Balkan South-East Corpora Aligned to English

Elena Paskaleva
Institute for Parallel Processing, Bulgarian Academy of Sciences
25A Acad. G.Bonchev St., Sofia 1113, Bulgaria
hellen@lml.bas.bg

Abstract

The paper describes the new corpus of 9 Balkan Southeast languages (BSEC) aligned to English with a volume of 3 Mio words for every language. A new aligning tool for batch alignment is created and some functional features of different aligners – constructed and used in IPP, BAS - are discussed. The general principles of the morphosyntactic annotations (used in INTERA, INTEX and Multext-East projects) available for the majority of BSEC languages are investigated for possible harmonization and future unification. The paper examines the principles and the main issues of the transition INTEX→MTE formats for Bulgarian and Russian.

Keywords: aligned corpora, aligners, morphosyntactic annotations, tagset standards.

1. Aligned corpora beyond "old" European languages

In the multilingual European world the development of computational data support for multilingual applications reveals specific geopolitical characteristics. It moved from test-bed compilations for comparative research through unified formats and common language material to the creation of multilingual corpora – parallel, aligned, annotated, classified, clustered, ontologically structured. This resource creation started with the "oldest" European languages and in the course of EU enlargement kept extending to the respective "new" ones. Outside this circle there are languages of Southeastern and Eastern regions that are not in the European political structures but nevertheless connected with the languages of the EC old members and new joiners by virtue of belonging to the same geographical area, language family or linguistic union. Yet, all these seemingly "poor relatives" have developed advanced LT tools, and have adopted common LR formats within previous joint projects, so that their joining a common NLP paradigm would anticipate their inclusion in the political paradigm of EU. In the project Bulgarian IST Centre of Competence in 21 Century (BIS-21++), (FP6-2004-ACC-SSA-2) the work package Multilinguality in Advanced Information Search foresees the creation of basic translation memory components for Greek-Bulgarian translation. The collaborative experience with a Greek partner ILSP Greece http://www.ilsp.gr/info_eng.html within a previous project INTERA [1] allowed us to take over the formats accepted for representation of aligned corpora. As a resource for producing a representative aligned Bulgarian-Greek corpus, the Greek partner proposed the portal known as *Balkan-News*, http://www.setimes.org. The languages presented in this site, as well as the processing tools developed for transforming it from multilingual web resources into aligned corpora, allowed us to start with the building of Parallel Balkan-South-East corpora (BSEC) aligned to English.

1.1 The WEB source of BSEC

The site www.setimes.com is defined by the institution maintaining it, the US ministry of Defense, as a news and information website covering the Balkans featuring current events, business, diplomacy, film, tourism, sports and science. It has been maintained since 2002 in four rubrics – News, Reportage, Features, Roundup in 10 languages – Albanian, Romanian, Greek, Turkish, Bulgarian, Macedonian, Serbian, Croatian, Bosnian, English (until 2007 also Serbian Cyrillic).

1.2 Volume of aligned texts

On the expectation of approximate equality of the volumes of language material, the differences are determined largely by the beginning of the processing of a given language. The initial site map from 01.2002 contained four languages – English, Greek, Albanian, and Serbian. In the following 10 months Bulgarian, Macedonian, Romanian and Turkish have been included, and after 5 more months – Croatian, while Bosnian has been added in 02.2007. Serbian-Cyrillic existed until 2007. Due to unequal start and non-translated texts there are "holes" in the compilation. This requires intensive manual pre- and post-editing of the material, which is not automated yet – cf. the difference in volumes in Table 1.

Table 1. Number of files and words –calculated for .html format

Languages	Files	Words
English	~15800	~3 456 000
Greek	~15790	~3 942 000
Albanian	~15791	~3 740 000
Serbian	~15799	~3 380 000
Bulgarian	~13000	~3 000 000
Turkish	~13000	~3 580 000
Romanian	~13000	~3 500 000
Macedonian	~13000	~ 3 250 000
Croatian	~10826	~2 800 000
Bosnian	~1900	~380 000

Obvious processing advantages are, on the one hand, the uniformity of text division and, on the other hand, the

thematic diversity of the texts – i.e. **simple** structure and **rich** lexicon.

1.3 I/O formats in processing the corpus

While producing BSE aligned corpus, the following text formats were encountered:

First, texts are extracted in HTML format with a spider procedure storing them in a file structure reflecting their groups and date (Y, M and D). For example, English groups *Newsbriefs* and *Features* for years 2002-2006 have 26589 files in 1606 directories, with minimal length of a file - a title and a single sentence. The result of the alignment is given in two formats – plain text and XML presentation in XCES format [2] as used in the INTERA project. The text output of the aligner for 26589 files produces triplets – Bulgarian, English and aligned text - and contains 1719 files, with given prerequisite option to the aligner to unite the results in portions of 20 files. Maximal length of the aligned file is 21300 words.

2. Aligning procedure

Linguistic Modeling Department of IPP–BAS has experience in constructing sentence aligners within European projects. For example, for processing English and Bulgarian texts, within GLOSSER [3] the sentence aligner called Mark Alister [4] was created with user friendly interface which includes possibilities for pre- and post-editing, as well as for visualization of I/O and intermediate results.

The volume and the nature of the present task required the construction of another type of sentence alignment; work which is more industrial, than scientific.

The difference between wholesale manufacture and tailor-made production when aligning corpora is analogous with producing any other product. Here, the accent is on the existence of obligatory input format of the processed material and the continuance of the production process in mass production.

Mark Alister aligner, as a tailor-made product, allows multiple text formats for each of the two input files. The phase of preprocessing – counting and equalization of the number of paragraphs and sentences is separate from the actual aligning, which allows intermediate regulation of text division entities with editing. Also, further correction of aligned units is allowed before the finalization of the product – see Figure 1. However, these very useful functions (betterments) of the program design and interface are possible only for creating separate products, like above mentioned single text files.

Considering the volume of BSEC, the inevitable industrial approach prefers a batch mode alignment of huge amounts of text as a single one time only process with no possibility of tuning and intervening in the pre-processing module or editing the final results. Errors (details of defects) may be signaled during processing, but they do not stop the aligning. For aligners with continuous production, news texts are especially useful. Their standardized format (short notices with simple text division structure) satisfies the requirement for uniformity of the input product.

To meet these requirements the aligner LORA (Language Objects Raw Analyzer) was developed – see Figure 2. It is based on the well-known Gale-Church algorithm for sentence alignment, written in C# MS.NET Framework 2.0. All its modules have been developed for adjusting the aligner to the respective language pair, taking into consideration textual and linguistic specificities [5].

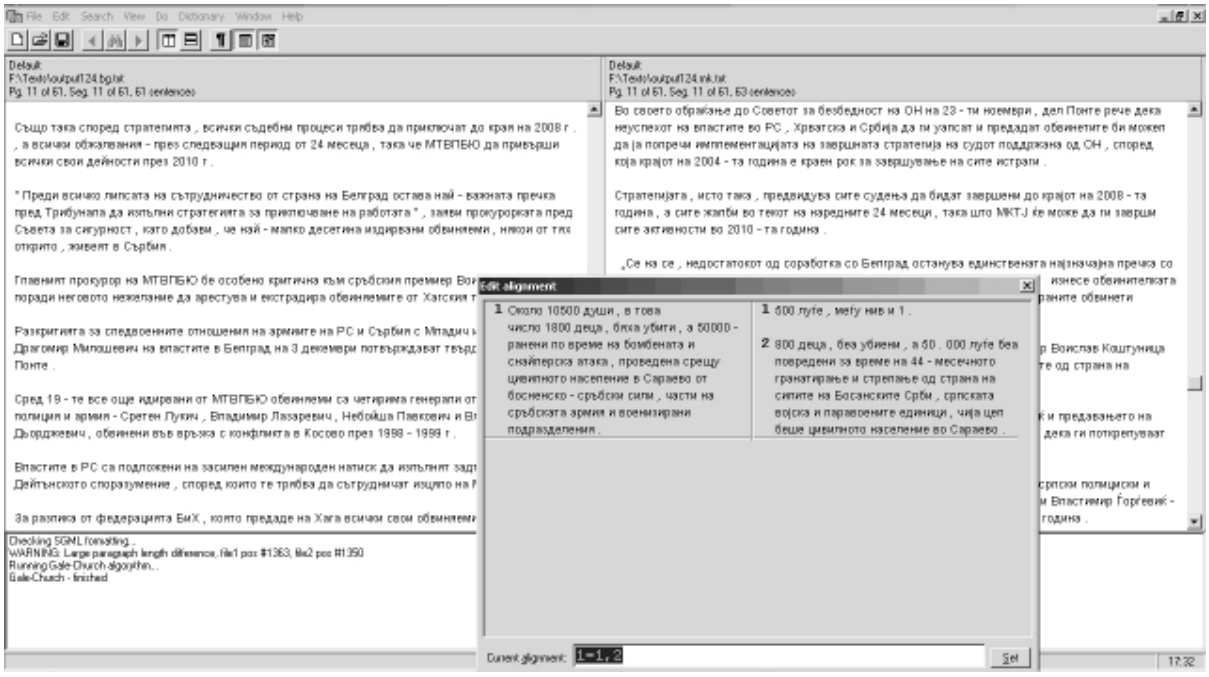

Figure 1.Mark Alister – single mode alignment with intermediary editing facilities

.

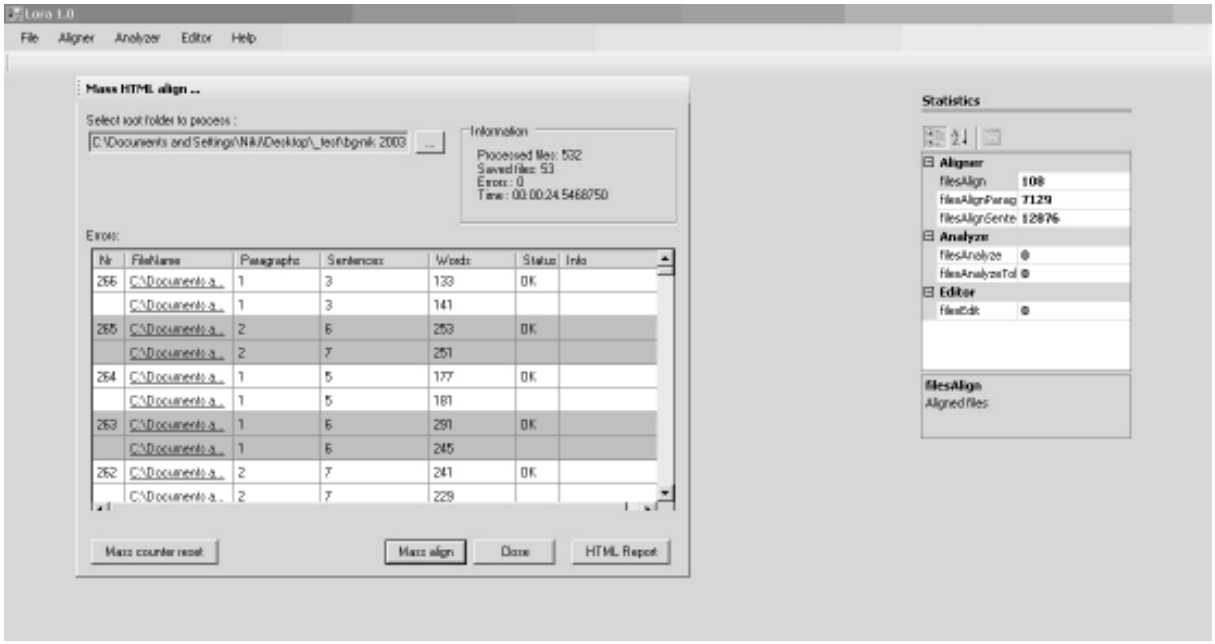

Figure 2. LORA – batch mode alignment with statistics of: processing time, number of files and text units

2.1 Language-specific components in the aligning procedure of LORA

In applications based on the Gale-Church algorithm, the main procedure is language independent, due to the nature of the basic units of statistical processing – symbols, sentences and paragraphs. From these objects, only the last one is really language independent. Being part of the alphabet of the given language, the symbols are related both to the writing system and the specific diacritics; therefore they are language dependent on graphemic level. There are three main writing systems in the BSEC languages – Latin, Cyrillic and Greek. Only two of these languages – Bulgarian and English – do not use diacritics. The encoding system at the website has changed during the years of its maintenance – e.g., Macedonian was encoded in ANSI until 04.2007 and in UTF-8 afterwards, which is accepted as common format. Such encoding differences are carried over to the alignment results and call for special *View* functions of the interface as well as for possibilities of *Auto-detect* or *Encoding* selection. The latter is used only in single mode, where the same encoding of both aligned files is required. As object of alignment, the sentence boundaries are also crucial for the work of the aligner itself. The linguistic objects entering the sentence splitting procedure are punctuation marks and abbreviations – the former are universal while the latter are language specific. We assume, based on statistical studies of Bulgarian text, that in choosing the role of the dot – sentence splitter or not, only two types of abbreviation are relevant for predicting the nature of the next string (sentence or word). These are the abbreviations immediately preceding a proper name (like *prof* and *mr*) and a number (like *fig*.or *p*). All the rest, like abbreviations of measure units, dates, etc., cannot determine unambiguously a sentence boundary. Therefore the lists with the two types of abbreviations for the respective language are part of the input settings of the language tuning of the system.

Reporting functions signal errors in the structure of the input files related to the number of paragraphs and sentences. Also statistics of the same parameters are derived. The processing modes include batch alignment as well as alignment of single files. In the latter case there is a *View* option for incoming files along with checking possibilities for alignment direction and encoding.

3. Tools for further processing of aligned BSEC corpora – availability and possibilities

Within the standards of INTERA project, where three of the BSEC languages participate (Bulgarian, Serbian and Greek), each one with its own text resources, aligned separately to English), the project deliverables were supplied in three types of representation according to the standards of XCES – *.doc*, *.tmx* and *.ana* [2]. The first two supply the XML representation of both files – separately and aligned. The third standard defines the form of grammatically annotated single-language files, and, as a result of the annotation, for every word form (<toc>) values for <orth>, <base> and <msd> are calculated, i.e. lemmatization and tagging is performed. The proposed

common tagset standard is the one used in Multext-East (MTE) project. Considering that there is no tagging tool for Bulgarian according to that standard (there are only values and single examples given in [6]), the procedure had to be performed in two steps – **a**. tagging with some other morphosyntactic format, and **b**. reformatting of already tagged text in MTE formats.

When matching those two tagsets (the first one adopted in Bulgarian Grammatical Dictionary), some differences appeared on a conceptual level, as well as in specific grammar values. The first could not be overcome because of the ideological differences between the two systems, but the latter seemed interesting for further work, for an update of the Bulgarian tagset in MTE family. (It was done later when reformatting the grammar dictionary – see 4.2). So, the family of languages with MTE annotation tagset was enriched with Bulgarian full-form dictionary of 80000 lemmas, and some cases in the first Bulgarian MTE annotation were revisited.

Along with alignment tools, which are language independent (excluding some stipulated components), the crucial module for creation and support of common NLP paradigm for Balkan and South-Slavic languages is the availability for every language of tools for grammatical annotation and tagging in unified format – linguistic and technological.

4. Prerequisites for creation of common annotation format and tagging tools for BSE languages

In the creation of common annotation formats for South-Slavic and Balkan languages, there was a considerable and continuous contribution by joint local European initiatives aimed at practical unification of NLP tools. TELRI and MTE [6] were the first with that intent, and in the last couple of years – as already mentioned – INTERA. The development of Bulgarian lexical resources is connected to BILEDITA project [7] (the first description of Bulgarian morphology with FST on INTEX system), to some separate comparative studies over existing tagsets for Bulgarian oriented towards international standards, see [12] and more details in 4.1, and to the lexical base of BALKANET project [8].

During the last 10 years, various annotation formats appeared and gradually there started to emerge an intersection between the tagsets mentioned, which were realized over more than one language. For example, such an intersection exists for Bulgarian and Serbian in INTEX [10] and MTE [6]. Within the Linguistic Modeling Department, such a transformation between two formats issued in two variants of the Russian grammar dictionary was performed over the Grammar dictionary of A.A.Zaliznjak [11].

The transformation of large grammatical dictionaries for Bulgarian and Russian was made possible with special DB tool designed for representation of full-form dictionaries in DELAF format of the INTEX system – see its description in [9]. It offers tools for transformation of lexical data, as well as annotation markers. The transition was performed after creating the rules for the transformation of morphosyntactic parameters of the first format to those of the second. An example for transformation of a part of a dictionary is shown below – see Figure 3.

вредную,вредный.A:fsa врезную,врезной.A:fsa временную,временной.A:fsa временную,временный.A:fsa вредного,вредный.A:msg:msa:nsg врезного,врезной.A:msg:msa:nsg c временного,временной.A:msg:msa:nsg временного,временный.A:msg:msa:nsg вредно,вредный.A:nsy временно,временный.A:nsy вреднейшими,вредный.A:spr временников,временник.N+NA+M:pg вредить,вредить.V+IPF+i:AI врезаться,врезаться.V+IPF+i:AI врезываться,врезываться.V+IPF+i:AI временить,временить.V+IPF+i:AI врезались,врезаться.V+IPF+i:APp врезались,врезаться.V+PF+i:APp	вредную,вредный.Afpfsa---c врезную,врезной.Afpfsa---c временную,временной.Afpfsa---c временную,временный.Afpfsa---c вредного,вредный.Afpmsg---c,Afpmsa---c,Afpnsg---c врезного,врезной.Afpmsg---c,Afpmsa---c,Afpnsg---c временного,временной.Afpmsg---c,Afpmsa---c,Afpnsg---c временного,временный.Afpmsg---c,Afpmsa---c,Afpnsg---c вредно,вредный.Afpns----n временно,временный.Afpns----n вреднейшими,вредный.Afs-p-i---c временников,временник.Ncmpg--n вредить,вредить.Vmi----------p врезаться,врезаться.Vmi----------p врезываться,врезываться.Vmi----------p временить,временить.Vmi----------p вредили,вредить.Vmis-p------p врезались,врезаться.Vmis-p------p

Figure 3. INTEX and MTE formats of Russian grammar dictionary

4.1 From INTEX to MTE format – peculiarities of transition

Comparative research over tagging systems started long ago in IPP BAS, being based on the genuine MSA of Bulgarian, developed in the Department, that is comparable with EAGLES and MTE annotation standards [12]. In these works they have only abstracted over the values of other tagsets, without any practical experiments for full transformation of tag values.

While transferring for one and the same language from one morphosyntactic annotation set (MSA 1) to another (MSA 2) in order to produce an annotated collection in the second format, two by-products have been manufactured, outside the original assignment;

1. a measure of equality for the methodologies of two annotation types,
2. possibilities for separate corrections in both formats, passed through the conversion mixer

When creating the transformation rules for a transition between those two MSA basic components – the objects of the transformation are

1. *syntax* of the group of annotation markers, attached to a single running word;
2. *content* of annotation parameters – as a range of a grammatical information and values of separate criteria

Syntax of the annotation markers. There are two basic differences between syntaxes of annotation of INTEX and MTE:

1. Position, number and length of tags:

- MTE, being a multilingual unified tagset, implies – for a concrete POS, equal number of tags, fixed position of tags, length of tag – 1 symbol
- INTEX allows different number of tags for one POS, only the relative position of tags is fixed, but not the absolute one, the length of the tag can exceed one symbol (especially for lexical ones – see below)

2. Internal structure of tag sequence:

- MTE – two elements are distinguished in tag structure – POS marker, written in capitals, and other features, not capitalized
- INTEX – two substructures are distinguished in tag structure – the lexical and the grammatical part of the tagset. The distinction is linguistically motivated:

  a. the first one characterizes the whole lemma, the invariant part of the features of the whole paradigm, and

  b. the second represents the individual features of every member of the paradigm.

The **a.** and **b.** are formally separated in the tag record:

N+M+H:sh – noun, masculine, human: singular, short article.

None of the known (to us) tagsets for Slavic languages makes such a distinction.

The content of annotation markers. It is defined as a number of values, attached to a single marker, as well as a common number of markers, creating the tag.

An example for checking according to the number of values criteria: is there a marker for aspect in the annotation of the verb, and if yes, what is its value – perfective, imperfective or dual, or just the first two. Within the last of these, dual verbs are marked as imperfective, having in mind the maximal range of the verb paradigm for that grammatical attribute, i.e. its non-marked position.

The basic differences in the number of markers in the tag, i.e. according to the range of grammatical information for both systems, are the following:

INTEX parameters of a tagset are much richer than those of MTE (at least, for the presumed structural possibility to separate parameters of a lemma). Within the local tagset for every language, the richness of the tagset depends also on the concept of the authors – their tendencies to record more characteristics in the annotation markers. These collide with the necessity of a small number of tags for the needs of statistical tagging. The range of grammatical parameters is defined – in its core, by:

a. necessity to differentiate the separate members of a paradigm – each one with a unique tagset, and after that, by

b. purpose of the annotation, for the following types of text processing (disambiguation, shallow parsing etc.).

These pragmatic requirements should prevent the temptation to record everything known about the word within the tagset of the annotation instrument. To illustrate how the authors of tagsets for new languages succumb to this temptation, see detailed table of tagsets in [1] to compare the annotation markers even in a common format. The number of markers, even in a common format and for closely related Slavonic languages, differs substantially.

4.2 Producing new grammar resources with conversion of existing ones

The scheme :

text → annotation in MSA1 → converting to annotated text in MSA2,

which was used for producing the Bulgarian annotated resources, could be improved substantially if the sole annotating instruments could be converted , in this case – the full-form dictionaries. This operation, for the Bulgarian grammar dictionary, was performed using a

special DB tool for compiling, update and correction of grammatical dictionaries (see its description in [9]). So, as a by-product of the assignment, a Bulgarian dictionary in MTE format was produced. The changes, aimed at the harmonization of the two tagsets, were mainly content oriented, not structural – adding some new values to the meanings of some basic categories (for example, for the gender of nouns a new value was introduced – common gender), and for some POS, the tagset was rebuilt according to the common principles of the unified INTERA tagset. This applies mainly to the pronouns, a famous problem (bottle neck) when creating a tagset for Slavonic languages. The tendency for conceptual depth of tagging leads to the discomforting implication of trying to achieve equality between the number of tagged elements and the number of tags – 309 pronouns were annotated with 296 unique tags. For them, the annotation scheme was compiled not through conversion, but starting from the beginning, following the principles presented in [6].[1]

[1] This is the place to mention our great astonishment at the lack of a marker for *owner person* (in the inventory of pronoun markers) with possessive pronouns for all MTE languages, yet with the notable exception of (the exotic) Hungarian – where this marker is used in both adjectival and nominal inflections. It is unclear why the authors of the standard saw no real difference in the values of that property for English *my*, *your* and *his*, quite obvious indeed even for this weak morphology. Even more perplexing is this lack for Czech pronouns *muj*, *tvuj* and *jeho* where these values are also missing. And quite unexplainable for a Slavic linguist is the absence of a category of *transitivity* for verbs while *aspect* is added in the updated version.

5. Common resources, common standards and instruments for Balkan, Slavic and South-Slavic languages – realities and intentions

The table below represents the relationship between the three groups of languages mentioned in the subtitle according to their NLP equipment. This is judged not independently, but according to the link with pre-existing text collections and unified processing standards.

A terminological remark is needed for the languages in the table. The notions – *Balkan*, *Slavic* and *Southeast* (B, S and SE) refer to languages of the Southeast part of Europe by:

a. geopolitical criteria, based mainly on their rank in the European integration,

b. strictly geographical criteria, and

c. linguistic typological criteria.

As an intersection of these three groups, we get:

B/SE/S – Bulgarian, Serbian, Macedonian,

B/SE – Greek, Romanian, Turkish,

SE/S – Croatian, Slovene,

S – Russian and Czech.

Outside this classification Hungarian and Estonian are mentioned because of their inclusion in common standards. English being also outside this group, figures separately – as a participant in all of the discussed common activities.

.

Table 2. Languages, aligned corpora, standards, tools

LANGUAGE	EC MEMBER	BSEC aligned	IRC ACQUIS	INTERA tagset	MTE tagset	ANNOTATING TOOLS
ALBANIAN		x				
ROMANIAN	x	x	x		x	x
GREEK	x	x	x	x	x	x
TURKISH		x				x
BULGARIAN	x	x	x	x	x	x
MACEDONIAN		x			x	
SERBIAN		x		x	x	x
CROATIAN		x			x	x
BOSNIAN		x				
SLOVENE	x		x	x	x	x
CZECH	x		x		x	x
RUSSIAN					x	x
HUNGARIAN	x		x		x	x
ESTONIAN	x		x		x	x
ENGLISH	x	x	x	x	x	x

6. Using the corpus for research and applications

Various research and application activities within the framework of the project BIS21++ have already used the following aligned corpora:

- Bulgarian-Greek – for preparing the basic material for translation memories.
- Bulgarian-English – for clusterization experiments [13].
- Bulgarian-Macedonian – for experiments with cognates, false friends and lexical closeness – ongoing work to be reported at the present workshop [14].

There is special research interest in using the corpus for measuring the closeness of related languages, which are well-represented in BSEC (Bulgarian, Serbian, Macedonian, Croatian, and Bosnian). A series of such studies and experiments have been performed for another pair of Slavic languages Bulgarian and Russian – and the methodology can successfully be applied to the BSEC material.

7. Conclusions and further work

A careful look at the table representation of the parameters of the afore-mentioned set of languages (aligned corpora and morphosyntactic formats) shows that – within these two collections (BSEC and ACQUIS) and with common annotation format (MTE) – Bulgarian, Greek and Romanian are the best represented languages. Uniting the rest of BSEC languages to this group will be easier in the domain of the morphosyntactic format, rather than the ACQUIS corpus. Pre-existing annotation instruments can be translated in a short time, but translation of millions of words takes a lot more time. The same applies to the Russian language, with his undisputable place in the typology of Slavic languages.

So, in our opinion, the realistic approach to the unification of resources for languages in this part of Europe, as well as of tools for their processing, is as follows: using already aligned corpora (the topic of this paper), the work continues with already unified MTE formats, which will draw Serbian, Croatian, and Slovene in the common paradigm according to the principle – "common corpora - common annotation scheme". This principle has already been established in Multext-East, but, unfortunately, for some languages, like Bulgarian, the text volume has been too small for the contemporary needs.

This unification must be preceded with a further harmonization of MTE annotation set – work, started successfully within INTERA project, and hopefully continued in future join local activities, over rich enough common language material.

Acknowledgements

Special thanks go to Stelios Piperidis, who showed me the way to the corpus in the WEB space jungle and to Nikolaj Genov, who invested all his energies in fulfilling the unrealizable linguist's hope for the Universal Aligner.

The work was partially funded by the project BIS-21++ at IPP, BAS via contract no. INCO-CT-2005 -016639.

References

[1] http://www.elda.org/intera

[2] TMX standard (Translation Memory eXchange), http://www.lisa.org/tmx/

[3] J.Nerbonne, L.Karttunen, E.Paskaleva, G.Proczeky, T.Roosmaa. Reading more in Foreign Languages, *Fifth Applied Natural Language Processing Conference*, April 1997, Washington, ACL, http://ucrel.lancs.ac.uk/acl/A/A97/A97-1020.pdf

[4] E.Paskaleva, St.Michov, Second Language Acquisition from Aligned Corpora. In: Proc. of the Int. Conf. Language Technology and Language Teaching, Groningen, The Netherlands, April 1997.

[5] N. Genov**.** LORA - a basic tool for creation and primary processing of multilingual Balkan text corpora aligned to English. Presentation on *BIS21++ Information days*, May 2007, Hissar

[6] MULTEXT-East Morphosyntactic Specifications. Version 3.0, May 10th, 2004, http://nl.ijs.si/ME/V3/msd/msd2.pdf

[7] BILEDITA - Bilingual Electronic Dictionaries and Inteligent Text Alignment http://www.cis.uni-muenchen.de/projects/ BILEDITA/leaflet_cover.html

[8] BalkaNet http://www.ceid.upatras.gr/Balkanet/

[9] E. Paskaleva. Compilation and validation of morphological resources. In Workshop on Balkan Language Resources and Tools (Balkan Conference on Informatics), pages 68–74, 2003

[10] INTEX http://www.nyu.edu/pages/linguistics/intex/

[11] А.А.Зализняк. «Грамматический словарь русского языка» (1977, 4-е изд. 2003)

[12] http://www.lml.bas.bg/PROJECTS/BG-EUstand/ eagles/index.html) http://www.lml.bas.bg/PROJECTS/BG-EUstand/.

[13] R. Alfred, E. Paskaleva , D. Kazakov , M. Bartlett**.** Hierarchical agglomerative clustering for cross-language Information Retrieval. *International journal of translation*, Vol. XX, 2007

[14] P.Nakov, V.Pacovski, E.Paskaleva. Extracting Translation Lexicons from Bilingual Corpora: Application to South Slavonic Languages. A common natural language processing paradigm for Balkan languages. Workshop, RANLP-2007, September, 2007