The BAF: A Corpus of English-French Bitext

Michel Simard

Laboratoire de Recherche Appliquée en Linguistique Informatique (RALI)
Université de Montréal
simardm@IRO.UMontreal.CA

Abstract

The BAF is a corpus of English and French translations, hand-aligned at the sentence level, which was developed by the University of Montreal's RALI laboratory, within the "Action de recherche concertée" (*ARC*) A2, a cooperative research project initiated and financed by the *AUPELF-UREF*. The corpus, which totals approximately 800 000 words, is primarily intended as an evaluation tool in the development of automatic bilingual text alignment method. In this paper, we discuss why this corpus was assembled, how it was produced, and what it contains. We also describe some of the computer tools that were developed and used in the process.

1 Introduction

The *BAF*[1]is a corpus of English and French *bitext*: it consists of pairs of English and French documents, which are translations of one another, and whose sentences have been aligned. The corpus was produced by researchers at the *CITI*, a Canadian government research laboratory, as part of their contribution to the "Action de recherche concertée" (*ARC*) A2, a cooperative research project initiated and financed by the *AUPELF-UREF*. The CITI's machine-aided translation program has since been handed over to the University of Montreal's *RALI* laboratory, where we continue working on the BAF. We have recently made version 1.1 of the corpus available to the research community. It can be obtained from the RALI, through its World Wide Web server, at URL `http://www-rali.iro.umontreal.ca`.

While such bilingual corpora are now quite common, what distinguishes the BAF from other bitexts is that the alignments were entirely done by hand. As a result, the corpus can be used as a reference to evaluate and compare the performance of various automatic alignment techniques.

The greater part of the corpus is made up of "institutional" texts: debates of the Canadian parliament (Hansards), court transcripts and UN reports; but we have also included some scientific, technical and literary documents. In all, the corpus contains approximately 400 000 words in each language.

In this paper, we discuss why this corpus was assembled, how it was produced, and what it contains. We also describe some of the computer tools that were developed and used in the process, and which we have also made available.

[1]The name *BAF* is the French acronym for *English-French Bitext*

2 Background

For some years now, the international scientific community has shown interest in automated techniques that reproduce a multilingual speaker's ability to align a text with its translation, i.e. to identify the correspondences that exist between the segments of the two texts. Members of our laboratory, the RALI, have been actively involved in this area since 1991. Our interest in this question stems from the conviction that accurate alignment methods are the required basis for a whole set of computer tools for human translators (Isabelle et al., 1993). The simplest example of such a tool is probably the *TransSearch* bilingual concordancing system (Simard et al., 1993), which allows a user to query a large archive of existing translations, in order to find ready-made solutions to specific translation problems. Such a tool has proved extremely useful not only for translators, but also for bilingual lexicographers (Langlois, 1996) and terminologists (Dagan and Church, 1994). More sophisticated applications based on alignment technology have also been the object of recent work, such as the automatic building of bilingual lexical resources (Melamed, 1996; Klavans and Tzoukermann, 1995), the automatic verification of translations (Macklovitch, 1996; Macklovitch, 1995), the automatic dictation of translations (Brousseau et al., 1995) and even interactive machine translation (Foster et al., 1997).

Enthusiasm for this relatively new field of work was sparked early on by the apparent demonstration that very simple techniques could yield almost perfect results. For instance, to produce sentence alignments, Brown et al. (Brown et al., 1991) and Gale and Church (Gale and Church, 1991) both proposed methods that completely ignored the lexical content of the texts, and relied almost entirely on the intuition that short sentences tend to translate into short sentences, while longer sentences tend to translate into longer ones. With simple programs in which this observation was encoded into a statistical model, both teams were able to achieve accuracy levels exceeding 98%.

However, it quickly became apparent that this kind of performance could not be obtained with just any type of text, and that in general, the level of success that can be expected from automatic text alignment programs is highly dependent on the specific pair of texts under consideration. The truth is that, while text alignment is mostly an easy problem, especially when considered at the sentence level, there are situations where even humans have a hard time making the right decisions. In fact, the argument could be made that, ultimately, text alignment is no easier than the more general problem of natural language understanding.

Substantial work remains to be done, therefore, if the alignment technology is to achieve its full potential. Given the number of groups working on this problem, there is a pressing need for tools and resources that make it possible to evaluate and compare the performance of the various methods proposed. One of the things that is required is a common testbed, in the form of reference alignment corpora. This is precisely what the BAF is meant to be.

3 Guidelines and Definitions

The first step in the process of building a corpus of hand-aligned bitext is to clarify what we understand by the term *alignment*. Essentially, this entails describing the objects that the alignment connects, and defining how the alignment connects them. Based on the answers to these questions, a set of *guidelines* can then be devised, which the human aligners will be instructed to follow when producing the alignments.

3.1 What is an Alignment?

A bitext alignment describes the relations that exists between a text and its translation. These relations can be viewed at various levels of granularity: between text divisions, paragraphs, sentences, propositions, words, even characters. While it would certainly have been interesting to produce finer-grain alignments, it was decided that the BAF would record correspondences at the level of sentences. This decision was based on a number of factors.

First, sentence-level alignments have so far proved very useful in a number of applications, which could be characterized as *high recall, low precision* applications, i.e. applications where it is more important to have all the answers to a specific question than to have only the "good" ones.

One example of such an application is bilingual lexicography. When a lexicographer is examining a bilingual concordance, with a view to mapping out the various meanings or contexts of use of a particular term or expression, he seeks exhaustivity. In other words, he is willing to tolerate a relatively high number of irrelevant or redundant examples ("noise"), in order to make sure that he doesn't overlook anything ("silence").

Automatic or machine-assisted translation verification is another such application. A system that does translation verification will look for specific translation errors, such as omissions on the part of the translator, the use of *faux-amis* (false cognates), inconsistent use of terminology, etc. If translation verification is anything like spelling or grammar checking, we can expect users to be ready to tolerate a fair amount of noise, just to make sure they don't miss out on glaring errors.

A final example is the automatic acquisition of information about translation, as was proposed in (Brown et al., 1993) as part of a project to build a machine translation system entirely based on statistical knowledge. While such ambitious projects now seem to have been abandoned, the statistical models at the heart of these projects are still around, for example in less ambitious interactive MT projects (Foster et al., 1997) and text alignment systems (Simard and Plamondon, 1996). Such statistical models need to be "trained" with large quantities of bitext. Intuitively, the ideal training material for this task would be bitext aligned at the level of words. Yet, because these models picture the translation process in an extremely simplified manner, reliable statistical estimates can nevertheless be obtained from much less precise data, such as pairs of sentences.

This explains why a lot of the research effort in this domain has so far focussed on sentence-level alignments. Of course, this is not to say that reference alignments at a finer level would not be a useful thing, in the contrary. Besides, a word-level alignment could be made to incorporate the sentence-level alignment as a by-product.

Unfortunately, producing such a thing as a word-level alignment turns out to be a much more difficult problem: while there is often a one-to-one correspondence between the sentences of a text and its translation, matters get a lot more complicated when we get down to the level of "words". The main reason is that, at this level, syntactic and stylistic constraints in the target language affect the content and structure of the translated text at least as much as does the source text. As a result, in order to accurately describe the complex relations that exist between the words of a text and its translation, we will likely need a fairly elaborate alignment scheme. Finally, it is clear that producing hand-made word-alignments for more than a few sentences is going to be a very costly proposition.

For all these reasons, we decided that it would be more appropriate initially to concentrate on sentence-level alignments. Furthermore, we decided to restrict ourselves to "non-crossing" alignments:

- An *alignment* is a parallel segmentation of the two texts, into an equal number of segments, such that the n^{th} segment in one text and the n^{th} segment in the other text are translations of one another.

We refer to such alignments as "non-crossing" because of the impossibility to explicitly account for *inversions*, i.e. situations where the order of sentences is not the same in the two texts. This type of alignment nonetheless covers the vast majority of situations encountered in real-life texts. Furthermore, this is the type of output that is actually produced by most existing sentence alignment programs.

3.2 What is a Sentence?

If we are going to align sentences, then obviously we must clarify what we understand by *sentence*: while most people have strong intuitions about what is a sentence and what is not, there is no universal definition of that notion. Before we set out on devising one, however, it should be noted that because the BAF is primarily intended to be used as a testbed for alignment methods, neither the exact definition, nor the actual segmentation of the text that results are crucially important: if the evaluation process focuses on alignment, the tested methods should all work on the same prior segmentation of the text. It is unlikely that a particular segmentation will favor one alignment method over another.

Therefore, our main concern in this regard was to come up with some guidelines for segmentation that would be both practical for the aligners and useful for the end-users of the corpus. We started out with something relatively straightforward, which we then expanded as needed. Essentially, these were the guiding principles:

- *A Sentence is a syntactically autonomous sequence of words, terminated by a full-stop punctuation.*

  The term *full-stop punctuation* naturally includes periods ('.'), exclamation marks ('!') and question marks ('?'), but we also admitted the possibility of a sentence ending with a colon (':') or semicolon (';'), as long as the sentence could stand on its own syntactically (this is what we mean by *syntactically autonomous*). In general, we consider that the symbol that explicitly marks the end of a sentence (if such a symbol exists) belongs to that sentence.

- *Titles are sentences.*

  This applies to chapter titles, section titles, table titles, figure titles, etc. even though these generally do not end with a full-stop punctuation.

- *Enumerators are sentences.*

  Any number, Roman or Arabic, or letter that appears in front of a title (chapter title, section title, etc.) or paragraph, is a sentence. This is also true of the "N.B." or "Note:" that precedes notes.

- *Items of an enumeration are sentences,*

  as is, of course, the "header" of the enumeration. This rule only applies when the items in the enumeration are separated by semicolons, or when the presentation clearly suggests that this is an enumeration, such as, for example, when all items appear on separate lines.

- *Each cell in a table is a sentence.*

  Some documents contained tables. In most cases, however, the formatting of the table was lost, and all that remained was the content of the cells, separated by arbitrary markers (for example, pairs of commas or vertical bars).

3.3 What is a translation?

Finally, we needed to provide the aligners with is some criteria for determining what constitutes a translation. In general, we found it satisfactory to say that segments of text A and B were translations of one another if they conveyed the same "ideas" or "concepts", at least to an acceptable point. The main practical problems we had to solve revolved around situations where the translation deviated from its usual "linear" progression.

First, there were the cases of *omissions* and *insertions*, i.e. situations where some segment in one text does not appear to have a corresponding counterpart in the other text. In these cases, we allowed for the existence of "empty" segments in the alignment. This way, a sentence that does not have an equivalent in the other text can be aligned with an empty segment.

There were some situations where we chose to ignore an omission (or insertion), for the benefit of recording a larger correspondence. This would happen, for example, if a single sentence A in one language was translated as two sentences A'_1 and A'_2, between which a third, untranslated sentence B' was interpolated. In this case, we would simply align A with the sequence $A'_1B'A'_2$, regardless of the fact that B' has no equivalent in A.

Then, there was the case of *inversions*. This happens when the order of the sentences is not the same in the source and translated texts. As mentioned earlier, our definition of alignment makes it impossible to explicitly account for inversions. Two different strategies were adopted, depending on the nature of the inversion.

For simple inversions, we opted for a strategy of "under-segmentation": when a pair of contiguous sentences AB appeared as $B'A'$ in the other text, we chose not to segment the texts after sentences A and B', but rather to keep A and B together within the same segment, and then do the same for B' and A'.

For more complex inversions, we usually chose to treat the inverted segments as omissions. For example, given some sequence of sentences $A_1A_2A_3...A_n$ translated as $A'_2A'_3...A'_nA'_1$, we would consider A_1 and A'_1 to be "omitted" segments (align them with empty segments), and then align A_2 with A'_2, A_3 with A'_3, etc. Although this was clearly not the correct way of aligning the texts, it was felt that in the end, such an alignment would be more "useful".

4 The Alignment Protocol

The definition of alignment given above suggested a very straightforward way of producing alignments by hand: read both texts in parallel, and segment them as you go along, in such a way that:

1. segment boundaries always coincide with sentence boundaries;
2. the n^{th} segment in one text and the n^{th} segment in the other are translations of one another;
3. segments are always as small as possible.

Of course, given the relative vagueness of the definitions of sentence and translation given above, it was clear that in many situations, arbitrary decisions would have to be made. Our human "aligners"[2] were instructed to be as consistent as possible, and when in doubt, to try to do the most "useful" thing. But even then, because of the repetitive nature of the task, errors had to be expected.

For these reasons, it was decided that all the texts would be aligned twice, each time by a different aligner. The resulting alignments would then be compared, so as to detect any discrepancies between the two. The aligners were then asked to conciliate these differences together. Because all of the BAF corpus was aligned by the same two aligners, this way of proceeding not only minimized the number of errors, it also ensured that both aligners had the same understanding of the guidelines.

In this regard, it is interesting to note that the vast majority of disagreements between the two aligners revolved around questions of sentence segmentation rather than questions of translational equivalence. Considering that, as discussed earlier, the actual segmentation on which the alignment is based is not crucially important, this suggests that it would probably have been a good idea to first have the texts segmented by a single person, and then have the aligners produce the alignments based on that segmentation.

5 The Documents

As mentioned earlier, the complexity of the alignment task is very dependent on the type of text. Using exclusively Hansard documents would probably have greatly simplified the task of assembling the BAF, not only because these texts are typically fairly easy to align, but also because they are widely available in plain-text format. However, we felt that concentrating one a single genre would make the corpus less useful as an evaluation resource. So, when selecting the pairs of documents that make up the BAF, we tried to include documents from various sources and of various genres.

Of course, the corpus is, to some extent, representative of the types of texts that are available in multilingual versions. For instance, it does not contain such things as newspaper articles or e-mail messages simply because such texts are usually not translated. On the other hand, documents produced by international organizations or governments of countries with two or more official languages are usually easy to find in multilingual versions, and so the BAF contains a lot of these. This is also true, although maybe to a lesser extent, of some technical documents, such as user's manuals, and of literary text. In general, scientific articles are not routinely translated, but as part of a Canadian government-owned research in-

[2]In effect, the "aligners" that produced all the alignments of the BAF were professional translators.

stitution, we found that we had access to a number of bilingual technical reports. This explains their presence in the BAF.

The documents that currently make up the BAF corpus are presented in Table 1.

6 File Formats

Right from the start, the BAF was intended to be used as an evaluation resource in the development of general-purpose alignment methods. For this reason, we were interested in "plain-text" documents, i.e. text files that were not tied to a specific word-processing program, and that contained no formatting or structural mark-up. In many cases, the documents were explicitly converted from proprietary formats (WordPerfect, FrameMaker, etc.) to "plain-text" format. In other cases, mark-up (SGML and others) was eliminated.

In its current version, the BAF corpus takes the form of a collection of computer files. Because the corpus is available in three different formats, several files correspond to each pair of documents.

1. **COAL format**: This is the format in which the alignments were originally produced. A pair of documents in COAL format consists of three distinct files: two plain text files, and an alignment file. The alignment file contains a sequence of pairs $[(s_1, t_1), (s_2, t_2), ..., (s_n, t_n)]$ where each pair corresponds to a *segmentation point*, expressed as a pair of *character offsets*. The interpretation of these numbers is straightforward: the segment of text in the first text file that starts at the s_i^{th} character and ends just before the s_{i+1}^{th} character corresponds to the segment of text in the second text file that extends between the t_i^{th} and t_{i+1}^{th} characters.

2. **CES format**: This SGML-based *Corpus Encoding Standard* was proposed jointly by Vassar College's Department of Computer Science and the Laboratoire Parole et Langage (LP&L) of The Centre National de la Recherche Scientifique (CNRS) in Aix-en-Provence, France (Ide et al., 1995). In this format, three files also correspond to each pair of documents: two text files (*CESANA* format) and an alignment file (*CESALIGN* format). The text files are enriched with SGML mark-up that uniquely identifies each sentence in the text. The alignments are then expressed as pairs of lists of sentence identifiers.

   One important difference between the COAL and CES formats is that the CES assumes a complete segmentation of the texts into sentences, which is made explicit by the mark-up. The COAL format does not make that segmentation explicit, and nothing guarantees that the segmentation that is implicit in the alignment is complete. For example, if one sentence in the English text is translated as two sentences in the French text, the boundary between the first and second French sentences will not appear in the COAL alignment.

   The CES version of the BAF was produced automatically from the COAL version. While performing the conversion, we "completed" the implicit segmentation, by using a number of sentence-boundary detection heuristics. A random sampling of the resulting sentences reveals that the segmentation is about 97.5% correct: less than 2.5% of the sentence boundaries in the corpus are incorrect, and less than 2.5% of the real boundaries are missing.

3. **HTML format**: This is a "visualization" format. Here, to a pair of documents corresponds a single HTML file, which can be loaded into any HTML viewer capable of displaying tables and colors.

7 Programs

To assist our human "aligners" in their work, we developed a number of computer programs, the most important of which is the *Manual* program, whose purpose is to visualize and manipulate alignments. *Manual* was implemented as a special "editing mode" in the well-known *Emacs* editor.

With the *Manual* program, the two texts to align are shown in two separate Emacs windows. To display the alignment itself, *Manual* uses colors: aligned segments of text (which we call *couples*) are displayed on same-color backgrounds. The user can easily navigate the texts in parallel, position the cursor as he wishes, perform searches, etc. But he cannot modify or edit the texts. Since we define an alignment as a parallel segmentation of the texts, what the *Manual* program will allow the user to do is specify where the texts should be segmented. All of this is done without physically altering the texts: the alignment is recorded separately, as a sequence of pairs of segmentation points.

To modify an alignment, *Manual* provides a number of editing functions. These basically allow for two types of actions: either *split* a couple, to produce a pair of couples, or *merge* two adjacent couples, to produce a single unit. Typically, to create an alignment from scratch, the aligner starts out with a sin-

Genre	Reference	Source
Institutional	Hansard - Canadian Parliamentary Proceedings. March 14, 1994	House of Commons publication service.
	Supreme court of Canada (1995). *Terrence Wayne Burlingham v. Her Majesty the Queen*	*Centre de recherche en droit public* of the Law Faculty of the University of Montreal.
	UN International Labor Organization (1985). *241st and 242nd Reports of the Committee on Freedom of Association*	*ECI Multilingual Corpus.*
	UN (1993). *Report of the Secretary-General on the Work of the Organization*	UN translation services.
Scientific	Geoffroy, Catherine (1994). *Les technologies de communication de l'information et les aîné(e)s.* CITI technical report	All these documents were available as technical reports at the CITI.
	Lapointe, François (1995). *Changement technologique et organisation du travail.* CITI technical report	
	Macklovitch, Elliott (1995), *Peut-on vérifier automatiquement la cohérence terminologique?* in **Actes des IVes journées scientifiques, Lexicommatique et Dictionnairiques**, Lyon, France	
	Simard, Michel (1995), *Réaccentuation automatique de textes français*, CITI technical report	
Technical	Xerox Corporation, *ScanWorX User's Guide*	*ECI Multilingual Corpus.* *Note*: these documents contain in appendix a relatively large glossary of terms, which could not be aligned, because the order of the entries is completely different in French and English.
Literary	Verne, Jules. *De la terre à la lune*	The original French version was obtained from the WWW site of the *Association des Bibliophiles Universels.* The English translation comes from the *Project Gutenberg.*

Table 1: BAF Documents

gle couple that covers the entirety of the two texts, and repeatedly applies "splitting" functions; "merging" functions are normally used to correct errors.

Other programs were developed to view, compare, concatenate, split, compute various statistics about and convert alignments to and from various formats. It's amazing just how many silly things you can do with alignments.

All of these programs are publicly available, and can be obtained from the RALI, through our World Wide Web server.

8 Conclusion and Future Work

We have described the BAF, a corpus of English and French translations, hand-aligned at the sentence level. The corpus, which totals approximately 800 000 words, is primarily intended as an evaluation tool in the development of automatic bilingual text alignment method. It is currently available from the RALI's Web site, at URL `http://www-rali.iro.umontreal.ca`. A number of computer tools for manipulating bitext alignments were also developed in the course of the project, the most important of which is the *Manual* program, which allows one to visualize and edit an alignment within the *Emacs* editor. All of these programs are also available on the RALI's Web site.

The BAF corpus has already been put to use: in the context of the AUPELF-UREF's ARC-A2, a friendly competition between alignment programs was organized. The BAF was used, along with a similar corpus, as a common testbed: the corpus files were submitted to each of the alignment programs, and the output of each was compared to the hand-made alignments. Performances were then measured, using various evaluation metrics of the kind proposed in (Isabelle and Simard, 1996).

Yet, while the corpus is already usable, there is still much room left for improvement. For instance, we did not initially believe it necessary to first produce a complete segmentation of the texts, and then describe correspondences with regard to this segmentation: in practice, both segmentation and alignment were performed in parallel. Although this way of proceeding was probably more economical, it had one major drawback: when two or more sentences translated to a single one, the internal segmentation was not recorded. As a result, the segmentation of the corpus in the BAF was partial. In version 1.1 of the corpus, the segmentation was completed by means of automatic segmentation methods. This automatic segmentation contains errors, and should be manually verified, something that we plan to do in the short term.

Also, the BAF does not record crossing alignments. While the vast majority of alignments are non-crossing, it is currently impossible to use the corpus to evaluate alignment methods that do crossing alignments. Correcting this situation would probably not be too costly, because once the segmentation of the text is verified, crossing alignments can be easily detected by examining exclusively alignment patterns that deviate from the usual "1 to 1".

Finally, we have recently begun experimenting with word-level alignments. At this level, the questions of "what to align?" and "how to align them?" become much more complex. In spite of these difficulties, we hope to have at least a small part of the BAF aligned at this level before the end of the ARC-A2 collaboration.

References

J. Brousseau, C. Drouin, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, and P. Plamondon. 1995. French Speech Recognition in an Automatic Dictation System for Translators: the TransTalk Project. In *Proceedings of Eurospeech 95*, Madrid, Spain.

Peter Brown, Jennifer C. Lai, and Robert Mercer. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of ACL-91*, Berkeley CA.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2).

Ido Dagan and Kenneth W. Church. 1994. Termight: Identifying and Translating Technical Terminology. In *Proceedings of ANLP-94*, Stuttgart, Germany.

George Foster, Pierre Isabelle, and Pierre Plamondon. 1997. Target-Text Mediated Interactive Machine Translation. *Machine Translation*, 21(1-2).

William A. Gale and Kenneth W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of ACL-91*, Berkeley CA.

Nancy Ide, G Priest-Dorman, and Jean Véronis. 1995. Corpus Encoding Standard. `http://www.cs.vassar.edu/CES/`.

Pierre Isabelle and Michel Simard. 1996. Propositions pour la représentation et l'évaluation des alignements de textes parallèles. `http://www-rali.iro.umontreal.ca/arc-a2/PropEval`.

Pierre Isabelle, Marc Dymetman, George Foster, Jean-Marc Jutras, Elliott Macklovitch, François Perrault, Xiaobo Ren, and Michel Simard. 1993. Translation Analysis and Translation Automation. In *Proceedings of TMI-93*, Kyoto, Japan.

Judith Klavans and Evelyne Tzoukermann. 1995. Combining Corpus and Machine-readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation*, 10(3).

Lucie Langlois. 1996. Bilingual Concordances: A New Tool for Bilingual Lexicographers. In *Proceedings of AMTA-96*, Montréal, Canada.

Elliott Macklovitch. 1995. TransCheck — or the Automatic Validation of Human Translations. In *Proceedings of the MT Summit V*, Luxembourg.

Elliott Macklovitch. 1996. Peut-on vérifier automatiquement la cohérence terminologique? *META*, 41(3).

I. Dan Melamed. 1996. Automatic Construction of Clean Broad-coverage Translation Lexicons. In *Proceedings of AMTA-96*, Montréal, Canada.

Michel Simard and Pierre Plamondon. 1996. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. In *Proceedings of AMTA-96*, Montréal, Canada.

Michel Simard, George Foster, and François Perrault. 1993. TransSearch : un concordancier bilingue. Technical report, CITI.