Institute for Parallel Processing
Bulgarian Academy of Sciences
(IPP-BAS)

Aligning the translations a possible strategy for creation of aligned corpora for South-Slavic languages

Veno Pachovski

Department of Mathematics and Informatics, University
"St.St. Cyril and Methodius"
Skopje, FYROM

BIS 21++ report No. 2006-02 March 2006

Acad. G. Bonchev, Bl. 25A, 1113 Sofia, Bulgaria

http://bis-21pp.acad.bg

Aligning the translations – a possible strategy for creation of aligned corpora

(for South-Slavic languages)

0. Introduction

Presently, the creation of large corpora of aligned texts for Multilanguage processing became a routine job. The main efforts were re-routed to the creation of common standards for their use and various computer applications. For basic European languages, the task is easier, because of:

a. Available net sources;

b. Well developed tools for processing of those languages;

c. Multilanguage warehouse of European structures' texts ;

Obviously, that is not the case with South-Slavic languages. The reasons are technological as well as geopolitical, and language specific as well.

1. Specific problems for compiling a south-slavic corpus of aligned texts

The **technological** issues of coding in those languages are concerned with the differences between code pages in Windows - a kind of official standard, not widely used for publishing purposes. Macedonian code page is among the last established among south-slavic languages and its layout is very different from the generally used font-based keyboard layout. Also, some publishers create their own font layout, which additionally confuses matters.

Geopolitical reasons narrow the flow of translated texts .

1. *For administrative and legal texts (basic source for Multilanguage corpora lately).* South-slavic countries are in various phases of joining the euro-structures, so there is also the difference in the quantity of translated texts and their availability.

2. *For literary texts. The volume of translated literary sources* between south-slavic languages is minimal and continuing to diminish.

Today, wealth of electronic texts is readily available directly from the publisher, but unfortunately, the great majority comes from English. So, this situation on the "market" of electronic texts of translations between south-slavic languages, forces another approach in building Multilanguage resources.

2. Aligning translations

The standard reading of the parallel aligned text presumes availability of the original and translated text - source/target language (SL and TL). The SL is usually the native language and this kind of research of the two texts measures the difference between **said** and **translated** (this distinction is not clear). However, when approaching the translations of the European documents, the SL is unknown.

Considering the aforementioned situation with translations between south-Slavic languages, the idea of constructing the corpora of SL-TL should be put aside for the time-being.

Pragmatically, the solution is to align the translations of the original in the third language, which should be popular enough. It is the only possible approach when building Multilanguage aligned corpora (see the choice of Orwell's *1984* for the Multext-East project and Plato's *Republic* for TELRI project). So, Lt1 → Lt2 direction, obviously supposes the existence of Ls → Lt1 and Ls→Lt2 pairs, which automatically leads to compiling three-language corpora with pivot (intermediary) text.

Consultations with Bulgarian, Serbian and Macedonian colleagues, working in the field, just highlighted one sad fact. Even if there are tens of translated texts between a pair of languages, a great majority of them probably won't exist for another pair. So the way to aligned south-Slavic corpora passes through Ls→Lt pairs. (See Fig.1 and the table below).

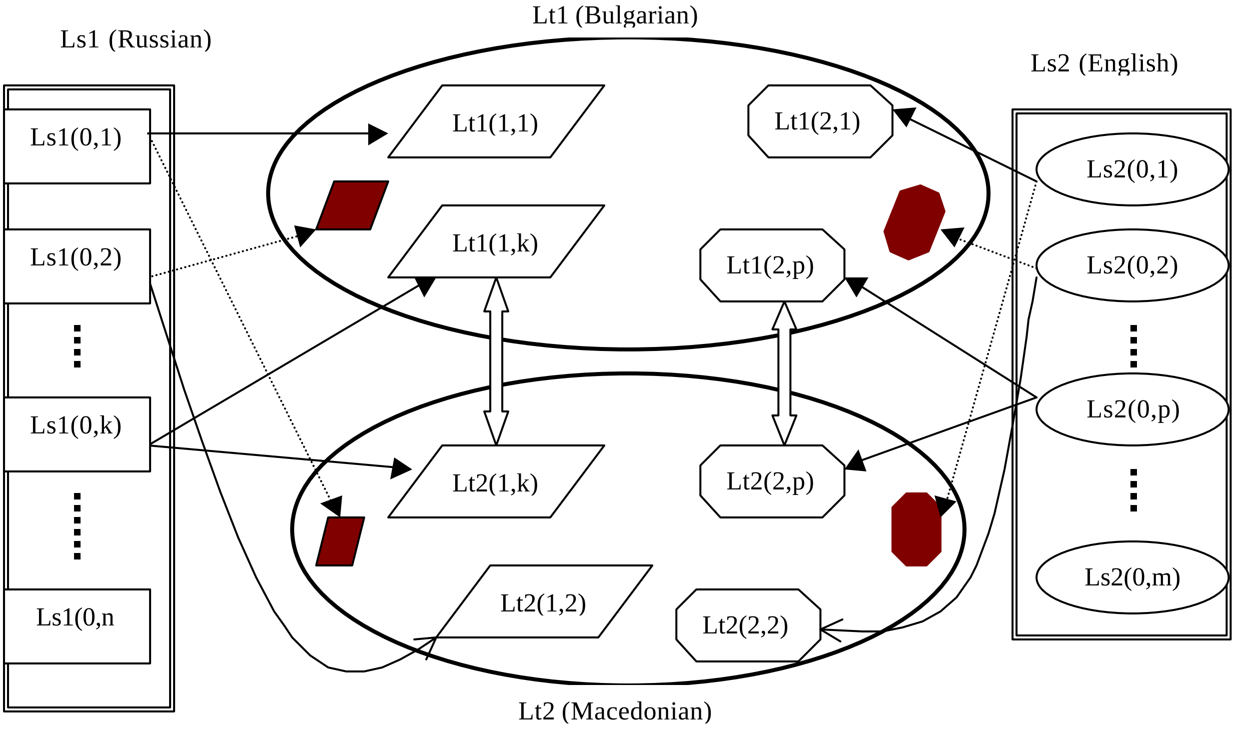


Fig. 1: Working towards aligned south-slavic corpora **(k,pÎ** ***N*****, n=|Ls1|, m=|Ls2|)**

Existing pairs (candidates for alignment)
Ls1(0,k) → Lt1(1,k) and Ls1(0,k) → Lt2(1,k)
Ls2(0,p) → Lt1(2,p) and Ls2(0,p) → Lt2(2,p)
⇨ Lt1(1,k) and Lt2(1,k);
⇨ Lt1(2,p) and Lt2(2,p);

Problems (one of the pairs is missing)
Ls1(0,1) → Lt1(1,1) but Ls1(0,1) → Lt2(**?**,**?**)
Ls2(0,1) → Lt1(2,1) but Ls2(0,1) → Lt2(**?**,**?**)
Ls1(0,2) → Lt2(1,2) but Ls1(0,2) → Lt1(**?**,**?**)
Ls2(0,2) → Lt2(2,2) but Ls2(0,2) → Lt1(**?**,**?**)

3. The selection of the pivot text

It is obvious that the pivot text should be chosen from well translated language, which narrows the choice to English, Russian and French.

The practical sources of electronic translations are:

1. Publishing houses
2. Digital libraries.

The first is more limited and quite often has copyright problems. The second is richer (the copyright problems are already partially solved by the fact of the Net publishing itself).

The selecting criterion is the net-availability of the intermediary pivot text.

For Russian and English, candidates for the pivot role, the approach to copyrighting in general is very different, which can be deduced from the available digital libraries. There are lots of web addresses, which contain full texts of Russian literature, available because of the non-standard Russian copyright law. In English, the available digital texts are dated 50 or more years ago.

So, as a starting point, the Bulgarian site - http://borislav.free.fr/mylib/ was chosen and the originals of those translations in Ls (Russian) were checked in www.lib.ru .

The next step is to check the availability of translations of the original in Ls for the other Slavic language.

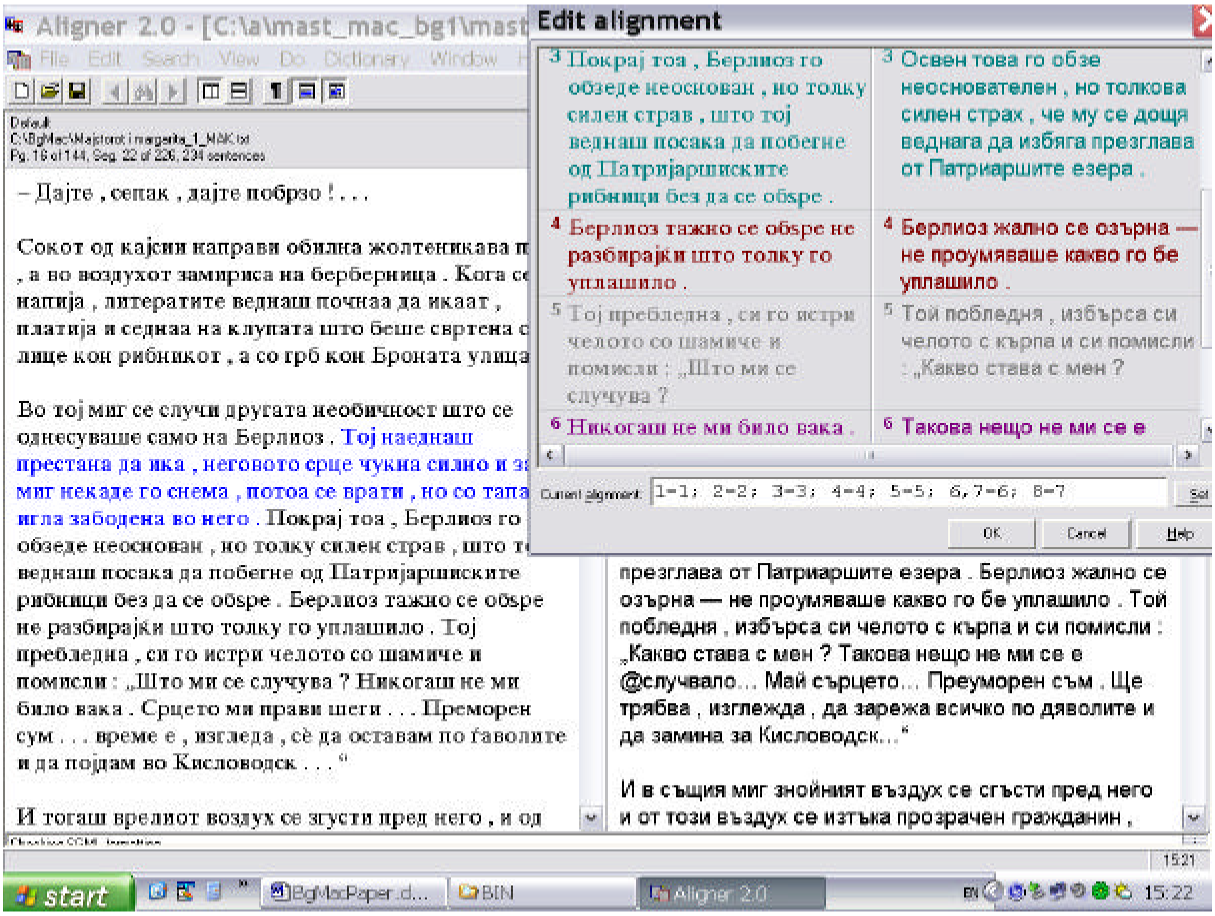
4. Experiment

a. *Sources*. The novel "Master and Marguerite" by Michail Bulgakov was chosen as a first brick in the building of the parallel translations for Bulgarian and Macedonian. The Bulgarian translation and the Russian original were downloaded from aforementioned net-libraries; the Macedonian was obtained courtesy of "Tabenakul", Skopje (translator Tania Urosevic). The size of each text is approximately 110 000 words.

b. *Instruments*. For producing the aligned corpora, *Mark Alister* aligner was used, based upon the *Gale-Church* algorithm. The phases of production were:

1. **manual pre-processing**, necessary for Gale-Church algorithm - the equal number of paragraphs in both texts. This is the hardest phase in the production because of the rather big differences between the formats used. There were additional difficulties with the processing of the Macedonian font which is non-standard.
2. **aligning with the Mark Alister aligner**, with a friendly *Delphi* interface, allowing the editing of the alignment results, as well as searching of selected text units.

The alignment was performed for the language pairs: Russian - Bulgarian, Russian - Macedonian and Macedonian - Bulgarian. (The order of alignment in the third pair is arbitrary).


Aligner 2.0 - [C:\a\mast_mac_bg1\mast
File Edit Search View Do Dictionary Window H
Edit alignment
OK
Cancel
Help
Set
start
BgMacPaper.d...
BIN
Aligner 2.0
EN
15:22
15:21

Default
C:\BgMac\Majstorot i margarita_1_MAK.txt
Pg. 16 of 144, Seg. 22 of 226, 234 sentences

– Дајте , сепак , дајте побрзо ! . . .

Сокот од кајсии направи обилна жолтеникава п
, а во воздухот замириса на берберница . Кога се
напија , литературите веднаш почнаа да икаат ,
платија и седнаа на клупата што беше свртена с
лице кон рибникот , а со грб кон Броната улица

Во тој миг се случи другата необичност што се
однесуваше само на Берлиоз . Тој наеднаш
престана да ика , неговото срце чукна силно и за
миг некаде го снема , потоа се врати , но со тапа
игла забодена во него . Покрај тоа , Берлиоз го
обзеде неоснован , но толку силен страв , што то
веднаш посака да побегне од Патријаршиските
рибници без да се обѕрe . Берлиоз тажно се обѕрe
не разбирајќи што толку го уплашило . Тој
пребледна , си го истри челото со шамиче и
помисли : „Што ми се случува ? Никогаш не ми
било вака . Срцето ми прави шеги . . . Преморен
сум . . . време е , изгледа , сè да оставам по ѓаволите
и да појдам во Кисловодск . . . "

И тогаш врелиот воздух се згусти пред него , и од

3 Покрај тоа , Берлиоз го обзеде неоснован , но толку силен страв , што тој веднаш посака да побегне од Патријаршиските рибници без да се обѕрe .

3 Освен това го обзе неоснователен , но толкова силен страх , че му се дощя веднага да избяга презглава от Патриаршите езера .

4 Берлиоз тажно се обѕрe не разбирајќи што толку го уплашило .

4 Берлиоз жално се озърна — не проумяваше какво го бе уплашило .

5 Тој пребледна , си го истри челото со шамиче и помисли : „Што ми се случува ?

5 Той побледня , избърса си челото с кърпа и си помисли : „Какво става с мен ?

6 Никогаш не ми било вака .

6 Такова нещо не ми се е

Current alignment: 1=1; 2=2; 3=3; 4=4; 5=5; 6,7=6; 8=7

презглава от Патриаршите езера . Берлиоз жално се
озърна — не проумяваше какво го бе уплашило . Той
побледня , избърса си челото с кърпа и си помисли :
„Какво става с мен ? Такова нещо не ми се е
@случвало... Май сърцето... Преуморен съм . Ще
трябва , изглежда , да зарежа всичко по дяволите и
да замина за Кисловодск..."

И в същия миг знойният въздух се сгъсти пред него
и от този въздух се изтъка прозрачен гражданин ,

Fig.2. Results of alignment and the editing function of the interface

5. Some results

The aligned texts corpora may generally be used:

1. as a necessary resource for computer applications, modeling the translation - experimental test-bed for statistical machine translation;

2. for various kinds of computer assisted translation - translation memory applications;

3. in statistical comparative research upon the pair itself. The last is especially interesting for close related languages, as is the triplet Russian-Bulgarian-Macedonian. Even further, comparative research into the degrees of closeness between **pairs** of languages is enabled. Obviously the distance RUS-BG is greater than BG-MAC distance;

4. in comparative lexical research on translation equivalents in a given pair . For close related languages, of particular importance can be the extraction of the so called *false friends* - similar grapheme sequences, but different meanings.

But, the concept of false friends is relevant only in an Ls-Lt pair, not in Lt1-Lt2 pair.

An example from our research. The Macedonian particle **Ï ÀÊ** appears 12 times in Chapter 1, and 10 of them does not have an equivalent in the corresponding Bulgarian text, because this particle is obviously used as a kind of strengthening, even parasitic element (similar to the Bulgarian **ÏÚÊ**). In 2 uses with temporary meaning (*again*) there is Bulgarian equivalent **ÎÒÍÎÂÎ**.

The lack of direct connection between lexical elements of two languages becomes clearer, if we look at the pair RUS-MAC. Here 7 of those 10 uses of Macedonian **ÏÀÊ**, are a direct translation of frequently used Russian particle **ÆÅ** with the same function.

So, in this case, there is no comparative research in MAC-BG pair, but the differences in translation pairs can be discussed. Namely, it can not be claimed that Macedonian translation is closer (in a word sense) to the original, but it can be said that in Bulgarian, there is no direct equivalent to Russian **ÆÅ**. In Macedonian, maybe **ÏÀÊ** is the functional equivalent to **ÆÅ**.

Even further, it would be very interesting to align two translations to the same language made by different translators or in different time periods.

6. Conclusions

On a conceptual level. The aligned translations of the third language original can not be viewed nor investigated as translations in the real sense of the word. They can be perceived rather as a test-bed îf comparable corpora, although they are far from that notion. Really deep investigation of translation mechanism is realizable only in the both pairs Ls-Lt1 and Ls-Lt2. Considering other applications of aligned corpora as cited above, the aligned translation represents an excellent resource base.

On a practical level. The building of the south-slavic aligned corpora should be done in two lines, following the Ts - English and Russian. The choice of Russian enriches the research, setting it in the sphere of Slavic languages and refining it from typological point of view. On the other hand, the choice of English is challenging in view of its grammatical difference from Slavic languages. Both sources should be investigated having in mind the limited access to the Ls resources.

The coordination of the accumulation of appropriated resources should start as soon as possible, for which this Conference offers a favorable ground.

REFERENCES:

1. Collecting resources:

TELRI – PECO/COPERNICUS 1202/40444 (Trans - European Language Resources Infrastructure Information)

MULTEXT-EAST – PECO/COPERNICUS (Mmultilingual Text Tools and Corpora for Central and Eastern European Languages)

2. Standardizing resources:

INTERA - E-Content (Integrated European language data Repository Area)

3. **William A. Gale and Kenneth W. Church**, A Program for Aligning Sentences in Bilingual Corpora, in *Computational Linguistics*, 19(1), 1993.

Acknowledgement. The work reported in this paper is partially supported by the project BIS-21++ funded by the European Commission in FP6 INCO via contract no.: INCO-CT-2005-016639.