A survey on parallel corpora alignment

André Santos[1]
Universidade do Minho

A parallel text is the set formed by a text and its translation (in which case it is called a bitext) or translations. Parallel text alignment is the task of identifying correspondences between blocks or tokens in each halve of a bitext. Aligned parallel corpora are used in several different areas of linguistic and computational linguistics research.

In this paper, a survey on parallel text alignment is presented: the historical background is provided, and the main methods are described. A list of relevant tools and projects is presented as well.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms: Algorithms, Languages, Human factors, Performance

Additional Key Words and Phrases: Alignment, Parallel, Corpora, NLP

1. INTRODUCTION

1.1 Aligned parallel texts

A parallel text is the set formed by a text and its translation (in which case it is called a bitext) or translations. Parallel text alignment is the task of identifying correspondences between blocks or tokens of each halve of a bitext. Aligned parallel texts are currently used in a wide range of areas: knowledge retrieval, machine learning, natural language processing, and others [Véronis 2000].

Bitexts are generally obtained by performing an alignment between two texts. This alignment can usually be made at paragraph, sentence or even word level.

1.2 Precision and recall

Like in many knowledge retrieval, natural language processing or pattern recognition systems, parallel text alignment algorithms performance is usually measured in terms of *precision*, *recall* [Olson and Delen 2008] and *F-score* [Rijsbergen 1979].

In this specific context, these concepts can be simply explained as follows: given a set of parallel documents D, there is a number C_{total} of correct correspondences between them. After aligning the texts with an aligner A, $T_{A(D)}$ represents the total number of correspondences found by A in D, and $C_{A(D)}$ represents the number of correct correspondences found by A in D. The precision $P_{A(D)}$ and recall $R_{A(D)}$ are then given by:

$$P_{A(D)} = \frac{C_{A(D)}}{T_{A(D)}} \tag{1}$$

[1]CeSIUM, Dep. Informática, Universidade do Minho, Campus de Gualtar, 4710-057 BRAGA
[1]Email: pg15973@alunos.uminho.pt
Universidade do Minho - Master Course on Informatics - State of the Art Reports 2011

MI-STAR 2011, Pages 117–128.

$$R_{A(D)} = \frac{C_{A(D)}}{C_{total}} \tag{2}$$

The *F-score* is the harmonic mean between precision and recall, obtained with the following equation:

$$F\text{-}score = \frac{2 * P * R}{P + R} \tag{3}$$

1.3 Alignment details

The alignment of texts may be performed at several levels of granularity. Usually, when the objective is to achieve low level alignment, like word alignment, higher level alignment is performed first, which allows to obtain improved results.

Independently of the type of alignment being performed, several outcomes must be considered for each correspondence found: the most common case is when a source text sentence corresponds exactly to a target text sentence (1:1). Less frequently, there are omissions (1:0), additions (0:1) or something more complex (m:n), usually with $1 \leq m, n \leq 2$.

An example of a sentence-level alignment is presented in Table I.

Table I. Extract of sentence-level alignment performed using Portuguese and Russian subtitles from the movie Tron.

Portuguese	**Russian**
A actividade do laser começará em 30 segundos.	Лазер будет включен через 30 секунд.
Ponham os óculos de proteção. Abandonem a área.	Наденьте защитные очки и покиньте главный зал
Deixe-me ver se nós temos a luz verde.	Интересно, будет ли на этот раз зелёное свечение...
Área do alvo protegida?	- Облучаемая область свободна?
Verificaram a segurança.	- Да , её уже проверили.
20 segundos	20 секунд.
- Parece bom.	- Вроде всё нормально.
- Deixe-o iniciar.	- Можно запускать.
Isto é o que nós temos esperado.	Включаю.

Through the years, several alignment methods have been proposed. The next section describes the evolution of the field and details the most relevant efforts to create gold standards and evaluate existing systems (in this case, gold standards are hand-performed alignments which were thoroughly checked and are considered "correct"; they are generally used to test and assess other systems). Section 3 describes the alignment process step-by-step, and Section 4 lists several of the most relevant tools and projects currently being developed or used. The last section presents some conclusions about this state-of-the-art and the corpora alignment area of knowledge itself.

2. BACKGROUND

Parallel text use in automatic language processing was first tried in the late fifties, but probably due to limitations in the computers of that time (both in storage space and computing power) and the reduced availability of large amount of textual data in a digital format, the results were not encouraging [Véronis 2000].

In the eighties, with computers being several orders of magnitude faster, a similar increase in storage capacity and an increasing amount of large corpora available, a new interest in text alignment emerged, resulting in several publications years later.

The first documented approaches to sentence alignment were based on measuring the length of the sentences in each text. Kay [1991] system was the first to devise an automatic parallel alignment method. This method was based on the idea that when a sentence corresponds to another, the words in them must also correspond. All the necessary information, including lexical mapping, was derived from the texts themselves.

The algorithm initially sets up a matrix of correspondence, based on sentences which are reasonable candidates: the initial and final sentences have a good probability to correspond to each other, and the remaining sentences should be distributed close to the matrix main diagonal. Then the word distributions are calculated, and words with similar distribution values are matched, and considered anchor points, narrowing the "alignment corridor" of the candidate sentences. The algorithm then iterates until it converges on a minimal solution.

This system was not efficient enough to apply to large corpora [Moore 2002].

Two other similar methods were proposed, the main difference between the two being how the sentence length was measured: Brown et al. [1991] counted words, while Gale and Church [1993] counted characters. The authors assumed that the length of a sentence is highly correlated with the length of its translation. Additionally, they concluded that there is a relatively fixed ratio between the sentences lengths in any two languages. The optimal alignment minimizes the total dissimilarity of all the aligned sentences. Gale and Church [1993] achieve the optimal alignment with a dynamic programming algorithm, while Brown et al. [1991] applied Hidden Markov Models.

In 1993, Chen [1993] proposed an alignment method which relied on additional lexical information. In spite of not being the first of its kind, their algorithm was the first lexical-based alignment which was efficient enough for large corpora. It required a minimum of human intervention – at least 100 sentences had to be aligned for each language pair to bootstrap the translation model – and it was capable of handling large deletions in text.

Another family of algorithms was proposed soon after, based on the concept of cognates. The concept of cognate may vary a little, but generally cognates are defined as occurrences of tokens that are graphically or otherwise identical in some way. These tokens may be dates, proper nouns, some punctuation marks or even closely-spelled words – Simard and Plamondon [1998] considered cognates any two words which shared the first four characters. When recognisable elements like dates, technical terms or markup are present in considerable amounts, this method works well even with unrelated languages, despite some loss of performance being

noticeable.

Most methods developed ever since rely on one or more of these main ideas – length based, lexical or dictionary based, or partial similarity (cognate) based.

2.1 ARCADE

In the mid-nineties, the amount of studies describing alignment techniques was already impressive. Comparing their performance, however, was difficult due to the aligners sensitivity to factors as the type of text used or different interpretations of what is a "correct alignment". Details on the protocols or the evaluation performed on the systems developed were also frequently not disclosed. A precise evaluation of the techniques would be most valuable; yet, there was no established framework for evaluation of parallel text alignment.

The ARCADE I was a project to evaluate parallel text alignment systems which took place between 1995 and 1999, consisting in a competition among systems at the international level [Langlais et al. 1998; Véronis and Langlais 2000]. It was divided in two phases: the first two years were spent on corpus collection and preparation, and methodology definition, and a small competition on sentence alignment was held; in the remaining time, the sentence alignment competition was opened to a larger number of teams, and a second track was created to evaluate word-level alignment.

For the sentence alignment track, several types of texts were selected: institutional texts, technical manuals, scientific articles and literature. To build the reference alignment, an automatic aligner was used, followed by hand-verification by two different persons, who annotated the text.

F-score values were based on the number of correct alignments, proposed alignments and reference alignments, and calculated for different types of granularity. For the sentence alignment track, several types of texts were selected for the competition: institutional texts, technical manuals, scientific articles and literature. The best systems achieved an F-score of over 0.985 on institutional and scientific texts. On the other hand, all systems performed poorly on the literature texts.

For the word track, the systems had to perform translation spotting, a sub-problem of full alignment: for a given word or expression, the objective is to find its translation in the target text. A set of sixty French words (20 adjectives, 20 nouns and 20 verbs) were selected based on frequency criteria and polysemy features. The results were better for adjectives (0.94 precision and recall for the best system) than for verbs (0.72 and 0.62 respectively). The overall results were 0.77 and 0.73 respectively for the best system.

The ARCADE project had a few limitations: the systems were tested by limited tasks which did not reflect their full capacity, and only one language pair (French-English) was tested. Nonetheless, it allowed to conclude, at the time, that the techniques were satisfactory for texts with a similar structure. The same techniques were not as good when it came to align texts whose structure did not match perfectly. Methodological advances on sentence alignment and, in a limited form, word alignment were also made possible, resulting in methods and tools for the generation of reference data, and a set of measures for system performance assessment. Additionally, a large standardized bilingual corpus was constructed and made available as a gold standard for future evaluation.

2.2 ARCADE-II

The second campaign of ARCADE was held between 2003 and 2005, and differed from the first one in its multilingual context and on the type of alignment addressed. French was used as the pivot language for the study of 10 other language pairs: English, German, Italian and Spanish for western-European languages; Arabic, Chinese, Greek, Japanese, Persian and Russian for more distant languages using non-Latin scripts.

Two tasks were proposed: sentence alignment and word alignment. Each task involved the alignment of Western-European languages and distant languages as well. For the sentence alignment, a pre-segmented corpus and a raw corpus were provided.

The results showed that sentence alignment is more difficult on raw corpora (and also, curiously, that German is harder than the other languages). The systems achieved an F-score between 0.94 and 0.97 on raw corpora, and between 0.98 and 0.99 on the segmented one. As for distant languages alignment, two systems (P1 and P2) were evaluated. The results allowed to conclude that sentence segmentation is very hard on non-Latin scripts, as P1 was incapable of performing it and the results for raw corpora alignment of P2 scored 0.421.

Evaluating word alignment presents some additional difficulties, given the differences in word order, part-of-speech and syntactic structure, discontinuity of multi-token expressions, etc which are possible to find between texts and its translations. Sometimes it is not even clear how some words should be aligned when they do not have a direct correspondence in the other language. Nevertheless, a small competition was held, in which the systems were supposed to identify named entities phrases translation in the parallel text.

The scope of this task was limited; yet, it allowed to define a test protocol and metrics.

2.3 Blinker project

In 1998, a "bilingual linker", Blinker, was created in order to help bilingual annotators (people who go through bitexts and annotate the correspondences between sentences or words) in the process of linking word tokens that are mutual translations in parallel texts. This tool was developed in the context of a project whose goal was to manually produce a gold standard which could be used to future research and development [Melamed 1998b]. Along with this tools, several annotation guidelines were defined [Melamed 1998a], in order to increase consistency between the annotations produced by all the teams.

The parallel texts chosen to align in this project were a modern French version and a modern English version of the Bible. This choice was motivated by the canonical division of the Bible in verses, which is common across most of the translations. This was used as a "ready-made, indisputable and fairly detailed bitext map".

Several methods were used to improve reliability. First, as many annotators as possible were recruited. This allowed to identify deviations from the norm, and to evaluate the gold standard produced in terms of inter-annotator agreement rates (how much the annotators agreed with each other).

Second, Blinker was developed with some unique features, such as forcing the

annotator to annotate all the words in a given sentence before proceeding (either by linking them with the corresponding target text word, or by declaring as "not translated". These feature aimed to ensure that a good first approximation was produced, and to discourage the annotators natural tendency to classify words whose translation was less obvious as "not translated".

Third, the annotation guidelines were created to reduce experimenter bias, and a monetary bonus was offered to the annotators who stayed closest to the guidelines.

The inter-annotator agreement rates on the gold standard were approximately 82% (92% if function words were ignored), which indicated that the gold standard is reasonably reliable and that the task is reasonably easy to replicate.

3. ALIGNMENT PROCESS

Despite the existence of several methods and tools to align corpora, the process itself shares some common steps. In this section the alignment process is detailed step-by-step.

3.1 Gathering corpora

As mentioned in the previous section, the lack of available corpora was one of the reasons appointed to why the earlier efforts made to align texts did not work out. Fortunately, with the massification of computers and globalization of the internet, several sources of alignable corpora have appeared:

– *Literary Texts.* Books published in electronic format are becoming increasingly common. Despite the vast majority being subject to copyright restrictions, some of them are not.

Project Gutenberg, for example, is a volunteer effort to digitize and archive cultural works. Currently it comprises more than 33.000 books, most of them in public domain [Hart and Newby 1997]. National libraries are beginning to maintain a repository of eBooks as well [Varga et al. 2005].

Sometimes it is also possible to find publishers who are willing to cooperate by providing their books as long as it is for research purposes only.

– *Religious Texts.* The Bible has been translated into over 400 languages, and other religious books are wide spread as well. Additionally, the Catholic Church translates papal edicts to other languages from the original Latin, and the Taizé Community website[1] also provides a great deal of its content in several languages.

– *International Law.* Important legal documents, such as the Universal Declaration of Human Rights[2] or the Kyoto Protocol[3] are freely available in many different languages. One of the first corpora used in parallel text alignment were the Hansards – proceedings of the Canadian Parliament [Brown et al. 1991; Gale and Church 1993; Kay 1991; Chen 1993].

– *Movie subtitles.* There are several on-line databases of movie subtitles. Depending on the movie, it is possible to find dozens of subtitles files, in several different languages (frequently even more than one version for each language). Subtitles are

[1] `http://www.taize.fr`
[2] `http://www.un.org/en/documents/udhr/index.shtml`
[3] `http://unfccc.int/kyoto_protocol/items/2830.php`

shared on the internet as plain text files (sometimes tagged with extra information such as language, genre, release year, etc) [Tiedemann 2007].

– *Software internationalization.* There is an increasing amount of multilingual documentation belonging to software available in several countries. Open source software is particularly useful for not being subject to copyright [Tiedemann et al. 2004].

– *Bilingual Magazines.* Magazines like National Geographic, Rolling Stone or frequent flyer magazines are oftenly published in other languages besides English, and in several countries there are magazines with complete mirror translations into English.

– *Websites.* Websites often allow the user to choose the language in which they are presented. This means that a web crawler may be pointed to this websites to retrieve all reachable pages – a process called "mining the Web" [Resnik and Smith 2003; Tsvetkov and Wintner 2010].

Corporate webpages are one example of websites available in several languages. Additionally, some international companies have their subsidiaries publishing their reports in both their native language and in their common language (usually English).

3.2 Input

The format of the documents to align usually depends on where they come from: while literary texts are often obtained either as PDF or plain text files; legal documents usually come as PDF files; corporate websites come in HTML; and movie subtitles either in SubRip or microDVD format.

Globally accepted as the standard format to share text documents, PDF files are usually converted to some kind of plain text format: either unformatted plain text, HTML or XML. In spite of being available several tools to perform this conversion, given that the final layout of PDF files is dictated by graphical directives, without the notion of text structure, frequently the final result of the conversion is less that perfect: remains of page structure, loss of text formatting and noise introduction are common problems. A comparative study of the tools available can be found in Robinson [2001].

HTML or XML files are less problematic. Besides being more easily processed, some of the markup elements may be used to guide the alignment process: for example, the header tags in HTML may be used to delimit the different sections of the document, and the text formatting tags as `<i>` or `<b>` might be preserved and used as well. Depending on the schema used, XML annotations can be an useful source of information too.

3.3 Document alignment

Frequently, a large number of documents is retrieved from a given source with no information provided of which documents match each other. This is typically the case when documents are gathered from websites with a crawler.

Fortunately, when there are multiple versions available, in different languages, of the same page, it is common to include in the name a substring identifying the language: for instance, `Portuguese`, `por` or `pt` [Almeida and Simões 2010].

These substrings usually follow either the ISO-639-2[4] or the ISO 3166[5] standards. This way, it is possible to write a script with a set of rules which help to find corresponding documents.

Other ways to match documents are to analyze their meta-information, or the path where the were stored.

3.4 Paragraph/sentence boundary detection

The higher levels of alignment are section, paragraph or, most commonly, sentence alignment. In order to perform sentence-level alignment, texts must be segmented into sentences. Some word-level aligners work based on sentence-segmented corpora as well, which allows them to achieve better results.

Splitting a text into sentences is not a trivial task because the formal definition of what is a sentence is a problem that has eluded linguistic research for quite a while (see Simard [1998] for further details on this subject). Véronis and Langlais [2000] give an example of several valid yet divergent segmentations of sentences.

A simple heuristic for a sentence segmenter is to consider symbols like `.!?` as sentence terminators. This method may be improved by taking into account other terminator characters or abbreviations patterns; however, this patterns are typically language-dependent, which makes it impossible to have an exhaustive list for all languages[6] [Koehn 2005].

This process usually outputs the given documents with the sentences clearly delimited.

3.5 Sentence Alignment

There are several documented algorithms and tools available to perform sentence-level alignment. Generally speaking, they can be divided into three categories: length-based, dictionary or lexicon based or partial similarity-based (see section 2 on page 119 for more information).

Generally, sentence aligners take as input the texts to align, and, in some cases, additional information, such as a dictionary, to help establish the correspondences.

A typical sentence alignment algorithm starts by calculating alignment scores , trying to find the most reliable initial points of alignment – denominated "anchor points". This score may be calculated based on the similarity in terms of length, words, lexicon or even syntax-tree [Tiedemann 2010]. After finding the anchor points, the process is repeated, trying to align the middle points. Typically, this ends when no new correspondences are found.

The alignment is performed without allowing cross-matching, meaning that the sentences in the source text must be matched in the same order in the target text.

3.6 Tokenizer

Splitting sentences into words allows to subsequently perform word-level alignment.

As with sentence segmenting, it is possible to implement a very simple heuristic to

[4] `http://www.loc.gov/standards/iso639-2/php/code_list.php`
[5] `http://www.iso.org/iso/country_codes/iso_3166_code_lists.htm`
[6] An example of such a sentence segmenter can be downloaded at `http://www.eng.ritsumei.ac.jp/asao/resources/sentseg/`.

accomplish this task by defining sets of characters which are to be considered either word characters or non-word characters. For example, in Portuguese, it is possible to define `A-Z`, `a-z`, `-` and `'` (and all the characters with diacritics) as belonging inside words, and all others as being non-word characters.

However, here too there are languages in which other methods must be used: in Chinese, for example, word boundaries are not as easy to determine as with Western-European languages. This imposes interesting challenges on how to align texts in some languages at word-level.

3.7 Word alignment

The word-alignment process presents some similarities with higher level alignments. However, this is a more complex process, given the more frequent order inversions, differences in part-of-speech and syntactic structure, and multi-word units. As a result, research in this area is less advanced than in sentence alignment [Véronis and Langlais 2000; Chiao et al. 2006].

Multi-word units (MWU) are idiomatic expressions composed by more than one word. MWUs must be taken into account because frequently they cannot be translated word-by-word; the corresponding expression must be found, whether it is also an MWU or a single word [Tiedemann 1999; 2004]. This happens frequently when aligning English and German, for instance, because of the German composed words such as Geschwindigkeitsbegrenzung (in English, speed limit).

3.8 Output

The output formats vary greatly. Generally, every format must include some kind of sentence identification (either by the offsets of its begin and end characters or by a given identification code) and the pairs of correspondences; optionally, additional information may also be included, like a confidence value, or the stemmed form of the word.

BAF, for example, is available in the following formats [Simard 1998]:

– *COAL format.* An alignment in this format is composed by three files: two plain text files containing the actual sentences and words originally provided, and an alignment file which contains a sequence of pairs $[(s1, t1), (s2, t2), ..., (sn, tn)]$. Each s_i^{th} segment in the source text has its beginning position given by s_i, and the end position given by $S_{i+1} - 1$. The corresponding segment in the target text is delimited by t_i and $t_i{+}1 - 1$.

– *CES format.* This format is also composed by three files: two text files in CESANA format, enriched with SGML mark-up that uniquely identifies each sentence, and an alignment file in CESALIGN format, containing a list of pairs of sentence identifiers. The results of ARCADE II were also stored under this format.

– *HTML format.* Not intended for further processing purposes. This is a visualization format, displayable in a common browser.

Other possible formats are translation memory (TM) formats [Désilets et al. 2008]. Translation memories are databases (in the broader sense of the word) that store pairs of "segments" (either paragraphs, sentences or words). TM store the source text and its translation in language pairs denominated translation units. There are multiple TM formats available.

4. CURRENT PROJECTS AND TOOLS

This section lists some of the most relevant projects and tools being developed or used at the moment.

4.1 NATools

NATools is a set of tools for processing, analyze and extract translation resources from parallel corpora, developed at Universidade do Minho. It includes sentence and word aligners, a probabilistic translation dictionary (PTD) extractor, a corpus server, a set of corpora and dictionary query tools and tools for extracting bilingual resources [Simoes and Almeida 2007; 2006].

4.2 Giza++

Giza++ is an extension of an older program, Giza. This tool performs statistical alignment, implementing several Hidden Markov Models and advanced techniques which allow to improve alignment results [Och and Ney 2000].

4.3 hunalign

hunalign is a sentence-level aligner built on top of [Gale and Church]'s algorithm written in C++. When provided with a dictionary, hunalign uses its information to help in the alignment process, despite being able to work without one [Varga et al. 2005].

4.4 Per-Fide

Per-Fide is a project from Universidade do Minho which aims to compile parallel corpora between **P**ortuguese and other six languages (**E**spañol, **R**ussian, **F**rançais, **I**taliano, **D**eutsch and **E**nglish) [Araújo et al. 2010]. This corpora includes different kinds of speech, such as literary, religious, journalistic, legal and technical. The corpora are sentence-level aligned, and include morphological and morphosyntactic annotations in the possible languages. Automatic extraction of resources is implemented, and the results are made available on the net.

4.5 cwb-align

Also known as easy-align, this tool is integrated in the IMS CWB Open Corpus Workbench, a collection of open source tools for managing and query large text corpora with linguistic annotations, based of an efficient query processor, CQP [Evert 2001]. It is considered as a standard *de facto*, given its quality and implemented features, such as tools for encoding, indexing, compression, decoding, and frequency distributions, a query processor and a CWB/Perl API for post-processing, scripting and web interfaces.

4.6 WinAlign

WinAlign is a commercial solution from the Trados package, developed for professional translators. It accepts previous translations as translations memories and uses it to guide further alignments. This is also useful when clients provide reference material from previous jobs. The Trados package may be purchased at `http://www.translationzone.com/`.

5. CONCLUSIONS

This paper presented the field of parallel corpora alignment. Its historical background and first development initiatives were described, and the alignment process was detailed step-by-step. A list of currently relevant projects and tools was also presented.

Several methods and algorithms for sentence-level text alignment have been devised and improved throughout the years. While the existing methods perform acceptably for well formatted and almost perfectly parallel texts, their performance decreases greatly when the bitexts present major deletions, inversions or uncleaned mark up. The quality of the sentence segmentation process also has a determinant role in the alignment.

Word level alignment is still at an early stage; being considerable more difficult. Despite its similarities with sentence alignment, the elevated number of inversions, deletions and the existence of multi-word units must be taken into account. Here too the tokenizer quality has a decisive influence on the final results.

There is a wide range of applications for aligned parallel corpora, specially in the areas of computational linguistics, lexicography, machine translation and knowledge extraction.

In order to improve the lower level alignments, we feel that there is the need for a structural aligner, capable of aligning big blocks of text (like sections or chapters), which would allow to divide the problem of aligning a given pair of texts into smaller subproblems of aligning each of their sections. This would also allow to detect and discard beforehand unmatched sections (a common problem when aligning literary texts, for example).

REFERENCES

Almeida, J. and Simões, A. 2010. Automatic Parallel Corpora and Bilingual Terminology extraction from Parallel WebSites. In *Proceedings of LREC-2010*. 50.

Araújo, S., Almeida, J., Dias, I., and Simões, A. 2010. Apresentação do projecto Per-Fide: Paralelizando o Português com seis outras línguas. *Linguamática*, 71.

Brown, P., Lai, J., and Mercer, R. 1991. Aligning sentences in parallel corpora. 169–176.

Chen, S. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 9–16.

Chiao, Y., Kraif, O., Laurent, D., Nguyen, T., Semmar, N., Stuck, F., Véronis, J., and Zaghouani, W. 2006. Evaluation of multilingual text alignment systems: the ARCADE II project. In *Proceedings of LREC-2006*. Citeseer.

Désilets, A., Farley, B., Stojanovic, M., and Patenaude, G. 2008. WeBiText: Building large heterogeneous translation memories from parallel web content. *Proc. of Translating and the Computer 30*, 27–28.

Evert, S. 2001. The CQP query language tutorial. *IMS Stuttgart 13*.

Gale, W. and Church, K. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics 19,* 1, 75–102.

Hart, M. and Newby, G. 1997. Project Gutenberg. `http://www.gutenberg.org/wiki/Main_Page`.

Kay, M. 1991. Text-translation alignment. In *ACH/ALLC '91: "Making Connections"Conference Handbook*. Tempe, Arizona.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *5*.

Langlais, P., Simard, M., Veronis, J., Armstrong, S., Bonhomme, P., Debili, F., Isabelle, P., Souissi, E., and Theron, P. 1998. Arcade: A cooperative research project on parallel text alignment evaluation.

Melamed, I. D. 1998a. Annotation style guide for the blinker project. *Arxiv preprint cmp-lg/9805004 cmp-lg/9805004*.

Melamed, I. D. 1998b. Manual annotation of translational equivalence: The blinker project. *Arxiv preprint cmp-lg/9805005 cmp-lg/9805005*.

Moore, R. 2002. Fast and accurate sentence alignment of bilingual corpora. *Machine Translation: From Research to Real Users*, 135–144.

Och, F. and Ney, H. 2000. Improved statistical alignment models. 440–447.

Olson, D. and Delen, D. 2008. *Advanced data mining techniques*. Springer Verlag.

Resnik, P. and Smith, N. 2003. The Web as a parallel corpus. *Computational Linguistics 29,* 3, 349–380.

Rijsbergen, C. J. V. 1979. *Information Retrieval*, 2nd ed. Butterworth-Heinemann, Newton, MA, USA.

Robinson, N. 2001. A Comparison of Utilities for converting from Postscript or Portable Document Format to Text. Tech. rep., CERN-OPEN-2001.

Simard, M. 1998. The BAF: a corpus of English-French bitext. In *First International Conference on Language Resources and Evaluation*. Vol. 1. Citeseer, 489–494.

Simard, M. and Plamondon, P. 1998. Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation 13,* 1, 59–80.

Simoes, A. and Almeida, J. 2006. NatServer: a client-server architecture for building parallel corpora applications. *Procesamiento del Lenguaje Natural 37*, 91–97.

Simoes, A. and Almeida, J. 2007. Parallel corpora based translation resources extraction. *Procesamiento del lenguaje natural 39*, 265–272.

Tiedemann, J. 1999. Word alignment-step by step. 216–227.

Tiedemann, J. 2004. Word to word alignment strategies. 212.

Tiedemann, J. 2007. Building a multilingual parallel subtitle corpus. *Proc. CLIN*.

Tiedemann, J. 2010. Lingua-Align: An Experimental Toolbox for Automatic Tree-to-Tree Alignment. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*.

Tiedemann, J., Nygaard, L., and Hf, T. 2004. The OPUS corpus–parallel and free. In *In Proceeding of the 4th International Conference on Language Resources and Evaluation (LREC*. Citeseer.

Tsvetkov, Y. and Wintner, S. 2010. Automatic Acquisition of Parallel Corpora from Websites with Dynamic Content. In *Proceedings of LREC-2010*.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. 2005. Parallel corpora for medium density languages. *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*.

Véronis, J. 2000. From the Rosetta stone to the information society. 1–24.

Véronis, J. and Langlais, P. 2000. Evaluation of parallel text alignment systems. Vol. 13. 369–388.