An evaluation of the Translation Corpus Aligner, with special reference to the language pair English-Portuguese

Diana Santos & Signe Oksefjell
SINTEF Telecom and Informatics & IBA, University of Oslo
Diana.Santos@informatics.sintef.no & Signe.Oksefjell@hia.no

Abstract

In this paper we describe the evaluation of a language-dependent aligner. We begin by introducing the alignment program, explaining why it would be interesting to evaluate it with particular emphasis on the language pair English-Portuguese. A short presentation of the corpus used to test the aligner is also given. We then describe three experiments that were performed in the evaluation process, presenting the results and discussing the methodology. The paper ends with a discussion of more general conclusions relative to an evaluation of this kind.

1. Introduction

Two criteria that are often employed in the evaluation of NLP programs are performance and usability. Another criterion, less frequently mentioned, is the adequacy of handling particular languages. The present study describes a set of experiments devised to perform such an evaluation.

Although researchers concerned with parallel corpus building and exploration will generally be happy to use a system available for their languages without evaluating it thoroughly, especially when the system is freely distributed – as is the case of the present system, the kind of work reported originates from two relevant concerns. The first one is about methodological aspects related to the development of NLP systems. The second concern is evaluation and comparison of products. In fact, there is a blatant lack of serious evaluation work of products and systems concerning the Portuguese language, which is a situation we have been trying to change in the project Computational Processing of Portuguese at SINTEF.[1]

The Translation Corpus Aligner (TCA) was developed in connection with the English-Norwegian Parallel Corpus (ENPC) project with the aim of automatically aligning English and Norwegian texts (see e.g. Hofland 1996, Hofland & Johansson 1998). Although the program was originally written for the language pair English-Norwegian, it has been further developed to handle other language pairs, including English-Portuguese. It includes a language-dependent component in the form of an anchor word list.

In the present paper we set out to evaluate the TCA for the language pair English-Portuguese. In particular, we want to

- investigate the effect of the anchor word list;
- compare the results of the program with and without the anchor word list;
- find out how much a proofreader has to check manually after alignment

In order to perform the evaluation, we used the English-Portuguese part of the ENPC, which currently includes 16 English texts, about 12,000 words each, that have translations into Portuguese.[2]

2. A short description of the TCA

The alignment program automatically matches original sentences with their translations. In the process of linking corresponding sentences, the program makes use of an anchor word list that contains word pairs of the languages in question. The one used here was not originally made for Portuguese and English, but was adapted from the English-Norwegian anchor word list. In the alignment process, a value is given to the combination of sentences based on matches in the word list. The program goes through the texts and reads chunks of sentences in each language, resulting in matrices in which the program seeks the highest values for a match between sentences. In addition to the anchor word list, these values depend on the number of characters within the s-units/sentences in each language, on special characters (such as ?!;, etc.), on proper names, and on cognates. One of the strengths of the program is that it does not require any preprocessing in the form of "hard regions", e.g. paragraph alignment, and therefore can get back on track after an alignment error (or in spite of a translation discrepancy).

To take an example, we could imagine the following chunks of text to be aligned:

English original (extract from Doris Lessing's *The Good Terrorist*)	Portuguese translation by Bernardette Pinto Leite
<s>She faced him, undefiant but confident, and said, "I wonder if they will accept us?"</s> <s>And, as she had known he would, he said, "It is a question of whether we will accept them."</s> <s>She had withstood the test on her, that bony pain, and he let her wrist go and went on to the door.</s>	<s>Ela encarou-o, sem desafio mas confiante, e perguntou:</s> <s>— Achas que nos aceitam?</s> <s>— E, conforme sabia que Jasper responderia, este retorquiu:</s> <s>— É tudo uma questão de nós os aceitarmos a eles.</s> <s>Alice resistira ao teste sobre a sua pessoa, à dor óssea, e ele largou-lhe o pulso e dirigiu-se para a porta.</s>

Figure 1. Texts to be aligned

We would expect that some words in the English extract would match some of the Portuguese words in the anchor word list; these matches would in turn enhance the possibility of the program linking the correct sentences. For the second English sentence in the extract above, for instance, we will get the following matches in the word list:[3]

and e
is, 's / é, está
question* / pergunt*, quest*
we nós
accept* aceit*
them / lhes, os, as

The TCA assigns a unique identification to each s-unit with its corresponding s-unit(s) in the translation. When the original sentence has only one corresponding sentence in the translation, we get a 1:1 correspondence. When the original sentence has been translated into two sentences, we get a 1:2 correspondence:

```
<s id=DL2.1.s16 corresp='DL2TP.1.s17 DL2TP.1.s18'>And, as she had known he would, he said, "It is a question of whether we will accept them."</s>

<s id=DL2TP.1.s17 corresp=DL2.1.s16>&mdash; E, conforme sabia que Jasper responderia, este retorquiu:</s>
```

```
<s id=DL2TP.1.s18 corresp=DL2.1.s16>&mdash; É tudo uma questão de nós os aceitarmos a eles.</s>
```

Each text has a unique code, normally starting with the authors' initials, e.g. a text by Doris Lessing has a code DLx. All texts will be referred to by their code.

3. Human intervention required

The program handles 1:1, 1:2, and 1:0 correspondences, i.e. one s-unit matching one, two or zero s-units in the translation; the latter two have to be checked manually. The remaining matrices, containing 1:1 correspondences only, are assumed to be correct and are not systematically checked. Hence, the file that is proofread only includes matrices that do not contain 1:1 correspondences throughout.

Table 1 shows the percentage of matrices that the proofreader has to check, an average of 48.8%, for the 16 English-Portuguese texts. This apparently discouraging percentage needs some explanation since it does not reflect the actual manual intervention that takes place. The picture becomes skewed simply because we immediately associate half of the matrices with half of the sentences in a text. This is not the case, however. Each matrix contains approximately 10-12 sentence pairs, and as will be shown in the matrix below (Figure 2), the proofreader will only have to investigate the s-units that are not 1:1 correspondences. This is to say that although the proofreader has to investigate almost half of the matrices, he will not have to investigate half of the text/s-units, but merely a small percentage.

Table 1. Number of matrices to check

Text	Matrices in output file[4]	Matrices to check	%
ABR1	120	78	65
AH1	131	98	74
AT1	113	30	27
BC1	94	39	41
DL1	90	54	60
DL2	139	100	72
FF1	88	76	86
JB1P	97	32	33
JB1PP	97	32	33
JH1	61	20	33
MA1	74	6	8
NG1	74	37	50
PDJ3	108	49	45
RDO1	143	52	36
ST1	128	87	68
WB1	87	12	14
Total	1,643	802	48.8

To take an example, three of the sentence pairs in the matrix in Figure 2 have to be manually checked, and if necessary, corrected. This can be seen from the low values given in the diagonal of the matrix below (top left to bottom right). Moreover, the numbers below the matrix reflect this; the English sentence 4, for instance, is said to correspond to 0, sentence 5 to Portuguese sentences 4+5, etc. It can be seen, then, that

sentence 4 in the English original is not linked to any sentence in the translation, and wrongly so. We will therefore have to match it to the translation – sentence 3 in Portuguese – manually, correcting the alignment to a 1:2 correspondence. The other two 1:2 correspondences found in the matrix did not have to be corrected.

```
              84  20  71  51  23  56  48 106 133  51
                1   2   3   4   5   6   7   8   9  10  Port. trans.
             ----------------------------------------------------
  1  97 I    6   0   1   2   0   0   0   2   1   1
  2  18 I    1   2   0   0   0   0   0   1   0   0
  3  58 I    2   0   0   1   0   1   0   2   2   0
  4   7 I    0   0   0   0   0   0   0   0   0   0
  5  78 I    2   0   1   4   3   1   2   3   1   0
  6  85 I    3   1   1   4   2   2   6   3   1   3
  7  96 I    2   2   1   3   1   2   1   7   2   1
  8 124 I    1   0   0   1   0   1   0   4   6   0
  9  51 I    1   0   1   2   1   2   2   1   1   2
 10 164 I    1   1   1   4   2   2   2   6   2   0
  •
Eng. orig.
            1,1 2,2 3,3 4,0 5,4+5 6,6+7 7,8 8,9 9,10
-------------------------------------------------------------------------
 1: <s>For her part she did not have to be told that she was wearing <hi
     rend=italic>her look</hi>, described by him as silly.</s>  (DL2.1.11)

 1: <s>Quanto a ela, sabia que estava com o <hi rend=italic>seu olhar</hi>,
     que ele descrevia como de aparvalhado.</s></p>  (DL2TP.1.12)
-------------------------------------------------------------------------
 2: <s>"Stop it," he ordered.</s>  (DL2.1.12)

 2: <p><s>&mdash; Pára &mdash; ordenou ele.</s>  (DL2TP.1.13)
-------------------------------------------------------------------------
 3: <s>His hand shot out, and her wrist was encircled by hard bone.</s>
(DL2.1.13)

 3: <s>Estendendo a mão, apertou com força o pulso da rapariga, causando-lhe
dor.</s></p>  (DL2TP.1.14)
-------------------------------------------------------------------------
 4: <s>It hurt.</s>  (DL2.1.14)


-------------------------------------------------------------------------
 5: <s>She faced him, undefiant but confident, and said, "I wonder if they
will accept us?"</s>  (DL2.1.15)

 4: <p><s>Ela encarou-o, sem desafio mas confiante, e perguntou:</s></p>
(DL2TP.1.15)
 5: <p><s>&mdash; Achas que nos aceitam?</s></p>  (DL2TP.1.16)
-------------------------------------------------------------------------
 6: <s>And, as she had known he would, he said, "It is a question of whether
we will accept them."</s></p>  (DL2.1.16)

 6: <p><s>&mdash; E, conforme sabia que Jasper responderia, este
retorquiu:</s></p>  (DL2TP.1.17)
 7: <p><s>&mdash; É tudo uma questão de nós os aceitarmos a eles.</s></p>
(DL2TP.1.18)
-------------------------------------------------------------------------
 7: <p><s>She had withstood the test on her, that bony pain, and he let her
wrist go and went on to the door.</s>  (DL2.1.17)

 8: <p><s>Alice resistira ao teste sobre a sua pessoa, à dor óssea, e ele
largou-lhe o pulso e dirigiu-se para a porta.</s>  (DL2TP.1.19)
-------------------------------------------------------------------------
 8: <s>It was a front door, solid and sure of itself, in a little side
street full of suburban gardens and similar comfortable houses.</s>
(DL2.1.18)
```

```
 9: <s>Era uma porta de entrada sólida, segura, situada numa ruazinha
secundária com jardins de subúrbios e casas semelhantemente
confortáveis.</s>  (DL2TP.1.20)
--------------------------------------------------------------------------
 9: <s>They did not have slates missing and broken windows.</s></p>
(DL2.1.19)

10: <s>Não lhes faltavam telhas nem tinham vidros partidos.</s></p>
(DL2TP.1.21)
--------------------------------------------------------------------------
```

Figure 2. Example of a matrix calculated by the Translation Corpus Aligner[5]

After the percentage of matrices to be checked had been calculated, the next step was to find out how many corrections one actually had to make. Table 2 gives the percentages of corrections that were made after alignment.

Table 2 presents, for each text, the number of s-units that need to be corrected after alignment with the anchor word list. Since not all s-units have actually been inspected by the proofreader (the percentage of matrices was shown in Table 1), the two rightmost columns give the number of s-units inspected manually, and the corresponding correction percentage.

Table 2. Number of corrected s-units after running the program with the anchor list

Text	Corrections	Total number of s-units	Corrected s-units (% of total)	Number of s-units in the matrices inspected[6]	Corrected s-units (% of inspected)
ABR1	33	1,139	2.9	740	4.4
AH1	42	1,263	3.3	934	4.5
AT1	19	1,102	1.7	297	6.4
BC1	25	893	2.8	366	6.8
DL1	11	855	1.3	513	2.1
DL2	61	1,307	4.7	941	6.5
FF1	48	713	6.7	613	7.8
JB1P	12	934	1.3	308	3.9
JB1PP	23	927	2.5	305	7.5
JH1	15	584	2.6	192	7.8
MA1	2	730	0.3	58	3.4
NG1	12	702	1.7	351	3.4
PDJ3	14	1,041	1.3	468	3.0
RDO1	42	1,396	3.0	502	8.4
ST1	50	1,204	4.2	818	6.1
WB1	5	727	0.7	101	5.0
Average			2.9		5.5

We see that the proofreader has to make changes in about 5.5% of the s-units inspected, which corresponds to about 2.9% of all s-units present in the corpus. Again it should be pointed out that the number of s-units found in the matrices that are inspected is considerably lower than the number of s-units actually inspected.

4. The importance of the anchor word list

We now proceed to evaluate the importance of the anchor word list, by comparing the amount of revision and modification required when running the alignment program with and without language specific information (that is, with an empty anchor list). In order

to estimate the differences, we ran the program with and without the anchor word list and then compared automatically the differences with the final (post-edited) version.

Table 3. Number of differences in the alignment of English-Portuguese texts

Text	No. of "skips"[7]	No. of differences (final version vs. raw version with anchor word list)	No. of differences (final version vs. raw version w/o anchor word list)	No. of differences (with vs. w/o anchor word list)
ABR1	8	33	139	126
AH1	3	44	111	88
AT1	4	27	38	39
BC1*	2	38		
DL1	2	14	31	30
DL2*	6	64		
FF1*	22	60		
JB1P*	3	18		
JB1PP	5	31	45	50
JH1	2	25	19	35
MA1	0	4	12	14
NG1*	0	15		
PDJ3	2	19	11	26
RDO1	2	67	72	65
ST1*	3	68		
WB1	8	6	14	9

The results are shown in Table 3. The texts marked with a star (six out of sixteen) could not be aligned without the anchor word list. Further, Table 3 shows the differences between the final, proofread versions of the ENPC texts and the versions with and without the anchor word list prior to revision.[8]

Due to the simplicity of the (programming) approach (only comparing the target part), missing s-units in the Portuguese translations, such as the (rare) example in Figure 3, were not identified:

```
6: <s>If we got married, we could no longer go back, whether we wanted to or
not.</s>  (ABR1.1.1.242)

 6: <s>Se nos casássemos, não poderíamos mais voltar, quer quiséssemos ou
não.</s>  (ABR1TP.1.247)
--------------------------------------------------------------------------
 7: <s>Neither he nor I.</s>  (ABR1.1.1.243)

--------------------------------------------------------------------------
 8: <s>He white; I coloured.</s>  (ABR1.1.1.244)

 7: <s>Ele era branco, eu mestiça.</s>  (ABR1TP.1.248)
--------------------------------------------------------------------------
```

Figure 3. Example of missing Portuguese translation, overlooked by the `compara_alinhamento.pl` program

It is clear that the English-Portuguese anchor word list does help the alignment program and reduces the number of changes to be made by the proofreader, though perhaps not as markedly as one might expect. However, the fact that 6 out of the 16 texts did not make it through alignment indicates that the program to a large extent depends on the anchor word list, and not only on the other factors mentioned in Section 2.

By using the anchor word list, the percentage of sentences to be corrected drops from 4 to 2% of all the sentences that are manually inspected (which, in turn, can be considered approximately one eighth of all sentences).[9] Again, let us stress that the anchor word list was not originally made for the language pair English-Portuguese. Neither of the anchor word lists was corpus driven; the original anchor list, for English-Norwegian, was manually encoded based on the intuition of the linguists working on the ENPC previous to the choice of the actual texts.[10]

5. Considering the anchor word list in detail

Would more attention to the English-Portuguese pair pay off? What would one do in order to optimize the performance for this language pair, and what would be the net gain? This is what we set out to test in the next step.

We have thus created a program which, for each English source text and its corresponding Portuguese translation, counts

- the number and percentage of the English anchor entries in the English text
    - compared to the entries in the anchor list
    - in terms of the s-units the matches refer to
- the number and percentage of the Portuguese anchor entries in the Portuguese text
    - compared to the entries in the anchor list
    - compared to the total number of target s-units
- the actual successful matches (i.e., the cases where both members of the anchor pair occur in a translation pair)
- the ratio of successful matches vs. all possible matches
    - in terms of the occurrences of the source member of the anchor list
    - in terms of the occurrences of the target member of the anchor list

In order to compute these numbers, several decisions have to be made:

First of all, the anchor word list is unwrapped from the source side, i.e., the cases of A,B / C are transformed into A / C and B / C, resulting in 1,022 pairs (from an original anchor list containing 882 lines). On the other hand, this is not done for the Portuguese side, since it is understood that E / F, G would succeed in either case. In Figure 4, we provide some illustration of what the unwrapping does:[11]

is, 's / é, está	's / ((é)|(está)) is / ((é)|(está))
English* / ingl*	English.* / ingl.*
became, becom* / torn*, volt*, fic*	became / ((torn.*)|(volt.*)|(fic.*)) becom.* / ((torn.*)|(volt.*)|(fic.*))
has, have, 've / tenho, tens, tem, temos, têm	has / ((tenho)|(tens)|(tem)|(temos)|(têm)) have / ((tenho)|(tens)|(tem)|(temos)|(têm)) 've / ((tenho)|(tens)|(tem)|(temos)|(têm))[12]
7*, seven / 7*, sete	7.* : ((7.*)|(sete)) seven : ((7.*)|(sete))

Figure 4. Original and modified anchor list

Then, pattern matching is done so that the matches are only counted in word contexts. Parts of words were not counted as successful matches (so *are* does not succeed in *mare*, for example), except when the pattern expression explicitly says so (like in *becom**).

In addition, the matching ignores case. This is justified due to the low probability of occurrence of non-capitalized instances of words that require capitalization (such as **english* or **inglaterra*).

The most important decision, as well as the one which may be less obvious, is that we only count 1:1 matches as being successful. That is, "successful pairs" in Table 4 below include only those in which the source member is found in a source s-unit which has **one** single sentence corresponding to it and in which, moreover, (one of) the target member(s) of the anchor word list is found. The reason for this restrictive computation of anchor list successes is twofold: If mappings of 1:2 or 1:3 were found,

- they would have been subjected to the revisor's consideration anyway
- it would be considerably more difficult to quantify both the success (should one use a pondered average of target s-units instead of natural numbers?) and the usefulness of the information in the anchor word list[13]

Finally, it should be noted that the anchor list included the entry *&mdash / &mdash*, which might not look specific to the English-Portuguese pair. However, some punctuation – and its translation – is language specific (Santos, 1998b), and therefore the anchor pair including *&mdash* was kept in the list for evaluation.[14]

5.1 Looking at the occurrences of anchor items in the corpus

Table 4 presents the quantitative results for each pair of texts. As far as the English matches are concerned, the first column displays how many members of the anchor list are actually present in the source text (the percentage relative to the total number of anchor pairs is given in the second column). The number of occurrences in the text of the source expressions of the anchor word list is given in the third column under "English matches". The number of different (target) s-units corresponding to some match of the English expressions in the source text is shown in the fourth column (i.e., more than one match of the same English expression per target s-unit is disregarded). As far as Portuguese matches are concerned, the corresponding information is provided: The next two columns give the number and percentage of target expressions of the anchor list which were found in the target texts, together with the number of actual matches. The most interesting information is found under the heading "Successful pairs", which displays how many times pairs in the anchor list were actually found in a translation pair, and what the percentage is relative to the simple occurrence of one element of the pair.

The success matches correspond roughly to a fourth of the source units in which they are found and to less than a tenth of the target units in which they appear. I.e., the relevance of the target expressions seems to be considerably lower than that of the source expressions. This can be explained by several distinct factors. First of all, the English terms were originally chosen with a view to finding a good translation correspondence with respect to Norwegian. This is not necessarily the case for the English-Portuguese pair. Then, the linguist adapting the list[15] failed to note that some target expressions were too general (matching, for example, unrelated and very frequent Portuguese prepositions – example in (4) below), which obviously diminishes the percentage of relevant pairs. Note that this does not necessarily diminish the program's performance, since there is no reason to suppose that the TCA looks for every occurrence in the target language, as we did here for evaluation purposes.

Table 4. Coverage of the anchor list for the English-Portuguese corpus

Text	English matches				Portuguese matches			Successful pairs		
	anchor list		s-units		anchor list		s-units	number	source	target
ABR1	695	68%	7,194	6,454	780	76%	50,226	3,585	49.8%	7.1%
AH1	655	64%	6,865	6,206	723	71%	51,147	2,981	43.4%	5.8%
AT1	624	61%	5,988	5,361	739	72%	48,585	3,177	53.1%	6.5%
BC1	700	68%	4,646	3,933	795	78%	38,381	2,500	53.8%	6.5%
DL1	609	60%	6,400	5,783	716	70%	45,860	3,328	52.0%	7.3%
DL2	623	61%	7,014	6,375	739	72%	48,279	2,761	39.4%	5.7%
FF1	690	68%	6,063	5,246	776	76%	39,095	2,505	41.3%	6.4%
JB1P	616	60%	6,056	5,434	737	72%	47,353	3,507	57.9%	7.4%
JB1PP	616	60%	6,063	5,437	758	74%	42,989	2,866	47.3%	6.7%
JH1	623	61%	5,040	4,417	743	73%	39,991	2,996	59.4%	7.5%
MA1	615	60%	5,448	4,833	735	72%	39,231	3,011	55.2%	7.7%
NG1	675	66%	6,536	5,857	746	73%	43,566	3,717	56.9%	8.5%
PDJ3	718	70%	7,632	6,910	803	79%	55,941	4,072	53.4%	7.3%
RDO1	538	53%	6,522	5,980	665	65%	43,813	3,880	59.5%	8.8%
ST1	692	68%	6,024	5,318	786	77%	49,826	2,781	46.1%	5.6%
WB1	629	62%	4,662	4,030	728	71%	32,463	2,675	57.4%	8.2%
Average	645	63%			748	73%			51.6%	6.6%

In any case, we find a significant number of matches, even if several pairs (at least around a fourth) do not appear in the files. If we divide the number of successes by the number of pairs whose source expression is present in the source text, we get on average 5 matches per pair. Later on we will look more closely at the significance of some elements of the anchor list.

Still, the number of successful pairs in Table 4 does not really offer a good illustration of the contribution of the anchor list, since they take each success independently of the place it occurs in. For example, for a particular s-unit there might be 10 successful matches with ten different pairs, while for another not a single one. This means that the values in the last two columns do not express the percentages of s-units that received positive match in the anchor list, but simply the percentage of occurrence of (source or target) expressions that have been useful for alignment.

So, we made our program also count the (target) s-units for which there was one or more matches for any anchor pair (i.e., the s-units which corresponded to success in finding both the source expression and the target expression). The program also provided the percentage of s-units in terms of the number of translation pairs in the file. These results, which are much easier to interpret, are displayed in Table 5: The first column shows the number of matching anchor-list pairs (after restructuring), and the percentage is given in column 2. The third column shows the number of s-units which had one or more successful matches in the anchor list, and column 4 lists the total number of translation pairs in each file (counted in the target file) for ease of reference. Column 5 is simply the result of dividing the value of column 3 by that of column 4, giving the percentage of translation pairs with successful matches in the anchor list.

We see from this table that the great majority of sentences have hits in the anchor pairs, which may indicate that, when a pair actually occurs in the text, it gives some positive result for the alignment process.

Table 5. Coverage of the successful anchor pairs in terms of anchor list and s-units

Text	Successes				
	anchor list		s-units	total	%
ABR1	563	55%	1014	1139	89%
AH1	488	48%	1069	1263	85%
AT1	496	49%	1030	1102	93%
BC1	557	55%	823	893	92%
DL1	492	48%	783	855	92%
DL2	482	47%	1120	1307	86%
FF1	512	50%	706	713	99%
JB1P	479	47%	882	934	94%
JB1PP	464	45%	879	927	95%
JH1	490	48%	572	584	98%
MA1	508	50%	711	730	97%
NG1	526	51%	653	702	93%
PDJ3	586	57%	970	1041	93%
RDO1	430	42%	1221	1396	87%
ST1	521	77%	870	1204	72%
WB1	515	50%	719	853	84%
Average	506	50%			91%

5.2 Looking at some anchor pairs in particular

We expect, however, that not all anchor pairs have the same import, and in fact we see from Table 5 that, on average, only half of them work for each text. That is why we wanted to take a closer look at the results.

From a rough inspection of the results, and along with more accurate pairs, the two following cases were easy to detect:

- systematic translations were forgotten (like the weapon sense of *arm** or the *artist** continuation of *art**)
- translation involves too much restructuring, due to language differences, so that lexical clues are not the right place to look (as is the case of *I* or *be*).

In order to look more closely at the results, we selected several frequent cases – pairs that were most successful (in absolute terms) – and looked at their distribution in the different texts. Table 6 shows the number of occurrences in the English source and those that corresponded to a match in the Portuguese translation. Texts in the Brazilian variant are identified in bold in the table.

Table 6 shows that some pairs are more useful than others, and also how this may vary for individual texts. Although we have to leave a detailed discussion of these results for another forum, we provide here some preliminary comments.

One case that is worth discussing is *I*/*eu*, which is interesting in that *eu* is considerably more rare in Portuguese (being a null subject language) than *I* in English. One can, however, notice that there is a considerable difference in utility (or translation correspondence) from text to text, and even note that, as expected, in the Brazilian variant there is a less marked tendency to dispose of the personal pronoun than in European Portuguese.

The opposite happens with dashes (coded *—*) in terms of relative frequencies: dashes are used much more frequently in Portuguese than in English. It is

interesting to argue for a different behaviour of the aligner in such cases, i.e., a different weighting of pairs related to criteria such as these. (The lower the frequency of the target item, the higher the probability that, being there, it is a translation of the source term.)

Table 6. The distribution of some frequent anchor pairs

pair	**AB**	**AH**	**AT**	BC	D1	D2	FF	**JB**	JB'	JH	MA	**NG**	PD	RD	ST	**WB**
&mdash	57	65	53	66	66	36	8	37	37	8	7	130	17	436	20	36
&mdash	247	454	343	305	236	507	12	47	130	36	33	127	210	445	389	234
	16	25	46	50	22	22	10	30	33	7	6	107	7	426	4	32
's	82	75	38	2	40	71	11	98	92	0	69	50	67	62	39	16
é, está	504	520	300	304	320	432	100	576	608	236	584	264	504	316	504	144
	30	35	12	2	21	27	3	55	57	0	33	29	34	33	15	11
I	238	171	73	84	43	115	39	316	316	11	217	110	124	322	80	205
eu	105	61	32	20	11	39	1	85	122	4	37	53	35	109	43	71
	87	40	19	17	8	19	1	62	104	3	32	43	22	94	30	63
be	65	92	41	21	65	66	64	54	54	48	38	61	80	24	61	19
ser, estar	172	116	132	60	152	120	84	172	248	160	116	172	172	68	220	60
	18	17	14	5	13	9	9	19	29	15	11	24	20	3	8	11
been	62	45	33	21	53	42	62	28	28	35	13	41	55	19	45	24
sido,	64	36	40	64	72	48	28	40	64	72	32	36	100	68	84	12
estado	13	0	7	7	10	4	5	4	11	10	4	5	18	6	7	2
could	29	19	48	18	33	41	45	25	25	54	13	33	68	46	32	21
podia.*, etc.	132	120	216	96	114	66	168	126	138	180	42	222	180	150	108	90
	12	10	28	8	7	6	18	10	9	24	6	20	25	19	9	10
is	65	51	14	41	29	47	25	60	60	53	130	25	45	19	35	21
é, está	504	520	300	304	320	432	100	576	608	236	584	264	504	316	504	144
	34	29	12	26	13	17	7	38	41	28	69	17	24	15	16	11
his	81	81	167	142	66	57	157	59	59	158	37	95	153	95	99	94
seu*,	220	272	488	364	264	212	268	124	364	360	136	488	496	272	208	224
dele*	19	12	64	52	24	13	37	7	23	61	12	44	60	25	15	26
there	52	54	42	23	47	44	43	38	38	47	64	58	51	79	23	31
lá, ali	144	88	84	32	124	120	28	172	68	36	80	112	140	180	152	68
	14	11	5	2	14	7	2	11	10	8	8	21	10	18	3	6

It is, however, also possible to see great variation without having either a significant frequency reduction or increase of the corresponding terms in the two languages, as the cases of *been* and *sido/estado* show. More detailed studies of the subtleties of the translation into Portuguese of *could* and of *there* can be found, respectively, in Santos (1998a) and Ebeling & Oksefjell (1998).

5.3 Looking at some instances of successful matches

One thing worth noticing is that, even though much of the data presented above could be explained with some linguistic ingenuity, it is not the case that all instances that were counted as successful matches are actually linguistically motivated. In fact, it was easy to detect three interesting situations from a crude inspection of the results:

First, even though the result might be correct, some of the resulting pairs "succeeded" without any sort of linguistic motivation; cf. (1) - (4):

(1) little / pequen.*?: *À sua frente, pelo lado esquerdo do caminho, ia um grupo de crianças, a mais velha a empurrar um carrinho, com duas crianças mais pequenas, uma de cada lado, agarradas ao varão.* (Translation of: Ahead of him, trudging along on the left of the path, was a little group of children, the eldest girl wheeling a pushchair with two smaller children, one each side of her, clutching the bars, PDJ3)

(2) be / ((ser)|(estar)): *Lá um **ser*** ('being', n.) *humano sempre será um estranho, jamais "pertencerá".* (Translation of: In there, one will always **be** a stranger, will never "belong", ABR1)

(3) look.*? / olh.*?: *Não há quebra-luz, apenas a lâmpada por cima que dá à minha cara um aspecto pálido e doentio, com **olheiras**.*(Translation of: There 's no shade on the light, just a bare bulb overhead, which makes my face **look** pallid and ill, with circles under the eyes, MA1.)

(4) couple.*? / par.*?: *Depois de alguns anos **para*** ('for') *se estabelecer, ele começa a escrever.* (Translation of: After a **couple** of years to settle down he begins to write again, ABR1).

Secondly, it was also noted that the use of the Kleene star, instead of a more rigid morphologically closed list, allowed for correct matches which would certainly not have been listed by a linguist (due to considerations of frequency), as is the case of (5)-(6). Incidentally, a particularly valuable feature of the anchor list format was the ease of specifying cross-categorial translations (i.e., part-of-speech change in the translation).

(5) buil.*? / constru.*?: *Terminada a obra, os **construtores** serão mortos.* (Translation of: Once the job is finished the **builders** are killed, ABR1.)

(6) crim.*? / crim.*?: *Então, perto do final de 1347, uma pequena frota de cerca de uma dúzia de galeras genovesas chegou à Sicília vinda de um lugar distante, talvez da **Criméia**, e em poucos dias a população de Messina começou a morrer às centenas.* (Translation of: Then, towards the end of 1347, a small fleet of about a dozen Genoan galleys arrived in Sicily from somewhere far away, perhaps the **Crimea**, and within a few days the people of Messina began to die in their hundreds, ABR1)

Thirdly, example (7) displays a match which works despite the linguistic data it is supposed to mirror and/or despite the inaccurate translation. In fact, *earlier* in (7) was translated simply by *cedo* ('early'), leaving out the comparison (a literal translation would use *mais cedo*), but the fact that one has chosen to simply list the absolute form made the anchor pair more useful.

(7) earli.*? / cedo: *Ela trabalhava mais que qualquer um, acordava **cedo*** ('early')*, era a que ia dormir mais tarde.* (Translation of: She worked harder than anybody else, got up **earlier**, came to bed long after the others, ABR1)

In conclusion, it was possible to detect several pairs that led to "successes" without having been thought of as pairs beforehand. In some cases this increases performance, in others it diminishes it. Without looking at every pair that counted as a success, it is difficult to estimate the actual percentage of these cases.

The overall performance of the aligner did not appear significantly diminished, though, since, as mentioned above, there can be many possible unrelated "successes" for each s-unit. This rather illustrates that the TCA's approach, while not 100% linguistically motivated, seems to compensate on accounts of both simplicity and robustness.

6. Concluding remarks

Even though it was already known from the beginning that the TCA had been of extreme help in aligning the English-Portuguese texts of the ENPC, we think that the evaluation performed was interesting.

On the one hand, it measures the actual consequences of using this particular program for a given language pair, which is something that had not been done before. It is then up to the individual user to decide whether s/he should use this tool in his or her project.

On the other hand, this investigation also gives origin to some reflections concerning the evaluation process proper and the role of language dependency in NLP tools:

- One of the undeniable conclusions of the present study is, in fact, that it is a time-consuming and complex task to evaluate a particular program, and that it is often necessary to create auxiliary tools to evaluate it. It is, therefore, understandable that, in most cases, tool developers have no time to implement a thorough evaluation of the tools they develop because they employ their time in improving them (adding more functionalities, for example), not evaluating them. It is, therefore, understandable that, in most cases, tool developers have no time to implement a thorough evaluation of the tools they develop. Furthermore, it is often unclear whether the time used in this kind of evaluation is worthwhile, given that the program has a satisfactory behaviour for the task it was meant to perform.
- While adapting a particular language-dependent tool for other typologically similar languages, it seems that performance degradation is rarely so severe that it prohibits reuse of the tool. It is, however, seldom quantified what the changes are, and how well the "new" tool works when adapted to the new language (in this case, new pair of languages). That is why the present study may be interesting beyond the particular tool and pair of languages considered.
- Our preliminary conclusion is that even language-dependent tools rely to a lesser extent on languages than one would expect. Therefore, it is often feasible (and useful from an engineering point of view) to adapt that program to similar languages without considerable loss of performance. But see Santos (in press) for a more detailed discussion of language dependency and its methodological consequences.

As to the particular work reported here, we note that there are two other ways in which we could proceed for a thorough evaluation of the TCA for the English-Portuguese language pair:

- The first way would be to experiment with the constitution of the anchor list, creating new anchor lists, including for instance only the 20 most frequent pairs, or only the translations of the most frequent English content words, and compare the performances obtained.
- The second way would be to apply new texts and their translations to the TCA, and see whether this would result in any appreciable decrease in performance.

Unfortunately, both tasks will have to wait for another occasion, since they amount to considerable work.[16] Another revealing experiment would be to mimick the whole evaluation process for the English-Norwegian pair and compare the results.

7. Acknowledgements

We are grateful to Knut Hofland for making his program widely available and for answering our questions about it. In addition, the present paper was considerably improved by Stig Johansson's comments on both structure and content. We also thank Ana Frankenberg-Garcia for commenting on a previous draft.

Notes

[1] See http://www.portugues.mct.pt/.

[2] One of the English texts has been counted twice since two translations of it have been included in the corpus; one into European Portuguese and one into Brazilian Portuguese. For a complete list of authors and texts in the corpus see the ENPC homepage at http://www.hf.uio.no/iba/prosjekt/. For a more detailed description of the ENPC, see Johansson et al. (1999).

[3] This is the original format of the TCA wordlist. "*" means truncation, "/ "separates source and target items, "," separates items inside each language.

[4] In other words, the total number of matrices in the text.

[5] The first line of underscored numbers gives the number of characters in each Portuguese sentence, the second line the sentence number. The vertical italicized lines give the sentence number of each English sentence and its length in number of characters.

[6] We assume that the percentage of s-units inspected is equal to the percentage of matrices shown, therefore computing the present numbers by multiplying the total number of s-units by the percentage of matrices to check (displayed in Table 1). This is an approximation, because the same s-unit can belong to two contiguous matrices.

[7] It is important to note that in the process described so far, so called skip-attributes had already been inserted in the text files. Skip attributes are commands for the program to "skip" a particular sentence, and they are one of the few tags and attributes added by the ENPC project to the basic recomendations put forward by the Text Encoding Initative (Sperberg-McQueen & Burnard, 1994). The purpose of skip-attributes (in either language) is to mark that a sentence has not been translated, or has been invented. Although this is a sort of human intervention, it is hard to quantify, because it is usually inserted in an interactive mode when and if the program does not manage to continue. Here we have chosen to disregard them.

[8] These differences were computed automatically with the help of a simple Perl program (`compara_alinhamento.pl`), running after the Unix command `diff` was invoked between the resulting files.

[9] Assuming that for each matrix to be observed, 3-4 sentences have to be read by the human reviewer, this means a fourth of all sentences present in the matrices. Then, only half the matrices are revised: ½*1/4=1/8

[10] Knut Hofland (p.c.) revised and improved the list, using a program that computed a threshold of usefulness for very frequent items, and removing those which would not help the alignment.

[11] The "." and the parentheses in the reformulation are simply the rendition of the same semantics in Perl syntax and have no significance for the success of the pattern matching.

[12] From the unwrapping we see clearly that, given that Portuguese has more verbal inflections than English (whereas the opposite happens to Norwegian), one should have written the pairs `has/tem` and `have, 've/tenho, tens, temos, têm` instead.

[13] For example, one would expect the presence of the target member of the pair in a sentence contiguous to the right one (the one in the final alignment) as a measure of the negative influence a particular word-pair would have. When one sentence is aligned with two sentences, one of which would quantify as success in terms of our computations, one could expect a favourable import from the anchor word list if the alignment is right, and the opposite if it is not right. So, a good measure of the anchor list doings would have to take into account the percentage of e.g. 1:2 alignments that were correct. To do this would most probably require mimicking the way the program works, which is outside the scope of the present paper.

[14] In fact, even "invariants" such as numbers can actually result in a different translation. Floors are counted differently in different languages: ground floor in Portuguese is Norwegian first floor; soccer's first division in Norway is second division in Portugal; a person measured in feet will not have a

corresponding height in meters, etc. And Figure 4 reminds us that *seven hundred* does not get translated into *sete centos* but into *setecentos* (as opposed to Norwegian, which has two words as well).

[15] Incidentally, one of the authors of the present paper (D.S.).

[16] In order to compare the performance of the system with a revised version, one has to manually proofread the results first. For the case of the ENPC we had access to the proofread files, which is obviously not the general case.

References

Ebeling, J. & Oksefjell, S. 1998. On the translation of English there-sentences into Norwegian and Portuguese. What does a translation corpus tell us?. In Ydstie, J. T. & Wollebæk, A.C. (eds.), *Working Papers in Applied Linguistics 4/98*, Oslo: Department of linguistics, Faculty of Arts, University of Oslo. 188-206.

Hofland, K. 1996. A program for aligning English and Norwegian sentences. In Hockey, S, Ide, N. & Perissinotto, G. (eds.), *Research in Humanities Computing*. Oxford: Oxford University Press. 165-178.

Hofland, K. & Johansson, S. 1998. The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In Johansson, S. & Oksefjell, S. (eds.), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi. 87-100.

Johansson, S, Ebeling, J. & Hofland, K. 1996. Coding and aligning the English-Norwegian Parallel Corpus. In Aijmer, K., Altenberg, B. & Johansson, M. (eds.). *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies in Lund* (Lund, 4-5 March 1994). Lund: Lund University Press. 87-112.

Johansson, S., Ebeling, J. & Oksefjell, S. 1999. English-Norwegian Parallel Corpus: Manual. Oslo: Department of British and American Studies, University of Oslo, http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html.

Santos, D. 1998a. Perception verbs in English and Portuguese. In Johansson, S. & Oksefjell, S. (eds.), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi. 319-342.

Santos, D. 1998b. Punctuation and multilinguality: Reflections from a language engineering perspective. In J. T. Ydstie & A.C. Wollebæk (eds.), *Working Papers in Applied Linguistics 4/98*, Oslo: Department of linguistics, Faculty of Arts, University of Oslo.138-160.

Santos, D. In press. Towards language-dependent applications, *Machine Translation* 14 (1999).

Sperberg-McQueen, M. C. M. & Burnard, L. (eds). 1994. *Guidelines for electronic text encoding and interchange. TEI P3*. Chicago & Oxford: Association for Computers and the Humanities / Association for Computational Linguistics / Association for Literary and Linguistic Computing.