Preparation and exploitation of bilingual texts

Dusko Vitas, Cvetana Krstev, Eric Laporte

Preparation and exploitation of bilingual texts

Duško Vitas
Faculty of Mathematics
Studentski trg 16, CS-11000 Belgrade, Serbia

Cvetana Krstev
Faculty of Philology
Studentski trg 3, CS-11000 Belgrade, Serbia

Éric Laporte
Institut Gaspard-Monge, Université de Marne-la-Vallée
5, bd Descartes, 77454 Marne-la-Vallée CEDEX 2, France

Introduction

A *bitext* is a merged document composed of two versions of a given text, usually in two different languages. An aligned bitext is produced by an alignment tool or aligner, that automatically aligns or matches the versions of the same text, generally sentence by sentence. A multilingual aligned corpus or collection of aligned bitexts, when consulted with a search tool, can be extremely useful for translation, language teaching and the investigation of literary text (Veronis, 2000). This is all the more true for a pair of languages such as Korean and French, for which few people are bilingual, and many literary translations involve pairs of translators. For such language pairs, retrieving solutions of previously resolved translation problems is an invaluable aid. In addition, multilingual corpora are in the core of some research in natural language processing (NLP), both in theoretical fields, such as contrastive linguistic and lexicography, and in applicative fields, such as translation, term extraction, or translation memories production.

The current methods of construction and exploitation of multilingual aligned corpora are essentially based on statistical models of text. In this article, we propose an enhancement of these methods with the use of lexical and grammatical resources. The open-source Unitex system is the main corpus processor that systematically makes use of lexicons and grammars for text exploration. This system can process one language at a time. We outline a project of extension of Unitex to the processing of bitexts.

The authors of this article are European, and their experience of bitexts stems from European projects. By handling bitexts involving Slavic languages, they had the opportunity to get familiar with two types of problems likely to occur with Korean-French bitexts: alphabet transliteration, and massive inflectional variation of words.

This article is organised as follows. In section 1, we define and exemplify the notions of bitext and alignment. Section 2 introduces the statistic-based approach to NLP and surveys methods of text alignment. In section 3 we present the linguistic-based approach, the Unitex system, and the potential contribution of this type of methods to the processing of multilingual corpora. Final remarks are presented in the conclusion[1].

[1] This research has been partially financed by the CNRS.

1. Basic notions: bitext, alignment

The word and the notion of *bitext* are attributed to Harris (1988). A bitext is composed of two versions of a given text, usually in two different languages. The two texts are assumed to be *semantically* equivalent, for instance, the original text and its translation. It is not necessary that the original text itself is included in a bitext: it can consist of various versions of one text in different languages, but also of different translations into one language of the same source text, or of closely connected source texts.

An aligned bitext is produced by a tool that automatically aligns or matches the versions of the same text, generally sentence by sentence. In general, the bitext construction proceeds in two main steps: in the first one, each text is separately segmented into instances of a given unit, and in the second one these units are aligned. The units are usually sentences, but they can also be larger, as paragraphs, or smaller, as words.

1.1. Markup

Our first example of a bitext is the Statute of the International Court of Justice, which has its seat in The Hague. The text of this Statute exists in the languages of several United Nation members. Since it is an international law document, it can be assumed that the text in all languages has almost exactly the same meaning. The relation between the two parts of a bitext can be illustrated by the sample of the English and French version displayed in Figure 1.

STATUTE OF THE INTERNATIONAL COURT OF JUSTICE Article 1 The International Court of Justice established by the Charter of the United Nations as the principal judicial organ of the United Nations shall be constituted and shall function in accordance with the provisions of the present Statute.	STATUT DE LA COUR INTERNATIONALE DE JUSTICE Article 1 La Cour internationale de Justice instituée par la Charte des Nations Unies comme organe judiciaire principal de l'Organisation sera constituée et fonctionnera conformément aux dispositions du présent Statut.

Figure 1. A raw bitext

The common methods of alignment of a bitext usually assume that before alignment both texts have been marked up, which means that the elements of its logical layout were explicitly and unambiguously annotated. If we use XML tags to tag the logical layout of our chosen texts, we will insert into these texts explicit information about the elements of their logical structure, as illustrated in Figure 2. The tags in this example mark the potentially equivalent units in a bitext.

<head>STATUTE OF THE INTERNATIONAL COURT OF JUSTICE</head> <head>Article 1</head> <seg>The International Court of Justice established by the Charter of the United Nations as the principal judicial organ of the United Nations shall be constituted and shall function in accordance with the provisions of the present Statute.</seg>	<head>STATUT DE LA COUR INTERNATIONALE DE JUSTICE</head> <head>Article 1</head> <seg>La Cour internationale de Justice instituée par la Charte des Nations Unies omme organe judiciaire principal de l'Organisation sera constituée et fonctionnera conformément aux dispositions du présent Statut.</seg>

Figure 2. A bitext with logical layout mark-up

From the texts marked up in this way, and using different methods, it is possible to effectively match the marked segments of one of the texts with the equivalent segments in the other. Our example of bitext, once aligned and represented in the TMX standard (http://www.lisa.org/standards/tmx/), would have the following form (for this presentation, we simplified this example by omitting some obligatory attributes):

```
...
<tu>
   <tuv xml.lang="EN" ><head>STATUTE OF THE INTERNATIONAL COURT OF
JUSTICE</head></tuv>
   <tuv xml.lang="FR"><head>STATUT DE LA COUR INTERNATIONALE DE
JUSTICE</head></tuv>
</tu>
<tu>
   <tuv xml.lang="EN"><head>Article 1</head></tuv>
   <tuv xml.lang="FR"><head>Article 1</head></tuv>
</tu>
<tu>
   <tuv xml.lang="EN"><p>The International Court of Justice established
by the Charter of the United Nations as the principal judicial organ of
the United Nations shall be constituted and shall function in accordance
with the provisions of the present Statute.</p></tuv>
    <tuv xml.lang="FR"><p>La Cour internationale de Justice instituée par
la Charte des Nations Unies comme organe judiciaire principal de
l'Organisation sera constituée et
fonctionnera conformément aux dispositions du présent Statut.</p></tuv>
</tu>
```

The text aligned in this way can be used in translation memories in some systems for Machine Aided Human Translation (MAHT), such as Trados, for instance. It can also be consulted by a concordancer.

1.2. Complex cases of correspondence

We will investigate the complexity of the alignment problem in general on a second example of an aligned bitext, which is a sample of Plato's *Republic* in English and in French, processed by the Vanilla aligner (Danielsson, Ridings, 1997):

```
(EN-d2p4seg1) "Not a bad guess," said I.
(FR-d2p3seg1) - Ta conjecture n'est pas fausse, dis-je.
------------------------------------------------------------------
(EN-d2p5seg1) "But you see how many we are?" he said.
(FR-d2p4seg1) - Et vois-tu combien nous sommes ? dit-il.
------------------------------------------------------------------
(EN-d2p6seg1) "Surely."
(FR-d2p5seg1) - Impossible de ne pas le voir !
------------------------------------------------------------------
(EN-d2p7seg1) "You must either then prove yourselves the better men or
stay here."
(FR-d2p6seg1) - Alors, dit-il, ou bien montrez-vous plus forts que les
hommes que voici; ou bien restez ici.
```

The codes in parentheses refer to the sentence number codes in the original texts. As opposed to the previous, straightforward example, this one illustrates several problems that can occur in the process of text alignment.

- Inserted clauses, such as *dis-je*, can be segmented as independent units, which makes the one-to-one correspondence between sentences impossible.
- The use of punctuation marks can significantly differ in texts forming a bitext, as the use of double quotes and long dashes shows in this sample.
- Alignment at word level may be difficult because of differences in word order but also in lexical choices, as in the EN-d2p6 - FR-d2p5 pair.
- Some fragments of a text may be missing in the other. For instance, the EN-d2p7 English sentence does not contain any equivalent of the French sequence *dit-il*.

When the two texts are written in different alphabets, this brings about additional complexity, and some aligners do not process such input. The following excerpts of Orwell's *1984* in Bulgarian, Hungarian, Serbian and English (Erjavec *et al.,* 1998; Krstev *et al.*, 2004a) illustrate both the problem of alphabets and the difference in the number of sentences.

```
<Obg.1.1.24.1>Вторият беше мъж на име O'Брайън, член на Партядрото, човек
на толкова отговорен и толкова висок пост, че Уинстън имаше само смътна
представа за важността му.

<Ohu.1.2.25.1>A másik személy egy O'Brien nevű férfi volt, a Belső Párt
tagja, aki valami fontos és titkos szolgálatot teljesített, de ennek
természetéről Winstonnak csak homályos sejtelme volt.

<Oshs.1.2.26.1>Čovek se zvao O'Brajen.<Oshs.1.2.26.2>Bio je član Uže
partije i zauzimao neki položaj toliko važan i udaljen da je Vinston imao
samo bledu predstavu o njegovoj prirodi.

<Oen.1.1.25.1>The other person was a man named O'Brien, a member of the
Inner Party and holder of some post so important and remote that Winston
had only a dim idea of its nature.
```

1.3. Formalization

The examples above suggest a formal model of an aligned bitext. According to this model, an aligned bitext is a relation *R* between portions of the two texts. Intuitively, this relation represents the *semantic equivalence* between text portions. The coarsest form of this relation connects the two texts as unsegmented units. For instance, we assume that the two integral texts of the *Statute* in section 1.1, or of Plato's *Republic* in section 1.2, are semantic equivalents.

Apart from this trivial relation, both texts *S* and *T* will be assumed to be segmented into smaller units – paragraphs, sentences, or words – in such way that the relation *R* is defined at a finer level than that of the integral texts. Let the relation *R* connect *n* portions of the text *S* with respective portions of *T* in the same order: we will write $s_1 R t_1, s_2 R t_2,..., s_n R t_n$, where *S* is the concatenation of $s_1, s_2,..., s_n$, and *T* of $t_1, t_2,..., t_n$. Some of the elements in the sequences $s_1, s_2,..., s_n$ and $t_1, t_2,..., t_n$ can be empty, or consist of a sequence of various units – paragraphs, sentences, words. In this case, a sequence of units is treated as the semantic equivalent of the corresponding sequence. Such sequences of corresponding units are called *blocks*. It is important to be aware of the possibility of empty sequences, since they describe the fragments missing either from *S* or *T*: in the case of a translation, such blocks represent the sequences that were dropped from the translation or inserted into it.

In general, the finer the initial segmentation into units, the better the quality of the eventual aligned bitext. For the *Statute* text of section 1.1, the result of the word alignment process can be easily imagined. However, the *Republic* example of section 1.2 shows that word-level alignment requires knowledge about different language levels – morphological, syntactic, etc.

2. Quantitative alignment methods

The development of approaches to natural language processing (NLP) in the last ten years is characterized by a sustained interest in the use of statistical models, in connection with the dynamical growth of the number of documents in digital form and to various demands to process them in short time and with a certain reliability. The motivation for this development line is mostly of an applicative nature.

Current methods of text alignment belong to this approach. They consist in general in two steps:

1. segmentation of text into sentences,
2. the alignment of the sentences.

We will not consider here the third step, word-level alignment (Brown *et al.*, 1993).

The methods of segmentation are applied to each of the two texts separately in order to determine the units from which the blocks will be built. In many cases units consist of sentences, but other kinds of units can be used as well. Thus, one of the familiar circularities of computational linguistics, namely the fact that sentences have to be marked before processing, though that processing itself will determine what the sentences are, is present in the alignment problem as well.

For some methods, more detailed tagging, e.g. with paragraphs or headings, is necessary (Bonhomme, Romary, 1995). Sentence tagging is performed in most alignment systems by some machine learning method (Palmer, Hearst, 1994), or through the principle of maximal entropy (Reynar, Ratnaparkhi, 1997).

Once sentences are tagged, sentence alignment is based either on statistical or geometrical methods. Pure statistical methods are based on the assumption that blocks are approximately proportional in length to their equivalents (lengths being expressed in numbers of characters).

Namely, a short sentence in *S* corresponds to a short sentence in *T*. The origin of this method is the Church-Gale index (Gale, Church, 1993) that establishes the lengths of blocks of sentences in correspondence: 1:1, 1:0, 0:1, 2:1, 1:2, 2:2. The Church-Gale method gives good results for texts in which 1:1 blocks prevail, such as law texts or technical documentation. The necessity to correct bitexts produced by this method was noticed by Wu (1994) on the results of Chinese-English text alignment experiments. Wu corrected the errors in segmentation and block formation by using lexical resources, such as a Chinese-English lexicon.

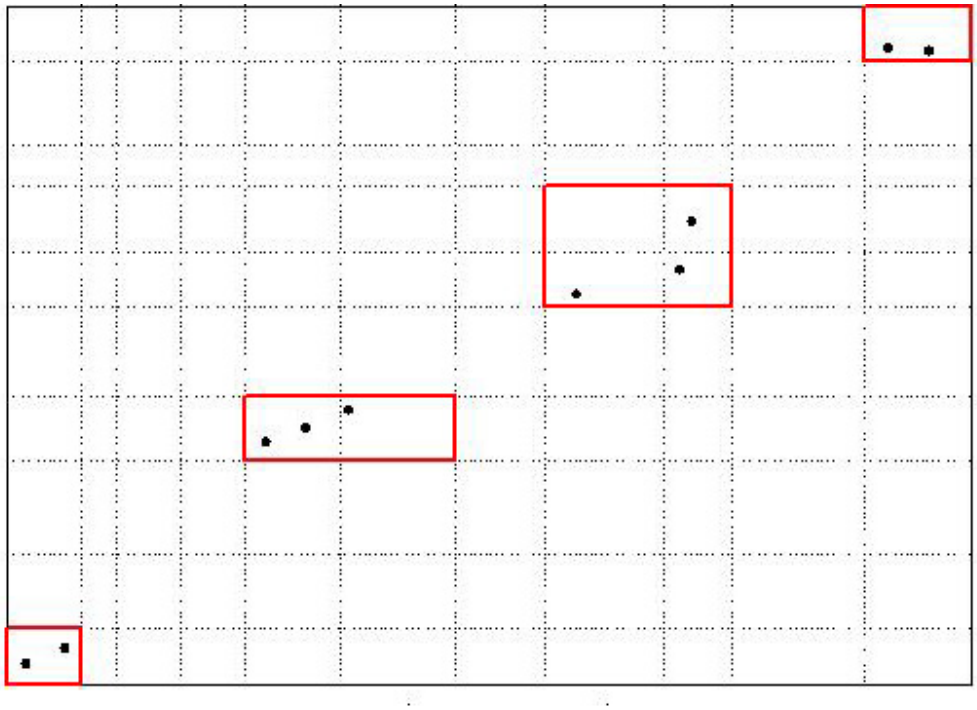


Figure 3. Bitext space: the framed parts represent the correspondences (from left to right) 1:1, 1:2, 2:2, 1:1. The dots represent the location of cognates in respect to the main diagonal.

In the cases when one constituent text is severely deformed, as a result of optical character recognition (OCR), poor paragraph tagging, differences between languages, or bad translation, the results obtained by the Church-Gale method are not valid. In such cases the geometrical approach (Melamed, 1996) is preferred. This approach is based on the definition of the bitext space as the Cartesian product $S \times T$ of texts *S* and *T* considered as sets of sentences. The pairs of sentences with approximatively equal length are identified. (In case of close languages written with the same alphabet, such as French and English, two matching sentences can actually have approximately the same number of characters; in the case of Korean and French, the 'standard' proportion between lengths of matching sentences can be defined by comparing the lengths of the two texts.) Several sentences $t_1,...,t_i$ from text *T* may *a priori* correspond to a sentence *s* from text *S*. The pairs $(s, t_1),...,(s, t_i)$, which are points in the bitext space, are usually represented in a dot-plot diagram (see Figure 3).

The alignment procedure then consists in defining the search band around the main diagonal which potentially contains the equivalent segments.

A correction to this method was introduced by Melamed (2001) who suggested, after Simard *et al.* (1992), that *cognates* should be used instead of lengths of sentences as indications of

correspondence in the bitext space. Cognates are words that in different languages have the same meaning and similar spelling. For instance, in the main heading of example 1, the cognates recognized by the Levenshtein distance[2] with a threshold of 1 are: (*statute, statut*), (*international, internationale*), (*court, cour*), (*justice, justice*). Thus, the alignment procedure takes into consideration not only the tags used during the segmentation phase, but also the cognates detected during the sentence alignment phase.

The use of cognates, though appealing, has serious drawbacks. Firstly, pairs of historically close languages, as English and French, have numerous cognates, but between Korean and French, cognates can be found mainly among borrowings and proper names. Secondly, inflection blurs similarities. For instance, the English noun *bank* (a financial institution) and the Serbian noun *banka* in the nominative singular would be cognates. However, the same Serbian noun in the dative singular is *banci*, which differs in two characters out of five from the English noun, so they are not cognates any more. In Korean, the graphically undelimited suffixes appended to nouns, verbs and adjectives will have the same effect. Thirdly, Korean and French use different alphabets, with several possibilities of transcription between them, even in the case of borrowings and proper names, which are sometimes regarded as obvious and most reliable cognates. In fact, they can be successfully used only for some language pairs. In the *1984* example, the personal name *O'Brien* differs too much in both Bulgarian and Serbian texts from the English original to be considered a cognate. Even more severe problems arise with multi-word units, e.g. *the bridge in Novi Sad* is in Serbian *novosadski most*, where *novosadski* is a relational adjective derived from the name of the city of Novi Sad. Finally, false friends are another danger in the identification of the cognates, for instance the English adjectives *actual* and *eventual* and the French adjectives *actuel* 'present' and *éventuel* 'potential'.

Further improvements of the statistically based methods use the *n*-gram structures of constituent texts in a bitext or resort to particular lexical resources (Barbu, 2004).

Statistically based alignment methods give good results for pairs of similar languages and for texts belonging to certain limited domains. However, their linguistic interpretation is not clear, and neither is their potential for other language pairs.

3. The contribution of linguistic methods to mutilingual corpus processing

In section 2, we mentioned the durable interest of the NLP community in statistical models and, in particular, in the application of this approach to text alignment.

An alternative development line is the continuation of the long-term research that started as early as mid-twentieth century and which tends to develop formal models that describe linguistic knowledge about concrete language systems (Gross, 72). This is a much more complex task. In the present state of the art, it does not meet directly and with the same effectiveness the demands posed by commercial applications. Yet, it already enables much more precise and profound text analyses. A few immediate applications, such as spell checking (Silberztein, 1997) and named entity recognition (Poibeau, 2003), are available in this framework, and others are expected in the future, either as pure applications of this approach, or of the hybrid approach that combines statistically based and linguistic based techniques.

[2] The Levenshtein distance between two strings is the minimal number of insertions, deletions and substitutions required to change one of them into the other.

The main difference between the statistical and linguistic approaches is observed in different ways the knowledge about languages is represented. In statistical models, this knowledge is implicit and hidden. When a model does not yield the expected results, there exist possibilities to alter it, but they are not sure to improve its performances. The linguistic based approach removes exactly this type of deficiency of statistical models, since the knowledge about the language is explicitly represented in some formal framework or theoretical model. In this approach, we develop and improve frameworks or models that allow for precise, comprehensive descriptions of different language systems. As a consequence, the knowledge about the language is explicitly represented, and it is possible to correct potential errors.

In this section, we survey existing and potential contributions of linguistic methods to multilingual corpus processing.

3.1. Unitex as a monolingual tool

The linguistic based approach to natural language processing involves the use of high quality language resources such as electronic lexicons and grammars. The manual construction and maintenance of such resources requires trained linguists and resource-management tools. Few systems in the world include both corpus-processing and resource-management functionality. The open-source Unitex system (Paumier, 2002) is one of them. An engineering-oriented counterpart of Unitex, Outilex, is under construction (Blanc *et al.*, 2006).

As a corpus processor, Unitex segments text into sentences, annotates words, locates linguistic patterns in text and produces lemmatised concordances. A lemmatised concordance of a text is a concordance in which the sequences identified in the text may contain inflected forms even though the user's query contains lemmas.

As a resource management tool, Unitex supports the generation of inflected-form lexicons from readable lemma lexicons, and the graphical edition of syntactic grammars.

Therefore, it is complementary to common statistic-based tools and practices.

Unitex can presently process more than 10 languages, including Korean, but in a monolingual way, i.e. separately. The language resources usable with this system, and in particular those distributed with it, are monolingual lexicons and grammars. Unitex fully supports Unicode, which is a prerequisite for multilingual text processing. This development direction seems natural having in mind the number of languages for which the Unitex resources were developed.

However, Unitex in its present form can be applied to bitext production, under the assumption that language resources, and primarily electronic lexicons, are available for the language pairs involved. In addition, with extensions to its software, Unitex would become a bilingual concordancer, i.e. support the production of concordances of aligned bitexts. In what follows, we examine these directions to a multilingual Unitex.

3.2. Segmentation

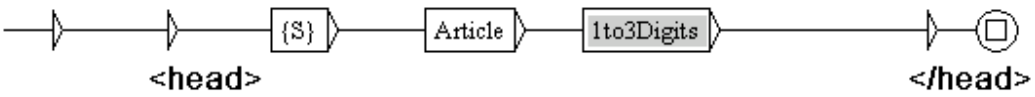
As stated, the first step in bitext production is the segmentation of both constituent texts. This is done in Unitex environment with sentence transducers[3] that detect the sentence borders with high precision. These transducers can output <*seg*> XML tags instead of *{S}* that is traditionally used by Unitex.

[3] A transducer is a resource that can be used to translate sequences into other sequences. Transducers can be graphically represented as graphs of the kind of that of Fig. 4, which inserts <*head*> and <*/head*> tags around sequences of the form *Article 1*.

We applied this method during the compilation of the English-Serbian aligned corpus of 2 million words (Gavrilidou *et al.*, 2005) in order to segment the Serbian and English texts. We used existing sentence transducers. Since such transducers take the form of graphically editable graphs, we could amend them. For instance, for the purpose of segmentation of legal texts, the sequence of abbreviations listed in the sentence graph for English for which the point is not to be interpreted as a full stop was enhanced by *Art., App., Arts., para., paras., Nos.*, etc.

Raw text obtained by conversion from a graphical format, such as pdf or ps, as a rule lacks the tags for logical layout. However, if the texts are uniformly presented, the similar graph approach can be used to recognize and accordingly tag other logical elements besides sentences, such as paragraphs, and sometimes headings, etc.

Paragraphs are almost always separated by some predefined space: an empty line, a hard line break followed by tabulator, etc. A very simple Perl-like expression can identify these sequences and insert paragraph tags <*p*>. The identification of headings is more complex: for instance, a heuristic rule such as 'a sequence of upper-case letters that is not followed by a full stop' can be used. In Example 1, articles can automatically be tagged by the Unitex graph of Figure 4.



COUR INTERNATIONALE DE JUSTICE <head>{S}Article 1</head> {S}La Cour internationale

Figure 4. A transducer for tagging article titles and an example of output

3.3. Detection of cognates

Unitex offers several opportunities of improvement of alignment results by the use of cognates. Recall that cognates are words with the same meaning and similar spellings.

First of all, it is possible to produce lists of candidate cognate pairs by comparing inflected-form lexicons of two languages. The comparison result will be a list of pairs that can be manually checked in order to obtain actual cognates. This procedure is compatible with transliteration rules able to neutralize differences between alphabets. For instance, in Serbian, strings *banka* and *банка* are entirely different when represented in Unicode, because they do not have a single character in common; nevertheless, they represent exactly the same word, written respectively in the Latin and in the Cyrillic alphabets. Thus, with transliteration rules, for instance, English *bank* and Serbian Cyrillic *банка* could be detected as candidate cognates. However, the transliteration of French words into the Korean alphabet is so complex that this approach is not likely to be very successful.

An approach to another cognate problem consists in establishing correspondences not only between isolated words, but between sets of inflected, or even derived, words. Our experiment with Gustave Flaubert's novel *Bouvard et Pécuchet* and its Serbian translation shows that two Serbian lemmas, *Buvar*.N and *Buvarov*.AdjPoss, correspond to the French name *Bouvard*, with as many as 20 different inflected forms (Vitas, Krstev, 2005):

Buvar; N: *Buvar+Buvara+Buvaru+Buvarom+Buvare*
Buvarov; AdjPoss: *Buvarov+Buvarova+Buvarovoj+Buvarovom+Buvarovog+*
Buvarovu+Buvarovih+Buvarovi+Buvarovim+Buvarove+Buvarovo

The extension of cognateness to lexical entries with attached information on inflectional and derivational variation, for instance *Bouvard*.N or [*Buvar*.N + *Buvarov*.AdjPoss], generates a large number of reliable cognate pairs that can be used during the analysis of bitext space.

Yet another approach to the identification of cognates is based on texts in which named entities are tagged and normalized by appropriate transducers. For instance, date transducers (Gross, 2002) can identify sequences that denote dates and normalize them into the Timex2 form (Ferro *et al.*, 2005), regardless of how they are represented in various languages and orthographic systems. For instance, the counterparts of French date *14 juillet 1789* are English *July 14th, 1789*, Serbian *14 juli 1789*, and Croatian *14 srpanj 1789*. Straightforward procedures of cognate identification fail to identify that *July*, *7*, and *srpanj* all correspond to French *juillet*, but appropriate transducers can do that easily through normalization.

3.4. Bitext concordancers

Aligned texts are generally considered as valuable resources, and tools that allow users to explore them through the production of concordances are most useful (Langlois, 1996). A bitext concordancer is a tool that produces concordances of a bitext. It searches one of the constituent texts for the user's query, and displays the occurrences found along with the corresponding segments in the other text. Several bitext concordancers have been developed recently: MultiConcord (Woolls, 1998), TransSearch (Macklovitch *et al.*, 2000), ParaConc (Barlow, 2002), TotalRecall (Wu *et al.*, 2003), Text-Searcher (Chujo *et al.*, 2005)... One of the best known, ParaConc, offers a variety of useful facilities: regular expression search, tag search, identification of potential translation equivalents, etc.

The operation of a bitext concordancer is simple. Texts furnished with tags for logical layout and segmented into sentences can be used as input to alignment systems, for instance XAlign (Bonhomme, Romary, 1995). The output of XAlign is internally represented in the form illustrated by the following example:

```
 <link targets="n5 n6" type="linking" id="l1" />
 <link targets="n1 x1" />
 <link targets="n2 x2" />
.....................................
 <link targets="l1 x5" />
 <link targets="n7 x6" />
```

This excerpt means that sentences 1 and 2 of text *n* (identified by *n1* and *n2*) are directly aligned with sentences 1 and 2 of text *x* (identified by *x1* and *x2*). Sentences 5 and 6 of text *n* (identified by *n5* and *n6*), however, form a block (identified by *l1*) which corresponds to sentence 5 of text *x*. With such a representation, a monolingual concordancer is easily extended into a system that displays a concordance of one of the constituent texts, and in parallel the corresponding segments in the other text, for instance in the form used in Example 1.

However, all existing bitext concordancers lack linguistic support. On the other hand, advanced concordancers with linguistic support such as Unitex allow for much more elaborate concordancing, but only on monolingual texts, since they do not presently process bitexts. Linguistic based concordancing has two main advantages.

First of all, word queries (also called lexical masks) can contain linguistic criteria: lemmas, parts of speech and other features that can be checked in the lexicons. Thus, in English, Unitex query `<rise>` also retrieves *rose* as a conjugated verb whose base form is *rise*. Similarly, `<N>`,

where `N` stands for *noun*, retrieves *roses* but not *raised*, which is only a verb. In French, the expression `<DET> <A> <N> <V:3>` applied to the 12th chapter of Jean Potocki's *Manuscrit trouvé a Saragosse* retrieves the following lines:

```
certain que l'air raréfié des hautes montagnes agit sur nos corps d'une manière
intéressait peu, et dès que la dernière femme était passée, il prenait le
cousines.</seg> <seg> Le vieux chef paraissait s'amuser de mon embarras.</seg
```

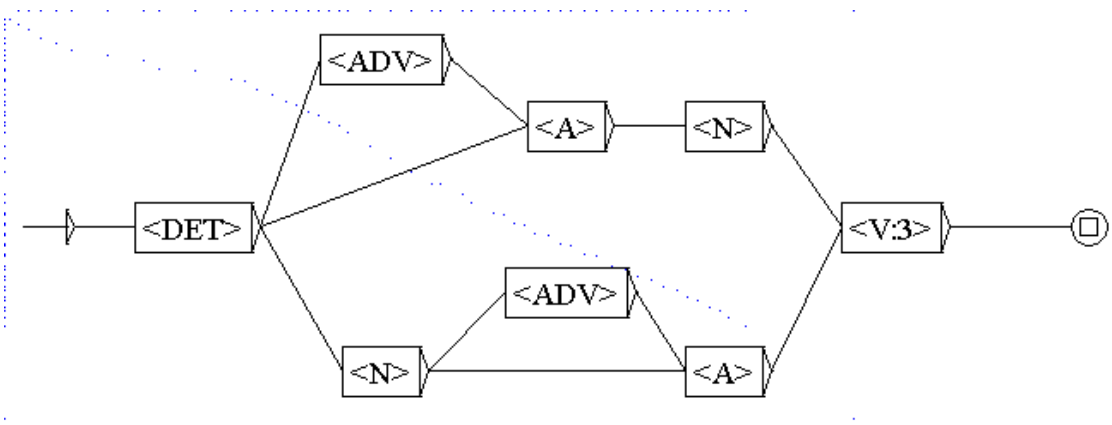


Figure 5. A graph for a syntactic pattern.

This feature is obviously useful for producing concordances of texts in an inflected language such as French and even more in an agglutinative language such as Korean. Unitex now performs morphological segmentation of Korean words (Berlocher *et al.*, 2006), which is required for the implementation of this feature of lexical masks.

Secondly, queries can be expressed in the form of graphs with the graph editor of Unitex. A graph can contain various parallel paths with variants of the linguistic pattern to be searched for, as exemplified by Figure 5 which allows for optional adverbs and for two positions of the adjective.

Advanced concordancing and bitext processing are by no means technincally incompatible. With an extension of the concordance-generation component, and with the French-Serbian bitext of *Le Manuscrit trouvé a Saragosse*, Unitex would display the following parallel French-Serbian concordance lines:

```
certain que l'air raréfié des hautes montagnes agit sur nos corps d'une manière
je da proredxeni vazduh na visokim planinama uticye na nasxe telo na poseban
-------------------------------------------------------------------------------
intéressait peu, et dès que la dernière femme était passée, il prenait le animao za
nxih, pa bi, cyim poslednxa zxena prodxe, odlazio u gostionicu
-------------------------------------------------------------------------------
cousines.</seg> <seg> Le vieux chef paraissait s'amuser de mon embarras.</seg
zgledalo je kao da se stari knez zabavlxa mojom neizvesnosxcxu.</seg></p>
```

3.5. Bitext concordancing and lexical resources

Bitext exploration interacts with other language resources, and the results of exploration can be used to further develop these resources. In order to investigate in this direction, a special module of the Workstation for Lexical Resources (WS4LR) has been developed (Krstev *et al.*, 2004b). WS4LR is another linguistic-based tool for corpus processing and language resource

management. It supports development and exploitation of wordnets (Miller *et al.*, 1990), bitexts, and electronic lexicons in the Dela format (Courtois, 1990). This tool does not align bitexts. It processes previously aligned bitexts in TMX format or in the XAlign output format.

We will illustrate the exploration of these lexical and textual resources by an example. You can search a Croatian-Serbian bitext of Jules Verne's *Le Tour du monde en quatre-vingts jours* for the occurrences of the Serbian verb *pokazati* 'show'. In the first step, the user can expand his query by including all the verbs of the same synset in the Serbian wordnet (Figure 6, upper left window). He can edit this list and delete all the lemmas that are not appropriate for his search; for instance, he can retain only *pokazati* and its imperfective counterpart *pokazivati*. In the next step, all the chosen lemmas can be automatically inflected (Figure 6, upper right window). In the final step, the user can initiate the search with all the generated words. Here, two options are available: the search can be extended to the whole bitext or limited to one of the constituent texts. The former option (Figure 6) is useful if the texts are in very close languages, which is the case for Croatian and Serbian, or in the same language, as in the case of two independent translations of the same original text. If the user chooses the latter option, he can request the identification of potential equivalents. This option requires the existence of wordnets for both languages and an interlingual index to synchronize them (Vossen, 1988).

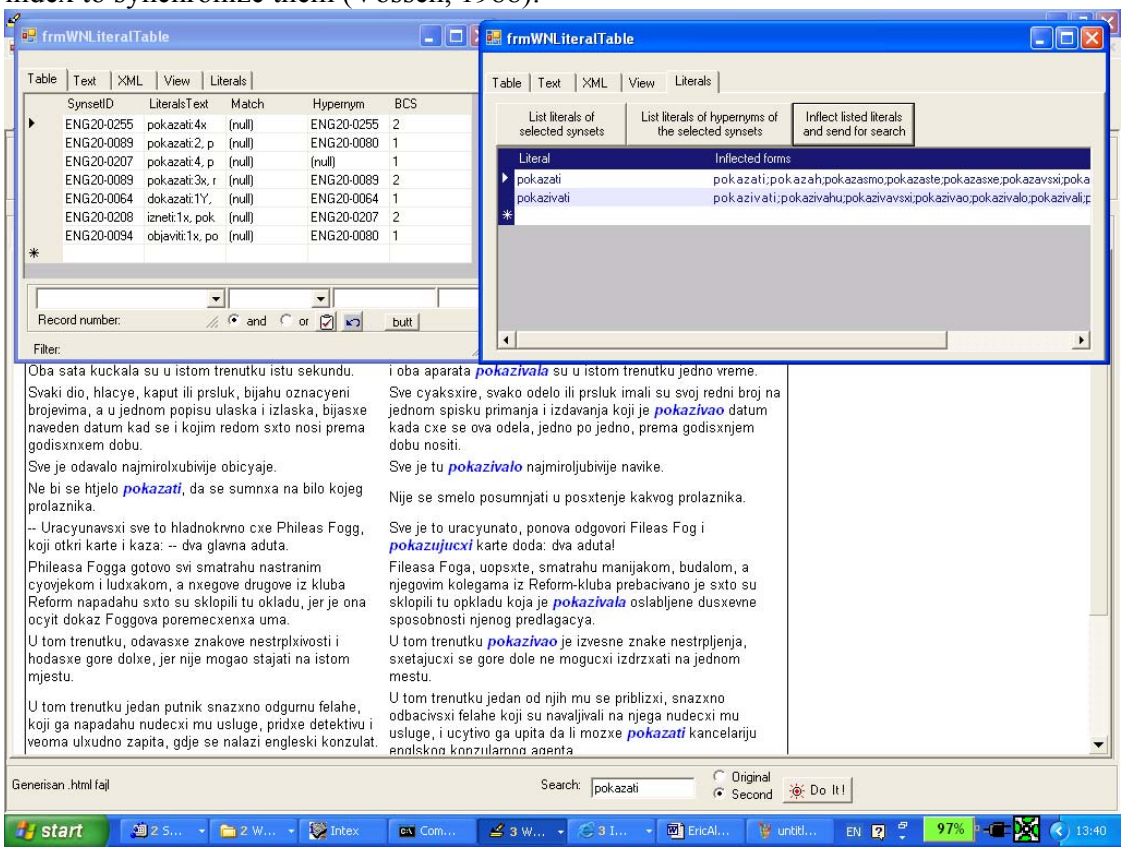


Figure 6. The module for the aligned text in WS4LR

The concordance obtained through this procedure is displayed in Figure 6 (lower window). It shows, for instance, that both verbs *pokazati* and *pokazivati* are considerably more frequent in Serbian than in Croatian. For some meanings, the verb *odavati,* used in the Croatian translation,

can also be used, which signals that it can be considered for inclusion in the Serbian wordnet, and this task can be accomplished by using the same tool.

Conclusion

The preparation and exploitation of bitexts are complex problems. For the French-Korean pair of languages, this complexity is illustrated by three particular aspects: the use of two distinct alphabets, the typological difference between an inflectional and an agglutinative language, and the small proportion of cognates in the two vocabularies.

We surveyed the main points of the state of the art in the preparation of bitexts, which essentially applies quantitative approaches. Linguistic approaches could improve both the preparation and the exploitation of bitexts. In particular, Unitex's linguistic-based, advanced methods of production of concordances are technically compatible with the mode of operation of existing bitext concordancers. Integrating a bitext-concordance functionality into Unitex would produce a bitext-exploration tool of an unprecedented quality. Applications include human translation, language teaching, investigation of literary text, and natural language processing, including enhancement of lexical resources.

References

Barbu, A.-M. (2004): A Positional Linguistics-Based System for Word Alignment. *TSD. LNAI* 3206, Springer, pp. 23-30.

Barlow, M. (2002): ParaConc. Concordance software for multilingual parallel corpora. In *Proceedings of Language Resources for Translation Work and Research*, pp. 20-24.

Berlocher, I., Huh, H.G., Laporte, E., Nam, J.S. (2006): Morphological annotation of Korean with Directly Maintainable Resources, Poster. In *Proceedings of LREC*, Genoa.

Blanc, O., Constant, M., Laporte, É. (2006): Outilex, plate-forme logicielle de traitement de textes écrits. In *Verbum ex machina. Proceedings of TALN, Cahiers du CENTAL* 2(1), Presses universitaires de Louvain, pp. 83-92.

Bonhomme, P., Romary, L. (1995): The Lingua Parallel Concordancing Project. Managing Multilingual Texts for Educational Purpose. In *Proceedings of Language Engineering*, Montpellier.

Brown, P., Della Pietra, S, Mercer, R. (1993): The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19(2), pp. 263-311.

Chujo, K., Nishigaki, Ch., Utiyama, M. (2005): A Japanese-English Parallel Corpus and CALL: A Powerful Tool for Vocabulary Learning. In *Selected Proceedings of FLEAT*, Brigham Young University, pp. 16-19.

Courtois, B. (1990): Un système de dictionnaires électroniques pour les mots simples du français, *Langue Française* 87, Paris: Larousse, pp. 11-22.

Danielsson, P., Ridings, D. (1997): *Practical Presentation of a "Vanilla" Aligner.* Presentation held at the TELRI Workshop in Alignment and Exploitation of Texts, Ljubljana; Research report, Department of Swedish, Göteborg University, GU-ISS-97-2, Språkdata (http://nl.ijs.si/telri/Vanilla/).

Erjavec, T., Lawson, A., Romary, L., Eds (1998): *East meets West. A Compendium of Multilingual Resources*, Mannheim: TELRI (CD-ROM).

Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G. (2005): *Tides. 2005 Standard for the Annotation of Temporal Expressions*, MITRE.

Gale, W., Church, K. (1993): A Program for Aligning Sentences in Bilingual Corpora. *Computational linguistics* 19(1), pp. 75-102.

Gavrilidou, M., Labropoulou, P., Monachini, M., Piperidis, S., Soria, C. (2005): Building Multilingual Terminological Resources. In *Proceedings of the Workshop on Language and Speech Infrastructure for Information Access in the Balkan Countries*, Borovets, Bulgaria, pp. 15-22.

Gross, M. (1972): *Mathematical Models in Linguistics*, Englewood Cliffs, New Jersey : Prentice-Hall.

Gross, M. (2002): Les déterminants numéraux, un exemple: les dates horaires. *Langages* 145, Paris: Larousse.

Harris, B. (1988): Bi-text. A New Concept in Translation Theory. *Language Monthly* 54, pp. 8-10.

Krstev, C., Vitas, D., Erjavec, T. (2004a): Morpho-Syntactic Descriptions in MULTEXT-East. The Case of Serbian. *Informatica* 28, pp. 431-436, Ljubljana: The Slovene Society Informatika.

Krstev, C., Vitas, D., Stanković, R., Obradović, I., Pavlović-Lažetić, G. (2004b): Combining Heterogeneous Lexical Resources. In *Proceedings of LREC*, vol. 4, pp. 1103-1106, Lisbon: ARTIPOL.

Langlois, L. (1996): *Bitexte, bi-concordance et collocation*. PhD thesis, University of Ottawa.

Macklovitch, E., Simard, M., Langlais, Ph. (2000): TransSearch. A Free Translation Memory on the World Wide Web. In *Proceedings of LREC*, vol. 3, pp. 1201-1208.

Melamed, I.D. (1996): A geometric approach to mapping bitext correspondence. In *Proceedings of EMNLP*, Philadelphia, pp. 1-12.

Melamed, I. D. (2001): *Empirical Methods for Exploiting parallel texts*. The MIT Press.

Miller, G., Beckwith, R., Fellbaum, Ch., Gross, D., Miller, K. (1990): Introduction to WordNet. An on-line lexical database. *International Journal of Lexicography* 3(4), pp. 235-244.

Palmer, D., Hearst, M. (1994): Adaptive sentence boundary disambiguation. In *Proceedings of ANLP*, pp. 78-83.

Paumier, S. (2002): *Manuel d'utilisation du logiciel Unitex.* IGM, Université de Marne-la-Vallée. http://www-igm.univ-mlv.fr/~unitex/manuelunitex.pdf

Poibeau, Th. (2003): The Multilingual Named Entity Recognition Framework. In *Proceedings of EACL*, pp. 155-158.

Reynar, J.C., Ratnaparkhi, A. (1997): A maximum entropy approach to identifying sentence boundaries. In *Proceedings of ANLP*, pp. 16-19.

Silberztein, M. (1997): The Lexical Analysis of Natural Languages. In *Finite-State Language Processing*, E. Roche and Y. Schabes (eds). MIT Press, pp. 175-203.

Simard, M., Foster, G., Isabelle, P. (1992): Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of TMI*, University of Montréal, pp. 67-81.

Véronis, J., Ed. (2000): *Parallel Text processing. Alignment and Use of Translation Corpora*. Dordrecht: Kluwer.

Vitas, D., Krstev, C. (2005): Structural derivation and meaning extraction. A comparative study on French-Serbo-Croatian parallel texts. In *Meaningful Texts. The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, G. Barnbrook, P. Danielsson, M. Mahlberg (Eds.), Birmingham: Univ. of Birmingham Press, pp. 166-178.

Vossen, P., Ed. (1988): *EuroWordNet. A multilingual database with lexical semantic network,* Dordrecht: Kluwer.

Woolls, D. (1998): Multilingual Parallel Concordancing for Pedagogical Use. In *Proceedings of Teaching and Language Corpora*, Oxford: Keble College, pp. 222-227.

Wu D. (1994): Aligning a parallel English-Chinese statistically with lexical criteria. In *Proceedings of ACL*, pp. 80-87.

Wu J.Ch., Yeh K.C., Chuang Th.C., Shei W.-Ch., Chang J.S. (2003): TotalRecall. A Bilingual Concordance for Computer Assisted Translation and Language Learning, In *Companion Volume to the Proceedings of ACL*, pp. 201-204.