# DISPARA, a system for distributing parallel corpora on the Web

Diana Santos

SINTEF Telecom & Informatics, Pb 124 Blindern, NO-0314 Oslo, Norway
Diana.Santos@sintef.no,
WWW home page: http://www.portugues.mct.pt/Diana/

**Abstract.** The main purpose of the present paper is to document the process of creating a parallel corpus available on the Web, thereby illuminating technical and design issues involved in such a project. By this we hope to gather more researchers to help with the building process, as well as boast considerably the number of users of the parallel corpus.
We start by noting that resource creation is far from a trivial process, and proceed by providing a brief introduction to COMPARA, the particular parallel corpus in connection with which the present system was developed, although with a view to achieve a general architecture. In the following sections we describe the incremental building process in DISPARA, emphasizing the reuse of software components, and discuss the Web interface. We conclude by discussing remaining work and emphasizing the importance of user feedback.

## 1 Introduction

Most people outside the corpus linguistics community have absolutely no idea about what is involved in making a corpus available (through the Web). In fact, there are quite a number of misconceptions regarding corpora [14]. Johansson [10] has convincingly argued that creating a corpus is like editing a book; still, many corpus users fail to distinguish between carefully edited corpora and text masses unjudiciously amassed.[1]

Describing thus in detail the process of creating one particular parallel corpus and several issues involved in its creation, distribution and protection can help users making such distinction, as well as teaching them to take a more critical look at what is loosely called "corpora".

By making public the computational work required to build this particular Portuguese-English paralell corpus, we also hope to foster practical collaboration from readers, in order to have a quicker development, as well as enlarge significantly the range of the corpus users.

We will not dwell here into the need and usefulness of parallel corpora, taking it for granted [21] [11] [2].

[1] Both can be useful, but for different purposes.

## 2 A Portuguese-English parallel corpus created from scratch

COMPARA, the corpus in connection with which the present system was developed, was created as a collaboration between two teams, Ana Frankenberg-Garcia's at ISLA and ours at SINTEF, both with considerable experience in the use of parallel corpora, albeit with different perspectives of use: the first mainly related to language teaching, translation training and error analysis, the second specifically concerned with contrastive linguistics and natural language processing.

This provided a rather broad range of interests for devising a corpus dissemination system. In order to profit from the two backgrounds, design and system issues have been prepared and planned together from the start.

One important motivation for the whole project was the belief that publically available resources for Portuguese linguistics and NLP would greatly benefit all people concerned with Portuguese processing; thus access to COMPARA is freely granted through the Web[2], and the DISPARA system was all along implemented with such a aim, as all other resources built under the framework of the Computatonal Processing of Portuguese project.[3]

## 3 The process of building a corpus to be used in the DISPARA system

The amount of work that goes into the seemingly simple process of creating a corpus may seem surprising for the reader. In this paper we will **only** be concerned with the actual processing, not copyright clearance and/or tracing, which was obviously successfully done beforehand. It is, however, instructive, both for prospective corpus builders (in order to plan realistically their work) and for corpus users to know what lies in a resource now available.

For COMPARA, there was a physical separation between the team concerned with the texts and the team concerned with the programs (henceforth the "text team" and the "programming team"), respectively located at Lisbon and Oslo. While this produced an obvious overhead in e-mail communication, it brought the benefit of structuring very rigorously each team's task.

### 3.1 Outline of the creation process

**Phase 1.** The first task was to get the texts (English and Portuguese) in electronic form. This was achieved either by using previous electronic editions (e.g. from Projecto Vercial [20], the Gutenberg Project [7], the Biblioteca Virtual do Estudante Brasileiro [3] or the English-Norwegian parallel corpus[9]), or – in the majority of cases – by scanning and proofreading the result.

[2] `http://www.portugues.mct.pt/COMPARA`
[3] `http://www.portugues.mct.pt/`

The texts were then manually aligned at paragraph level with the help of standard editing facilities of word processor software. Markup concerning titles, foreign words and phrases, and other typographical emphasis was added (see [5] for the algorithm used) whenever subsentential parts of a sentence were highlighted in the original text. Obvious typos were corrected and marked, and translator's footnotes were physically brought to the place to which they refer, and marked up accordingly.

**Phase 2.** The texts were then sent to the programming team, which separated both texts into sentences (using the tools developed in another project dealing with Portuguese corpora [17]), and aligned them at sentence level using EasyAlign (version 1.0), a tool pertaining to the IMS Corpus Workbench [4].

In order to make the best use of the aligner while not losing any information already manually encoded in the text, two things had to be catered for, previous to invoking it: First, the markup should not be taken into consideration when aligning – this is solved by creating, for each specific pair of texts, intermediary corpora in which markup is encoded as structural attribute. Second, translators' notes have to be physically extracted and saved in an external file in order not to confound the automatic aligner, which employs text length as one decision criterion.

**Phase 3.** The result of the automatic alignment is then sent to be revised by the text team, in order to make it conform to the 1-to-any policy followed [6] and thereby defining alignment units (**ua**s). Information about sentence reordering, sentence addition or dismissal, and type of complex alignment, is inserted in this pass (always in the translation **only**). Examples of reordering, addition and complex sentence splitting are displayed in figures 1 to 3.

File EBDL1T2fra.en:

```
917. <p><s> I shrugged.
918. <p><s> `` I don't know.
```

File EBDL1T2fra.po:

```
917. <p><s2> <reord resp=trans place=1> -- respondi, dando de ombros.
</reord>
918. <p><s2> -- Não sei. <place 1>
```

**Fig. 1.** Markup inserted in alignment units in case of reordering

**Phase 4.** The files are then sent back to the programming team, which (automatically) performs the following subtasks: create pairs of texts using the manually revised alignment, insert alignment type and alignment id, put translators' notes in the right place, and create special structural attributes to deal with reordering.

We chose to process automatically the notes, even though it might have been less time-consuming, in the first versions of the corpora at least, to have edited them manually in place after the rest of the processing. Automatic processing

File PBMA3fra.po:

```
591. <p><s> Mamãe é boa demais; dá-lhe atenção demais.
592. <p><s> Parece até que chorou.
593. <p><s> - José Dias?
```

File PBMA3fra.en:

```
591. <p><s2> Mamma's too good; she pays him too much attention:
592. <p><s1/2+1> there were even tears...  <add>  Who cried? </add>
593. <p><s> José Dias?
```

**Fig. 2.** Markup inserted in alignment units in case of addition

File PPMC1fra.po:

```
648. <p><s> Era um homem simples.
649. <p><s> Limitava-se a cumprir.
```

File PPMC1fra.en:

```
648. <p><s2> He was a simple man;
649. <p><s2> he limited himself to complying.
```

**Fig. 3.** Markup inserted in alignment units in case of sentence merging (1-1/2)

means that the program has to recover the right position and add the text of the note (preserved during phase 2) inserted as value of an attribute. Due to particularities of the underlying corpus encoding system, the text of the notes itself had to be neutralized for all the "dangerous" characters that might disrupt the normal behaviour of the corpus building tools (tokenization, etc.), and then it had to be restored later so that users had access to the right contents.

The question of reordering, although even rarer (at the date of writing this paper, 5 cases against 97 translation notes, with 14 pairs of texts fully processed), was still more complicated to automate, since it required access to a larger text span than the other cases dealt with. In fact, a reordering of alignment units can only be seen if displaying more than one **ua**, and one of the basic design principles of DISPARA was that it was **ua**-based. The solution found was to create automatically a special structural attribute, called **ur** (unit of reordering) from the original markup (attributes <reord n> and <place n>), encompassing the **ua**s containing these two tags. This provided half the solution, namely for the cases when one was looking at translated text (because only in the translation one can talk about reordering). To be able to recover the right size **also** in the source text, one more programming trick had to be used, namely the automatic creation of a program effecting several substitution commands to be run immediately after, in order to insert the corresponding **ur**s in the source part(s).

Two reciprocally aligned corpora are finally created, together with automatic documentation (counting tokens, types, and **ua**s), namely the (parallel) files Contents.html and Conteudo.html, one in each language.

For each new pair of texts added to the corpus, the Web interface must be accordingly manually updated, to allow the specific selection of that particular text pair.

**Phase 5.** The text team proceeds, then, already through the Web interface, by testing whether the automatic labelling of the 1-to-many cases conforms to their interpretation of what a sentence in the target language should be. In fact, there are cases far from clear-cut; opinions may differ between the original program writer and the text team; and, most importantly, the programs rely mainly on punctuation clues only, not necessarily present in all cases.[4] Every case where it is considered necessary to change the automatic judgement is written down in a special text file[5], which is sent to the programming team.

**Phase 6 (repeatable).** The two corpora are then created once more, taking into consideration any `div` files present, and which superseed the alignment type automatically found. Any remaining errors or problems, whenever discovered, provoke changes in the files after the manual revision of the sentence alignment, and subsequent corpus creation is done again, as many times as necessary.

### 3.2 Some comments on reuse

The whole process was designed obeying the following principle: make use of as many automatic tools as were available, minimizing human workload, though still giving the linguist the last word.

Paragraph alignment is done manually because it appeared an extremely simple task – few paragraph changes, and because the linguist would still have to browse through the whole text so that no gross errors appeared, while at the same time inserting the few markup information tags and translation notes in the right place.

Sentence separation inside each paragraph is done automatically reusing the sentence separation programs developed for huge Portuguese corpora in the context of the AC/DC project. Although these programs were developed for Portuguese, they were easily adapted to English literary text (remember, in any case, that there is also manual revision on this subject).

Sentence alignment is automatically performed by an off-the-shelf tool (Easy Align), buffered by a pre-processing stage, implemented in order to remove potential problems to the aligner, and a post-processing phase, added to get the result in a linguist-friendly version for subsequent revision.

Marking up, for each alignment unit, the kind of alignment type was also implemented as automatically as possible, by using (again) the sentence separation module, in all cases where the target part of the alignment unit was not

[4] The very same programs have been evaluated in the context of CETEMPúblico, a large newspaper corpus of text from the daily newspaper PÚBLICO [18]; considerable disagreement about the linguistic definition of sentence has also been met with in the Floresta Sintá(c)tica project [1], leading to significant revision [16].

[5] A `div` file, containing the right alignment kind to be used in COMPARA, one per line for each **ua** that requires change, e.g. 156: 1-1.

composed of parts of sentences (which had been previously manually encoded by the text team in phase 3, see figure 3). As mentioned above, this was still subject to scrutiny in phase 6 (and eventual change in following versions of the corpus).

So, the reuse policy is clear: Everything that can be done automatically is automated, but the corpus's creators, with their linguistic sensitivity, have always the last word and may require modifications. These modifications are **not** performed by manual edition, though. They are rather encoded and supported in the building process, in order to prevent having to perform the same modification more than once. This may bring a programming overhead compared to other corpus projects, but we believe it the right way to go.[6]

## 4 Resulting Web service

One thing is the process used to create a resource; another thing is the way this resource can be put to use or interrogated (generally described as its user interface). Still another thing are actual patterns of use of a given resource.

In the previous section we described the process of creating the kind of resource the system was designed to build, namely a parallel corpus. The present section is concerned with explaining some of the fundamental options of the service that was implemented on top of the resource: both some that are visible in user options and others which are implicit in the system's behaviour.

An important issue is the clear-cut distinction between a search condition and an output option, even when they are closely related.

Take the interface choices "Search for translation notes" and "Hide/show translation notes". Although, in most cases, one can expect that the user will select one particular combination (if she is looking for notes, then will want to see them), the two options are, from a design point of view, independent. In fact, one may interrogate the corpus looking for alignment units that have translation notes attached, without being interested in the specific content of the translation notes. Or, conversely, one may want to see the translation notes if the search (which may well have obeyed other criteria) happens to find **ua**s with them associated.

The same is true about alignment properties, associated with both search and display options: One may want to have a closer look at the alignment kind, by selecting all cases different from 1-1, or in which the translator added or removed whole sentences. Or, one may just want to see the alignment type when looking for non-related questions, such as inquiring whether sentences with perception verbs trigger more sentence restructuring in translation than other constructions (as hinted in [12] for the English-Portuguese pair).

Finally, another obvious example contrasting search and display modes, is either selecting specific parts of the corpus, or looking at the whole corpus, and

[6] Note that the reused tools themselves are not static, and may have upgrades that require rebuilding the corpus – it would be an overkill to do manual modifications every time one would want to recreate the corpus.

display the distribution according to source. [7] A little thought should be enough to convince the reader of the different purposes of these two interrogation modes.

Another interesting property of DISPARA is the use of **one** original sentence as definitory of alignment unit. This was motivated by the requirement that DISPARA should handle several translations of the same original, which, in order to be easily compared, had to have the same source side. As far as we know, this is one original feature of our system (one which actually did bring about a lot of extra work since the aligner used – and we believe most aligners – have result categories like 2-1 and 2-3 and 2-2[8]). Note, on this particular, the definition of alignment unit: one has a 1-1 alignment unit even if the target sentence has been moved (reordered) to another place.

DISPARA also provides an original display capability, which we believe very useful for contrastive studies, termed a "quantitative wrapup". Basically, given a parallel search in both languages, i.e., a search specifying conditions on **both** the first and the second language[9], a quantitative wrapup presents the number of hits obeying only the first condition, the number of hits obeying both, and the number of hits conforming to the second restriction. This is meant to give a sketch of the "translation match" between one word, expression or construction in one language and another word, expression or construction in the other. Some examples are shown in Table 1:

**Table 1.** Examples of quantitative wrapup, in COMPARA's version 1.1

| Search | Quantitative Wrapup |
|---|---|
| (translation from P to E) enquanto - while | 46 20 51 |
| (All P to E) enquanto - while | 178 92 162 |
| (translation from E to P) while - enquanto | 111 73 132 |
| (All E to P) while - enquanto | 162 93 178 |
| espert.* - clever | 8 2 7 |
| clever - espert.* | 7 2 8 |
| home - casa | 113 92 284 |
| home.* - casas* | 127 98 313 |
| casa - home | 284 98 113 |
| casa - home* | 284 104 127 |
| thinks* OR thoughts* - pens.+ | 470 173 277 |

[7] In fact, as soon as we have more texts so that such questions really make sense, we should add the following distribution functions (corresponding to search options already implemented): by date; by date of translation; by variant; by variant of translation; by text type (translated or original).

[8] An interesting question, which we will have to leave to another paper, is how to evaluate different aligners with different output behaviour.

[9] Depending on the search direction, the first may be the source language and the second the target, or vice-versa.

The above data seems to imply that *enquanto* is a pretty good translation for *while*, and vice versa; as far as adjectives are concerned, the mismatch between *clever* and *esperto* is, on the other hand, obvious. We can also appreciate that *casa* is much more general than *home*, while *home* is an almost perfect match for (one of the senses of) *casa*. Finally, we see a rather bad match of the verbs *think* and *pensar* and their corresponding nouns.

DISPARA itself is parallel in both languages – error messages, titles, numbers, etc., appear in the language of the interface used. Since this is done in a modular way, one could actually add further languages for interaction as well, which might come in handy when new corpora in other languages are added.

## 5 Concluding remarks

Although we consider the system already useful (see results presented with fewer data in [19]), we are aware that a lot may still remain to be done. This will mostly be discovered through actual use of the system, both directly and by observing others using it.

For example, one of the first and probably most common sources of error by unexperienced users was their attempt to use phrases without conforming to the IMS-CWB syntax. As soon as we discovered this through browsing the logs, we added such information right above the input box, so that users were informed before erring.

Other detected (and now solved) problem was the lack of interaction between the quantitative wrapup facility and the selection of parts of the corpus. In previous versions, only the columns referring to the source text were restricted, not the target (thus yielding uninterpretable numbers).

Other initially too loose definition of the semantics of the interface had to do with variant selection. We assumed, although not stated it, that variant choice concerned exclusively the (language of the) original text. This choice had the unwelcome effect of permitting a user to choose an empty corpus: For example looking at patterns from English to Portuguese, from originals to translations only, if one selected e.g. Brazilian Portuguese. Now, the semantic interpretation of this choice in the interface has been changed: Depending on the search direction, variant is contextually interpreted to refer either to original or to translation.

It is not an easy task to detect mismatches between user expectations and system behaviour. Although we are still in the preliminary stages of a log-based usability study of DISPARA, expecting to uncover somehow the kinds of users the system has nowadays, and the kinds of queries and problems they face, we are aware that an extremely important source of feedback is actual contact with users, which we strongly encourage.

Some things, however, are obviously missing. One is tagging or parsing of both sides. This is not in our short-term plans, though, because there is no free tagger or parser for Portuguese that we could run on the COMPARA texts (and most of them, due to the very narrow copyright terms, cannot even be sent to a third party for parsing purposes).

We are also aware that an ideal display of results has not been achieved yet for the cases of more than one translation of a given original: Ideally, when displaying concordances, the source text would only appear once, with the two (or more) translations on the target side, if the user query concerned simply the source. Likewise, each source text should only count as one in the distribution numbers. On the other hand, if the user specified a pair of conditions (as in the previous examples of quantitative wrapups), then each pair of texts should count as an individual item, and should be counted as such, as well as displayed. (This is the current behaviour of the system as of today, no matter the kind of query.) Such a change in display and counting behaviour based on the form of the query implies, however, rather drastic changes to the whole architecture, and therefore no decision on this issue has been taken yet.

Finally, answering a reviewer's question, we do not envisage an "extension" of DISPARA to multilinguality, due to the lack of any real theoretical or practical basis for such a political concept[15, 13]. But we do expect to integrate other bilingual corpora for translations from or to languages different from English, since the whole system design does not hinge on the particular Portuguese-English pair.

## Acknowledgements

I am obviously indebted to Ana Frankenberg-Garcia, without whom this project would not have existed, and thank her not only for initiating it but for her truly collaborative spirit, effectiveness and enthusiasm.

I also thank Stefan Evert for invaluable support concerning EasyAlign and the IMS Corpus Workbench as a whole. DISPARA owes a lot to the outstanding capabilities and features of this corpus encoding system.

## References

1. Afonso, Susana, Bick, Eckhard, Haber, Renato, Santos, Diana: "Floresta Sintá(c)tica": a treebank for Portuguese. In Proceedings of LREC2002 (to appear)
2. Baker, Mona: Corpora in translation studies: an overview and some suggestions for future research. Target **7** (1995) 223–243
3. Biblioteca Virtual do Estudante Brasileiro. `http://www.bibvirt.futuro.usp.br/acervo/literatura/literatura.html`
4. Christ, O., Schulze, B., Hofmann, A., Koenig, E.: The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. IMS, University of Stuttgart, March 8, 1999 (CQP V2.2)
5. Frankenberg-Garcia, Ana, Santos, Diana: Introducing COMPARA, the Portuguese-English parallel translation corpus. In S. Bernardini & F. Zanettin (eds.), Corpora in translator training. Manchester: St. Jerome (to appear)
6. Frankenberg-Garcia, Ana, Santos, Diana: Apresentando o COMPARA, um corpus português-inglês na Web. Cadernos de Tradução, Universidade de São Paulo (to appear)
7. Project Gutenberg. `http://www.promo.net/pg/`

8. Johansson, Stig, Ebeling, Jarle, Hofland, Knut: 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 (Lund, 4-5 March 1994). Lund: Lund University Press, 1996, 87–112
9. Johansson, Stig, Ebeling, Jarle, Oksefjell, Signe: English-Norwegian Parallel Corpus: Manual. Oslo: Department of British and American Studies, University of Oslo, 1999, `http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html`
10. Johansson, S.: ICAME - Quo Vadis? Reflections on the Use of Computer Corpora in Linguistics. Computers and the Humanities **28** (1995) 243–252
11. Johansson, S., Oksefjell, S. (eds.): Corpora and cross-linguistic research: theory, method, and case studies. Amsterdam: Rodopi, 1998
12. Santos, Diana: Bilingual alignment and tense. Proceedings of the Second Annual Workshop on Very Large Corpora (Kyoto, 4 August 1994), 129–141
13. Santos, Diana: Punctuation and multilinguality: Reflections from a language engineering perspective. In Jo Terje Ydstie & Anne C. Wollebak (eds.), Working Papers in Applied Linguistics **4/98**, Oslo: Department of Linguistics, Faculty of Arts, University of Oslo, 1998 138–160.
14. Santos, Diana: Disponibilização de corpora através da WWW. In Marrafa, P. & Mota, M.A. (eds.), Linguística Computacional: Investigação Fundamental e Aplicações. Lisboa: Colibri, 1999, 323–346
15. Santos, Diana: Toward language-specific applications. Machine Translation **14**, 1999, 83–112.
16. Santos, Diana: Resultado da revisão do primeiro milhão de palavras do CETEMPúblico. 2000. `http://cgi.portugues.mct.pt/treebank/RevisaoMilhao.html`
17. Santos, Diana, Bick, Eckhard: Providing Internet access to Portuguese corpora: the AC/DC project. In Gavriladou et al. (eds.), Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000, ELRA: 205–210
18. Santos, Diana, Rocha, Paulo: Evaluating CETEMPúblico, a free resource for Portuguese. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL2001). ACL, 2001, 442–449
19. Tsang, Vivian, Stevenson, Suzanne: Automatic Verb Classification Using Multilingual Resources. In Walter Daelemans and Rémi Zajac (eds.), Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL-2001). 2001, 30–37
20. Projecto Vercial. `http://www.ipn.pt/opsis/litera/index.html`
21. Véronis, Jean (ed.): Parallel Text Processing. Dordrecht: Kluwer Academic Publishers, 2000