Portuguese-English word alignment: some experiments

Diana Santos[1], **Alberto Simões**[2]

[1] SINTEF ICT, Pb 124 Blindern,
N-0314 Oslo, Norway (diana.santos@sintef.no)
[2] Departamento de Informática, Universidade do Minho,
Campus de Gualtar, Braga, Portugal (ambs@di.uminho.pt)

Abstract

In this paper we describe some studies of Portuguese-English word alignment, focusing on (i) measuring the importance of the coupling between dictionaries and corpus; (ii) assessing the relevance of using syntactic information (POS and lemma) or just word forms, and (iii) taking into account the direction of translation. We first provide some motivation for the studies, as well as insist in separating type from token anlignment. We then briefly describe the resources employed: the EuroParl and COMPARA corpora, and the alignment tools, NATools, introducing some measures to evaluate the two kinds of dictionaries obtained.

We then present the results of several experiments, comparing sizes, overlap, translation fertility and alignment density of the several bilingual resources built. We also describe preliminary data as far as quality of the resulting dictionaries or alignment results is concerned.

1. Word alignment and its evaluation

Word alignment is a relatively well-known technology for both statistical and example based machine translation (SMT and EBMT), but there are still several open research questions involved.

As already pointed by others, the term "word alignment" has been used to depict two different tasks, which we will call in the present paper "(word) type alignment" and "(word) token alignment".

The former has also been called "bilingual lexicon acquisition" (from parallel aligned corpora) by (Karlgren and Sahlgren, 2005).

The latter, the process of establishing translation relationships between words belonging to two chunks of text has been termed "bilingual word alignment" by (Dagan et al., 1993); or just "bilingual alignment" by (Mihalcea and Pedersen, 2003).

Instead of *alignment* – which conveys, at least originally, the idea of order preservation –, *correspondence* would be a more suitable designation, even if one did not separate beween type and token. For example, (Moore, 2001) uses the more adequate description of "translation relationships among words", and (Melamed, 2000) has moved to "translation equivalence among words", but unfortunately it was the phrase "word alignment" that won.

In any case, we are here interested in looking in more detail into these two tasks and their evaluation: we are not arguing here that they are unrelated, but that they should be conceptually separated to be adequately evaluated.

Word-type models are usually obtained from processing word-tokens, so there is often, though not always, a relevant connection between word-token co-occurrences in a bi-text and the word-type bilingual dictionary that the method gives origin to.

In this paper, we want to test empirically some of the assumptions involved in (type and token) word alignment, as well as study the interdependence of the two in the case of our tools. To fix terminology, we use the *term probabilistic translation dictionary* (PTD) to denote the result of word type alignment, and the term *word correspondences* to refer to the specific relationships between specific words in bi-texts. The set of all word correspondences will be called a *token dictionary*.

We will provide a brief literature review of the evaluation methods for both kinds of tasks:

For type-alignment, the standard methods employ coverage (how many of the words in the corpus occur in the dictionaries) and accuracy (manual investigation of a sample of entries, compared to published dictionaries or checked in the corpus).

For token alignment, (Melamed, 1998) created a golden standard using the Bible for the French English pair. However, consideration of his detailed instructions uncovered deep disagreement with our own views of word alignment. For example: *to the land of honey and milk*, although phrasally related to the French pronoun *y* in a particular passage, does not seem to us to be a good enough reason for making *land-y*, *honey-y*, and *milk-y* links, as Melamed does (for the sake of maximal annotation). We doubt whether so much disagreement about what a correct/usable/interesting word token aligner should do does not preclude indeed a comparison with this work. In fact, if there is such a wide disagreement as to what a token link should be, maybe it is completely vacuous to compare word alignment systems.

A detailed comparison of two kinds of word aligners and the different assumptions used by each, namely the works by Moore and Melamed can be appreciated more fully in (Santos, 2007).

Tasks like word spotting (Véronis and Langlais, 2000), which can be agreed upon, are probably the only ones liable to comparative evaluation. Other tasks such as support for bilingual lexicographers, and more informed translation browsers, could be also produce some indirect evaluation results, provided it were possible to quantify user satisfaction in using those systems.

2. Research questions

Some of the issues we are interested in are:

- To assess the quality of word type alignment, what is the importance of the underlying corpus in order

to build a good Probabilistic Translation Dictionary (PTD)?

- How relevant are the issues of balance, textual genre and translation direction for building a PTD?
- Likewise, if one uses a PTD based on a particular corpus to do word token alignment of a different corpus, how much degradation is it to be expected as compared to using the "proper" PTD?
- Generalizing, can we extrapolate this last question to provide a measure for comparing bilingual corpora? Most work so far has concentrated on monolingual corpus comparison (Kilgarriff, 2001).

Another relevant subject for machine translation is a measure of the word vs. phrase correspondence that a particular word alignment can produce. This can be considered an approximation to the translation difficulty of the particular language pair, and also provide a ceiling for what can be expected of aligment, or correspondence.
In this paper, we will however only deal with the following questions:

2.1. How important is text type?

It is common to read that there are huge differences in translation practice according to text type, but there is hardly any measure or even quantitative support for this claim. One might partially pin down creativity in translation using the measures "number of different translations per word" and "number of no-one-to-one-translations used".
Apriori we would predict that creative text types (and creative translation practices) would rank higher in both of those measures.
We will therefore propose a quantitative analogue of "translation fertility" based on PTDs and investigate how sensitive it is to translation direction, and especially text type.
It is possible that we can also pinpoint which lexical or grammatical areas are more sensitive to these distinctions. From an MT developer point of view, it is highly relevant to measure which part – if any – of the translation between two languages is common to different text types, and which one is more variable.

2.2. Is translation direction relevant?

Interestingly, significant differences in machine translation depending on the translation direction have been reported, but they often remain unexplained, and in any case not explicitly quantified.
For example, (Way and Gough, 2005) mention that the English-French and the French-English direction behave differently as far as their comparison of EBMT and SMT systems is concerned, while (Talbot, 2005) states that "models trained with German as the source language tend to have significantly lower AER than those with English".
Also (Koehn, 2005) notes this in connection with the EuroParl languages: "some languages are more difficult to translate into than from", and mentions that translation into morphologically rich languages, as opposed to into English, has been comparatively neglected.

2.3. How important is syntactic analysis for word alignment?

What – if any – is the impact on word alignment of proper name recognition, multiword detection, lemmatization and POS labelling is something that we want to assess, by creating different alignment dictionaries according to these different possibilities. (Choueka et al., 2000) used lemmatized versions of their English-Hebrew parallel texts, but did not compare with non-lemmatized versions.

3. Resources employed

3.1. COMPARA

COMPARA (Frankenberg-Garcia and Santos, 2003) is a large human-edited parallel corpus, whose sentence alignment, sentence separation, lemmatization and POS tagging have been human revised (the two last so far for Portuguese only) (Santos and Inácio, 2006).
COMPARA[1] contains 75 fiction texts and their published translations, corresponding to approximately 1.5 million words in each language (English and Portuguese). Several varieties of both languages (Portuguese, Brazilian, African, American, British, South African...) are included (Santos and Frankenberg-Garcia, 2007).

3.2. EuroParl

We have also used EuroParl (Koehn, 2002; Koehn, 2005) aligned for the same language pair. Europarl is publically available[2] and V2 contains more than one million translation units, with around 30 million words in each language. In contradistinction to COMPARA, Europarl's sentence alignment has not been manually revised, nor is translation direction known (in fact many of the translation units may not even be directed translated between English and Portuguese, coming from other source languages). It belongs to another genre (transcription of parliamentary debates); and language varieties of both Portuguese and English are restricted to European (and even Europese) parlance.

3.3. NATools

NATools [3] is an open source workbench for parallel corpora processing developed upon the Twente-Aligner (Hiemstra, 1996), which is based on an expectation maximization (EM) algorithm. Its tool set includes a sentence aligner, a probabilistic translation dictionary (PTD) extractor (Simões and Almeida, 2003) (= a word type aligner), and a word (token) aligner.

3.3.1. Probabilistic translation dictionary creation (`nat-create`)

The algorithm employed to create the PTDs has been described in detail elsewhere (Simões, 2004). Basically, from a set of aligned sentences `nat-create` creates a list of entries, to each it is associated its frequency and the probability of the several translations, see example below:

[1] http://www.linguateca.pt/COMPARA/
[2] http://www.statmt.org/europarl/
[3] http://natools.sourceforge.net/

```
** europe ** 42853 occs

       europa: 94.71 %
     europeus:  3.39 %
      europeu:  0.81 %
     europeia:  0.11 %

** stupid ** 180 occs

     estúpido:  17.55 %
     estúpida:  10.99 %
    estúpidos:   7.41 %
      avisada:   5.65 %
      direita:   5.58 %
      impasse:   4.48 %
      ocupado:   3.75 %
```

Let us just present the assumptions (or limitations) of the work reported in this paper: (i) sentences longer than 500 words were ignored by the algorithm; (ii) a fixed maximum of eight translations per entry was postulated; and (iii) capitalization was removed.

We then used two different strategies to build PTDs for this paper: a) no filtering at all; b) only keeping as dictionary entries words occurring more than 5 times in the source language (SO>5), and maintaining only those translations that corresponded to a probability higher than 5% (TP>5%).

So far, this program only creates 1-1 alignments, but uses the tokenization provided at the input. This means one can cheat the system by sending already glued multiword chunks (which cannot be discontinuous), as described below.

3.3.2. Token alignment (`nat-chunker`)

In token alignment, we use the extracted dictionaries to align terms across sentences. First, we create a matrix of translation probabilities (from the PTD) between all words in the two sentence-aligned sentences, then we apply a smoothing algorithm, followed by application of some pattern matching rules, to deal with usual word order changes across the two languages. The result that maximises the number of cells put into correspondence wins. Figure 1 shows an example of a (rather good) alignment matrix from EuroParl.

	discussion	about	alternative	sources	of	financing	for	the	european	radical	alliance	.
discussão	**44**	0	0	0	0	0	0	0	0	0	0	0
sobre	0	**11**	0	0	0	0	0	0	0	0	0	0
fontes	0	0	0	**74**	0	0	0	0	0	0	0	0
de	0	3	0	0	**27**	0	6	3	0	0	0	0
financiamento	0	0	0	0	0	**56**	0	0	0	0	0	0
alternativas	0	0	**23**	0	0	0	0	0	0	0	0	0
para	0	0	0	0	0	0	**28**	0	0	0	0	0
a	0	1	0	0	1	0	4	**33**	0	0	0	0
aliança	0	0	0	0	0	0	0	0	0	0	**65**	0
radical	0	0	0	0	0	0	0	0	0	**80**	0	0
europeia	0	0	0	0	0	0	0	0	**59**	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	**80**

Figure 1: Word alignment matrix.

The result of `nat-chunker` is a set of alignment pairs such as shown on table 1.

discussão	discussion
sobre	about
fontes de financ. alternativas	alternative sources of financing
para	for
a	the
aliança radical europeia	european radical alliance

Table 1: Extracted aligned pairs.

In the previous example, a very clean and easy piece chosen for illustrative purposes, there is one 4:4 and one 3:3 chunk alignment. In general, there are often 2:2 or 3:1 or 10:5 alignments where one cannot go further down pinpointing which corresponds to which. This is a consequence of the linguistic differences between the languages, a ceiling to what word alignment can meaningfully accomplish.

Measuring this ceiling is unfortunately outside the scope of the present paper.

Now, what we really want for our purposes here is the individual alignments among words, so we built `nat-mkTokenDic`, which, from the output of `nat-chunker`, identifies the possible translation relationships between words. All instances of one-to-one word pairs will count as token-word correspondences. So, for example *discussão : discussion* and *para : for* would be gathered by this procedure, but not *europeia : european*.

The result of this process for the whole corpus we call a token dictionary; i.e., simply the accumulation of all translations a given word had in the particular parallel text which was aligned. For practical purposes, these token dictionaries are stored in the same format of PTDs, the difference being that the numbers associated with the target language words reflect relative frequencies in a particular parallel corpus and not probability estimations of any sort.

`nat-chunker` also produces a composite measure of its performance in a particular corpus, by computing the average size of chunks obtained (ASC), as well as providing the distribution of chunk size (i.e., how many 1-1, 1-2, 2-5 etc.).

Let us in any case note that token dictionaries are currently a proper subset of PTDs, i.e., there is no mechanism to infer new word alignments, such as from *torneira esverdeada partida - broken greenish tap* and the entries *torneira - tap* and *partida - broken* to obtain the new alignment *esverdeada - greenish*. This would considerably improve the (token) alignment process and would allow us to get more interesting results, but has to be left for future work.

3.3.3. Dictionary comparison (`nat-compareDicts`)

NATools also includes a program that compares dictionaries: globally, by providing the number of identical, overlapping and intersective entries in the two dictionaries, and per word, in an interactive mode, allowing human inspection of the two full entries. Due to the similar format of the two kinds of dictionaries, this program can also be applied to the comparison of a PTD and a token dictionary, or to two token dictionaries.

3.3.4. Describing dictionaries (`nat-descDict`)

We have developed a new module for NATools which computes the following characateristics for a PTD or a token dictionary.

- translation fertility (TF): the average number of translations found in a dictionary (PTD or token dictionary);
- type or token alignment density (tyAD or toAD): this is the ratio of aligned tokens in a parallel text after token alignment (applies therefore only to token dictionaries), either counting different forms that got aligned (over all different forms in the target side of the corpus): tyAD; or counting the number of all forms that got aligned over all forms of the target side of the corpus

4. Experiments

4.1. PTD overlap

We have built several PTDs from EuroParl and COMPARA, which we proceed to describe now.

Using the full corpora, the resulting (unfiltered) PTDs include about 137,000 and 71,000 entries respectively for the Portuguese language (and 87,000 and 45,000 for English). If we consider the filtered ones, which should be more reliable estimators of the translation relationship, we get sizes of 47,000 and 16,000 entries for EuroParl and COMPARA. Numbers concerning the size of these and other PTDs are shown together in Table 7 below.

As to overlap between entries of the two PTDs, 44% of the Portuguese words from COMPARA are not in the EuroParl PTD, while 71% of the Portuguese entries from EuroParl are not in the COMPARA PTD. Interestingly, for English almost identical figures are obtained: 42% of COMPARA entries are absent from EuroParl, and 70% of Europarl entries are not in the PTD obtained from COMPARA)

As to the filtered dictionaries, which should contain the most frequent words, 31% of the entries in COMPARA are not in the EuroParl filtered DTD, and 76% of the entries in the EuroParl DTD are not in the COMPARA PTD, in Portuguese. For English, the figures are 30% and 68% respectively.

Explanation for the significant number of missing entries in each PTD becomes clear by considering the most frequent words in each PTD that are not to be found in the other, in Tables 2 and 3 (for the unfiltered PTDs).

In fact, considering Table 2, most words of the COMPARA-only column are proper nouns, two (*idéia* and *moça*) are forms belonging to the Brazilian variety (and therefore not to be found in EuroParl), and the final one, *titi*, is a familiar way to call one's aunt, clearly out of place in a parliamentary context. The words of the Europarl-only set are clearly political (and even European political) terms, very improbable to appear in fiction that spans the 1800s and the 1900s. The entries in Table 3 show a similar pattern (in fact the COMPARA lists even have several common terms, corresponding to proper names which remain untranslated into the other language).

COMPARA		EuroParl	
305	raimundo	37645	estados-membros
259	frances	19173	directiva
257	persse	18876	deputada
218	moça	9932	legislação
196	idéia	8198	comissária
195	zapp	6633	orçamental
192	brodsky	6458	cimeira
189	simão	6323	euro
184	lu	5369	jurídica
183	swallow	5309	euros
181	estácio	5229	relatora
179	judy	4439	estado-membro
178	titi	3953	coesão
163	sophy	3931	reforço

Table 2: Top occurring Portuguese words that are not common to the two dictionaries.

COMPARA		EuroParl	
298	raimundo	16056	rapporteur
268	frances	7069	implementation
251	persse	6470	euro
185	rummidge	5592	president-in-office
180	zapp	5375	fisheries
176	brodsky	5022	tabled
157	sophy	4984	legislative
152	estacio	4926	eur
146	gina	4049	sustainable
141	vic	3837	organisations
139	rubião	3830	cohesion
136	lizzie	3710	coordination
133	leaned	3647	implemented
131	murmured	3401	intergovernmental

Table 3: Top occurring English words that are not common to the two dictionaries.

4.2. The influence of genre

While this clearly illustrates the difference between the two domains, we want to know how much word alignment becomes degraded by using the same or different genre PTDs. We have therefore created two new corpora from each corpus: COMPARA reduced (`CMPred`) has 90% of COMPARA contents (excluding void 1-0 alignment units); `CMPtst` includes the remaing 10% for testing purposes.

Then we have also created a PTD based on an Europarl subcorpus with roughly the same size as COMPARA, named `EuroParlred2`.[4] Because we wanted to make the two corpora as comparable as possible, the smaller one had to be the determining one in terms of size. We got also a corresponding `EuroParltst` with the same size in alignment units of `CMPtst`. Comparing the sizes of the resulting unfiltered PTDs, it is clear that the subject diversity is higher in COMPARA than in EuroParl, as should be expected. But

[4] In fact, we have also reduced EuroParl to the same number of randomly chosen sentence units as `CMPred` (90%), obtaining `EuroParlred`, for other experiments.

this also means that when it comes to the filtered ones EuroParl is able to preserve more entries.
Then, we word-aligned the two test corpora, using the corresponding PTD and the "other corpus" PTD. Results in term of token dictionary size are in Table 4: the first column corresponds to the corresponding PTD, the second to the other one.

	COMPARA		EuroParl	
PT	17 433	12 348	17 486	6 570
EN	12 271	9 371	10 927	4 853

Table 4: Size of token dictionaries created after alignment: with dictionaries based on same and other genre, trained on same size corpora.

Results show what we expected, namely that changing genre degrades alignment, and they allow us to measure how much: the size of the PTDs is reduced to 71% and 76% in COMPARA in Portuguese and in English, while for EuroParl the corresponding numbers are as low as 38% and 44%.
This shows that a fiction dictionary seems to be far more unsuitable to align political discourse than the opposite (all things being equal: size of training and test materials).
The differences between languages are not very significant here, although English degrades slightly less in both corpora. This is most probably due to a higher number of different forms in Portuguese (e.g. adjectives have 4 different forms vs. one in English, verbs have more than 70 forms).

4.3. Comparing content

Size may not say much about the quality or accuracy of a given PTD, so we performed a comparison of the (full) PTDs in the colour domain, reusing the extensive lists of colour words appearing in COMPARA that have been manually compiled by the COMPARA team and which are publicly available (Silva et al., 2008). The Portuguese list includes 420 forms denoting colour (218 lemmas), while the English list includes 491 forms (see (Santos et al., 2008) for more details).
We have then computed, for all elements of the lists which stood as entries in the PTDs, how often the best translation was also a colour, and how often at least one of the translations was also a colour. Results are shown in Table 5, where we have used the two versions of COMPARA (forms and lemmas) to see if using lemmas would fare better for Portuguese.
Before we proceed to analyse in some detail these results, it is important to note that colour in politics or in daily life (fiction) has quite a different scope, so it would be foreseeable that the two corpora would provide a different picture altogether of the colour domains in the two languages.
In fact, colour as denoting political affiliation seems to be by far the most frequent meaning in EuroParl, with *green* (3630 occurrences) and *greens* (1173) clearly outperforming the next colour words, namely *white* 2065 and *black* (745) (and note that many White's are proper names in EuroParl). For comparison, note that *blue sky* and *brown hair* are some of the top collocations for colours in COMPARA and probably in general for fiction texts in the two languages.

PTD	Best	Any
CMP	172/382 (45%)	271/382 (71%)
	115/455 (25%)	162/455 (36%)
CMPlemma	97/185 (52%)	125/185 (68%)
	108/455 (24%)	139/455 (31%)
EuroParl	47/112 (42%)	70/112 (62%)
	24/96 (25%)	37/96 (39%)

Table 5: Colour correspondence in the (unfiltered) PTDs: first line, Portuguese, second line, English.

In any case, what we got was unexpected, in that English results lie significantly lower than those for Portuguese. The only explanation is that the threshold of a maximum of 8 translations is much more damaging in the English to Portuguese direction (because Portuguese has more forms) than in the Portuguese to English direction, but while using lemmas should in principle solve or aliviate this problem, it does not (results for `CMPlemma` are even worse).
Further investigation of this issue is thus in order, using the manual annotation of COMPARA (in both languages) to distinguish between colour meaning of a word or another meaning of a homograph.
In any case we proceeded to analyse the corresponding token dictionaries, whose numbers are displayed in Table 6, since the results from the PTDs, as they stand, are not encouraging regarding quality of the overall PTDs created.

PTD	Best	Any
CMP	184/308 (60%)	199/308 (65%)
	111/259 (43%)	123/259 (47%)
CMPlemma	101/152 (66%)	108/152 (71%)
	100/260 (38%)	113/260 (43%)
EuroParl	54/102 (53%)	63/102 (62%)
	31/79 (39%)	33/79 (42%)

Table 6: Colour correspondence in the token dictionaries: first line, Portuguese, second line, English.

The results are definitely better, which brings some hope that the word token alignment may be helpful. Note that not necessarily a colour term should be translated by a colour term. (Some colour terms have other, sometimes even more proeminent, meanings, e.g. *silver*, *orange* or *rose*). Still, the translation/alignment seems to work much better from Portuguese to English, no matter the genre.
Let us also report a previous evaluation of EuroParl PTDs in the Portuguese to English direction, presented in (Simões, 2008). Filtering was done such as keeping only words with frequency higher than 50, and translation probability above 20%. Then 1000 pairs of (entry-word,translation) were randomly sampled and manually evaluated, yielding 85% correctness.
It is difficult to compare to the present results, given that most colour terms would not survive the filtering used.

Corpus	PT-EN		EN-PT	
	Size	TF	Size	TF
COMPARA	71 767	4.77	45 463	3.84
	16 586	4.35	13 734	3.87
CMPred	68 292	4.68	43 603	3.78
	15 410	4.35	12 862	3.88
CMPtst	21 710	4.51	16 256	3.88
	2 631	3.50	2 635	3.47
EuroParl	137 607	5.54	87 511	4.47
	47 220	4.29	30 333	3.77
EuroParlred	137 008	5.57	87 128	4.42
	46 986	4.36	30 191	3.78
EuroParlred2	48 133	5.28	29 742	3.94
	14 547	4.69	10 592	4.10
EuroParltst	18 243	5.73	12 475	4.38
	3 948	4.24	3 488	4.21
EtoP	48 722	3.89	33 623	3.23
	10 538	3.65	9 010	3.49
PtoE	48 230	3.89	29 733	3.45
	8 977	3.57	8 088	3.63
CMPmwe	73 649	4.79	45 429	3.84
	17 144	4.34	13 720	3.88
CMPpos	37 193	4.75	45 431	3.36
	11 628	4.35	13 722	3.39
CMPmweprop	76 478	4.75	45 429	3.99
	16 956	4.35	13 718	3.98
CMPlemma	32 417	4.80	45 429	3.21
	10 935	4.41	13 721	3.28

Table 7: Sizes and translation fertilities of the several PTDs. For each, we present first the unfiltered and then the filtered one.

4.4. Translation direction

To study the import of translation direction – to be more precise, the influence of building PTDs based on texts of only one translation direction –, we created two dictionaries `EtoP` and `PtoE`, obtained by partitioning COMPARA in two parts (original English, and original Portuguese).

We have then measured, as in the EuroParl-COMPARA comparison, the partial overlap and the degradation after alignment with the corresponding part or the other, displayed in Table 8. For comparison purposes, we also used the full PTD created with COMPARA and the one with EuroParl to token word align these corpora.

PTD	language	EtoP	PtoE
same	PT	30 276	20 019
diff	PT	19 798	10 778
same	EN	19 646	12 960
diff	EN	14 279	7 853
full CMP	PT	39 339	38 688
full CMP	EN	25 101	22 429
EuroParl	PT	21 817	19 673
EuroParl	EN	16 124	13 952

Table 8: Size of token dictionaries for the two sections of COMPARA, aligned with several different PTDs.

Note that it is probable that a considerable fraction of EuroParl originated in the English to Portuguese direction, which would predict that using EuroParl would be more successful to align the EtoP subcorpus. But the difference was not high. On the contrary, what was remarkable in the results was that we observed a general marked degradation for the PtoE part. In fact, dictionary size halves for that corpus if we use the other direction PTD (both in Portuguese and in English), compared to a milder degradation for EtoP. This is hard to explain. Given our data so far, we have to accept, or at least not reject, that it comes from harder texts and not necessarily from any language difference. Given that there is no overlap between the English-speaking authors of COMPARA and the Portuguese-speaking ones, for all purposes the two sub-corpora are different and `PtoE` may be more difficult to align.

We can also see from the table that more data (on which the PTD is built) increases performance, but that (in this case) the other genre cannot in most cases supplant the same. (So 60 million words of EuroParl are hardly better than 1.5 million words in COMPARA.)

4.5. Grammatical analysis

Finally, we tried to check whether linguistic processing would improve any of the two kinds of word alignment, by adding several different additional pieces of information (in the Portuguese side only) and comparing the performances. The first change (changed corpus) was recognizing multiword proper names and joining them as one token (`CMPproper`), the second was recognizing multiword expressions from a Portuguese point of view and joining them in one token (`CMPmwe`), and doing both (`CMPmweprop`). It is not obvious that either of these strategies would help, since no corresponding processing is so far done on the English side. (This is work that will have to be done.)

On the other hand, the question of using lemmata instead of word forms may be more promising, given that Portuguese is morphologically richer than English, especially in what verbs are concerned. `CMPlemma` does just that, while `CMPpos` enriches the lemma with the additional POS.

Although it can be claimed that a parallel processing should be done to both languages for ideal results, we think it is interesting to investigate adding this information to just one side as well. [5]

However, apart from presenting their sizes in Table 7, further automatic comparison between these different dicionaries is hampered by the fact that they have different units and therefore it is hard to come up with a meaningful measure. We have nevertheless measured the size of the corresponding token dictionaries, under the assumption that a larger token dictionary size means better performance, displayed in Table 9.

The results show that using frequent MWEs and proper names as entries increases slightly the lexicon as well as TF.

More interesting, though, is the influence that lemmatization has in Portuguese (as source): while it reduces (natu-

[5] Given that languages are different, the most comparable processing would probably use lemma+PAST in Portuguese and no lemmatization at all for English.

Corpus	PT-EN				EN-PT			
	Size	TF	tyAD	toAD	Size	TF	tyAD	toAD
COMPARA	57 198	1.82	1.45	0.03	33 074	2.19	1.59	0.02
CMPred	55 584	2.70	2.20	0.04	32 548	3.97	3.45	0.03
CMPtst	17 433	1.84	1.48	0.10	12 271	2.41	1.82	0.07
EuroParl	115 327	6.97	5.84	0.00	68 090	10.45	8.13	0.00
EuroParlred	115 365	6.93	5.82	0.00	67 421	10.49	8.11	0.00
EuroParlred2	41 551	4.35	3.75	0.02	22 358	6.46	4.86	0.01
EuroParltst	17 486	2.48	2.37	0.06	10 927	3.02	2.64	0.04
EtoP	30 276	2.36	1.47	0.03	19 646	3.16	1.85	0.02
PtoE	20 019	1.83	0.76	0.03	12 960	2.62	1.14	0.02
CMPmwe	60 480	2.73	2.24	0.04	34 209	4.13	3.11	0.02
CMPmweprop	62 681	2.67	2.19	0.04	34 363	4.25	3.22	0.02
CMPlemma	27 348	4.25	3.59	0.02	33 196	3.62	2.64	0.02
CMPpos	30 903	3.90	3.24	0.02	33 087	3.72	2.71	0.02

Table 9: Size, translation fertility and alignment density of the several token dictionaries.

rally) the number of the entries (the size of the dictionary, both PTD and token dictionary), it keeps the translation fertility to the same level. In English as source, the fact that the Portuguese translation is lemmatized does not give so good results, and it even decreases TF.

We note that there is a marked asymmetry in the need or interest of lemmatization or other linguistic processing depending on the source language, that should be investigated in more detail. Possibly depending on the language pair, some features should be merged and others not.

Except for the linguistically analysed corpora (lemma and pos) both translation fertility and alignment density are higher for the English-Portuguese dictionaries (which also have fewer entries). However, token alignment density seems to correlate negatively with type, and we have to investigate the matter more thoroughly.

5. Concluding remarks

In addition to develop a set of particular tools to compare and assess bilingual dictionaries, available to the research community, we have also formulated a set of questions that we attempted to answer, as well as suggested a number of evaluation measures to be employed to characterize these objects.

Starting by these latter, we defined *translation fertility* to characterize a kind of parallel text, or genre, or corpus, by the average number of translation candidates in the PTD. (Strictly speaking, it is the parallel corpus that is being given that measure; but if it is big and representative enough we may talk about the kind of text or genre instead.)

We used *alignment density* to characterize the ratio of aligned tokens in a parallel text after token alignment. While this may be in the first place a measure of the quality of the aligner, note that, assuming an ideal aligner, this measure would be higher for more creative text types than for very predictable ones, in that new translations (or uses) of terms would pop up with higher frequency.

We have suggested yet another creativity measure related to the average size of word correspondences after word-token alignment (ASC), but have not computed it here because it reduces trivially to alignment density in our case, given that, currently, there is no refinement done by our token word aligner, nor any improvement of token dictionaries based on 1-to-N correspondences. It will turn out to be genuinely different and more relevant as soon as PTDs, and token dictionaries, include not just 1-to-1 translations, but N-to-M as well, as linguistically appropriate.

Summing up the partial answers gathered to our research questions, very briefly: (a) we did measure the influence of training in other corpora and training with other genre, (b) we noted – once again as former researchers before us – that the translation direction is relevant; and (c) found out that use of other units of analysis (such as lemmas or lemma-pos sets) brings mixed results.

In fact, it may well that be the most obvious conclusion of this study was that there was a fundamental flaw in the PTD creation design (the absolute limit of eight translations instead of a relative limit based on probability mass), and that to produce linguistically motivated tools and results this has to be changed.

In any case, the result of our work, the tools and the dictionaries, are publicly available for inspection in `http://natools.sf.net/` and therefore readers are welcome to make their own investigations or measures, for example with other term lists or semantic domains.

Acknowledgments

This work was done in the scope of the Linguateca project, contract no. 339/1.3/C/NAC, jointly funded by the Portuguese government and the European Union. We thank José João Dias de Almeida for relevant comments during the development of these tools.

6. References

Yaakov Choueka, Ehud S. Conley, and Ido Dagan. 2000. A comprehensive bilingual word alignment system. In Jean Véronis, editor, *Parallel Text Processing, Alignment and Use of Translation Corpora*, pages 69–96. Dordrecht: Kluwer Academic Publishers.

Ido Dagan, Kenneth W. Church, and William A. Gale. 1993. Robust bilingual word alignment for machine

aided translation. In *ACL Workshop on Very Large Corpora: Academic and Industrial Perspectives (Columbus, OH, Junho 1993)*, pages 1–8.

Ana Frankenberg-Garcia and Diana Santos. 2003. Introducing COMPARA, the portuguese-english parallel translation corpus. In Silvia Bernardini Federico Zanettin and Dominic Stewart, editors, *Corpora in Translation Education*, pages 71–87. St. Jerome Publishing.

Djoerd Hiemstra. 1996. Using statistical methods to create a bilingual dictionary. Master's thesis, Department of Computer Science, University of Twente.

Jussi Karlgren and Magnus Sahlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Journal of Natural Language Engineering, Special Issue on Parallel Texts*, 11(3):327–341.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

Philipp Koehn. 2002. EuroParl: a multilingual corpus for evaluation of machine translation. Unpublished, http://www.iccs.inf.ed.ac.uk/ pkoehn/publications/europarl.pdf.

Philipp Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *MT Summit*.

I. Dan Melamed. 1998. Manual annotation of translational equivalence: The Blinker project. Technical report, Dept. of Computer and Information Science, University of Pennsylvania, IRCS Technical Report #98-07.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond (Edmonton, Canada, May 2003)*, pages 1–10.

Robert C. Moore. 2001. Towards a simple and accurate statistical approach to learning translation relationships among words. In *Proceedings of Workshop on Data-driven Machine Translation, 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 79–86. Association for Computational Linguistics, Morristown, NJ.

Diana Santos and Ana Frankenberg-Garcia. 2007. The corpus, its users and their needs: a user-oriented evaluation of COMPARA. *International Journal of Corpus Linguistics*, 12(3):335–374.

Diana Santos and Susana Inácio. 2006. Annotating COMPARA, a grammar-aware parallel corpus. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik, and Daniel Tapias, editors, *5th International Conference on Language Resources and Evaluation (LREC'2006) (Genoa, Italy, 22-28 May 2006)*, pages 1216–1221.

Diana Santos, Rosário Silva, and Susana Inácio. 2008. What's in a colour? studying and contrasting colours with COMPARA. This volume.

Diana Santos. 2007. Evaluation in natural language processing, 6–17 August. European Summer School on Language, Logic and Information (ESSLI 2007).

Rosário Silva, Susana Inácio, and Diana Santos. 2008. Documentação da anotação relativa à cor no COMPARA, March. First version: 27 November 2007. Current version available at: http://www.linguateca.pt/COMPARA/-DocAnotacaoCorCOMPARA.pdf.

Alberto M. Simões and J. João Almeida. 2003. Natools – a statistical word aligner workbench. *SEPLN*, 31:217–224, Sep.

Alberto Manuel Brandão Simões. 2004. Parallel corpora word alignment and applications. Master's thesis, Escola de Engenharia - Universidade do Minho.

Alberto Manuel Brandão Simões. 2008. *Extracção de Recursos de Tradução com base em Dicionários Probabilísticos de Tradução*. Ph.D. thesis, Escola de Engenharia, Universidade do Minho, Braga, Portugal.

David Talbot. 2005. Constrained EM for parallel text alignment. *Natural Language Engineering*, 11(3):263–277.

Jean Véronis and Philippe Langlais. 2000. Evaluation of parallel text alignment systems: The ARCADE project. In Jean Véronis, editor, *Parallel Text Processing, Alignment and Use of Translation Corpora*, pages 369–388. Dordrecht: Kluwer Academic Publishers.

Andrew Way and Nano Gough. 2005. Comparing example-based and statistical machine translation. *Natural Language Engineering*, 11(3):295–309.