Multi-Language Word Alignments Annotation Guidelines

(version 0.9)

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

Spoken Language Systems Laboratory (L^2F)
R. Alves Redol, 9 – 1000-029 Lisboa, Portugal

{javg,joana,lcoheur,dcaseiro}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt/resources/translation/

May 25, 2008

This documents contains detailed guidelines for performing word-alignment annotations. These guidelines where proposed in [4] they are base on the guidelines described in [8] for Spanish/English, with some changes and refinements that are described in section 4. Another source of information that was used were the alignments guidelines defined on the Blinker project [9] for English/French and the guidelines defined in [7] for Czech/English, although the later two differ in some general principles.

The goal of this manual is to reduce as much as possible the ambiguity in the process of aligning bilingual texts. These guidelines were used to align bi-texts between all combinations of Portuguese (PT), Spanish (ES), French (FR) and English (EN) taken from the first 100 sentences of the common test set of the Europarl corpus [5].

Using this guidelines six gold alignments sets were built. These are 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, given that the existing and freely available ones are based on the Hansard corpus [10] and on the Bible [9]; 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.

1 Background

The concept of word alignment, introduced in [1] 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.

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 because 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 as *possible*, meaning that the words are a translation of each other in some contexts.

Although the main use of word alignments is statistical machine translation, directly on a translation system as originally proposed in [1], as a primary resource for phrase based machine translation [11] or syntax based machine translation [3], other applications of word alignments have been suggest in recent literature such as annotations' projections or extraction of bilingual lexica.

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 [9, 10, 8, 7] have been suggested.

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

[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/`.

2 Golden Collection

This section describes the corpus used to build the golden collection, the tool used to annotate the sentences, and the annotation process.

2.1 Corpus

We used the publicly available Europarl Corpus [5] 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 [6], 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.

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

2.2 Annotation Tool

Annotations were performed by using the annotation and visualization tool implemented by Chris Callison-Burch (University of Edinburgh) [2]. 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 1 shows a screenshot of the tool.

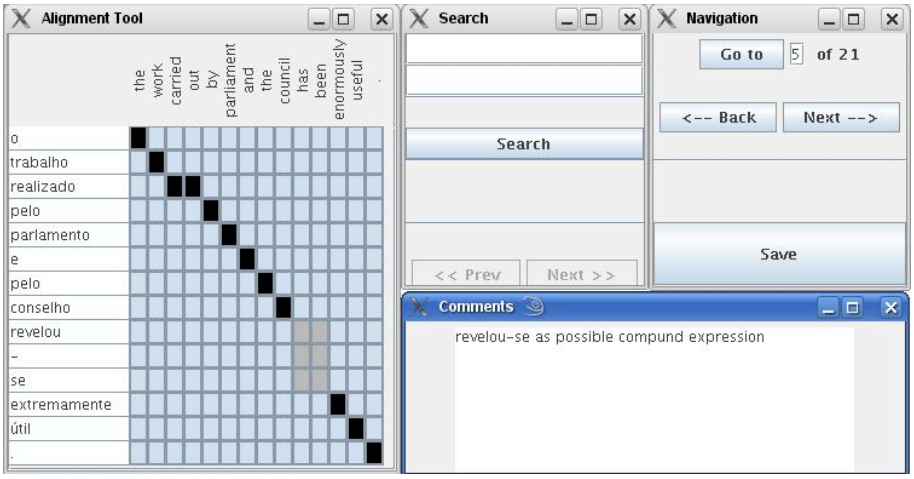

Figure 1: Screenshot of the tool. On the right: alignment for a given pair. Black is a Sure alignment. Grey is a Possible alignment. On top the search interface and the navigation toolbar. Left bottom, the comment window.

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

2.3 Annotation Process

Our starting point were the guidelines developed in [8] 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 this guidelines were produced.

The process begun by annotating the first 20 sentences of each language pair using the existing guidelines [8] by two annotators (h_1, h_2). During this first phase new guidelines were created with 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 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. Each annotator was given three sets of 20 sentences.

Table 3 resumes the annotation procedure.

At the end annotator h_4 reviewed all 100 sentences sets for all language pairs to correct existing differences due to guidelines changes or specialization.

	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 (figure 2).

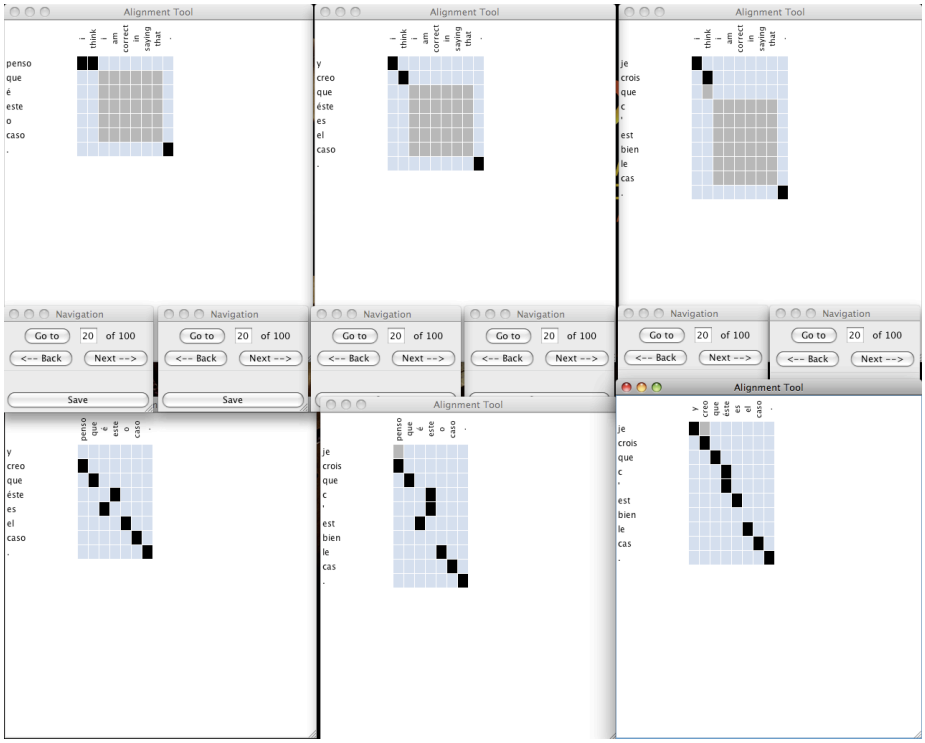

Figure 2: Final revision of the 100 sentences was done in turn for all language pairs.

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 Sure alignment in two languages but on a third language they only align as a Possible.

2.4 Examples

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. Giving the nationality of the authors, the favorite pair is English–Portuguese.

As described, the tool used to annotate the sentences presents each pair of sentences as a matrix with clickable squares. Similarly, in the remainder of this document, examples will be given with a similar representation.

Each (part of) sentence pair is represented by a matrix cell.

Non-aligned word-pairs are represented by a dot (to ease reading).

Aligned word-pairs are represented by filled squares.

Full dark blue box represent Sure alignments.

Empty light blue box represent Possible alignments.

Figure 3 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*".

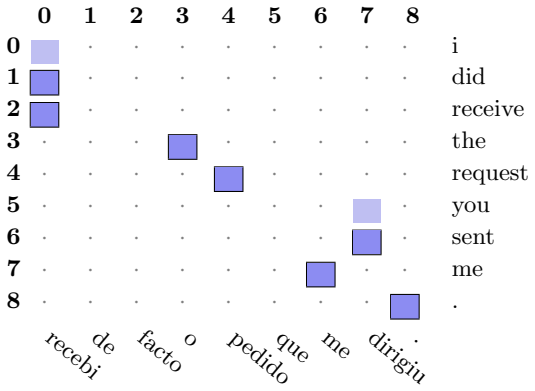

Figure 3: Word alignment between a Portuguese and an English sentence. Full dark blue box indicates a sure alignment point. Empty light blue box represents a possible alignment point.

There were systematically occurring cases for which it would be helpful to distinguish between *strong* and *weak* alignments. Also, it would be desirable, to encode gender and number variation, which is a relevant information for the Latin languages.

3 General Guidelines

As defined in [10] there are two types of alignments. Sure alignments (S-alignments) and possible alignments (P-alignments). However the meaning of the alignments is slightly different that in the previous work. an S-alignments is used when a translation is possible in every context, word-by-word, or compound expressions that are always interchangeable. 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 in the original work. 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.

The annotator should annotate as much as possible while at the same time don't aligned words or phrases (groups of phrases) just because they have the same semantic meaning. Its correct to leave words unaligned if they are incorrectly translated.

Regarding incorrect translation the rule is to always leave it unaligned. If a phrase has no relation with its translation although it might be the same in that special context it should be left unaligned. When it represents a construction that can be in some contexts be translated by the other it should be aligned as possible.

Regarding missing words or phrases in one translation. If these words are close class words they should be aligned, depending to the context, as a possible alignment to the respective head word. This is really frequent due to different writing styles or language differences. If the missing words are used as to give efface to a particular aspect or are just absent on the other language they should not be aligned.

4 Differences from EPPS Guidelines

The guidelines were heavily based on the guidelines defined in [8]. However we made some refinements on some of them since we felt their were ambiguous, and some changes since we didn't agree on their policy. Some changes include:

- When annotating compound expression, their definition is ambiguous on when using S-alignment blocks, P-alignment block, or P-alignment blocks with some S-alignment links.
- The guidelines for aligning passive vs active translation were underspecified.
- The decision to align determiners in compound nouns were also underspecified.
- The guideline to align determiners in enumerations say that the determiner should be unaligned. We prefer to follow the general rule and align it with the head of the enumeration.

5 Special Cases

We detail those cases that were annotated in different ways in the two initial stages, and those that raised discussion. We also detail the decisions we made.

5.1 Contractions

Contractions as a general rule should be aligned as sure alignment to the uncontracted words in the other language. Contractions may appear in various situations: In the case that the contraction is explicitly translated in the other language, either as a contraction or as several words translating the different parts of the contraction it should be aligned as sure alignments. Figure 4 shows an example where the contraction in $_{PT}$"*da* → *de* + *a*" is aligned with both sure alignments to the corresponding alignments *de* → *of* and *a* → *the*.

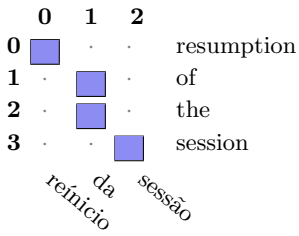

Figure 4: Direct contraction example

In the case that none of the parts of the contraction appear in the other language, it should be linked as possible to the head element of the contraction. Figure 5 shows an example where there is a missing element in the contraction.

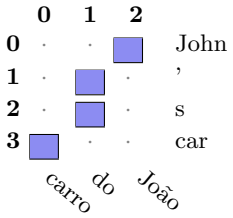

Figure 5: Contraction linked with head element example

In the case that only one part of the contraction appears in the translation it should be linked as sure to the part that appears. Figure 6 shows such an example where the contraction in Portuguese $_{PT}$"*das* → *de* + *as*" has only one part translated *de* → *of*.

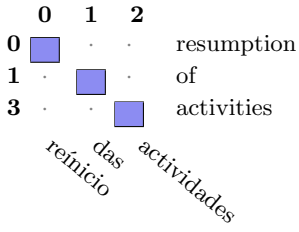

Figure 6: Missing contraction element in translation example

5.2 Compound expressions

Compound expressions should be aligned in different ways according to their resemblance: If the compound expression can be translated as a whole in all contexts it should be aligned with a block of S-Alignments, figure 11 shows an example on this.

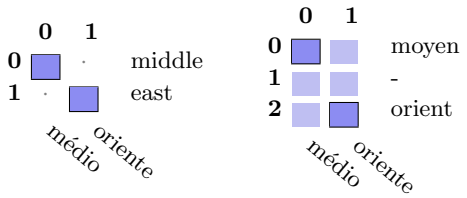

Figure 7: Fixed expression translation: word pairs as a sure alignments. To consider the "-", a surrounding possible block is needed.

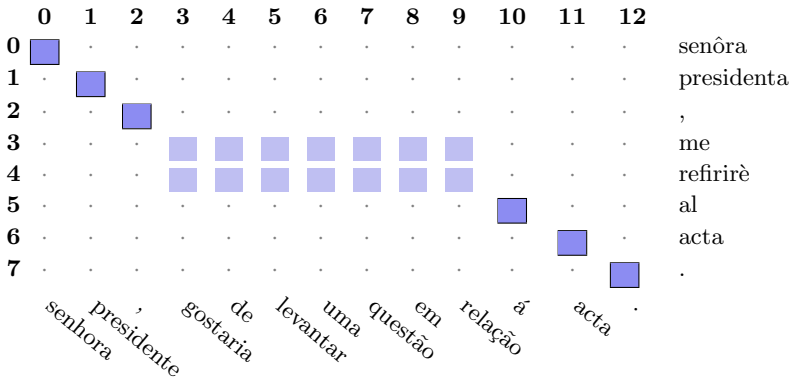

Figure 8: Example of big block of a possible compound expression

However is inside this possible compound expression there are parts that are a translation of each other those should be linked with a possible block like in the example on figure 9.

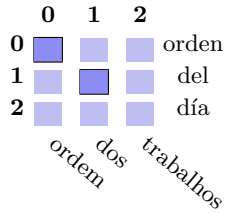

Figure 9: Possible compound term with sure links

5.3 Verb Constructions

5.3.1 Auxiliary Verbs

Auxiliary verbs that only appear in one language should be aligned with sure links to the verb in the other language (figure 10).

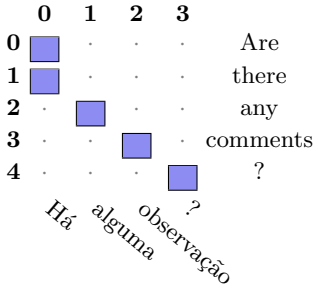

Figure 10: Auxiliar Verb Example

5.3.2 Verbs and Personal Pronouns

If a personal pronoun if present in both language then they should be treated as a different group of the verb and linked together with sure alignment. Otherwise if the personal pronoun is only available in one language it should be connected with a possible link to the corresponding verb in the other language (figure **??**).

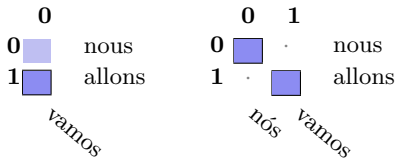

Figure 11: Personal pronouns translations.

5.3.3 Verbs Followed by a preposition

If the preposition changes the meaning of the verb (phrasal verb) it should be treated together with verb as a minimal unit. Otherwise the rule for prepositions apply.

5.3.4 Passive voice

In the case of passive voice all elements that are translation of each other should be aligned together following the specific rules. If some elements are not present in the translation like the subject then it should be left unaligned. Figure 12 shows an example of an *active* → *passive* where the subject ($_{PT}$"*o parlamento*") is not translated and should be left unaligned.

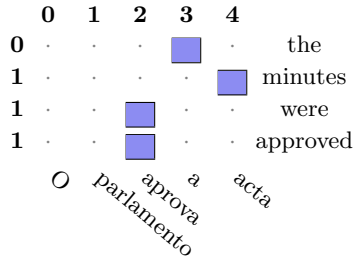

Figure 12: Passive example with missing subject

5.4 Noun Constructions

5.4.1 Determiner vs quantifier – *Specialization*

Different languages may uses different construction with nouns. for instance use a determiner or a quantifier that mean the same thing in that particular context. In this case they should be align as possible. Figure 13 shows an example where the quantifier $_{EN}$"*all*" as the same meaning as the determiner $_{PT}$"*as*" in this particular context.

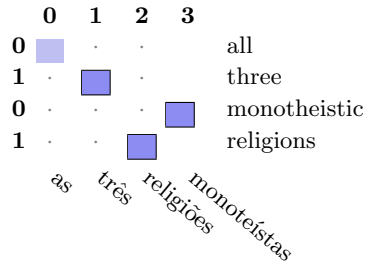

Figure 13: Missing determiner on noun composition

5.4.2 Noun complement construction

When a determiner or preposition is used with a name in a language and not in the other it should be linked with a possible alignment to the corresponding part of the name 14.

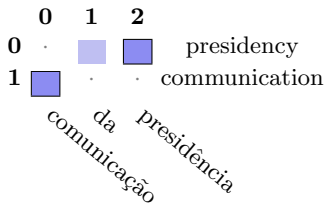

Figure 14: Missing determiner on noun composition

5.4.3 Noun vs. Adjective

5.5 Words Not or Incorrectly Translated

Words that are not translated in one language should not be translated, unless when the are grammar words which follow on the categories explained in other guidelines. In Figure 15 the Spanish expression is different from the Portuguese containing the words $_{ES}$“*del período*” this should be left unaligned.

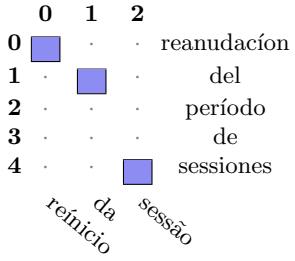

Figure 15: Words that don’t appear in one side of the translation

If the translation of a word means semantically a similar thing but is incorrect in every context it is left unaligned. In Figure 16 the word $_{PT}$“*semestre*” which means $_{EN}$“*semester*” is translated as $_{EN}$“*autumn*”. This word should not be translated.

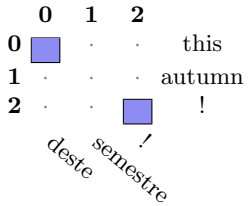

Figure 16: Incorrect translation in every context

5.6 Punctuation Marks

Use sure when they are translated as the same symbol and possible when a different symbol is used but means the same in the current context (figure 18).

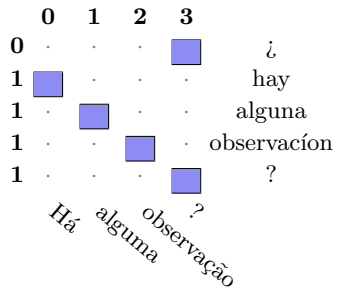

Figure 17: Spanish question mark as indivisible symbol

Spanish question mark symbols should be considered as a indivisible token 17

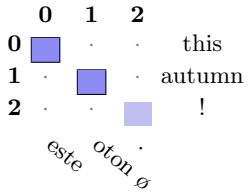

Figure 18: Wrong Punctuation

5.7 Repetitions

In the case one of the translation contains a repetition of the same phrase only the all the instances of the repetition should be aligned.

5.8 Pronouns

If not the same should be marked as possible.

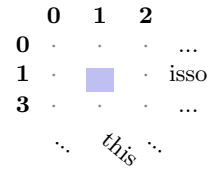

Figure 19: Different Pronoun translation

Because its an incorrect translation that can however be used.

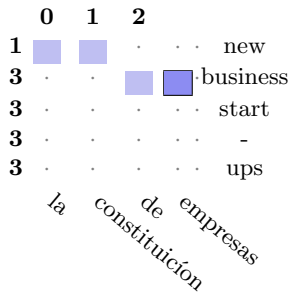

Figure 20: Different Pronoun translation

5.9 Numerals

Numbers like $_{EN}$"*Two hundred and five*" $_{PT}$"*duzentos e cinco*" should be considered indivisible units and be linked as a whole S block. Special case of compound nouns.

5.10 Date and Time

Follows same rules as fixed expressions. Figure **??** shows an example of a constant difference between dates description in Portuguese and English where in Portuguese determiners are used. This should be aligned with a possible link.

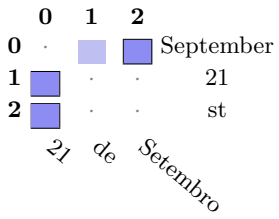

Figure 21: missingDetDate

Different expressions in different languages should be aligned as a block:

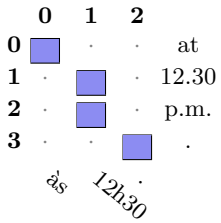

Figure 22: Different hour expressions

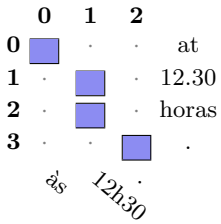

Figure 23: Different hour expressions

5.11 Acknowledgments and Thanking

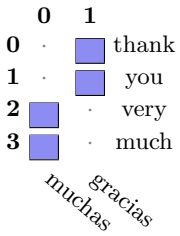

Figure 24: Different hour expressions

6 Language-specific Phenomena

6.1 General

6.1.1 Gender and Number variation

6.1.2 Clustering of Languages

6.2 Portuguese

6.2.1 Parenthetical Commas

6.3 English

6.3.1 Possessives

6.4 Spanish

6.4.1 Question Marks

6.5 French

6.5.1 Negation Constructions

References

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

[2] Chris Callison-Burch, David Talbot, and Miles Osborne. 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, 2004.

[3] Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 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 2004.

[4] João Graça, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro. Building a golden collection of parallel multi-language word alignments. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, LREC, Marrakech, Morocco, May 28-30 2008. LREC 2008.

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

[6] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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, 2003. Association for Computational Linguistics.

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

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

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

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

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