Archives of Control Sciences
Volume 15(LI), 2005
No. 4, *pages 529–540*

Massive multi lingual corpus compilation: Acquis Communautaire and totale

TOMAŽ ERJAVEC, CAMELIA IGNAT, BRUNO POULIQUEN and RALF STEINBERGER

Large, uniformly encoded collections of texts, corpora, are an invaluable source of data, not only for linguists, but also for Language Technology tools. Especially useful are multilingual parallel corpora, as they enable, e.g. the induction of translation knowledge in the shape of multilingual lexica or full-fledged machine translation models. But parallel corpora, esp. large ones, are still scarce, and have been, so far, difficult to acquire; recently, however, a large new source of parallel texts has become available on the Web, which contains EU law texts (the Acquis Communautaire) in all the languages of the current EU, and more, i.e. parallel texts in over twenty different languages. The paper discusses the compilation of this text collection into the massively multilingual JRC-Acquis corpus, which is freely available for research use. Next, the text annotation tool "totale", which performs multilingual text tokenization, tagging and lemmatisation is presented. The tool implements a simple pipelined architecture which is, for the most part, fully trainable, requiring a word-level syntactically annotated text corpus and, optionally, a morphological lexicon. We describe the MULTEXT-East corpus and lexicons, which have been used to train totale for seven languages, and the application of the tool to the Slovene part of the JRC-Acquis corpus.

Key words: multilingual corpora, EU languages, multilingual linguistic analysis, tokenisation, part-of-speech tagging, lemmatisation

1. Introduction

Large, uniformly encoded collections of texts and their translations - parallel multilingual corpora - ([10], [1], [3], [9], [8]) are a prime resource for the development of multilingual language technologies. Serving as training datasets for inductive programs, they can be used to learn models for machine translation, cross-lingual information retrieval, multilingual lexicon extraction, sense disambiguation, etc. The value of a parallel corpus grows with the following characteristics:

T. Erjavec works at the Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia, http://nl.ijs.si/et/, e-mail: tomaz.erjavec@ijs.si. C. Ignat, B. Pouliquen and R. Steinberger work at the European Commission – Joint Research Centre, I - 21020 Ispra (VA), Italy, http://www.jrc.it/langtech/, e-mail: Firstname.Lastname@jrc.it

The work presented in this paper was in part supported by the EU projects PASCAL and ALVIS, and by the first author's stay at the JRC.

Received 25.10.2005

- *Size*: larger corpora give not only statistically more reliable counts, but also reveal phenomena that are completely lacking in smaller samples.
- *Number of languages*: the utility here grows quadraticaly with the number of languages, as each language can be paired with any other. While bi-lingual corpora usually contain at least one 'major' language, larger multilingual collections will also contain pairings of less common languages, where such a resource is of great value (Maltese-Finish for example).
- *Linguistic annotation*: can be used as a normalisation step on the raw text, hence reducing the complexity (search space) for the LT task; or for enabling multiple knowledge of the text (e.g. morphosyntactic tags, collocations, predicate-argument structure) to be exploited.
- *Semantic annotation*: refers to the classification of documents (or their parts, e.g. words) into some hierarchy of concepts, which can be used to access the data (the Semantic Web paradigm)

This paper discusses the compilation of a large, massively multilingual corpus, where each document is classified according to a rich ontology. The corpus is freely available for research purposes. First experiments have also been performed on sentence alignment on the corpus and in annotating it with word-level syntactic information.

The rest of this paper is structured as follows: Section 2 introduces the EU ACQUIS text collection and the steps performed in turning it into an XML encoded corpus, the JRC-Acquis; current experiments in sentence alignment are also presented. Section 3 describes the text annotation tool totale, a trainable program, which performs multilingual text tokenization, tagging and lemmatisation; we explain the architecture of the program, the MULTEXT-East dataset used to train totale for seven languages and report on using the tool on the Slovene portion of the ACQUIS. Finally, Section 4 gives the conclusions and discusses future work.

2. The EU ACQUIS parallel corpus

The core EU law, variously known as the Acquis Communautaire, is comprised of 8 to 13 million running words of texts depending on the language. This collection of documents, some dating back to the 1950s, has been for a while translated into the eleven languages of the 'pre-enlargement' EU. For the last six years, the candidate countries have been translating them into their languages - this was one of the conditions to enable their accession to the EU. This process has by now been mostly completed, and, what is more, the complete set of documents has been recently made available in HTML on the Web.[1]

[1] http://europa.eu.int/

Language	Number of texts	Number of characters	Number of words	Average length of texts	Average number of words
Czech	6,304	47,380,160	7,310,147	7,515	1,159
Danish	8,099	70,526,322	10,330,345	8,708	1,275
German	8,149	83,845,850	11,628,856	10,289	1,427
Greek	8,003	**84,232,323**	13,073,101	**10,525**	**1,633**
English	**8,183**	72,363,833	12,007,560	8,843	1,467
Spanish	8,121	80,669,741	13,201,129	9,933	1,625
Estonian	7,009	53,194,338	6,751,386	7,589	**963**
Finnish	7,774	69,268,332	7,999,785	8,910	1,029
French	8,134	78,464,509	**13,113,163**	9,646	1,612
Hungarian	5,506	49,798,572	6,596,073	9,044	1,197
Italian	8,176	78,116,731	12,093,677	9,554	1,479
Lithuanian	6,073	48,221,853	6,461,944	7,940	1,064
Latvian	7,545	58,130,835	8,239,245	7,704	1,092
Maltese	**5,041**	**39,988,877**	6,574,607	7,932	1,304
Dutch	8,167	78,864,983	12,049,749	9,656	1,475
Polish	6,552	55,441,985	7,636,388	8,461	1,165
Portugese	8,088	79,323,159	13,067,222	9,807	1,615
Slovak	5,551	41,379,372	**6,191,172**	7,454	1,115
Slovene	7,772	57,852,722	9,133,019	**7,443**	1,175
Swedish	7,877	72,898,994	10,998,571	9,254	1,396

Tab. 1 Size of the corpus; minimum and maximum values in each column are in bold

Such a text collection is unprecedented in terms of size, the number of languages involved and access, being freely available on the Web.[2] Furthermore, each of the texts has also been manually classified according to the EUROVOC thesaurus,[3] a large multilingual 'ontology' being used for manual document classification by various European parliaments and other organisations, including the European Parliament and the European Commission. A corpus compiled from this text collection could thus be exploited not only for machine translation research, but also for "Semantic Web" experiments in, say, automatic document classification [11], or cross-lingual document similarity [12]. It is for these reasons that we proceeded with compiling the ACQUIS corpus, as explained in the remainder of this section.

The first version of the JRC-Acquis corpus contains 20 languages, 146,000 texts and 194 million running words. There are 5,000 to 8,000 texts per language, with each text being an average of 1,000 to 1,600 words in length (Table 1). To our knowledge, the JRC Collection of the Acquis Communautaire is the only parallel corpus of its size available in so many languages. To further research on Language Technology, esp. for

[2]A corpus based on a similar text collection, EUROPARL (http://www.isi.edu/ koehn/europarl/, [9]), contains 29 million words of original and translated debate transcripts from the European Parliament. Although it contains more text per language than does the ACQUIS, latter contains more languages, and is indexed with EUROVOC descriptors.

[3]http://europa.eu.int/celex/eurovoc/

the less well studied languages, the JRC-Acquis corpus is available for downloading at http://wt.jrc.it/lt/acquis/.

2.1. Compiling the corpus

The process of compiling the corpus consisted of the following steps:

1. *downloading* the texts: the interface enables locating the texts via their CELEX ID (unique identifier given for every EU official document); the copying was then a matter of querying over these IDs for all the languages; however, not all documents (IDs) are translated into each language, so the size of the various language parts varies;

2. *language identification* on the documents: for a few percent of documents, text purportedly in one language is in fact untranslated English text - such cases are not made part of the corpus;

3. *wrapper induction*: the texts can be usefully decomposed into the title, body of the text, the signature (e.g. "*Done at Brussels, 24 September 1989, for the commission*", etc.), and annexes (containing tables or lists of codes, usually not translated in all languages). It is the body that will contain most of the 'useful' text, yet the back-matter can comprise a considerable portion of the documents. These divisions were identified by Perl regular expressions over the texts, and the resulting "level 0" corpus was stored as XML;

4. *linguistic annotation* of the texts: sentence, word and punctuation tags were added to the corpus, and the words given their context disambiguated lemma and morphosyntactic attributes; this processing, so far only for a limited number of the language components of the corpus, was performed by the program totale, described in the Section 3;

5. *paragraph alignment*: paragraphs were given IDs, and (initial) alignment files made over language pairs of documents; the current experiments are described below.

2.2. Alignment

We have performed an experiment in language independent paragraph alignment of the English-Slovene pair, using the Vanilla aligner [6]. This aligner implements dynamic time warping by comparing the character counts of possibly aligned sentences [4]. The aligner is given the two files split into hard regions, which have to match among the files (in our case each document text corresponds to one hard region), and soft regions which are aligned 0-1, 1-0, 1-1, 2-1, 1-2, and 2-2. Soft regions are typically sentences, but in our case paragraphs, which do, however, tend to be rather short corresponding to one or two sentences or even partial sentences. An evaluation of the results showed that:

- The alignment is complicated by the fact that some English documents on the Web are not the versions that served as the source for the translation, e.g. they are a later/previous version with some ammendments. The size of the amendments in terms of text percentage is usually not that large, but it does raise the error rate of the aligner significantly.
- The number of 1-1 links among the paragraphs is approx 90%. As these links are highly reliable, this means that, with an added heuristic or two, it would be simple to achieve (almost) 100% precise alignments at the cost of sacrificing approximately one fifth of the text, i.e. settling for 80% recall. This still leaves ample text for the aligned corpus.
- It would be relatively easy to introduce a pre-processing step that would take into account enumeration tokens (e.g. 1), a),...) and declare them as the hard regions for the aligner. This would most likely significantly localise and thus reduce the alignment errors.

3. Multilingual tokenisation, tagging, and lemmatisation

Corpora can be annotated with various linguistic annotation, such as syntactic structure, anaphora and their referents, terms, names, etc., but the basic steps for all such annotations are the following:

1. *tokenisation*, which identifies words and punctuation in the text;
2. *part-of-speech tagging*, which assigns context-disambiguated word-level syntactic descriptions to the words, e.g. determines that the Slovene 'gledata' is a verb in the second person dual present tense indicative;
3. *lemmatisation* (or stemming) which assigns the base (uninflected) form to a word, e.g. 'gledati'.

We have developed a tool, named totale that performs the above steps in a multilingual setting. The main feature of the program is that both of the more complex, i.e. language specific and knowledge intensive modules of totale (2. & 3.) are learning programs, i.e. they induce the model of a language from pre-annotated data (corpus and lexicon) and are robust, i.e. they know how to deal with unknown words, a must for any application dealing with unrestricted text. The program is written in Perl and implements a simple pipe-lined architecture, where plain Unicode (UTF-8) text is first tokenised, the word tokens (word-forms) then tagged with the appropriate morphosyntactic description (MSD), and the word-forms, given their MSD, lemmatised to arrive at the canonical form of the word. The architecture of the program is given in Fig. 1. The program can produce output in several formats, in particular in tabular form or encoded in TEI-compliant XML.

In Fig. 2 we give a sample invocation of the program. The tabular output consists of four columns: the first lists the tokens as they appear in the input text; the second

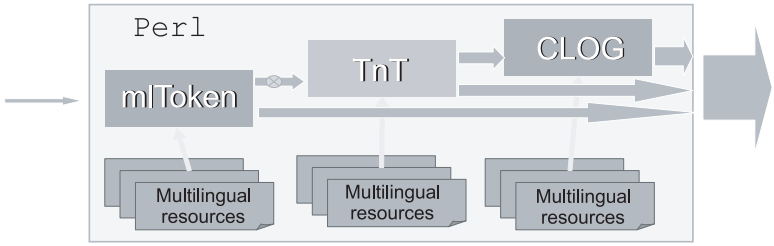

Figure 1. Architecture of the totale annotation tool: the three modules (tokeniser, tagger, and lemmatiser) combine their output to produce the output

```
$ totale -l en
Doctor, can you help?
^D
 <TEXT>
Doctor TOK doctor Ncfs
,  PUN
can  TOK can Voip
you TOK you Pp2
help TOK help Vmn
? PUN_TERM
  <S/>
  </TEXT>

$ totale -l sl -f xml
Kapucini in zdravniki so se pojavili na vseh koncih in krajih.
^D
<text>
<w lemma="kapucin" ana="Ncmpn">Kapucini</w>
<w ana="Ccs">in</w>
<w lemma="zdravnik" ana="Ncmpn">zdravniki</w>
<w lemma="biti" ana="Vcip3p--n">so</w>
<w ana="Px------y">se</w>
<w lemma="pojaviti" ana="Vmps-pma">pojavili</w>
<w ana="Spsl">na</w>
<w lemma="ves" ana="Pg-mpl----a">vseh</w>
<w lemma="konec" ana="Ncmpl">koncih</w>
<w ana="Ccs">in</w>
<w lemma="kraj" ana="Ncmpl">krajih</w>
<c type="TERM">.</c>
<s/>
</text>
```

Figure 2. Output of totale; the first processes English text and outputs it in tabular format, the second Slovene text, with much richer morphology, and outputs it in TEI/XML

contains the token type or the tag marking the end of the sentence or other recognised structure; the third the lemmas of the words; and the fourth their MSDs. The second example invocation shows that the program can also produce XML formatted output. The program is not extremely fast, i.e. it processes about 5,000 words per minute. This is partially due to the system architecture of file-mediated sequential processing, but mostly the fault of the lemmatisation module, which needs to load and use thousands of rules and exceptions encoded as if-then-else rules. The program is available for on-line experimentation at http://nl2.ijs.si/analyze/.

3.1. The tokenisation module

The multilingual tokenisation module mlToken is written in Perl, and, in addition to splitting the text input string into tokens has also the following features:

- Assigns to each token its token type. The types distinguish not only between words and punctuation marks but also mark digits, abbreviations, left and right splits (i.e. clitics, e.g. 's , enumeration tokens (e.g. a)) as well as URLs and email addresses.
- Marks end of paragraphs, and end of sentence punctuation, where sentence internal periods are distinguished from sentence final ones.
- Preserves (subject to a flag) the inter-word spacing of the original document, so that the input can be reconstituted from the output - this consideration is important when several tokenisers are applied to a text, either for evaluation or production purposes.

The model for our tokeniser was mtseg, the tokeniser (and segmenter) developed in the MULTEXT project [5] as with mtseg, mlToken also stores the language dependent features in resource files, in the case of mlToken of abbreviations and split/merge patterns. In the absence of a certain language resource, the tokeniser uses default resource files - in order to achieve best results, however, resource files for a language have to be written - this task is helped by having pre-tokenised corpora for the language.

```
študiji:7 Ncfda:1 Ncfdn:1 Ncfsd:1 Ncfsl:1 Ncmpi:2 Ncmpn:1
šum:5  Ncmsa--n:1 Ncmsn:4
šume:3  Ncmpa:2 Rgp:1
šumeče:9 Afpfpa:1 Afpfpn:1 Afpfsg:3 Afpmpa:1 Afpnsa:1 Afpnsn:1 Rgp:1
šumečo:3 Afpfsa:2 Afpfsi:1
šumi:5  Ncmpi:1 Ncmpn:1 Vmip3s--n:2 Vmmp2s:1
…
Px------y:2226
 Vcps-sma:4
  Vmps-sma:2
  Rgp:2
        Vcip3s-n:794
                Vcps-sma:2
                Vcip3s-n:1
                ,:72
                Aopmsn:2
```

Figure 3. The language resources for the TnT tagger: lexicon with wordform and ambiguity class (with frequencies) and 1,2,3-grams of MSDs (so, the 3-gram Px——y Vcps-sma Vmps-sma, corresponding to the reflexive pronoun followed by copula and main verb in a certain form appears twice)

3.2. The tagging module

For tagging words in the text with their context disambiguated morphosyntactic annotations we used a third-party tagger, namely TnT [2], a fast and robust tri-gram tagger. TnT is freely available for research purposes (but distributed only in compiled code for Linux), has an unknown-word guessing module, and is able to accommodate the large morphosyntactic tagsets that we find in various EU languages.

The tagger uses two resources, namely a lexicon giving the weighed ambiguity class for each word and a table of tri-grams of tags with weights assigned to the uni-, bi-,

and tri-grams; examples from the Slovene lexicon and the n-gram table are given in Fig. 3. Both resources are acquired from a pre-annotated corpus. The automatically induced lexicon can also be expanded with previously available lexicons.

3.3. The lemmatisation module

Automatic lemmatisation is a core application for many language processing tasks. In inflectionally rich languages, such as Slovene, assigning the correct lemma (base form) to each word in a running text is not trivial, as, for instance, nouns inflect for number and case, with a complex configuration of endings and stem modifications. The problem is especially difficult for unknown words, as word-forms cannot be matched against a morphological lexicon.

For our lemmatiser we used CLOG ([13], [7]), which implements a machine learning approach to the automatic lemmatisation of (unknown) words. CLOG learns on the basis of input examples (pairs word-form/lemma, where each MDS is learnt separately) a first-order decision list, essentially a sequence of if-then-else clauses, where the defined operation is string concatenation. The learnt structures are Prolog programs, but in order to minimise interface issues we made a converter from the Prolog program into one in Perl. In the final instance the usage for determining the lemma is simply the result of the function call *$lemma = lemmatise($msd,$wordform)*; This function then calls the appropriate rule-set, which transforms the input wordform into its lemma. We give in Fig. 4 an example of an induced rule for the Slovene MSD denoting the feature structure PoS:Adjective, Type:qualificative Degree:comparative, Gender:feminine, Number:dual, Case:accusative.

```
$sub{'Afcfda'}='SUB_afcfda';
sub SUB_afcfda {
 my $w  = $_[0]; my $lem;
 if ($w=~/^(.*)svetlej#353i$/){$lem=$1."svetel"}
 elsif ($w=~/^(.*)polnej#353i$/){$lem=$1."poln"}
 elsif ($w=~/^(.*)b#353i$/) {$lem=$1."b"}
 elsif ($w=~/^(.*)elej#353i$/) {$lem=$1."el"}
 elsif ($w=~/^(.*)ivej#353i$/) {$lem=$1."iv"}
 elsif ($w=~/^(.*)anej#353i$/) {$lem=$1."an"}
 elsif ($w=~/^(.*)kej#353i$/) {$lem=$1."ek"}
 elsif ($w=~/^(.*)tej#353i$/) {$lem=$1."t"}
 elsif ($w=~/^(.*)i#382ji$/) {$lem=$1."izek"}
 elsif ($w=~/^(.*)enej#353i$/) {$lem=$1."en"}
 elsif ($w=~/^(.*)rej#353i$/) {$lem=$1."er"}
 elsif ($w=~/^(.*)nej#353i$/) {$lem=$1."en"}
 else {$lem="???"}
 return $lem;
}
```

Figure 4. An induced lemmatisation rule in Perl for the Slovene MSD: PoS:Adjective, Type:qualificative Degree:comparative, Gender:feminine, Number:dual, Case:accusative.

3.4. MULTEXT-East resources

The main feature of totale is that it is multilingual and trainable for new languages, as the models for tagging and lemmatisation are induced from data. However, in order to make the tool useful, we first have to obtain such data, namely morphosyntactically

annotated corpora and lexicons. It is an added advantage if the multilingual training resources all follow the same guidelines for tagset and corpus annotation design.

The MULTEXT-East language resources, a multilingual dataset for language engineering research and development, first developed in the scope of the EU MULTEXT-East project, have now already reached the 3rd edition [8]. MULTEXT-East is a freely available standardised (XML/TEI P4, [14]) and linked set of resources, and covers a large number of mainly Central and Eastern European languages. It includes the EAGLES-based morphosyntactic specifications, defining the features that describe word-level syntactic annotations; medium scale morphosyntactic lexicons; and annotated parallel, comparable, and speech corpora. The most important component is the linguistically annotated corpus consisting of Orwell's novel "1984" in the English original and translations.

For training totale we used resources for Czech, English, Estonian, Hungarian, Romanian, Serbian, and Slovene. The MULTEXT-East mtseg resource files were used as sources for the mlToken resource files; the annotated corpus for training the TnT tagger; and the lexicons to improve the performance of the tagger and for training the CLOG lemmatiser. While training the tagger on this data is very fast, training the lemmatiser is much more process intensive, as each MSD is learned separately - so, for Slovene or Czech, this meant leaning around 1,000 different classes for a language, and the training time is measured in days.

Corpus elements		Corpus word types		Lexicon	
<text>	7,771	<w>	15,934,003	Entries	381,068
<signature>	7,683	#IMPLIED	14,393,953	Wordforms:	221,876
<annex>	3,658	DIG	1,036,076	Lemmas:	154,241
<P>	1,063,577	ENUM	331,426	MSDs:	970
<c>	2,865,307	ABBR	159,022		
<w>	15,934,003	MW	11,048	Corpus size	144 MB

Tab. 2 The Slovene portion of the ACQUIS: the number of different XML elements in the corpus; number of words (type = plain, digit, enumeration, abbreviation, multi-word unit); and size of MULTEXT lexicon, with number of all entries, of different word-forms, lemmas and morphosyntactic descriptions

3.5. The annotated Slovene ACQUIS

In this section we report on linguistically annotating the Slovene part of the corpus with totale. To process the corpus we wrote a wrapper Perl program that, for each file:

- extracted all the text from the XML document (all <P>s except first, which is the – often untranslated – official document name),
- piped the text to totale -l sl -f XML,
- substituted the contents of original <P>s with the totalised ones,
- validated the result against a DTD.

The size of the Slovene portion of the corpus and its vocabulary identified via the annotations is given in Table 2. The Corpus part gives the tag counts of the XML files; we

can see that the corpus has about 1 million paragraphs and 16 million words; of these 14 million are "normal" words. On the basis of these, a MULTEXT type lexicon was produced, where each entry consists of the triplet wordform/lemma/MSD. The corpus yields 380,000 such entries, with 220,000 distinct wordforms, and 150,000 lemmas; there are almost one thousand different MSDs used in the corpus.

```
8 rafinacija rafinacija Ncfsn
2 rafinacije rafinacija Ncfpa
40 rafinacije rafinacija Ncfsg
2 rafinacijol5 rafinacijol5 Mc---d
26 rafinaciji rafinacij Npmpn
9 rafinaciji rafinacija Ncfsl
17 rafinacijo rafinacija Ncfsa
```

Figure 5. An induced lemmatisation rule in Perl for the Slovene MSD: PoS:Adjective, Type:qualificative Degree:comparative, Gender:feminine, Number:dual, Case:accusative.

We also performed a preliminary evaluation of the results on the basis of this lexicon. Fig. 5 gives a stretch from the lexicon of a lemma unknown to the system, "*rafinacija*", with erroneous analyses crossed out. One error (line fifteen) is to do with the tokenisation, or, rather, with the poor quality of the HTML original. Lemmatisation is wrong once (but, unfortunately in 26 cases); the error originates in the incorrect MSD assignment, which specifies the noun as masculine plural nominative, where it is in fact feminine and singular locative. Finally, there is one 'minor' error, in line 2, where the tagger assigns the plural number, where it was in fact singular.

A more longitudal evaluation suggests that the greatest problem with annotated corpus is, in fact, not the quality of lemmatization per se, but rather the lacking support for identification of foreign words, and better handling of proper names, abbreviations and enumerations. Of course, the derived resource, the lexicon, can be rather easily cleaned-up of such noise, and can then serve as the interface between the corpus and more semantically oriented resources.

4. Conclusions and further work

The paper has presented the JRC-Acquis corpus, and the linguistic annotation tool totale. The corpus could become a significant new resource for research on multilingual language technologies and is freely available for research purposes at http://wt.jrc.it/lt/acquis/. The paper described the content and compilation steps which lead to the first version of this corpus. Further work will involve in part promoting the corpus, and, most likely, expansion of the corpus with new languages, and further processing steps on the corpus, e.g. higher quality alignment, linguistic processing for more languages, etc.

The other contribution of the paper is the discussion of the text annotation tool totale, which performs multilingual tokenisation, tagging and lemmatisation. The program is has been currently trained for seven languages and extensively tested on Slovenian. For

totale, we would like to extend the range of languages that it supports, and improve the models for existing ones. This of course would involve more training resources (lexicons and annotated corpora) but also the improvement of the underlying architecture. For example, by doing multi-pass processing through the texts the initial annotation could serve to construct a lexicon, this would be cleaned (automatically, with heuristics, or manually) and then used to re-annotate the text at a much higher precisions.

References

[1] S. ARMSTRONG, M. KEMPEN, D. MCKELVIE, D. PETITPIERRE, R. RAPP and H. THOMPSON: Multilingual Corpora for Cooperation. *Proc. 1st Int. Conf. on Language Resources and Evaluation*, ELRA, Paris, (1998), 579-980.

[2] T. BRANTS: TnT-A Statistical Part-of-Speech Tagger. *Proc. 6th Applied Natural Language Processing Conf.*, Seattle, WA, USA, (2000), 224-231.

[3] L. DIMITROVA, T. ERJAVEC, N. IDE, H.-J. KAALEP, V. PETKEVIČ and D. TUFIS: MULTEXT-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. *Proc. COLING-ACL'98*, Montreal, Quebec, Canada, (1998).

[4] W. GALE and K.W. CHURCH: A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, **19**(1), (1993), 75-102.

[5] P. DI CRISTO: MtSeg: The Multext multilingual segmenter tools. MULTEXT Deliverable MSG 1, Version 1.3.1. CNRS, Aix-en-Provence, http://www.lpl.univ-aix.fr/projects/multext/MtSeg/, (1996).

[6] P. DANIELSSON and D. RIDINGS: Practical Presentation of a "Vanilla" Aligner. TELRI Newsletter No. 5, Institute fuer Deutsche Sprache, Mannheim, http://nl.ijs.si/telri/Vanilla/ (1997).

[7] T. ERJAVEC and S. DŽEROSKI: Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, **18**(1), (2004), 17-41.

[8] T. ERJAVEC: MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *4th Int. Conf. on Language Resources and Evaluation*, ELRA, Paris, France, (2004), 1535-1538.

[9] P. KOEHN: Europarl: A Multilingual Corpus for Evaluation of Machine Translation. http://people.csail.mit.edu/people/koehn/publications/europarl/, (2002).

[10] T. MCENERY, A. WILSON, P. SANCHEZ-LEON and A. NIETO-SERRANO: Multilingual Resources in European Languages: Contributions of the CRATER Project. *Literary and Linguistic Computing*, **12**(4), (1997).

[11] B. POULIQUEN, R. STEINBERGER and C. IGNAT: Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. *Proc. of the Workshop Ontologies and Information Extraction at (EUROLAN'2003)*, Bucharest, Romania, (2003).

[12] B. POULIQUEN, R. STEINBERGER and C. IGNAT: Automatic Linking of Similar Texts across Languages. *In: Recent Advances in Natural Language Processing III*, John Benjamins Publishers, Amsterdam, 2004.

[13] S. MANANDHAR, S. DŽEROSKI and T. ERJAVEC: Learning Multilingual Morphology with CLOG. *Proc. of Inductive Logic Programming, 8th Int. Workshop ILP-98*, (Lecture Notes in Artificial Intelligence 1446), Springer-Verlag, Berlin, (1998), 135-144.

[14] C.M. SPERBERG-MCQUEEN and L. BURNARD (EDS.): Guidelines for Electronic Text Encoding and Interchange, the XML Version of the TEI Guidelines. The TEI Consortium, http://www.tei-c.org/, (2002).