Procesamiento del Lenguaje Natural, nº39 (2007), pp. 265-272 recibido 18-05-2007; aceptado 22-06-2007

Parallel Corpora based Translation Resources Extraction

Alberto Simões
Departamento de Informática
Universidade do Minho
Braga, Portugal
ambs@di.uminho.pt

José João Almeida
Departamento de Informática
Universidade do Minho
Braga, Portugal
jj@di.uminho.pt

Resumen: Este artículo describe NATools, un conjunto de herramientas de procesamiento, análisis y extracción de recursos de traducción de Corpora Paralelo. Entre las distintas herramientas disponibles se destacan herramientas de alineamiento de frases e palabras, un extractor de diccionarios probabilísticos de traducción, un servidor de corpus, un conjunto de herramientas de interrogación de corpora y diccionarios y así mismo un conjunto de herramientas de extracción de recursos bilingües.
Palabras clave: corpora paralelos, recursos bilingües, traducción automática

Abstract: This paper describes NATools, a toolkit to process, analyze and extract translation resources from Parallel Corpora. It includes tools like a sentence-aligner, a probabilistic translation dictionaries extractor, word-aligner, a corpus server, a set of tools to query corpora and dictionaries, as well as a set of tools to extract bilingual resources.
Keywords: parallel corpora, bilingual resources, machine translation

1 Introduction

NATools is a package with a set of tools for parallel corpora processing. It includes tools to help parallel corpora preparation, from sentence-alignment and tokenization, to full probabilistic translation dictionary extraction, word-alignment, and translation examples extraction for machine translation.

Follows a list with some of the available tools:

- a simple parallel corpora sentence aligner based on the algorithm proposed by (Gale and Church, 1991) and in the Vanilla Aligner implementation by (Danielsson and Ridings, 1997);
- a probabilistic translation dictionary (Simões and Almeida, 2003; Simões, 2004) extractor based on PTD Extractor based on work by (Hiemstra, August 1996; Hiemstra, 1998);
- a parallel corpora word-aligner (Simões and Almeida, 2006a) based on probabilistic translation dictionaries;
- NatServer (Simões and Almeida, 2006b), a parallel corpora server for quick concordances and probabilistic translation dictionary querying;
- a set of web clients to query parallel corpora using NatServer;
- tools for machine translation example extraction (Simões and Almeida, 2006a) based on probabilistic translation dictionaries and alignment pattern rules;
- A full C and Perl API for quick parallel corpora tools prototyping;
- a StarDict generation software;
- support for `Makefile::Parallel` (Simões, Fonseca, and Almeida, 2007), a Domain Specific Language for process parallelization (to take advantage of multi-processor machines and/or cluster systems).

This paper consists of three main sections. The first one explains how NATools helps preparing parallel corpora. Follows a section on querying parallel corpora both using a corpora server and using web interfaces. The third section is about using NATools for parallel resources extraction like translations examples.

2 Parallel Corpora Preparation

To create and make available a parallel corpora is not a simple task. In fact, this process does not depend just on the compilation of parallel texts. These texts should be processed in some different ways so it can be really useful. Important steps include the text tokenization, sentence boundaries detection and sentence alignment (or translation unit alignment). NATools include (and depends) on tools to perform these tasks.

2.1 Segmentation and Tokenization

While NATools does not include directly tools for segmentation and tokenization, it depends on `Lingua::PT::PLNbase`[1], a Perl module for based

[1] `http://search.cpan.org/dist/Lingua-PT-PLNbase`.

ISSN: 1135-5948

segmentation and tokenization for the Portuguese language. While it was developed with the Portuguese language in mind, through the time more and more support for Spanish, French and English has been incorporated. Thus, after installing NATools you will have access to the Perl module directly or using NATools options for segmentation and tokenization.

2.2 Sentence Alignment

The NATools sentence aligner uses the well known algorithm by (Gale and Church, 1991). Work is being done to include some clue-align (Tiedemann, 2003) information into the original algorithm, taking advantage of numbers and other non-textual elements in sentences in addition to the basic sentence length metrics.

While Gale and Church algorithm is known for not being robust enough for big corpora with big differences in number of sentences, the truth is that it works for most available corpora.

Also, note that NATools do not force the user to use the supplied sentence-aligner (or tokenizer). For instance, we are using *easy-align* from IMS-CWB (Christ et al., 1999) to perform sentence alignment on big corpora. Unfortunately *easy-align* is not open-source and the used algorithm is not described in any paper, but it uses not only the base length metrics but also uses other knowledge like bilingual dictionaries to perform better alignment.

2.3 Corpora Encoding

This is the only required step on using NATools. It performs the corpora encoding and creates auxiliary indexes for quick access. Two lexicon indexes are created (one for each language), mapping an integer identifier for each word. The corpora is codified using these integer values, and indexes for direct access by word and sentence are created. There are other tools to index corpora. Examples are Emdros (Petersen, 2004) and IMS-CWB (Christ et al., 1999). While the first one is freely available, it is intended for monolingual corpora. In the other hand, IMS-CWB is not open software.

2.4 Probabilistic Translation Dictionaries Extraction

This process extracts relationships between words and their probable translations. Some researchers (Hiemstra, August 1996) call this word-alignment. Within NATools, we prefer to call it probabilistic translation dictionaries (PTDs).

There are other tools like Giza++ (Och and Ney, 2004) that perform word-alignment directly from parallel corpora, but that is not our approach. Our dictionaries map for each word in a language, a set of probable translations on the other language (together with an translation probability). Follows a simple example of a PTD:

```
** europe (42853 occurrences)

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

** stupid (180 occurrences)

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

Note that although the first three entries for the *stupid* word have low probabilities, they refer to the same word with different inflections: masculine singular, feminine singular and masculine plural.

The algorithm based on Twente-Aligner (Hiemstra, August 1996; Hiemstra, 1998) was fully reviewed and enhanced, and was added support for big corpora (Simões, 2004). The version included in NATools supports arbitrary size corpora (only limited by disk space), and can be run on parallel machines and clusters.

NATools probabilistic dictionary extraction is being used for bilingual dictionary bootstrapping as presented by (Guinovart and Fontenla, 2005).

3 Querying Parallel Corpora

To make parallel corpora available for querying is not easy as well. After the encoding process described on section 2.3, there is the need for a server to help searching and querying the encoded corpora. Thus, NATools includes its own parallel corpora server.

3.1 NatServer: A Parallel Corpora Server

NATools includes NatServer, a socket-based program to query efficiently parallel corpora, corpora n-grams (bigrams, trigrams and tetragrams) and probabilistic translation dictionaries. It supports multiple corpora with different language pairs.

Given the modular implementation of NatServer, the C library can be used for other software and namely for NATools Perl API (Application Programmer Interface). This makes it easy for any software choose at run-time if it will use the socket server or access locally the encoded corpora. This is specially important for intensive batch tasks where the socket-based communication is a big over-head regarding performance.

NatServer is also being prepared to be responsible of the server part of Distributed Translation Memories (Simões, Guinovart, and Almeida, 2004),

a WebService to serve translators with external translation memories.

3.2 Query Tools

Linguistics and translators make heavy use of parallel corpora and bilingual resources. Meanwhile, they use simple applications or web interfaces. There are parallel corpora available for querying in the web like COMPARA (Frankenberg-Garcia and Santos, 2001; Frankenberg-Garcia and Santos, 2003) or Opus (Tiedemann and Nygaard, 2004), and they are quite used. Thus, it is important to provide mechanisms to make our parallel corpora available in the Web as well.

NATools include a set of web tools for concordances with translation guessing (see figure 1) and probabilistic translation dictionary browsing (see figure 2).

The web interface lets the user swap between concordances and dictionaries in an easy way, as well as to check corpora details (description, languages, sizes and so on).

4 Parallel Resources Extraction

NATools main objective was not to be a final-user software package, but instead, be a toolbox for the researcher that uses parallel corpora. Thus, research is being done using NATools and some of resulting applications are being incorporated in the toolbox. The probabilistic translation dictionaries presented in section 2.4 by themselves are useful parallel resources. They were presented earlier because they are crucial for querying correctly NATools corpora.

4.1 Terminology Extraction

(Och, 1999; Och and Ney, 2004) describes methods to infer translation patterns from parallel corpora. In our work we found out that to describe translation patterns and apply them to parallel corpora gives interesting results: bilingual terminology.

Translation patterns describe how words order change when translation occurs. For instance, we can describe a simple pattern to describe how the adjective swaps with the substantive when translating from Portuguese to English as[2]:

$$\mathcal{T}(A \cdot B) = \mathcal{T}(B) \cdot \mathcal{T}(A)$$

A bit complicated pattern:

$$\mathcal{T}(P \cdot de \cdot V \cdot N) = \mathcal{T}(N) \cdot \mathcal{T}(P) \cdot of \cdot \mathcal{T}(V)$$

is presented on figure 3 visually. NATools includes a Domain Specific Language (DSL) to define these patterns in a easy way. The last example shown can be written as "`P "de" V N = N P "of" V`".

[2]Note that letters on these patterns do not have any special meaning. They are just variable names.

	alternative	sources	of	financing
fontes		X		
de			Δ	
financiamento				X
alternativas	X			

Figure 3: Translation Pattern example.

Although these patterns can be inferred from parallel corpora most of them can be defined manually quite faster and with good results. Figure 4 show some extracts from terminology extracted. Each group is preceded by the rule. Numbers before the terminology pairs are the occurrence counter for that pair.

Note that the examples are the top five in number of occurrences. Although they are all good translations and they can all be considered terminology, this does not apply to all the extracted examples. Meanwhile, the DSL lets add morphological constrains and Perl predicates to the pattern. With these constrains it is quite easy to remove from the extracted entries those which are not terminology.

We did a massive test of terminology extraction using EuroParl (Koehn, 2002) Portuguese:English corpus. Table 1 shows some statistics on number of patterns extracted[3].

Total number of TUs	1 000 000
Number of processed TUs	700 000
Number of patterns found	578 103
Number of different patterns	139 781
Number of filtered patterns	103 617

Table 1: Terminology extraction statistics.

Table 2 shows the occurrence distribution by some patterns. The third column is a simple evaluation of how many patterns are really terminology and are correct. Evaluation was done with three samples: the 20 patterns with more occurrence, the 20 patterns with lower occurrence, and 20 patterns in the middle of the list.

4.2 Word Alignment and Example Extraction

While Word Alignment and Example Extraction are different tasks, the base algorithm used in NATools is the same. The word alignment is done for each pair of translation units creating a matrix of

[3]The number of translations units processes is not equal to the total number of translations units because at the time these statistics were reported the process did not have finished.

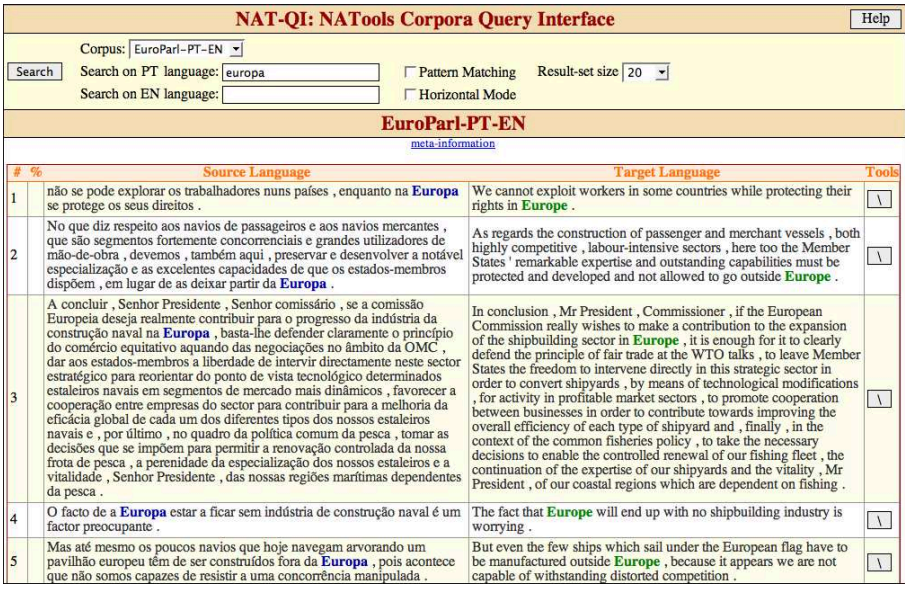

Figure 1: Concordances interface.

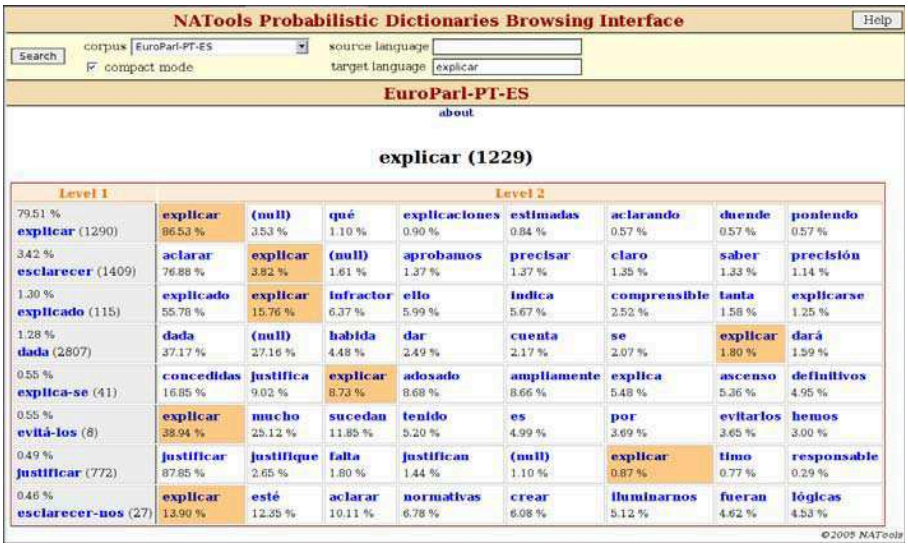

Figure 2: PTDs query interface.

Pattern	Occur.	Quality
A B = B A	77 497	86%
A *de* B = B A	12 694	95%
A B C = C B A	7 700	93%
H *de* D H = H D I	3 336	100%
A B C = C A B	1 466	40%
P *de* V N = N P *of* V	564	98%
P *de* T *de* F = F T P	360	96%

Table 2: Patterns occurrences by type, and respective quality.

translation probabilities as shown on figure 5. In this matrix one can see direct translations between word and some marked patterns. As these patterns are hopefully terminology, we are considering them as a term, and as such, aligning it all with another term. From this matrix we can extract the real word-alignment between these two translation units.

For the example in the figure, it would be extracted the alignments: *discussão:discussion*, *sobre:about*, *fontes de financiamento alternativas:alternative sources of financing*, *para:for*, *a:the*, *aliança radical europeia:european radical alliance*.

The truth is that single word translations are already present on the probabilistic translation dictionaries, and thus there is no advantage on extracting the word-to-word relation.

The alignment matrix can also be used to extract examples. If we join sequences of words (or terms) and their translations, a set of word sequences can be extracted (examples). Again, for the matrix shown, we can extract more relationships, like *discussão sobre:dicussion about*, *sobre fontes de financiamento alternativas:about alteran-*

```
A B = B A
    14949 comunidades europeias              | european communities
    12487 parlamento europeu                 | european parliament
    11645 comunidade europeia                | european community
    10055 união europeia                     | european union
     7705 jornal oficial                     | official journal

P "de" V N = N P "of" V
      134 comunicação de acusações alterada | revised statement of objections
       55 comunicação de acusações inicial  | initial statement of objections
       49 tribunal de justiça europeu       | european court of justice
       45 fontes de energia renováveis      | renewable sources of energy
       41 período de tempo limitado         | limited period of time

A "de" B = B A
     3383 medidas de execução                | implementing measures
     2754 comité de gestão                   | management committee
     1163 plano de acção                     | action plan
     1050 certificados de importação         | import licences
     1036 sigla de identificação             | identification marking
```

Figure 4: Bilingual terminology extracted by Translation Patterns.

	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 5: Word-alignment matrix.

tive sources of financing, *fontes de financiamento alternativas para:alternative sources of financing for*, *para a:for the*, *a aliança radical europeia:the european radical alliance*. This process can be repeated, resulting in bigger examples. This step is important to generate more examples occurrences and thus give more importance for those with bigger occurrence.

Figure 6 shows some examples extracted using this methodology. These examples can be consolidated (summed accordingly with their occurrence count) and be used for machine translation or computer assisted translation.

4.3 Example Generalization

Based on work from (Brown, 2000; Brown, 2001), we are incorporating generalization algorithms into NATools. One simple generalization is the detection of numbers, hours and dates. Follows some examples generalized using this technique.

```
399 às hour              hour
187 orçamento de year    year budget
136 int euros            eur int
135 int euros            eur int
127 directiva de year    year directive
 51 orçamento year       year budget
 46 int de setembro      september int
 31 partir de year       year onwards
 29 convenção de year    year convention
 26 eleições de year     year elections
 25 período year-year    year-year period
 25 int dólares          usd int
 24 relatório de year    year report
```

Although these patterns can be useful they are not as interesting as if could create place-holders for words. If we analyze similar entries in the examples listing we can find entries differing just in a few words like the following example.

```
  2 povo português   portuguese people
  2 povo paraguaio   paraguayan people
  2 povo nigeriano     nigerian people
  2 povo mexicano       mexican people
  2 povo marroquino    moroccan people
  2 povo mapuche         mapuche people
  2 povo indígena    indigenous people
  2 povo holandês          dutch people
  2 povo húngaro       hungarian people
  2 povo hmong             hmong people
```

This can be generalized creating automatically a class for the differing words (in this case we used gentilic). Given two different classes with a big number of similar members we can join them expanding the initial number of examples.

```
raw examples
           protocolo para prevenir        | protocol to prevent
           , reprimir e punir o           | , suppress and punish
           tráfico de pessoas             | trafficking in persons
           e em particular de             | , especially
           mulheres e crianças            | women and children

consolidated examples
   35736   tendo em conta                 | having regard
   11304   tratado que institui           | treaty establishing
   10335   das comunidades europeias      | of the european communities
    8789   institui a comunidade europeia | establishing the european community
    8424   e , nomeadamente               | and in particular
    8224   , a comissão                   | , the commission
    8142   redacção que lhe foi dada pelo | amended by
    7352   à comissão                     | to the commission
    7072   a comissão das                 | the commission of
    6870   pela comissão                  | for the commission
    6540   todos os estados-membros       | all member states
    6400   pela comissão                  | by the commission
    6379   considerando que ,             | whereas ,
    5409   regulamento é obrigatório      | regulation shall be binding
    5400   adoptou                        | has adopted this
```

Figure 6: Translation examples.

```
povo  X: gentilic(X)      T(X) people
governo  X: gentilic(X)   T(X) govern
```

4.4 StarDict generation

Although we are in the Internet era, there are a few people without Internet access at home, or working offline on a laptop. For these people, to access the online query system is not possible. Specially for non computer-science researchers, there is important to make dictionaries and some concordances available easily.

Figure 7: StarDict screen-shot.

With this in mind we created a tool to generate StarDict (Zheng, Evgeniy, and Murygin, 2007) dictionaries with probabilistic translation dictionary information and for each possible translation a set of three concordances.

```
use NAT::Client;

$client = NAT::Client->new(
     crp => "EuroParl-PT-EN");

$client->iterate(
  { Language => "PT" },
  sub {
    my %param = @_;

    for $trans (keys %{$param{trans}}) {
      if ($param{trans}{$trans} > 0.1) {
        $concs = $client->conc({
                    concordance => 1},
                    $param{word}, $trans);
        $stardict{$param{word}}{$trans}
                    = $concs->[0];
    }}});

print StarDict($stardict);
```

Figure 8: Perl code to create a StarDict dictionary.

This tool was also an exercise to see how versatile the NATools API was. The basic structure of the dictionary to be translated to StarDict can be created using just some lines of Perl code (see figure 8).

The process is done iterating over all the entries in the probabilistic translation dictionary. For each entry we grab concordances for each probable translation (with association above 10%).

5 Conclusions

While a lot of work needs to be done within NATools, most for efficiency, being open-source makes it easier. Any researcher can contribute with code, submit bugs reports, and get some support freely.

The whole NATools framework proved to be robust enough for different sized corpora. It was tested with Le Monde Diplomatique (PT:FR) (Correia, 2006), JRC-Acquis (PT:ES,PT:EN,PT:FR) (Steinberger et al., 2006) and EuroParl (PT:ES,PT:EN:PT:FR) (Koehn, 2002). All these corpora are available for querying in the Internet.

NATools include some other small tools not described in this paper. For instance, there is a set of small tools that grew up as experiences and where maintained in the package as tools to compare probabilistic translation dictionaries, tools to rank (or classify) translation memories accordingly with their translation probability, and others.

Acknowledgment

Alberto Simões has a scholarship from Fundação para a Computação Científica Nacional and the work reported here has been partially funded by Fundação para a Ciência e Tecnologia through project POSI/PLP/43931/2001, co-financed by POSI, and by POSC project POSC/339/1.3/C-/NAC.

References

Brown, Ralf D. 2000. Automated generalization of translation examples. In *Eighteenth International Conference on Computational Linguistics (COLING-2000)*, pages 125–131.

Brown, Ralf D. 2001. Transfer-rule induction for example-based translation. In Michael Carl and Andy Way, editors, *Workshop on Example-Based Machine Translation*, pages 1–11, September.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann, and Esther König, 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. Institute for Natural Language Processing, University of Stutgart, March.

Correia, Ana Teresa Varajão Moutinho Pereira. 2006. Colaboração na constituição do corpus paralelo Le Monde Diplomatique (FR-PT). Relatório de estágio, Conselho de Cursos de Letras e Ciências Humanas — Universidade do Minho, Braga, Dezembro.

Danielsson, Pernilla and Daniel Ridings. 1997. Practical presentation of a "vanilla" aligner. In *TELRI Workshop in alignment and exploitation of texts*, February.

Frankenberg-Garcia, Ana and Diana Santos, 2001. *Apresentando o COMPARA, um corpus português-inglês na Web*. Cadernos de Tradução, Universidade de São Paulo.

Frankenberg-Garcia, Ana 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*. Manchester: St. Jerome Publishing, pages 71–87.

Gale, William A. and Kenneth Ward Church. 1991. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184.

Guinovart, Xavier Gómez and Elena Sacau Fontenla. 2005. Técnicas para o desenvolvemento de dicionarios de tradución a partir de córpora aplicadas na xeración do Dicionario CLUVI Inglés-Galego. *Viceversa: Revista Galega de Traducción*, 11:159–171.

Hiemstra, Djoerd. 1998. Multilingual domain modeling in twenty-one: automatic creation of a bi-directional lexicon from a parallel corpus. Technical report, University of Twente, Parlevink Group.

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

Koehn, Philipp. 2002. EuroParl: a multilingual corpus for evaluation of machine translation. Draft, Unpublished.

Och, Franz Josef. 1999. An efficient method for determining bilingual word classes. In *the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 71–76.

Och, Franz Josef and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30:417–449.

Petersen, Ulrik. 2004. Emdros — a text database engine for analyzed or annotated text. In *20th International Conference on Computational Linguistics*, volume II, pages 1190–1193, Geneva, August.

Simões, Alberto and J. João Almeida. 2006a. Combinatory examples extraction for machine translation. In Jan Tore Lønning and Stephan Oepen, editors, *11th Annual Conference of the European Association for Machine Translation*, pages 27–32, Oslo, Norway, 19–20, June.

Simões, Alberto and J. João Almeida. 2006b. NatServer: a client-server architecture for building

parallel corpora applications. *Procesamiento del Lenguaje Natural*, 37:91–97, September.

Simões, Alberto, Rúben Fonseca, and José João Almeida. 2007. Makefile::Parallel dependency specification language. In *Euro-Par 2007*, Rennes, France, August. **Forthcoming**.

Simões, Alberto, Xavier Gómez Guinovart, and José João Almeida. 2004. Distributed translation memories implementation using webservices. *Procesamiento del Lenguaje Natural*, 33:89–94, July.

Simões, Alberto M. and J. João Almeida. 2003. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224, September.

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

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy, 24–26 May.

Tiedemann, Jörg. 2003. Combining clues for word alignment. In *10th Conference of the European Chapter of the ACL (EACL03)*, Budapest, Hungary, April 12–17.

Tiedemann, Jörg and Lars Nygaard. 2004. The opus corpus - parallel & free. In *Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 26–28.

Zheng, Hu, Evgeniy, and Alex Murygin. 2007. Stardict. Software and documentation homepage, StarDict, `http://stardict.sourceforge.net/`, January.