AN ALIGNED ENGLISH-SERBIAN CORPUS

Cvetana Krstev

University of Belgrade, Faculty of Philology

Duško Vitas

University of Belgrade, Faculty of Mathematics

Abstract

In this paper we present a special kind of multilingual corpora know as aligned corpora that are the focus of different research interests, especially in the domain of natural language processing. We present various steps in building such a corpus: a selection of texts, their pre-processing, the alignment process, and checking the results. Possible exploitations of such a corpus are presented, especially those that use corpora processing tools supported by language dependent modules. We illustrate these basic steps involved in the preparation and exploitation of an aligned corpus by an aligned English/Serbian corpus that consists of the complete works of Jane Austen. Finally, we give a brief overview of the free software used in all these steps.

Introduction

A multilingual corpus is a special kind of a corpus in which usually more then one language is involved. When talking about multilingual corpora, one usually thinks about parallel or aligned corpora that consist of several semantically equivalent texts that are aligned to a recognizable level that can be a paragraph, a sentence, a phrase or a word. Usually it is built around one or more original texts and their translations in one or more languages. Corpora of this kind are the focus of various different types of research, especially in the domain of natural language processing (NLP) (e.g. building translation memories, statistical machine translation, Wordnet-based sense disambiguation, annotation transfer, etc.) (Tufiş, 2006). The envisaged use of an aligned corpus determines the selection of texts, their size, the level of alignment and the method used for alignment.

We should mention here that aligned corpora are not the only type of multilingual corpora. Sometimes a comparable corpus is used that does not necessarily consist of semantically equivalent texts but rather texts that belong to the same domain, same period of time, etc. Such a corpus is easier to prepare and although it has less potential it can still prove useful in some NLP applications based on statistical methods. They are also used for very close languages for which translated texts are difficult to find, e.g. Serbian, Croatian and Bosnian, Czech and Slovak, etc. On the other hand, an aligned corpus need not be a multilingual corpus; namely, it can consist of two or more independent translations of a same text in one language. We will give some examples of such corpora in the next section.

In this paper we will talk mainly about preparation and use of a multilingual corpus aligned to the sentence or sub-sentence level: alignment to the paragraph level cannot be very useful for any application while alignment to the word level is used for rather specific applications.

Some Remarkable Examples of Aligned Corpora

One early attempt to develop the multilingual aligned text was the production of an aligned Plato's *Republic* as part of the f TELRI project (Erjavec et al. 1998). In this aligned version 16 languages were involved: Bulgarian, Croatian, Czech, English, Finnish, French, German, Hungarian, Latvian, Lithuanian, Polish, Romanian, Russian, Serbian, Slovak, and Slovenian. This multilingual text actually consisted of 16 bitexts – pairs of aligned texts – because all translations were aligned with the English version. The alignment at the sentence level was performed automatically and the obtained results were manually checked. This experience showed that the alignment task was by no means an easy one, especially for this text with a complex logical layout. Namely, the differences of translations into some of the languages were significant, and for some languages it was difficult to establish which version was used as the basis of translation, so the alignment process did not yield convincing results.

The multilingual aligned text was later produced as part of the project MULTEXT-East for Orwell's *1984*, which proved to be a more suitable text and was since used in many applications (Erjavec 2004). In this aligned version, 12 languages were involved: Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Resian, Romanian, Russian, Serbian, and Slovene. This multilingual text also consisted of 12 bitexts with all the translations aligned with the English version. The alignment at the sentence level was performed automatically and the obtained results were manually checked. The following examples illustrate the Serbian-English bitext with the first three sentences of the novel.

> <Oshs.1.2.2.1>Bio je vedar i hladan aprilski dan; na časovnicima je izbijalo trinaest.
> <Oen.1.1.1.1>It was a bright cold day in April, and the clocks were striking thirteen.
>
> <Oshs.1.2.2.2> <B>Vinston Smit</B>, brade zabijene u nedra da izbegne ljuti vetar, hitro zamače u staklenu kapiju stambene zgrade <I>Pobeda</I>, no nedovoljno hitro da bi sprečio jednu spiralu oštre prašine da uđe zajedno s njim.
> <Oen.1.1.1.2> <B>Winston Smith</B>, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of <B>Victory Mansions</B>, though not quickly enough to prevent a swirl of gritty dust from entering along with him.
>
> <Oshs.1.2.3.1>Hodnik je zaudarao na kuvani kupus i stare otirače.
> <Oen.1.1.2.1>The hallway smelt of boiled cabbage and old rag mats.

The corpus was morphosyntactically annotated for most of the languages which represented its additional value. This means that a lemma was assigned to each simple word in the text, together with a string of morphosyntactic codes representing a word's part of speech and other grammatical categories pertaining to it. For all languages the morphosyntactical annotation was performed automatically, while the disambiguation and double checking were done manually. All languages used a coherent set of morphosyntactic features and their values which allowed the wide-spread use of this corpus in many NLP applications. The following example illustrates the first sentence of the Serbian translation morphosyntactically annotated.

```
<p id="Oshs.1.2.2" >
<s id="Oshs.1.2.2.1" >
<w lemma="biti" ana="Vmps-sman-n---p">Bio</w>
<w lemma="jesam" ana="Va-p3s-an-y---p">je</w>
```

```
<w lemma="vedar" ana="Afpms1n">vedar</w>
<w lemma="i" ana="C-s">i</w>
<w lemma="hladan" ana="Afpms1n">hladan</w>
<w lemma="aprilski" ana="Aopmp1">aprilski</w>
<w lemma="dan" ana="Ncmsn--n">dan</w>
<w lemma="na" ana="Sps-">na</w>
<w lemma="časovnik" ana="Ncmsa--n">časovnicima</w>
<w lemma="jesam" ana="Va-p3s-an-y---p">je</w>
<w lemma="izbijati" ana="Vmps-snan-n---e">izbijalo</w>
<w lemma="trinaest" ana="Mc---l">trinaest</w>
```

Another corpus is being built around Verne's novel "Around the World in 80 Days" as part of an independent project undertaken at the Faculty of Mathematics and the Faculty of Philology, University of Belgrade (Vitas et al. 2008). In this aligned version, 16 languages are already included: Bulgarian, Croatian, English, French, German, Greek, Hungarian, Italian, Macedonian, Polish, Portuguese, Romanian, Russian, Serbian, Slovene, Spanish, while other are in preparation, among them Chinese. The alignment was automatically established at the sub-sentence level for each language and the original version and the obtained results were manually checked and corrected with the aim of obtaining a one-to-one correspondence between all segments. Thus, 240 bitexts were automatically obtained for each language pair. The structure of this corpus is illustrated by one sentence of the novel in some of the languages involved.

<seg lang="fr"><s id="Verne80days.n569">Vous savez que cette formalité du visa est inutile, et que nous n'exigeons plus la présentation du passeport?</s></seg>

<seg lang="sr"><s id="Verne80days.n569">**Vi znate da je ova formalnost viziranja izlišna i da se više ne traži pokazivanje isprava?**</s> </seg>

<seg lang="bg"><s id="Verne80days.n569"> Знаете ли, че тази формалност с паспортите е безполезна и че ние вече не изискваме да представяте паспортите си?</s></seg>

<seg lang="en"><s id=" Verne80days.n569">**You know that a visa is useless, and that no passport is required?**</s></seg>

<seg lang="gr"><s id="Verne80days.n569">Ξέρετε ότι αυτή η τυπική διαδικασία της βίζας δεν είναι αναγκαία και δεν απαιτείται πλέον η εμφάνιση του διαβατηρίου;</s></seg>

<seg lang="sl" ><s id="Verne80days.n569">Ali vam je znano, da je ta formalnost vidiranja nepotrebna in da ne zahtevamo več predložitve potnega lista ?</s></seg>

<seg lang="ro"><s id=" Verne80days.n569">ţiţi cã formalitatea vizei e inutilă şi că noi nu mai cerem prezentarea paşaportului.</s>

For all the languages the morphosyntactical annotation was performed automatically using electronic dictionaries of a corresponding language, the disambiguation was done semi-automatically, while the final checking was done manually. The morphosyntactic annotation is similar to the one used for the *1984* corpus, although morphosyntactic codes are different, as can be seen from the Serbian example.

```
{Vi,vi.PRO+PrsMB:py1r}
{znate,znati.V2+Imperf+Tr+Iref+Ref:Pyp}
{da,da.CONJ}
{je,jesam.V575+Imperf+It+Iref+Aux:Pzsi}
{ova,ovaj.PRO+ProA+Demon:fs1g}
{formalnost,formalnost.N:fs1q}
{viziranja,viziranje.N+VN:ns2q}
{izlišna,izlišan.A:aefs1g}
{i,i.CONJ}
{da,da.CONJ}
{se,se.PAR}
{više,više.ADV}
{ne,ne.PAR+Neg}
{traži,tražiti.V51+Imperf+Tr+Iref:Pzs}
{pokazivanje,pokazivanje.N+VN:ns1q}
{isprava,isprava.N:fp2q}?
```

This corpus proved to be a very useful resource and was used in many applications, especially, because of its specific content suitable for named entity processing. The same research group at the Faculty of Mathematics and the Faculty of Philology has also compiled a larger multilingual French/Serbian corpus that contains mainly French classic novels translated into Serbian. In this corpus, some Serbian/Serbian bi-texts were produced in the cases where more than one translation existed for the same French text (Voltaire's *Candide* and Jan Potocki's *The Manuscript Found in Saragossa*).

A very different multilingual corpus is the so-called JRC-Acquis, publicly released in May 2006 (Steinberger et al., 2006). It consists of the total body of the European Union (EU) law applicable in the EU member states. This collection of legislative texts changes continuously and currently comprises texts written between the 1950s and 2008 in all the languages of the EU member states (22 languages). It consists of more then 460,000 documents with over 1 billion words in total, almost 50 million words per language. This corpus was less minutely prepared then the aforementioned multilingual corpora; however, due to its remarkable size and the number of languages involved it was successfully used in many applications, especially those based on statistical methods, e.g. statistical machine translation. Serbian is not represented in this corpus because *The Acquis Communautaire* has not yet been translated to Serbian.

The Acquisition of the Jane Austen Corpus

Since our French/Serbian literary corpus reached a considerable size, our next goal is to prepare a similar English/Serbian literary corpus. The first issue that has to be considered before undertaking this task is the selection of texts for the corpus. The decision depends on several issues. First of all, the availability of texts in both languages in the digital form was established. However, it is not enough that texts exist in the digital form, because one has to be sure how reliable the source of the text is, e.g. whether it is clear on what edition the digital text is based, was it corrected after scanning or re-typing, whether the translation is of good quality, etc. The copyright questions have to be considered too, because some authors and publishing houses would not allow the use of their texts, not even for research purposes. Multilingual literary corpora are in general much smaller in size than, for instance, corpora of legislative or newspaper texts. Because of that, and because of the envisaged use, this kind of corpora is prepared with much care which is usually time-consuming and labour-demanding.

After pondering upon these questions, we have decided to start our project with the preparation of a Jane Austen corpus that would consist of her six major novels. We were able to obtain the English texts in the digital form easily: we have downloaded them from *The Republic of Pemberley* web site.[1] The copyright questions are not an issue for these novels. An additional value of this choice lies in

[1] www.pemberley.com/ (Your haven in a world programmed to misunderstand obsession with things Austen)

the fact that because of the popularity of these novels translations in many languages exist that can be added to this corpus in the future. We did not expect to find, on the web or elsewhere, Serbian translations of Jane Austen's novels or other English classics in the digital form. All novels where thus re-typed by students at the Department of Library and Information Sciences as their practical work as part of the introductory course "Information Literacy". Typing errors were spotted and subsequently corrected using Serbian electronic morphological dictionaries (see next sections).

The Pre-processing of the Jane Austen Corpus

In order to apply some program for automatic text alignment, additional pre-processing of texts is necessary. Namely, texts have to be transformed into XML documents with some rudimentary tags marking chapters (or other divisions) (`<div>`), headings (`<head>`), paragraphs (`<p>`) and sentences (or other recognizable text segments) (`<seg>`). These XML tags reflect the basic tree-like logical composition of most text types. The English version obtained from the Pemberley site was in the HTML format with paragraph tags already inserted; however, other HTML tags responsible for the visualisation by web browsers had to be removed. In the Serbian translations, paragraph tags were inserted by the students who re-typed them.

Sentence tags were inserted both in English and Serbian texts automatically using the **Unitex** corpus processing software and its finite-state transducers (graphs).[2] Although finite-state transducers emerge form the mathematical formal language theory, their use is very natural and can be successfully mastered by users without a strong mathematical background. A user can easily express straightforward rules for detecting sentence boundaries in a linear paragraph representation. One such rule is

[2] Unitex is a corpus processing system, based on automata-oriented technology. With this tool, one can handle electronic resources such as electronic dictionaries and grammars and apply them. One can work at the levels of morphology, the lexicon and syntax. (http://www-igm.univ-mlv.fr/~unitex/)

that a full-stop followed by an upper-case letter marks the end of a sentence. However, since the same character '.' is used for many purposes, exceptions need to be defined, for instance, an upper-case letter following an occurrence like "Mrs." will not mark the sentence boundary. XML tags for sentence boundaries `<seg>` and `</seg>` were automatically inserted in the processed text. The advantage of this software is that a user can easily formulate necessary transducers for each language involved, because they are obviously language dependent. The following example shows the first paragraph of the novel *Northanger Abbey* in the Serbian translation after the automatic addition of sentence tags.[3]

```
<div n="1">
<head>GLAVA PRVA</head>
<p><seg>Niko ko je imao prilike da vidi Katarinu Morlend u njenom detinjstvu ne
bi pomislio da se rodila da bude junakinja.</seg><seg> Njene životne prilike, narav
njenog oca i majke, njen vlastiti karakter i talenat, sve to skupa bilo je podjednako
protiv nje.</seg><seg> Njen otac bio je sveštenik, ni zanemaren ni siromašan,
veoma dostojanstven čovek iako mu je ime bilo Ričard - i nikada nije bio
lep.</seg><seg> Bio je u priličnoj meri nezavisan, povrh toga imao je dvostruki
dobar prihod - i nije uopšte bio sklon tome da zatvara u kuću svoje
kćeri.</seg><seg> Njena majka bila je jednostavna žena, praktična i neposredna,
uvek dobro raspoložena i, što je najznačajnije, dobre fizičke konstitucije.
</seg><seg>
```

The Alignment of the Jane Austen Corpus

The XAlign system[4] was used for the alignment process. The strategy used by this program is, basically, to attempt to align in the first step the tree structure of texts within bitexts (encoded by XML-tags), and then, in the second, to align segments as the smallest logical units. The important feature of this system is that it does not alter the two input files that contain texts in two different languages: it produces another XML file that records the links between the segments.

Establishing one-to-one relations on the segment level was set as the goal of the alignment process. This type of text alignment of bitexts required an intensive

[3] Džejn Osten, Nortengerska opatija, Narodna knjiga, Beograd, 1976, Translation: Smiljana and Nikola Kršić

[4] http://led.loria.fr/download/source/Xalign.zip

manual control of the output of the XAlign system, using the attached concordancer[5] that shows the aligned pairs. In this way, the missing segments or the inconsistencies between the source text and its translations were also identified. Strictly speaking, one-to-one relations between all segments are not necessary, but they are desirable if new languages are to be added in the future. All errors were corrected and new segment tags inserted in the original input files. This re-iterated process lead to 100% correct one-to-one alignment. The following example shows an instance when a Serbian sentence was split in two segments in order to obtain a one-to-one correspondence.

<seg> She had three sons before Catherine was born;**</seg><seg>** and instead of dying in bringing the latter into the world, as anybody might expect, she still lived on - lived to have six children more - to see them growing up around her, and to enjoy excellent health herself.</seg>	<seg>Pre nego što se Katarina rodila imala je već tri sina,**</seg> <seg>** i umesto da umre kada je nju donela na svet, kao što bi se moglo očekivati, ona je živela i dalje - živela je da dobije još šestoro dece, da ih gleda oko sebe kako rastu a da sama sačuva izvanredno zdravlje.</seg>

Following this procedure, the novel *Northanger Abbey* was segmented in 5003 segments in both languages. A locally made piece of software, developed at the Faculty of Mathematics, named ACIDE (Aligned Corpora Integrated Development Environment) wraps the XAlign aligner and concordancer in a user-friendly environment; in addition it allows the production of several useful output formats (Utvić et al., 2008). One of them is the HTML format for easy web browsing and the other is the TMX format used for translation memories.[6]

```
<tu>
        <prop type="Domain">Jane Austen: Northanger Abbey</prop>
        <tuv xml:lang="EN" creationid="n7 "creationdate="20091203">
        <seg>and instead of dying in bringing the latter into the world, as anybody
might expect, she still lived on - lived to have six children more - to see them
growing up around her, and to enjoy excellent health herself. </seg>
        </tuv>
        <tuv xml:lang="SR" creationid="n7 " creationdate="20091203">
```

[5] http://led.loria.fr/download/source/concordancier.zip

[6] TMX (*Translation Memory eXchange*) is a XML-based standard for the exchange of translation memories. It was developed in 1998 by OSCAR (*Open Standards for Container/Content Allowing Re-use*), a subgroup of LISA (*Localization Industray Standard Organization*).

```
        <seg>i umesto da umre kada je nju donela na svet, kao što bi se moglo
očekivati, ona je živela i dalje - živela je da dobije još šestoro dece, da ih gleda oko
sebe kako rastu a da sama sačuva izvanredno zdravlje. </seg>
        </tuv>
</tu>
```

The Exploitation of the Jane Austen Corpus

Once an aligned corpus is prepared, it can be exploited in many different ways. The possible uses fall, basically, into two categories. The corpus can be exploited by some program for various purposes: search for translation equivalents, machine learning for statistical machine translation, terminology extraction, etc. On the other hand, the corpus can be exploited by a human user as part of some linguistic or literary research. We will talk in this section only about the second type of exploitation.

One obvious use of an aligned corpus is its linear reading “under the microscope” in order to find the strong or week points in the translated text, which can be useful, for instance, in the classroom. The week points can be found probably in every translation; we have spotted some and among them the following example (segment 1785 in *Northanger Abbey*).

but perhaps it was because they were habituated to the finer performances of the London stage, which she knew, on Isabella's authority, rendered everything else of the kind "quite horrid."	ali možda je to bilo zbog toga što su bili naviknuti na bolje predstave londonske scene, na kojoj su se, kako je znala iz Izabelinog uveravanja, davali svakakvi "užasni" komadi.

However, the whole effort invested in producing an aligned text would be excessive if the text is eventually read in a linear order. The already mentioned corpus processing system, Unitex allows a more sophisticated exploitation by enabling the production of multilingual concordances. These concordances can be produced by querying with keywords in one language or another with the aim of obtaining a literal match between the keywords and the corresponding text. However, Unitex offers much more. Namely, a text processed by this system is subjected to pre-processing by electronic morphological dictionaries of the language involved. Electronic dictionaries provide useful information about each word from the analyzed text: its lemma, grammatical information concerning its usage, syntactic and semantic information. This added information largely enhances the possibilities of querying the aligned text.

Electronic dictionaries using the same format, known as the LADL format,[7] were developed for many languages, English (Monceaux 1995; Chrobot et al. 1999) and Serbian among others (Krstev 2008). The pre-processing of texts by electronic dictionaries produces dictionaries of texts; small parts of these dictionaries for the English and Serbian version of *Northanger Abbey* are given in the following example. The upper row shows small excerpts of the produced dictionaries of simple words while the lower row gives some examples of the produced dictionaries of compounds or multi-word units.

a,.DET+Dind:s a,.N:s abatement,.N:s abbey,.N+Conc:s abbeys,abbey.N+Conc:p abhor,.V:W:P1s:P2s:P1p:P2p:P3p abhorrent,.A abilities,ability.N:p able,.A abode,.N:s abode,abide.V:K:I1s:I2s:I3s…	barem,.ADV barem,.PAR barometra,barometar.N:mw2q barona,baron.N+Hum:ms4v baronima,baron.N+Hum:mp3v baš,.PAR bašti,bašta.N:fp2q Bat,.N+NProp+Top+Gr+UK:ms4q bede,beda.N+Ek:fs2q bede,bediti.V+Imperf+Tr+Iref:Pzp bedeme,bedem.N:mp4q
card-room,.N+XN+z1:s cast down,.A+z1 Christian name,.N+XN+z1:s circle of friends,.N+NPN+z1:s circulating library,.N+XN+z1:s cold meat,.N+XN+z1:s cold sweat,.N+XN+z1:s come back,.N+XN+z1:s common sense,.N+XN+z1:s country walk,.N+XN+z1:s	dnevni boravak,.N+Comp:ms4q dnevnoj sobi,dnevna soba.N:fs3q dve stotine sedamdeset šest,.NUM godišnje doba,.N+Comp:ns5q gore-dole,.ADV+C+Ek kod kuće,.ADV+C kućnom haljinom,kućna haljina.N:fs6q laku noć,.INT+Pozdrav+C lovačkim psima,lovački pas.N:mp3v navrat-nanos,.ADV+C

The XAlign format of aligned texts has been recently incorporated into the Unitex system which enables the use of the electronic dictionaries provided with Unitex for querying the aligned texts. We will illustrate this with some small examples. Our first query will be very simple: it consists of one English adjective

[7] Named after the laboratory in which the model was developed under the guidance of Maurice Gross - *Laboratoire d'Automatique Documentaire et Linguistique*.

dull, and the search is performed throughout the English text, aligning the retrieved segments with the corresponding Serbian segments. Four occurrences were retrieved and the aligned segments show that this adjective was translated into Serbian by using three different adjectives: *dosadan*, *glup*, *utučen*, while in the last segment the noun *gnjavaža* was used instead. The actual software window with the obtained aligned concordances is shown in Figure 1 at the end of this paper.

I tell Mr. Allen, when he talks of being sick of it, that I am sure he should not complain, for it is so very agreeable a place, that it is much better to be here than at home at this dull time of year.	Kad gospodin Alen počne da govori o tome da je već do guše sit svega, ja mu govorim da ne mogu da se potužim na ovo prijatno mesto i da je mnogo bolje biti ove nego kod kuće u ovo dosadno godišnje doba.
The rest of the evening she found very dull;	Ostatak večeri provela je na najgluplji mogući način.
"Do not be so dull, my dearest creature," she whispered.	- Nemojte biti tako utučeni, najdraža - šapnula joj je.
and yet I often think it odd that it should be so dull, for a great deal of it must be invention.	A ipak često mislim da je čudnovato što je sve to takva gnjavaža, kad je većina svih tih stvari sigurno izmišljena.

This example is, however, very simple because it does not, actually, use the produced dictionaries of texts. A slightly more sophisticated query <acknowledge> does not search for all occurrences of the string *acknowledge* but rather for all occurrences of all the inflected forms of the verb *acknowledge*, because that is what angular brackets are used for. The search for the text that fits this pattern retrieved 16 occurrences of this verb in three different forms; three examples are given in the following table.

Here Catherine secretly acknowledged the power of love;	Tu je Katarina tajno odala priznanje snazi ljubavi,
"I think Mr. Morland would acknowledge a difference.	- Ja mislim da gospodin Morlend dobro zna u čemu je razlika.
With a look of much respect, he immediately rose, and being introduced to her by her conscious daughter as "Mr. Henry Tilney," with the embarrassment of real sensibility began to apologize for his appearance there, acknowledging that after what had passed he had little right to expect a welcome at Fullerton, …	Ovaj je odmah ustao sa izrazom velikog poštovanja i pošto ga je njena prisebna ćerka predstavila kao "gospodina Henrija Tilnija", počeo se sa zbunjenošću koja je odavala osećajnog čoveka izvinjavati zbog svog dolaska, priznajući da posle svega što se dogodilo ima veoma malo prava da očekuje da je dobrodošao u Fulertonu…

Obviously, the use of such lemma patterns is even more useful when the search is performed throughout the Serbian text because of the high inflection rate in Serbian. For instance, a search with the pattern <dosadan> that corresponds to the adjective *dosadan,* that was used as one of translations for the English adjective *dull,* retrieved besides *dosadan* also forms *dosadne*, *dosadnog*, *dosadno*, *dosadni* – a total of 12 occurrences. By looking in the opposite direction we can see that one and the same Serbian adjective was used as a direct translation of the English adjectives: *dull*, *insipid*, *teasing*, *tired*, *impertinent*, *odious* and *tedious*, and that it also indirectly corresponds to the noun *tediousness* and a verb *vex*.

Search patterns can be even more complex since all grammatical, syntactic and semantic codes and marks from electronic dictionaries of texts can be used in them. For instance, the pattern <A+Pos+NProp+Hum> retrieves in the Serbian text all possessive (+Pos) adjectives (A) derived from proper names (+NProp) used for humans (+Hum). This pattern retrieves various occurrences, such as *Katarininom*, *Tejlorovoj*, *Izabeline*, *Sofijino*, *Henrijev*, *Džonov* and many others, all used as translations of possessive forms like *Cathrine's life* or prepositional phrases like *hand of Isabella*.

This patterns can be combined in order to form a more complex query; for instance, if we are looking for "third conditional" sentences, we can set the pattern (should+would) <have.V:W> <V:K> that retrieves all three word sequences in the English text, where the first word is either *should* or *would*, followed by the infinitive (W) of the verb *to have*, followed by the past participle (K) of any verb (V). Three examples of the obtained results are shown below:

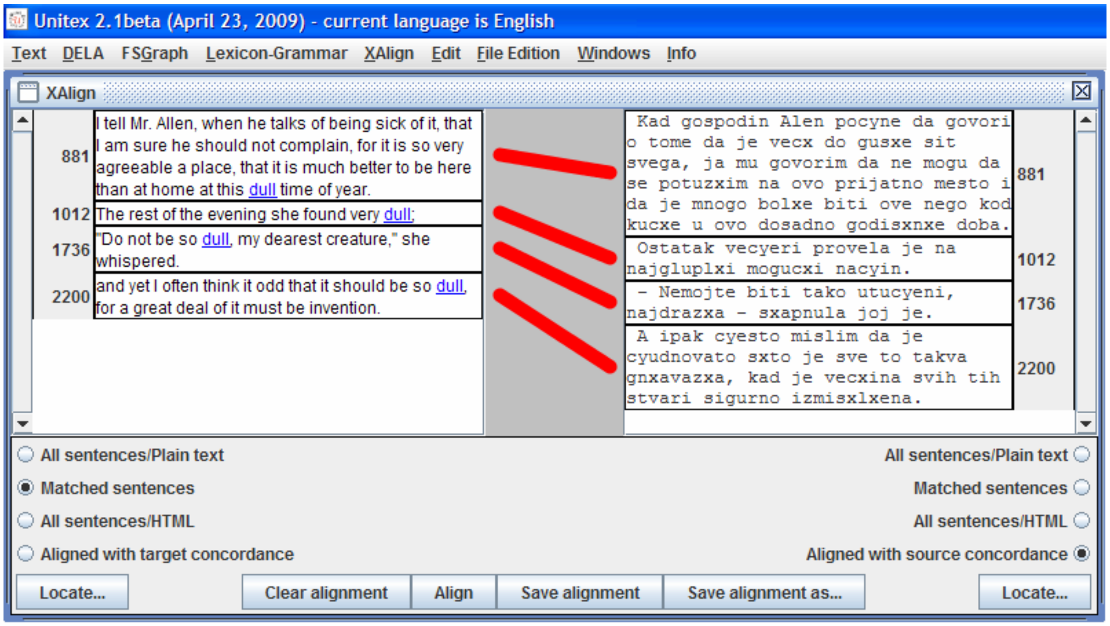
No one who had ever seen Catherine Morland in her infancy would have supposed her born to be an heroine.	Niko ko je imao prilike da vidi Katarinu Morlend u njenom detinjstvu ne bi pomislio da se rodila da bude junakinja.
That is exactly what I should have guessed it, madam," said Mr. Tilney, looking at the muslin.	- To je upravo koliko sam mislio, gospođo - reče gospodin Tilni, posmatrajući muslin.
it looked very showery, and that would have thrown me into agonies!	Izgledalo je da će udariti pljusak i došlo mi je da poludim!

Much more complex queries can be formulated by combining the basic patterns into search graphs.

Conclusion

For the preparation and exploitation of the aligned Jane Austin corpus we did not use one single piece of proprietary software. The version 2.1 of Unitex is freely distributed under the terms of the Lesser General Public License (LGPL), while XAlign can be used from Unitex, ACIDE or independently.

At present, only *Northanger Abbey* has been fully aligned and corrected, while the alignment of *Mansfield Park* is under way. Each aligned text can be used separately, before the completion of the whole task. The use of the aligned corpus, as well as the electronic dictionaries of Serbian for research purposes can be negotiated with the authors. The aligned corpus does not have to be used with Unitex; since ACIDE produces several different standard output formats, it can be used with any other software that supports them.



881	I tell Mr. Allen, when he talks of being sick of it, that I am sure he should not complain, for it is so very agreeable a place, that it is much better to be here than at home at this dull time of year.	Kad gospodin Alen pocyne da govori o tome da je vecx do gusxe sit svega, ja mu govorim da ne mogu da se potuzxim na ovo prijatno mesto i da je mnogo bolxe biti ove nego kod kucxe u ovo dosadno godisxnxe doba.	881
1012	The rest of the evening she found very dull;	Ostatak vecyeri provela je na najgluplxi mogucxi nacyin.	1012
1736	"Do not be so dull, my dearest creature," she whispered.	– Nemojte biti tako utucyeni, najdrazxa – sxapnula joj je.	1736
2200	and yet I often think it odd that it should be so dull, for a great deal of it must be invention.	A ipak cyesto mislim da je cyudnovato sxto je sve to takva gnxavazxa, kad je vecxina svih tih stvari sigurno izmisxlxena.	2200

○ All sentences/Plain text — All sentences/Plain text ○
● Matched sentences — Matched sentences ○
○ All sentences/HTML — All sentences/HTML ○
○ Aligned with target concordance — Aligned with source concordance ●

Figure 1 The XAlign window in the Unitex software. The search pattern was *dull* initiated by the left-hand side button "Locate…"

References

Chrobot, A. et al. (1999). Dictionnaire Electronique DELAC anglais : noms composés, rapport technique n°59, LADL, Université Paris 7.

Erjavec, T., A. Lawson and L. Romary (Eds.) (1998). *East Meets West – A Compendium of Multilingual Resources*. TELRI Association e.V.,IdS, Mannheim.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proc. of the 4th LREC Conference*, pp. 1535 - 1538, ELRA, Paris.

Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic Dictionaries*, Faculty of Philology, Belgrade.

Monceaux, A. (1995). Le dictionnaire des mots simples anglais : mots nouveaux et variantes orthographiques, rapport technique IGM 95-15, Institut Gaspard Monge, Université de Marne-la-Vallée

Steinberger, R. et al. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: *Proc. of the 5th LREC Conference*, Genoa, Italy, 22-28 May, 2006, pp. 2142-2147, ELRA

Tufiş, D. (2006). Cross-lingual Knowledge Induction from Parallel Corpora. In: *Formal Approaches to South Slavic and Balkan Languages*, Sofia, Bulgaria, 18-20 October 2006.

Utvić, M., R. Stanković and I. Obradović (2008). *Integrisano okruženje za pripremu paralelizovanog korpusa*. In: *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, pp. 563-578, LitVerlag, Muenster.

Vitas, D. et al. (2008). "*Tour du monde* through the dictionaries", Actes du 27eme Colloque International sur le Lexique et la Gammaire, L'Aquila, 10-13 septembre 2008, eds. M. Constant et al, pp. 249-256, Universite Paris-Est, Institut Gaspard-Monge.