}$
- *Strong Disagreement* - $\frac{2*(|I_{s-0}|+|I_{0-s}|)}{|A_1|+|A_2|}$

Where:

- $|A_x|$ means the number of element in the gold set produced by annotator X;
- $|I_{x-y}|$ means the number of points in the intersection of alignment of type X and Y (sure, possible, null) on each set.

Notice that if we collapse the Strong/Weak distinction we get the metric used on previous manual word alignment projects. In fact we also present results on this metric in the following section to make a direct comparison with the results from previous work.

4.1. Inter-Annotation Agreement

In order to evaluate the developed guidelines we performed two inter-annotator agreement evaluations on different parts of the final gold sets. The first evaluation was performed using the first version of the guidelines in the following way: each one of the four annotators (h_1 to h_4) was asked to read the guidelines. In this phase two of the annotators (h_3 and h_4) had no experience with annotation and had never seen the guidelines before. Then each annotate sentences 21 to 40 of three different language pairs. This produced two distinct annotations for each language pair. Table 4 presents the results of this evaluation using the metric presented above.

After performing this evaluation the annotators gathered and analyzed the differences between their annotations. This led to a refined version of the guidelines and to a corrected version of sentences 21-40 which where then considered final versions. It is interesting to mention that most of the errors found on this phase were related with different interpretation of the current guidelines. For instance, most of the Sure/Empty pairs were related to a misleading indication of the guidelines concerning the annotation of compound nouns. Some annotators considered that a compound noun should always be annotated as a sure block, while others consider that this was only the case if the meaning of the compound noun could not be subdivided into the parts, in which case only the parts would be aligned to each other. Another recurrent problem was the alignment of a noun as a sure or possible alignment. It was not clear in the guidelines how they should handle these cases. This led to an heuristic where one uses the synonyms of each word on both languages. If they have a big intersection one uses sure, if it's just on that specific case then one uses possible. Using the improved guidelines each annotator was given three different language pairs – sentences 41 to 60. The same evaluation process was carried out, to produce the final annotator agreement. Table 5 reports the results of this evaluation.

Although there were improvements over the intermediate evaluation, we were not completely satisfied with the final results. The first thing we notice was that this alignment set contained much bigger and harder sentences than the previous one. After this evaluation, the annotators gathered again and corrected their differences. The resulting annotations where considered final. The guidelines were refined. Whenever there was a question about how an example should be done, a new rule was derived and the example was added to the guidelines. Nevertheless our average inter-annotator agreement was of 91.6% if we only count alignment points (no sure and possible distinction), which is in line with the results from previous projects. The Czech-English project (Melamed, 1998; Kruijff-Korbayová et al., 2006) that used the same metric was of 93%.

We did not perform another evaluation after this last improvement although we expect the results to be even better.

5. Some Statistics

The overall percentage of sure alignments versus possible alignments is a relevant characteristic of the gold set, since the common used metrics for word alignments tend not to penalize the absence of possible alignments. For reference purpose, the percentage of sure alignment for the Hansard (Och and Ney, 2000) corpus of English/French is 28.6% while for the EPPS corpus (Lambert et al., 2005) of English/Spanish is 69%. One possible explanation for this huge difference is that in the Hansards corpus the semantics of the possible alignments was broader. Possible alignments where used when two annotators disagreed on what an alignment point should be. Both in EPPS and in our corpus the semantics is not the same.

Mean Fertility measures the amount of words that have fertility bigger than one (align to more than one word). Fertility is a known difficulty for word alignment models, so serves as a good indicator of the difficulty of the corpus. We measure fertility by counting the total number of words that have fertility bigger than one over the total number of words.

Finally the last indicator we used for corpus difficulty is the average distance to diagonal. This is related with the reordering of words from one language to the other.

Table 6 shows some statistics of the 100 alignments for each language pair.

	%Sure	Mean Fertility	Avg. Dist. Diag.
EN-PT	56%	1.3	0.23
EN-ES	58%	1.3	0.21
EN-FR	72%	1.4	0.25
PT-ES	67%	1.2	0.17
PT-FR	77%	1.3	0.22
ES-FR	79%	1.3	0.20

Table 6: Gold Corpus statistics

6. Conclusions and Future Work

In this paper we described the experience of building a golden collection of word alignments over 6 different language pairs (all combinations between Portuguese, English, French and Spanish), where word alignment of each language pair is made over the first 100 sentences of the common test set from the Europarl corpora, corresponding to 600 new annotated sentence pairs. During the development of the gold sets, a detailed manual containing guidelines for multi-language annotation was produced and a new metric for inter-annotator agreement was proposed.

As both the golden collection and the guidelines are available on-line, we hope that this effort can be the seed for the development of new methods for word alignments using various languages, and that people find the guidelines useful and help extending them with examples for specific languages.

Notice that, although the golden collection is small, it can be used to evaluate different techniques for producing word alignments, or as a development set for supervised decoding methods or semi-supervised word alignment models.

As future work, we intend to continue increasing the golden corpus with more alignments for each language pair and with different languages.

	$\|A_1\|$	$\|A_2\|$	SA	WA	WD	SD
EN-PT (h_1,h_4)	269	196	67.0	11.2	15.3	6.5
EN-ES (h_1,h_3)	271	314	73.2	13.3	7.9	5.6
EN-FR (h_2,h_4)	332	256	61.9	19.0	9.3	9.7
PT-ES (h_1,h_3)	260	259	78.6	8.9	6.7	5.8
PT-FR (h_2,h_4)	331	260	73.8	10.2	9.1	6.9
ES-FR (h_2,h_3)	324	349	75.8	11.0	3.3	10.0
Average			71.7	12.3	8.6	7.4
Undifferentiated Average			84.0		16.0	

Table 4: Results of the inter-annotator intermediary evaluation.

	$\|A_1\|$	$\|A_2\|$	SA	WA	WD	SD
EN-PT (h_1,h_4)	287	316	82.9	6.6	9.5	1.0
EN-ES (h_1,h_3)	277	272	80.9	5.8	10.6	2.7
EN-FR (h_2,h_4)	262	281	80.8	10.0	5.0	4.2
PT-ES (h_1,h_3)	298	307	86.9	6.3	4.8	2.0
PT-FR (h_2,h_4)	273	268	86.5	7.0	4.1	2.4
ES-FR (h_2,h_3)	290	305	87.1	9.4	0.4	3.2
Average			84.2	7.5	5.7	2.6
Undifferentiated Average			91.6		8.4	

Table 5: Results of the inter-annotator final evaluation.

Acknowledgments

João Graça is supported by a fellowship from Fundação para a Ciência e Tecnologia (SFRH/BD/27528/2006). Joana Paulo Pardal is supported by a fellowship from Fundação para a Ciência e Tecnologia (SFRH/BD/30791/2006).

7. References

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

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, Barcelona.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL-04)*, Boston, USA, May.

João Graça, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro. 2008. Multi-language word alignments annotation guidelines. Technical Report TR/38/2008, L^{2}F – INESC-ID Lisboa/IST, Lisboa, Portugal, April.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X*, Phuket, Thailand.

Ivana Kruijff-Korbayová, Klára Chvátalová, and Oana Postolache. 2006. Annotation guidelines for Czech-English word alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1256–1261.

Patrik Lambert, Adrià De Gispert, Rafael Banchs, and José B. Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. In *Language Resources and Evaluation, Volume 39, Number 4*, pages 267–285.

I. Dan Melamed. 1998. Annotation style guide for the Blinker project. Technical report, IRCS.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association For Computational Linguistics*, Hong Kong.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Compututational Linguistics*, 30(4):417–449.