BUILDING LANGUAGE RESOURCES AND TRANSLATION MODELS FOR MACHINE TRANSLATION FOCUSED ON SOUTH SLAVIC AND BALKAN LANGUAGES

Dan Tufiş[1], Svetla Koeva[2], Tomaž Erjavec[3], Maria Gavrilidou[4], Cvetana Krstev[5]

[1]Research Institute for Artificial Intelligence, Romanian Academy, 13, Calea 13 Septembrie, 050711, Bucharest, Romania, tufis@racai.ro
[2]Institute for Bulgarian Language, Bulgarian Academy, 52, Shipchenski prohod, 1113, Sofia Bulgaria, svetla@ibl.bas.bg
[3]Jožef Stefan Institute, Jamova cesta 39, SI-1000, Ljubljana, Slovenia, tomaz.erjavec@ijs.si
[4]Institute for Language and Speech Processing, 6, Artemidos, GR15125, Marousi, Greece, maria@ilsp.gr
[5]University of Belgrade, 16, Studentski trg, 11000, Belgrade, Serbia, cvetana@poincare.matf.bg.ac.yu

ABSTRACT

The paper presents the results of a small and short-term SEE-ERA.net project the purpose of which was to investigate the feasibility of machine translation (MT) research and development for several South Slavic and Balkan languages. For these languages MT systems are scarce and for some of them even non-existent. We argue that by investing efforts in building appropriate language resources, the current technology can be successfully used for a quick development of acceptable MT prototypes, easy to further extend to working systems. The paper describes the parallel corpus compiled in the scope of the project, concentrating on its composition, format, and linguistic analysis. Word-alignments automatically derived from the annotated parallel corpus are also discussed. The paper concludes with direction for further work.

Introduction

Since the seminal work of the IBM group in statistical word-based translation (Brown et al., 1993), new methodologies (memory-based, phrased-based, syntax-based etc.) and techniques (reification, factorization) emerged in multilingual data-driven approaches to machine translation. Yet, several studies underlined the idea that the quality of data to be fed into any machine learning system is of a crucial importance and cannot be compensated by using mass raw multilingual data. In spite of numerous attempts to construct MT systems entirely based on raw parallel data, the evaluations showed that although useful and encouraging results can be obtained in a short period of time, the translation quality can hardly be further improved by increasing the volume of data. The ongoing EuroMatrix project[1] started from this finding and adopted a very promising hybrid approach, combining the strength of rule-based and statistical machine translation and exploiting more and more linguistic knowledge[2]. The Factored Translation Models (Koehn & Hoang, 2007) allow for exploiting, where available, different levels of linguistic pre-processing: lemmatization, part-of-speech tagging, chunking, parsing, word-sense disambiguation, etc. For most of European languages there exist already tools for ensuring the basic pre-processing steps required for a factored translation approach. In fact, with current MT technologies (Och & Ney 2000; 2003; Koehn et al. 2007) which, to a large extent, are language independent, the development of large enough and high quality training data became the critical part of an MT development project.

In this paper we present some results of a small and short-term SEE-ERA.net project[3], the main objective of which was to provide necessary linguistic and technological resources that will foster machine translation RTD for South Slavic and Balkan languages. The partners in the project were from Bulgaria, Greece, Romania, Serbia and Slovenia. Some partners harmonized the objectives of this project with the objectives of other local or bilateral running projects and the project thus includes Czech, French and German as additional languages. Although the project officially ended in July 2008, we hope that this preparatory phase will be followed by another concerted action for further enhancing and exploiting the multilingual resources that have been created.

The Multilingual Data

The Acquis Communautaire is the total body of European Union (EU) law applicable in the EU Member States. This collection of legislative text changes continuously and currently comprises texts written between the 1950s and 2008 in all the languages of EU Member States. Thus, the Acquis Communautaire is a collection of parallel texts in the following

[1] http://www.euromatrix.net/

[2] For a nice demo with Moses MT, which is the basis for the EuroMatrix MT development, see http://demo.statmt.org/webtrans/.

[3] http://dcl.bas.bg/ssbc/home.html

22 languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene and Swedish. A significant part of these parallel texts have been compiled by the Language Technology group of the European Commission's Joint Research Centre at Ispra into an aligned parallel corpus, called JRC-Acquis [6], publicly released in May 2006. In November 2007, the European Commission's Directorate General for Translation (DGT) and the Joint Research Centre (JRC) have made available a multilingual Translation Memory (DGT-TM) of the Acquis Communautaire in the above mentioned official European Union languages. These unique language resources[4] are among the few available parallel corpora containing the languages we were interested in: Bulgarian, Greek, Romanian, Slovene plus Czech, English, French, and German (called further SEE-ERA.net Administrative Corpus - SEnAC). This resource does not yet exist for Serbian, and for that reason an additional resource, based on (called further SEE-ERA.net Literary Corpus - SEnLC), has been compiled.

SEnAC Corpus Construction and Encoding

From the entire JRC-Acquis, which uses the same identifiers (Celex numbers) for the same documents (trailed with the language code); we selected all the documents available in all our target languages. This resulted in a list of 1204 files per language. Since we have noticed several errors in the sentence alignments of the original JRC-Acquis corpus, we re-aligned the 1204 files for Bulgarian, Czech, French, Greek, German, Romanian, Slovenian against the corresponding files in English, using RACAI's SVM sentence aligner (Ceauşu et al. 2006). From the XX-EN aligned sentences, we retained only the 1-1 alignment pairs (more than 99% on average of the total alignments) and each partner had the responsibility to check and correct, if necessary, the sentence alignment. We are not aware of any alignment error for the retained 1-1 XX-EN sentences[5]. Finally we merged the alignments into one XML document, containing 60,389 translation units, each containing one sentence translated in 9 languages, as exemplified in Figure 1.

```
<tu id="3936">
  <seg lang="bg">
      <s id="31985L0337.n.83.1">
          Резултатите от консултациите и информацията , събрана съгласно членове 5,6 и 7 , трябва да
          се вземат предвид при процедурата по издаването на разрешението.</s></seg>
  <seg lang="cs">
      <s id="31985L0337.n.83.1">
          Informace shromážděné podle článků 5 , 6 a 7 musí být brány v úvahu v povolovacím řízení.</s> </seg>
  <seg lang="de">
      <s id="31985L0337.n.85.1">
          Die gemäß den Artikeln 5 , 6 und 7 eingeholten Angaben sind im Rahmen des
          Genehmigungsverfahrens zu berücksichtigen.</s></seg>
  <seg lang="el">
      <s id="31985L0337.n.85.1">
          Οι πληροφορίες που συγκεντρώνονται δυνάμει των άρθρων 5 , 6 και 7 πρέπει να λαμβάνονται υπόψη
          στα πλαίσια της διαδικασίας για τη χορήγηση αδείας.</s></seg>
  <seg lang="en">
      <s id="31985L0337.n.84.1">
          Information gathered pursuant to Articles 5 , 6 and 7 must be taken into consideration in the
          development consent procedure.</s></seg>
  <seg lang="fr">
      <s id="31985L0337.n.83.1">
          Les informations recueillies conformément aux articles 5 , 6 et 7 doivent être prises en considération
          dans le cadre de la procédure d'autorisation.</s></seg>
  <seg lang="ro">
      <s id="31985L0337.n.83.1">
          Informaţiile culese conform art. 5 , 6 şi 7 trebuie să fie luate în considerare în cadrul procedurii de
          autorizare.</s></seg>
  <seg lang="sl">
      <s id="31985L0337.n.83.1">
          Informacije , zbrane skladno s členi 5 , 6 in 7 , se morajo upoštevati v postopku za pridobitev soglasja
          za izvedbo.</s></seg>
```

[4] http://langtech.jrc.it/JRC-Acquis.html

[5] The sentence aligner took advantage of the specific structure of the corpus, and besides the usual sentence delimiters (period, semi-colon, exclamation mark, etc.) we took into account the hard line breaks. This is why, besides proper sentences, the alignments contain pairs of section titles (e.g. ("Article 1" ; Articolul 1), or pairs of dates or locations.

```
</tu>
```

Figure 1: A translation unit from the 9-language parallel corpus

This corpus was then tokenized, tagged and lemmatized by each partner. The tagsets used for all languages (except Bulgarian and German) were compliant with the MULTEXT specifications, for the most part with the MULTEXT-East specifications Version 3[6] (Erjavec 2004) (see http://nl.ijs.si/ME/V3/msd/). The Table 1 shows some statistics concerning the result of the pre-processed corpus:

Language	No. of tokens	Avg no. of tokens/sentence
BG	1436925	23.79
CS	1238981	20.51
DE	1314441	21.76
EL	1469642	24.33
EN	1466912	24.29
FR	1527241	25.29
RO	1422995	23.56
SL	1271011	21.04

Table 1: Statistical data on the compiled parallel corpus

After tokenization, tagging and lemmatization, this annotation was added to the XML encoding of the parallel corpus. Depending on the available processing tools for different languages, additional information could be added to each language-specific segment of a translation unit. Figure 2 shows the representation of the Romanian segment of the translation unit displayed in Figure 1.

```
<tu id="3936">
...
        <seg lang="ro">
                <s id="31985L0337.n.83.1">
                        <w lemma="informaţie" ana="Ncfpry">Informaţiile</w>
                        <w lemma="culege" ana="Vmp--pf">culese</w>
                        <w lemma="conform" ana="Spsd">conform</w>
                        <w lemma="art." ana="Yn">art.</w>
                        <w lemma="5" ana="Mc">5</w>
                        <c>,</c><w lemma="6" ana="Mc">6</w>
                        <w lemma="şi" ana="Crssp">şi</w>
                        <w lemma="7" ana="Mc">7</w>
                        <w lemma="trebui" ana="Vmip3s">trebuie</w>
                        <w lemma="să" ana="Qs">să</w>
                        <w lemma="fi" ana="Vasp3">fie</w>
                        <w lemma="lua" ana="Vmp--pf">luate</w>
                        <w lemma="în" ana="Spsa">în</w>
                        <w lemma="considerare" ana="Ncfsrn">considerare</w>
                        <w lemma="în_cadrul" ana="Spcg">în_cadrul</w>
                        <w lemma="procedură" ana="Ncfsoy">procedurii</w>
                        <w lemma="de" ana="Spsa">de</w>
                        <w lemma="autorizare" ana="Ncfsrn">autorizare</w>
                        <c>.</c>
                </s></seg>
...
 </tu>
```

Figure 2: A linguistically analysed sentence in a language-specific segment of a translation unit

[6] http://nl.ijs.si/ME/V3/msd/

SEnLC Corpus Construction and Encoding

One reason that we have chosen Jules Verne's novel is that this text is available in digital form for many of the languages that we were interested in. Moreover, for the majority of these languages lexical resources exist in the same format, which enables comparable processing of the text in different languages. Translation of the novel in sixteen languages have been acquired, namely: French, English, German, Spanish, Portuguese, Italian, Romanian, Russian, Serbian, Croatian, Bulgarian, Macedonian, Polish, Slovenian, Hungarian and Greek. Not all of these texts have yet been aligned; alignment was done for the five Balkan languages, French original and English.

In the preparatory phase each translation was marked in accordance with the TEI-standard in XML, and the title (<head>), paragraph (<p>) and "sentence" (<seg>) were included as units of text logical layout. Before alignment, each text was transformed to the TEI-conformant format[7]. The XAlign system[8] was used for the alignment process. Starting from the French version, the goal of the alignment was to establish 1:1 relations on the segment level (<seg> tag) with all other languages. In order to achieve this goal segments had to be further divided. So, the total number of segments in all texts is 4409. This type of text alignment of bitexts required an intensive manual control of the output of the XAlign system. In this way, the missing segments or the inconsistencies between the source text and its translations were also identified.

```
<tu id="n569">
<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">
          tiţi cã formalitatea vizei e inutilă şi că noi nu mai cerem prezentarea paşaportului.</s>
```

Figure 3: One sentence from a 7-language corpus of Verne's novel: French original, English as a hub language, and five South Slavic and Balkan languages

The total number tokens in this text in French is 71,793, while the total number of unique tokens (types) is 9,433 (ratio 7.6). The figures for other languages are different, e.g. for Serbian the total number of tokens is 58,722, while the total number of types is 12,733 (ratio 4.6), for Bulgarian the total number of tokens is 58,678, while the total number of types is 11,217 (ratio 5.2), while for Greek the total number tokens is 68,615, and the total number of types is 11,809 (ratio 5.8).

For the present, all the language versions of this corpus for which DELA type lexical resources exist were tagged, but disambiguation has been done for Serbian and Bulgarian. The initial tagsets were those used in the corresponding lexical resources, but they were latter mapped into MULTEXT-East specifications (Krstev et al. 2004). After tagging and lemmatization, this annotation information was added to the XML encoding of the parallel corpus. Figure 4 shows the representation of the Serbian segment of the translation unit displayed in Figure 3:

[7] http://www.tei-c.org/index.xml

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

```
<tu id="n569">
        <seg lang="sr">
                <s id="Verne80days.n569">
                        <w lemma="vi" ana="Pp2-pn">Vi</w>
                        <w lemma="znati" ana="Vm-p2p-an-n---p">znate</w>
                        <w lemma="da" ana="C-s">da</w>
                        <w lemma="jesam" ana="Va-p3s-an-y---p">je</w>
                        <w lemma="ovaj" ana="Pd-fsn">ova</w>
                        <w lemma="formalnost" ana="Ncfsn--n">formalnost</w>
                        <w lemma="viziranje" ana="Ncnsg--n">viziranja</w>
                        <w lemma="izlišan" ana="Afpfs1">izlišna</w>
                        <w lemma="i" ana="C-s">i</w>
                        <w lemma="da" ana="C-s">da</w>
                        <w lemma="se" ana="Q-">se</w>
                        <w lemma="više" ana="Rgp">više</w>
                        <w lemma="ne" ana="Q-">ne</w>
                        <w lemma="tražiti" ana="Vm-p3s-an-n---p">traži</w>
                        <w lemma="pokazivanje" ana="Ncnsn--n">pokazivanje</w>
                        <w lemma="isprava" ana="Ncfpg--n">isprava</w>
                        <c>?</c>
                </s></seg>
</tu>
```

Figure 4: A tagged and a lemmatized sentence from the Serbian version of Verne's novel

Word-Alignment of the SEnAC

Based on the pre-processing discussed in the previous section, we built, using GIZA++ (Och & Ney 2003) 8 unidirectional translation models (EN-RO, RO-EN, EN-BG, BG-EN, EN-SL, SL-EN, EN-GR, GR-EN). The processing unit considered in each language was not the wordform but the string formed by its lemma and the first two characters of the associated morphosyntactic tag (e.g. for the wordform "informaţiile" we took the item "informaţie/Nc"). We used for each language 20 iterations (5 for Model 1, 5 for HMM, 1 for THTo3, 4 for Model3, 1 for T2To4 and 4 for Model4). We did not include Model 5 nor Model 6 as we noticed a degradation of the perplexities. Given the formulaic language used by the Acquis-Communautaire documents, the perplexities of the resulting language models were encouraging, and range from 13.07 (RO-EN) to 19.88 (EN-BG). Based on these models we word-aligned the bitexts using the iterative high precision COWAL aligner (Tufiş et al. 2006). As described in (Tufiş et al. 2006), translation pairs prescribed by each unidirectional translation model were unconditionally included in the alignment skeleton. The rest of the links were established in the subsequent iterations of the aligner. The training corpora SEnAC, the alignments and the perplexities for each translation model are available on the project's site. Additionally, an alignment viewer and editor (see Figure 5) was implemented for allowing the visualization and correction of the alignments with the purpose of further fine-tuning the translation models.

At the time of this writing, the translation model for the RO and EN has been used for some preliminary experiments in translation. The results are very encouraging and in Table 2 we provide a translation example (from Romanian into English), as produced by our translation system prototype and by Google's MT. The input text is the following:

articolul 1
utilizând vehicule înmatriculate pe teritoriul unei părţi contractante , care datorită construcţiei şi echipamentelor lor sunt adecvate pentru a transporta mai mult de nouă persoane , inclusiv conducătorul auto , şi sunt destinate acelui scop ; în înţelesul prezentului acord , serviciile internaţionale înseamnă serviciile care tranzitează teritoriul a cel puţin două părţi contractante . în înţelesul prezentului acord , termenul teritoriul unei părţi contractante , acoperă , din punctul de vedere al comunităţii economice europene , acele teritorii unde se aplică tratatul de instituire a comunităţii respective şi în condiţiile stabilite în tratatul respectiv .

Our translation	Google translation
article 1 using vehicles registered in the territory of a contracting party , which because of the construction and their equipment are appropriate to carry more than nine persons , including the driver , and are intended for that purpose ; within the meaning of this agreement , the services of the international means services in transit through the territory of at least two contracting parties . within the meaning of this agreement , the term the territory of a contracting party , shall cover , from the point of view of the european economic community , those territories which shall apply to the treaty establishing the community in question and under the conditions laid down in the treaty in question .	Article 1 using vehicles registered in the territory of a contracting party, which due to their construction and equipment are adequate to carry more than nine persons including the driver, and are intended for that purpose; the meaning of this Agreement, international services means services transiting the territory of at least two contracting parties. the meaning of this Agreement, the term territory of a contracting party, cover, in terms of European economic community, those areas where applicable treaty establishing the community and the conditions laid down in the treaty.

Table 2: An example of our translation using the SEnAC RO-EN translation model vs. Google's MT

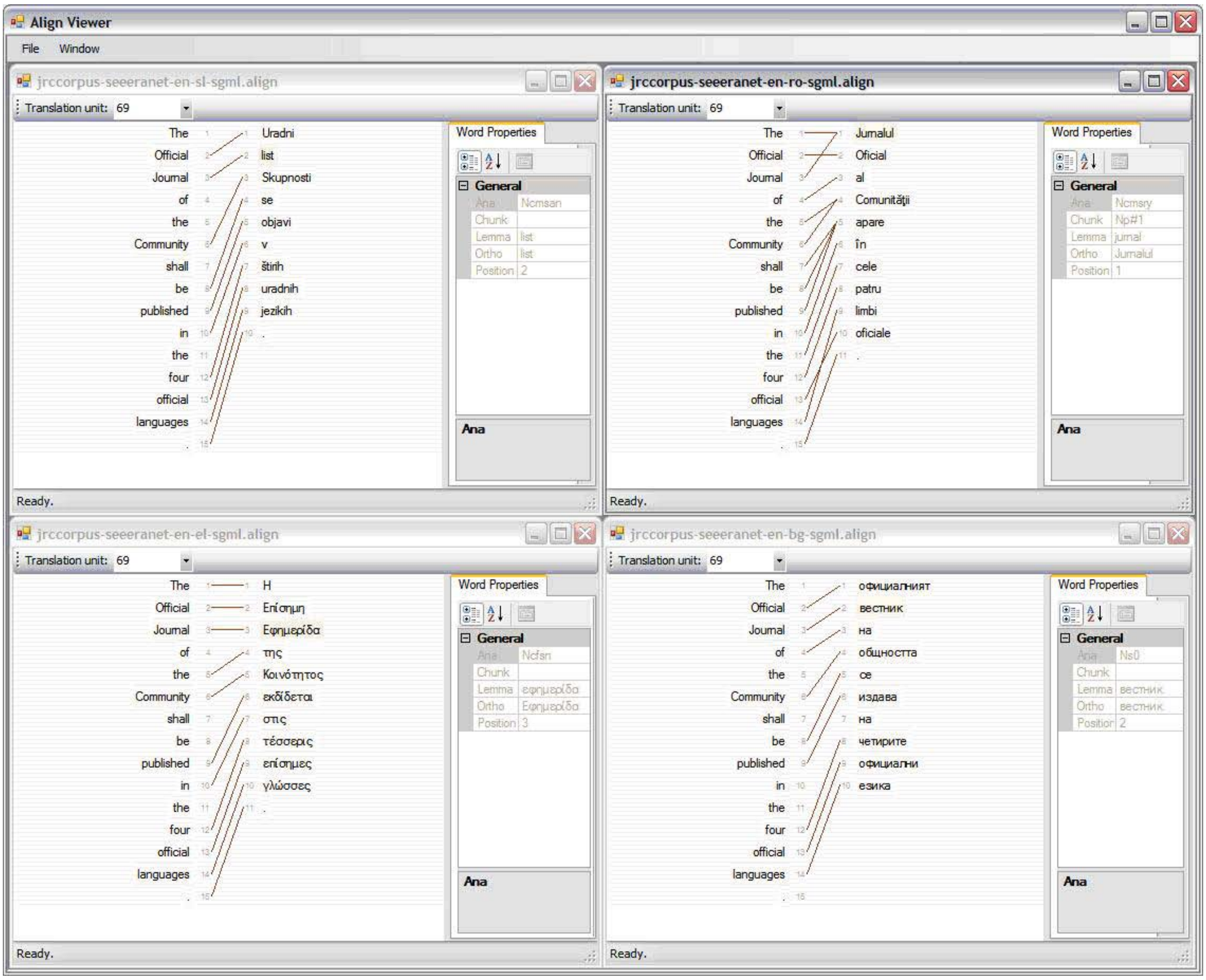
Align Viewer
File Window
jrccorpus-seeeranet-en-sl-sgml.align
Translation unit: 69
The Official Journal of the Community shall be published in the four official languages .
Uradni list Skupnosti se objavi v štirih uradnih jezikih .
Word Properties
General
Ana Ncmsan
Chunk
Lemma list
Ortho list
Position 2
Ana
Ready.
jrccorpus-seeeranet-en-ro-sgml.align
Translation unit: 69
The Official Journal of the Community shall be published in the four official languages .
Jurnalul Oficial al Comunităţii apare în cele patru limbi oficiale .
Word Properties
General
Ana Ncmsry
Chunk Np#1
Lemma jurnal
Ortho Jurnalul
Position 1
Ana
Ready.
jrccorpus-seeeranet-en-el-sgml.align
Translation unit: 69
The Official Journal of the Community shall be published in the four official languages .
H Επίσημη Εφημερίδα της Κοινότητος εκδίδεται στις τέσσερις επίσημες γλώσσες .
Word Properties
General
Ana Ncfsn
Chunk
Lemma εφημερίδα
Ortho Εφημερίδα
Position 3
Ana
Ready.
jrccorpus-seeeranet-en-bg-sgml.align
Translation unit: 69
The Official Journal of the Community shall be published in the four official languages .
официалният вестник на общността се издава на четирите официални езика
Word Properties
General
Ana Ns0
Chunk
Lemma вестник
Ortho вестник
Position 2
Ana
Ready.

Figure 5: The alignment of a translation unit (no. 69) in four bitexts

Future work

The presented work is still in progress. We plan to conduct experiments with the other models for translating from the XX language into English. The other direction (EN-XX) will be also considered. We plan to extend the experiments for other language pairs present in SEnAC corpus It is obvious that due to the same level of annotation we could experiment with any of the XX-YY language pair in SEnAC corpus, but we are equally interested in studying the effect on translation models building by using word and phrase alignments derived from a pivot/hub language alignment. We described in (Tufiş & Koeva 2007) a system for automatically deriving from pivot alignments PIVOT-X1 and PIVOT-X2 word alignments the X1-X2 alignment.

Future work will address the more challenging task on building translation models from the SEnLC literary corpus and provided adequate data will be available, experiments with other South East and Balkan languages.

References

Brown et al. 1993. Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert J. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2): 263–311.

Ceauşu et al. 2006. Alexandru Ceauşu, Dan Ştefănescu, Dan Tufiş. 2006. Acquis Communautaire sentence alignment using Support Vector Machines. In *Proceedings of the 5th LREC Conference*, Genoa, Italy.

Erjavec 2004. Erjavec, T. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation*, LREC'04, pp. 1535 - 1538, ELRA, Paris.

Krstev et al. 2004. Cvetana Krstev, Duško Vitas, Tomaž Erjavec. Morpho-Syntactic Descriptions in MULTEXT-East - the Case of Serbian, in Informatica, No. 28, pp. 431-436, The Slovene Society Informatika, Ljubljana.

Koehn & Hoang 2007. Koehn Philipp, and Hieu Hoang. *Factored Translation Models*. EMNLP.

Koehn et al. 2007. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

Och & Ney 2000 2000. Franz J. Och, Herman Ney. Improved Statistical Alignment Models. In *Proceedings of the 38th Conference of ACL*, Hong Kong: 440-447.

Och & Ney 2003. Franz J. Och, Herman Ney. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1):19-51

Steinberger et al. 2006. Ralph Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş.. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, 2006, pp.2142-2147

Tufiş et al. 2006. Dan Tufiş, Radu Ion, Alexandru Ceauşu, Dan Ştefănescu. Improved Lexical Alignment by Combining Multiple Reified Alignments. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006), Trento, Italy, 3-7 April, 2006, pp. 153-160.

Tufiş & Koeva 2007. Dan Tufiş, Svetla Koeva: "Ontology-supported Text Classification based on Cross-lingual Word Sense Disambiguation". In Francesco Masulli, Sushmita Mitra and Gabriella Pasi (eds.). Applications of Fuzzy Sets Theory. 7th International Workshop on Fuzzy Logic and Applications, WILF 2007, Camogli, Italy", LNAI 4578, Springer-Verlag Berlin Heidelberg, 2007, pp. 447-455