Contents

Abstract

In this technical report we present some guidelines defined during ReTraTos project for lexical alignment of Brazilian Portuguese, Spanish and English parallel texts. Parallel texts and their aligned version play an important role in many Natural Language Processing (NLP) applications, such as: transfer rule learning for machine translation (ReTraTos project's goal), Example-Based Machine Translation (EBMT), Statistical Machine Translation (SMT), bilingual lexicography, and word sense disambiguation, among others. By using these guidelines lexically aligned parallel corpora can be built following well-defined standards and avoiding, in this way, a lot of ambiguities inherent in the alignment process. The corpora and guidelines produced in this work can be used in future projects for building NLP tools and resources.

Chapter 1

Guidelines for lexical alignment

1.1 Introduction

This technical report presents the annotation style guide for lexical alignment of Brazilian Portuguese (PT), English (EN) and Spanish (ES) parallel texts has been developed during and for ReTraTos project.[1] Lexically aligned parallel texts have an important role in many Natural Language Processing (NLP) applications, such as: automatic transfer rule learning for machine translation [Carl, 2001, Menezes and Richardson, 2001] – ReTraTos project's goal –, Example-Based Machine Translation (EBMT) [Somers, 1999], Statistical Machine Translation (SMT) [Ayan et al., 2004, Och and Ney, 2000], bilingual lexicography [Melamed, 1996, Gómez Guinovart and Sacau Fontenla, 2004], and word sense disambiguation [Gale et al., 1992], among others.

In the last years, several automatic lexical alignment methods have been proposed in the literature [Hiemstra, 1998, Och and Ney, 2000, Ayan et al., 2004, Wu and Wang, 2004] achieving from 71% to 92% of precision and from 62% to 88% of recall according to the method and the language pairs involved, among other factors. However, to calculate alignment precision and recall automatically it is necessary to build a reference corpus, that is, a set of parallel texts which were lexically aligned by a human specialist and, hence, are considered to be correct. Precision and recall are calculated comparing automatically generated alignments with reference corpus.[2]

So, this technical report presents the main guidelines defined to the manual annotation process of alignments between single words and multiword units (sets of two or more words)

[1]We thank FAPESP, CAPES and CNPq for financial support.

[2]Precision is given by the number of correct source and target tokens (words, numbers, etc.) – in the intersection between proposed (by the automatic method) and reference alignments – per the number of proposed source and target tokens; while recall is given by the number of correct source and target tokens per the number of source and target tokens on reference alignment.

aiming at building PT-ES and PT-EN reference corpora. Other projects that have also defined their annotation style guides for lexical alignment are ARCADE[3] and Blinker [Melamed, 1998], in which this technical report is strongly based on. The PT-ES and PT-EN reference corpora built following these guidelines have been used to evaluate alignments generated by LIHLA lexical aligner implemented during ReTraTos project.[4]

Guidelines for the manual lexical alignment of 14 PT-ES parallel texts were obtained following a three-step process:

1. First, two native speakers of PT with reasonable knowledge of ES separately annotated all tokens (words, special characters and numbers) in the 14 PT-ES parallel texts;

2. Then, the two annotated corpora were compared and divergences were discussed and adjusted. An annotator agreement ratio of 95% between the two annotators – calculated as the number of identical alignments per the total number of produced alignments – was obtained considering only identical alignments, that is, partial agreements were considered as discrepancies.

3. A final version of the lexical-aligned corpus as well as the guidelines used to create it were compiled.

Following the guidelines built for PT-ES, 10 PT-EN parallel texts were lexically aligned aiming at verifying if these guidelines could also be applied to this new pair of languages. Furthermore, new guidelines were written to cover some specific cases of PT-EN. So, in this process 15,396 PT-ES tokens (7,236 PT tokens and 8,160 ES tokens) and 15,900 PT-EN tokens (7,631 PT tokens and 8,269 EN tokens) were aligned.

This technical report is organized as follows. The first section (1.2) shows the notation used in this document to describe the guidelines for manual lexical alignment process. Section 1.3 presents the general guidelines that can be applied to both PT-ES and PT-EN parallel texts; sections 1.4 and 1.5, respectively, present the specific guidelines for each pair of languages. Some examples of disagreements between the two annotators of PT-ES parallel texts are shown in section 1.6 and, finally, some concluding remarks are presented in section 1.7.

[3]ARCADE annotation style guide is available at: http://www.up.univ-mrs.fr/veronis/arcade/arcade1/2nd/word/guide/index.html.

[4]See NILC's web-site for more information about projects and resources developed by this group: http://www.nilc.icmc.usp.br.

1.2 Notation

In this section we present some considerations about the notation used in this document.

- **Lexical alignment category**

  The lexical alignment category is indicated by a sequence $l : d$, where l and d are positive numbers that stands for the number of source (left side) and target (right side) tokens involved in the alignment. Although the most frequent alignment category is $1 : 1$ (in which one source token is translated exactly as one target token), other categories such as omissions ($1 : 0$ or $0 : 1$) or those involving multiword units are also possible. Examples of alignments involving multiword units are: expansions ($n : m$, with $n < m; n, m \geq 1$), contractions ($n : m$, with $n > m; n, m \geq 1$) and unions ($n : n$, with $n > 1$).

- **Alignment**

  An alignment is indicated, in the scope of this document, as a sequence of one or more source tokens (or just the word NULL), a character "⇔" and a sequence of target tokens (or NULL). When the position in which a token occurs in the text is important to understand when a rule should be applied, the position of this token is indicated by a number followed by a character ":" before the token. For example, in the alignment $1{:}X \Leftrightarrow 2{:}Y$, the token X occurring at position 1 of the source text is aligned with the token at position 2 (Y) in the target text.

- **NULL**

  The special word NULL indicates an omission alignment ($1 : 0$ or $0 : 1$) and should be used in the alignment of a token without a translation in the parallel text.

- **Grammatical categories**

  Aiming at creating more general rules the grammatical category of a token can be indicated in the rule's definition. So, the grammatical categories used in this document are the following: `PREP` (preposition), `VERB` (verb), `AUX` (auxiliary verb), `SUBS` (noun), `ART` (article), `PRON` (pronoun).

- **Characters with a special meaning**

  [ ] - Brackets are used to indicate that the tokens between them are optional. So, for example, the sequence [`PREP`]`VERB` indicates a verb which may or may not be preceded by a preposition.

_ - "_" between two tokens indicates that they are concatenated and, hence, are found in the same word. For example, the sequence PREP_ART indicates that a preposition and an article are found together in the same word as in PT word "*dos*" or the ES word "*al*".

| - "|" indicates a logic or, that is, one (and only one) of the tokens separated by "|" would occur. For example, [PREP|ART] indicates that if there is a token (the brackets indicate that the token is optional) it will be a preposition or an article, but never both of them.

+ - "+" indicates an union of tokens to form a $n : m$ alignment with n and/or $m > 1$, that is, an alignment involving one or more multiword units. For example, it would be necessary to put together the EN preposition "*of*" and the EN article "*the*" to set the alignment between them and the PT PREP_ART "*da*", that is, a $2 : 1$ alignment (at EN-PT direction).

1.3 General guidelines

The manual annotation of alignments between single words and multiword units can be performed following these 3 steps:

For each token (word, number, special character):

1. Look for the best token in the other side (parallel text) whose meaning is the same as the token which is being aligned. If an one-to-one correspondence is found in both directions (source-target and target-source) set a $1 : 1$ alignment, otherwise go to step 2.
2. Look for the minimal set of tokens in the other side with the same meaning of the token which is being aligned, creating a $1 : n$ (or $n : 1$) alignment. However, if it is necessary to join more than one token in both sides to give the same meaning then go to step 3.
3. Try to set alignments between sub-pieces in both sides while preserving two-way equivalence. However, if it is not possible

`to set several alignments or there is no guarantee of semantic equivalence, set just one` $n:m$ `alignment between the` n `tokens in one side and the` m `tokens in the other side.`

So, some general rules can be derived from the sequence of steps showed previously.

R1. Set the minimal alignment while preserving the meaning in both directions (source-target and target-source).

Example (`PT-ES`):

1. "*sem ultrapassar o*" and "*que no irán más allá del*"

   In this case it is possible to identify two alignment blocks *sem+ultrapassar⇔que+no+(va)+más+allá+de* and *o⇔el*, but since there is no way to separate the preposition from the article in the word "*del*", the two blocks should be aligned together resulting in a single alignment:
   sem+ultrapassar+o⇔que+no+irán+más+allá+del

2. "*operadores de direito*" and "*legal workers*"

   In this case two alignment blocks can be identified:
   operadores⇔workers
   de+direito⇔legal

R2. Set as detailed a correspondence as possible; however, if there is any doubt about how to set the alignments between sub-pieces, a single alignment involving all the tokens should be set.

Example (`PT-EN`):

1. "*estimula*" and "*is capable of stimulating*"

   In this case it is not possible to set an alignment different from:
   estimula⇔is+capable+of+stimulating

2. "*ao que tudo indica*" and "*or so everything indicates*"

   In this case it is not easy to identify the correspondences at the first glance, but it is possible to identify two 1 : 1 alignments: one between "*indica*" and "*indicates*" and another between "*tudo*" and "*everything*". There is also a weaker correspondence between "*ao que*" and "*or so*". So, a possible choice is to set the following alignments:

ao+que⇔*or+so*

tudo⇔*everything*

indica⇔*indicates*

R3. An alignment between a token and NULL (an omission case) should be set when the token seems to have no correspondence and it also can not be joined to their preceding or following tokens because it is not essential to their meaning and, sometimes, it would be harmful to the alignment already established for its neighbors.

Example (PT-EN):

1. "*o pau-brasil*" and "*brasilwood*"

   In this case the PT definite article "*o*" should not be joined to the alignment between "*pau-brasil*" and "*brazilwood*" because it would harm the correspondence already established between these tokens. So, it is better to set two alignments:

   o⇔NULL

   pau-brasil⇔*brazilwood*

2. "*rapidamente*" and "*con maior rapidez*"

   In this example the ES sequence "*con rapidez*" have the same meaning as the PT word "*rapidamente*" and the word "*maior*" do not have a correspondence in the other side, so, two alignments should be set:

   rapidamente⇔*con+rapidez*

   NULL⇔*maior*

R4. Special characters should be aligned in a way that minimizes the number of "crossing" links. It is possible to align different characters to each other and also special characters and words, for example "*e*⇔," and "*y*⇔;".

Example (PT-ES):

1. "*a forma mais branda, a cutânea,*" and "*la forma cutánea, la más benigna,*"

   In this case the best alignments between the characters "," are obtained aligning them in the same order that they occur.

Following we present some cases which need more attention during the alignment process.

1.3.1 Alignment between words which differ in gender and/or number

Words which differ in gender and/or number can be aligned if and only if the meaning is preserved.

Examples (PT-EN):

- *metro*⇔*meters*
- *comuns*⇔*common*
- *florestas*⇔*forest*

Examples (PT-ES):

- *no*⇔*en+la*
- *estoque*⇔*existencias*
- *total*⇔*totales*

1.3.2 Alignment between determiners

Determiners can be aligned even if they do not belong to the same grammatical category since their role in parallel sentences are the same.

Examples (PT-EN):

- *os*⇔*those* (ART⇔PRON)
- *uma*⇔*its* (ART⇔PRON)

Examples (PT-ES):

- *o*⇔*ese* (ART⇔PRON)
- *o*⇔*Este* (ART⇔PRON)

1.3.3 Preposition and article found in the same word

When, in one side, there is a concatenation of PREP_PRON or PREP_ART in the same word and, in the other side, there is just one of these parts (PREP or PRON or ART) a 1 : 1 alignment is set even if just a partial correspondence exists because it is not possible to separate the tokens which occur together.

Examples (PT-EN):

- *da*⇔*of* (PREP_ART⇔PREP)
- *desse*⇔*this* (PREP_PRON⇔PRON)
- *dos*⇔*the* (PREP_ART⇔ART)

Examples (PT-ES):

- *dos*⇔*de* (PREP_ART⇔PREP)
- *o*⇔*al* (ART⇔PREP_ART)
- *no*⇔*en* (PREP_ART⇔PREP)

However, it can NOT be set a $1:n$ or $n:1$ alignment involving PREP, ART and PRON when it is possible to create a $1:1$ alignment involving them (PREP⇔PREP or ART⇔ART or PRON⇔PRON). In this case, the other parts should remain unaligned.

Examples (PT-EN):

- Don't do: *de*⇔*of+the*

  Do: *de*⇔*of* e NULL⇔*the*

Examples (PT-ES):

- Don't do: *de*⇔*de+la*

  Do: *de*⇔*de* e NULL⇔*la*

Furthermore, when there is a concatenation PREP_PRON or PREP_ART in one side and, in the other side, the two tokens can be found separately (PREP and PRON or PREP and ART), these tokens should to be joined to set a $1:n$ (or $n:1$) alignment.

Examples (PT-EN):

- *pela*⇔*by+the* (PREP_ART⇔PREP+ART)
- *dos*⇔*of+those* (PREP_ART⇔PREP+PRON) (see rule 1.2.2)
- *desse*⇔*of+this* (PREP_PRON⇔PREP+PRON)

Examples (PT-ES):

- *ao*⇔*a+lo* (it is possible) (PREP_ART⇔PREP+ART)
- *desse*⇔*de+ese* (PREP_PRON⇔PREP+PRON)
- *na*⇔*en+su* (PREP_ART⇔PREP+PRON) (see rule 1.2.2)

1.3.4 Relative pronouns

When the alignment involves relative pronouns, the longest one should be chosen.

Examples (PT-EN):

- *as+quais*⇔*which*
- *que*⇔*which*
- *em+que*⇔*where*
- *de+que*⇔*than*

Examples (PT-ES):

- *as*⇔*las*

  quais⇔*cuales*
- *a*⇔*a*

  que⇔*la+cual*
- *en*⇔*em*

  que⇔*los+que*
- *do+que*⇔*que*

1.3.5 Noun phrase X Verb phrase

When in one side there is a noun phrase and in the other side a verb phrase, all tokens in these phrases (`PREP`, `ART`, `SUBS`, `AUX`, `VERB`) should be joined to set just one 1 : 1 alignment.

Examples (`PT-EN`):

- *ao+vôo*⇔*to+fly*
- *à+eliminação*⇔*to+putting+down*

Examples (`PT-ES`):

- *ao+vôo*⇔*a+volar*
- *tratar*⇔*el+tratamiento*

1.3.6 Phrasal and prepositional verbs

A preposition should be joined to the verb in one or both sides if and only if it is part of the verb's semantics. So, phrasal verbs in English should never be split even if it is necessary to join a preposition to a verb in the other side.

Examples (`PT-EN`):

- *cuida+do*⇔*is+looking+to*

  Because: *cuida+de*⇔*look+to*
- *realizadas*⇔*carried+out*
- *ativa*⇔*sets+off*
- *fazem+parte*⇔*make+up+part*
- *deparam+com*⇔*come+across*

Examples (`PT-ES`):

- *fazem+parte*⇔*formam+parte*
- *é*⇔*consiste+en*

1.3.7 Main and auxiliary verbs

Auxiliary verbs should not be joined to the main verb in the other side if that main verb also has auxiliaries attached.

Examples (PT-EN):

- *não⇔not*

  tenha⇔has

  chegado⇔reached

- *haviam⇔had*

  sido⇔been

  contaminados⇔contamined

- *vem⇔has+been*

  aumentando⇔increasing

Examples (PT-ES):

- *não⇔no*

  tenha⇔ha

  chegado⇔llegado

- *vem⇔ha+ido*

  aumentando⇔en+aumento

However, when there are auxiliary verbs in one side but not in the other, the auxiliaries have to be joined to the main verb to be aligned with the main verb on the other side. The same can be applied to other particles (different from auxiliary verbs) that are part of verb's semantics, mainly in passive (see the last examples).

Examples (PT-EN):

- *acompanharão⇔will+accompany*
- *serão⇔will+be*
- *identificou⇔has+identified*
- *testa⇔is+testing*
- *não⇔not*

  surte⇔does+produce

- *subiram⇔were+launched*
- *se+reduzir⇔to+be+reduced*

Examples (PT-ES):

- *vão+integrar⇔integrarán*
- *iniciam⇔dan+inicio*
- *testa⇔está+probando*
- *registraram⇔han+registrado*
- *subiram⇔se+lanzaron*
- *vem+sendo+aplicado⇔se+lo+está+aplicando*

1.3.8 Verb + "*se*"

When the token "*se*" in `PT` or `ES` form part of the verb's semantics it should be joined to the verb (otherwise it has to remain unaligned). Furthermore, when the token "*se*" is concatenated to the verb in `ES`, it should be joined to the verb in `PT` (see the last `PT-ES` examples).

Examples (`PT-EN`):

- *torna+-+se⇔becomes*
- *vão⇔are+going+to*

  se+integrar⇔join

Examples (`PT-ES`):

- *cuida⇔se+encarga*
- *deparam⇔se+encuentram*
- *se+reduzir⇔reducirse*
- *se+tornar⇔convertirse*

1.3.9 Compound noun

When there is a noun in one side and in the other side there are several essential tokens to give the same meaning as this noun, then all these tokens should be joined to set a single alignment with the noun even if a one-to-one correspondence involving just two tokens already give the expected meaning. This have to be done because the tokens which seem to have no correspondence are essential to complement noun's semantics in that language.

Examples (`PT-EN`):

- *redes+de+arrasto⇔dragnets*
- *batimentos+cardíacos⇔heartbeats*
- *namorado⇔namorado+sandperch*
- *batata⇔potato+-+fish*
- *Bauru⇔the+town+of+Bauru*

- *Minas⇔the+state+of+Minas+Gerais*
- *endemias⇔endemic+diseases*
- *médias⇔medium+sized*

  e⇔and

  grandes⇔large+sized

  cidades⇔cities

Examples (`PT-ES`):

- *cherne⇔cherna+pinta*
- *batata⇔pez+ batata*
- *Minas⇔Minas+Gerais*
- *Maranhão⇔estado+de+Maranhão*
- *site⇔sitio+en+internet*
- *sudeste⇔región+sudeste*

1.3.10 "Frozen" expressions

Frozen expressions that are unique in one language or in the other should be linked as a whole, even if it is necessary to join several tokens in both sides. Furthermore, frozen expressions should be aligned even when they are not contiguous (as the last `PT-EN` example showed below).

Examples (`PT-EN`):

- *ao+redor+da⇔around+the*

  Since: *ao+redor+de⇔around*
- *de+acordo+com⇔according+to*
- *ao+lado+de⇔next+to*
- *tal+qual⇔just+like*
- *a+partir+dessa⇔following+this*

  Since: *a+partir+de⇔following*
- *segundo⇔according+to*
- *em+conjunto⇔as+a+whole*
- *na+casa+dos⇔in+the+region+of*
- *por+meio+de⇔by+means+of*
- *por+ora⇔for+the+time+being*
- *em+curso⇔under+way*
- *Comunidade+Européia⇔European+Union*

- 65:*tanto+*69:*quanto*⇔94:*both+*97:*and*

  66:*a*⇔NULL

  67:*leishmaniose*⇔96:*leishmaniasis*

  68:*tegumentar*⇔95:*tegumentary*

  70:*a*⇔98:*the*

  71:*visceral*⇔99:*visceral*

Examples (PT-ES):

- *de+acordo+com*⇔*conforme*
- *de+acordo+com*⇔*según*
- *ao+lado+de*⇔*junto+a*
- *tal+qual*⇔*a+la+manera+de*
- *a+partir+dessa*⇔*con+base+en+esta*

  Since: *a+partir+de*⇔*con+base+en*
- *além+de*⇔*al+margem+de*
- *abaixo*⇔*[por+]debajo*
- *em+conjunto*⇔*conjuntamente*
- *na+casa+dos*⇔*alrededor+de*
- *embora*⇔*pese+a[+que]*
- *em+vez+de*⇔*en+lugar+de*
- *por+meio+de*⇔*a+través+de*
- *mas*⇔*sino+también*

The longest alignment should be preferred, that is, if it is possible (according to a bilingual lexicon) to set an alignment with more than one token while preserving two-way equivalence, these tokens should be joined even when just some pieces already give the expected meaning.

Examples (PT-EN):

- *desde[+que]*⇔*ever+since*
- *mais[+de]*⇔*over*
- *novamente*⇔*[once+]again*

Examples (PT-ES):

- *apenas*⇔*[tan+]solo*
- *sozinho*⇔*[por+sí+]solo*
- *de*⇔*[a+partir+]de*

- *de⇔[por+parte+]de*
- *via⇔[por+]vía*
- *até[+mesmo]⇔incluso*
- *hoje⇔hoy[+en+día]*
- *como⇔[tal+]como*

OBS.: Proper noun

However, when the tokens under consideration form part of a proper noun or if there is a clear 1 : 1 correspondence between all sub-pieces, you should mark only the sub-piece (see R1). For example:

- Don't do: *Belo+Horizonte⇔Belo+Horizonte*

  Do: *Belo⇔Belo* and *Horizonte⇔Horizonte*
- Don't do: *Rio+de+Janeiro⇔Rio+de+Janeiro*

  Do: *Rio⇔Rio* and *de⇔de* and *Janeiro⇔Janeiro*

1.3.11 Other alignments involving multiword units

In addition to the examples of alignments involving multiwords presented above, several others are possible and should be created considering the general rules presented at the beginning of this section.

Examples (`PT-EN`):

- *dominó⇔dominoes*

  que+tomba⇔falling
- *mosquito⇔mosquito*

  transmissor⇔that+transmits

  causador⇔that+causes
- *apoio⇔support*

  dos+técnicos⇔technical
- *equipe⇔team*

  baiana⇔from+Bahia

Examples (`PT-ES`):

- *esquecidas⇔que+estaban+olvidadas*
- *prejudicando⇔que+prejudica*
- *emergenciais⇔de+emergencia*
- *consorciada⇔por+vía+de+consorcios*

- *de+moradores+de+bairro⇔vecinales*
- *decisórios⇔de+decisión*

1.3.12 Referring expressions

It is possible to align two tokens that are not translations of each other if they have the same role in parallel texts and can not be aligned to any other tokens in the given context.

Examples (PT-EN):

- *as+plantas⇔they*
- *os+pesquisadores⇔they*

Examples (PT-ES):

- *nesses⇔de+dichos*
- *peixe⇔especie*

1.3.13 Sequence of tokens repeated in just one of the two sides

Sometimes, a piece of text is repeated in one side but not in its translation. In this case, all instances of that repeated piece of text should be aligned to the single translation in the other side. Therefore, an ID which occurs more than once in different alignments does not indicate an alignment involving multiword units but several possible alignments involving the token with this ID.

Examples (PT-EN):

- 5:*do*⇔**5**:*of*+**6**:*the*
  nariz⇔*nose*
  ,⇔,
  8:*da*⇔**5**:*of*+**6**:*the*
  boca⇔*mouth*
  e⇔*and*
  11:*da*⇔**5**:*of*+**6**:*the*
  garganta⇔*throat*

Examples (PT-ES):

- 5:*do*⇔**5**:*de*+6:*la*
  nariz⇔*nariz*
  ,⇔,
  8:*da*⇔**5**:*de*+9:*la*

boca⇔*boca*

e⇔*y*

11:*da*⇔**5**:*de*+12:*la*

garganta⇔*garganta*

1.4 Specific rules for PT-ES

1.4.1 Verb + *la,lo,las,los,le,*etc.

In Spanish it is possible to join some particles to the verb when it occurs in infinitive, imperative or continuous form. Hence, when a Spanish verb is joined to some particles, to the verb on the other side it should be joined as many tokens as necessary to preserve the two-way equivalence.

Examples (PT-ES):

- *regular+essa+resposta*⇔*regularla*
- *tornam*⇔*continúan+tornándolo*
- *armazená+-+lo*⇔*almacenarlas*

1.5 Specific rules for PT-EN

1.5.1 Preposition between nouns

It is quite frequent to set an omission alignment for the preposition occurring between nouns when translating from PT into EN following rule R3 (a PREP should not to be added to their neighbours alignment).

Examples (PT-EN):

- *vapor*⇔*vapor*

  de⇔NULL

  água⇔*water*
- *Instituto*⇔*Institute*

  de⇔NULL

  Pesca⇔*Fishing*
- *testes*⇔*tests*

  em⇔NULL

  campo⇔*field*
- *semente*⇔*seed*

  do⇔NULL

pau+-+brasil⇔brazilwood

1.5.2 Possessive

The English possessive particle – *'s* or just *'* – should be aligned to the preposition (possible concatenated with an article) which plays the same role in Portuguese text, if there is such a preposition.

Examples (PT-EN):

- *do⇔'s*

  organismo⇔organism
- *dos⇔'*

  organismos⇔organisms
- *Alzheimer⇔Alzheimer*

  NULL⇔*'s*

1.5.3 Verb with and without subject

If the subject of a verb in PT can be found just in the verb suffix, then the subject in EN (if it is explicitly defined) have to be joined to the verb in the alignment even if the subject and the verb are not contiguous.

Examples (PT-EN):

- *atravessaram⇔they+have+crossed+out*
- *Vimos⇔We+saw*
- *duravam⇔they+would+last*
- *fossem⇔they+were*
- 5:*às*+6:*vezes*⇔6:*sometimes*

  7:*tenho*⇔5:*I*+7:*have*

However, if the subject can not be found explicitly or even in the verb suffix, then just the verbs should be aligned one with each other.

Examples (PT-EN):

- NULL⇔*it*

  destruir⇔destroys
- NULL⇔*when*

  NULL⇔*they*

  seguidos⇔followed

- NULL⇔*it*

  podendo⇔*can*

1.6 Disagreements

In this section we present some examples of different alignments set by the two PT-ES annotators since they disagree in 5% of the cases.

Table 1.1: Examples of disagreements between annotators

Annotator A	**Annotator B**
a+exemplo+do⇔*como+por+ejemplo+el*	*a*⇔*por* *exemplo*⇔*ejemplo* *do*⇔*como+el*
menos⇔*de+menor* *nobre*⇔*calidad*	*menos+nobre*⇔*de+menor+calidad*
tornar⇔*convierta* NULL⇔*en*	*tornar*⇔*convierta+en*
de+modo⇔*en+forma*	*de*⇔*en* *modo*⇔*forma*
deu⇔*verificó* *apenas*⇔*en+tan+solo*	*deu*⇔*verificó+en* *apenas*⇔*tan+solo*
demonstrar⇔*hacer* *publicamente*⇔*[75]:público*	*demonstrar+publicamente*⇔*hacer+público*
como⇔*al+modo*	*como*⇔*al+modo+de*
lei⇔*ordenanza+municipal*	*lei*⇔*ordenanza* NULL⇔*municipal*
virou⇔*se+ha+convertido+en*	NULL⇔*se* *virou*⇔*ha+convertido+en*
NULL⇔*en* NULL⇔*medio* NULL⇔*a* *num*⇔*un*	*num*⇔*en+medio+a+un*

1.7 Conclusions

Lexical alignment between single words and multiword units is a hard task since, frequently, it is not possible to set an one-to-one alignment between all tokens in parallel texts due to divergencies between languages and also non-literal translations.

So, in this technical report, we present the main guidelines on manual lexical alignment process carried out during ReTraTos project aiming at building PT-ES and PT-EN lexically

aligned reference corpora.

By using these guidelines it was possible to build reference corpora based on well-defined lexical alignment standards avoiding, in this way, a lot of possible ambiguities. The resulting corpora, together with these guidelines, can be used in future projects for building Natural Language Processing tools and resources.

Bibliography

[Ayan et al., 2004] Ayan, N. F., Dorr, B. J., and Habash, N. (2004). Multi-Align: Combining linguistic and statistical techniques to improve alignments for adaptable MT. In Frederking, R. E. and Taylor, K. B., editors, *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, number 3265 in Lecture Notes in Artificial Inteligence (LNAI), pages 17–26. Springer-Verlag Berlin Heidelberg.

[Carl, 2001] Carl, M. (2001). Inducing probabilistic invertible translation grammars from aligned texts. In *Proceedings of CoNLL-2001*, pages 145–151, Toulouse, France.

[Gale et al., 1992] Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1992)*, pages 101–112, Montreal, Canada.

[Gómez Guinovart and Sacau Fontenla, 2004] Gómez Guinovart, X. and Sacau Fontenla, E. (2004). Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Procesamiento del Lenguaje Natural*, 33:133–140.

[Hiemstra, 1998] Hiemstra, D. (1998). Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In Coppen, P. A., van Halteren, H., and Teunissen, L., editors, *Proceedings of the 8th CLIN meeting*, pages 41–58.

[Melamed, 1996] Melamed, I. D. (1996). Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA-1996)*, pages 125–134, Montreal, Canada.

[Melamed, 1998] Melamed, I. D. (1998). Annotation style guide for the Blinker project, version 1.0.4. Technical Report 98-06, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA.

[Menezes and Richardson, 2001] Menezes, A. and Richardson, S. D. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, pages 39–46, Toulouse, France.

[Och and Ney, 2000] Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 440–447, Hong Kong, China.

[Somers, 1999] Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157.

[Wu and Wang, 2004] Wu, H. and Wang, H. (2004). Improving domain-specific word alignment with a general bilingual corpus. In Frederking, R. E. and Taylor, K. B., editors, *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, number 3265 in Lecture Notes in Artificial Inteligence (LNAI), pages 262–271. Springer-Verlag Berlin Heidelberg.