Slovene-English Datasets for MT

Tomaž Erjavec

Dept. of Intelligent Systems
Institute "Jožef Stefan"
Ljubljana, Slovenia

Abstract

Advances in machine translation are becoming increasingly dependent on the availability of large scale language resources, in particular parallel corpora. The talk presents Slovene-English language resources that were developed as datasets for translation studies and machine learning programs. Three parallel datasets are introduced: the MULTEXT-East multilingual word-annotated corpus, the IJS-ELAN Slovene-English parallel corpus, and the Concede English-Slovene dictionary fragment and lexical database.

1. Introduction

While machine translation systems will, for a while to come, utilise hand-crafted lexica and rules, there is an increasingly important strand of research which attempts to develop methods that can automatically induce lexica and rules from language data, in particular from bilingual parallel language corpora. Such corpora must be developed separately for each language pair; it adds to their usability if they are additionally marked up for linguistic information, encoded in a standard manner and made widely available. The value of aligned corpora is well attested in practice, with several recent European projects devoted to producing them, e.g. MLCC [2] (nine EU languages), Crater [3] (Spanish, French and English), or ENPC [4] (English-Norwegian).

Slovene is a South-Slavic language with approximately 2 million speakers. The language is characterised, as other Slavic languages, with a complex inflectional system and free word order. So far, no serious attempts have been made to develop MT systems or dedicated MT resources for Slovene. In the last few years, however, various EU projects in language technologies have enabled the development of parallel Slovene language resources, which could prove useful for experiments in MT resource acquisition.

The parallel resources developed at our institute share a number of common characteristics. So far, we have concentrated on the Slovene-English language pair, as this combination can prove the most immediately useful. All the resources have been encoded in SGML, in particular in various parameterisation of the Text Encoding Initiative Guidelines, TEI P3 [5]. Such a standardised encoding makes the resources more portable between platforms and applications and ensures their longevity. Finally, we have made every effort to ensure easy access to the resources; any restrictions on their availability stem only from copyright on the original data. All this makes the datasets suitable for reuse for a wide variety of target applications.

In the paper we concentrate on three datasets. Section 2. introduces the multilingual MULTEXT-East 100.000 word morphosyntactically annotated corpus and associated resources. We discuss the new release of the corpus, which has been significantly corrected and re-encoded from the original publication in 1998. Section 3. presents the Slovene-English IJS-ELAN 1 million word corpus, specifically targeted as a dataset for experiments in automatic bi-lingual terminology extraction. We discuss the current state of the corpus, and mention the forthcoming new version, which will be word-annotated for morphosyntax. Section 4. mentions the English-Slovene dictionary fragment, which we have converted into a lexical database. Finally, section 5. gives some conclusions and directions for further research.

2. The MULTEXT-East dataset

The MULTEXT-East project (Multilingual Text Tools and Corpora for Eastern and Central European Languages) was a spin-off of the EU MULTEXT project. The project ran 1995-97 and developed language resources for six languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, and, additionally, for English as the "hub" language of the project. It also adapted existing tools and standards to these languages. The main results of the project [6] were an annotated multilingual corpus and morpholexical resources for the languages in question.

The development of the morpholexical resources proceeded in three stages. First, harmonised morphosyntactic descriptions (MSDs) were developed for the languages of the project. These are used to describe a wordform as being, say, a proper inanimate masculine noun in singular accusative. The second stage was in building the actual lexica, which cover the lexical stock of the corpus collected in the project. Finally, the developed lexica were used to annotate a portion of the MULTEXT-East corpus with context-disambiguated MSDs and lemmas.

One of the objectives of MULTEXT-East has been to make its resources freely available for research purposes. In the scope of the EU TELRI (Trans European Language Resources Infrastructure) concerted action, the complete results of TELRI and MULTEXT-East have been released on a CD-ROM [7]. Since the release of the CD-ROM, the resources have been used in a number of experiments. In the course of this work, errors and inconsistencies were discovered in the specifications and in the data, which were subsequently corrected. This lead the partners' to consider releasing a new version of the corrected resources, which had been recently completed. The rest of this section explains the contents and structure of this new release; more information is available on the home page of the project, at http://nl.ijs.si/ME/.

2.1 The Morphosyntactic Specifications

The syntax and semantics of the MULTEXT-East MSDs are given in the morphosyntactic specifications of the project. These specifications have been developed in the formalism and on the basis of specifications for six Western European languages of the EU MULTEXT project and in cooperation with EAGLES, the Expert Advisory Group on Language Engineering Standards.

The MULTEXT-East morphosyntactic specifications contain, along with introductory matter, also:
- the list of defined categories (parts-of-speech)
- common tables of attribute-values
- language particular tables

Of the MULTEXT-East categories, Slovene uses Noun (N), Verb (V), Adjective (A), Pronoun (P), Adverb (R), Adposition (S) (these include prepositions and postpositions; Slovene uses only prepositions), Conjunction (C), Numeral (M), Interjection (I), Residual (X) (for unknown words), Abbreviation (Y), and Particle (Q).

The common tables of the specification give for each category a table defining the attributes appropriate for the category and the values defined for these attributes. They also define which attributes/values are appropriate for each of the MULTEXT-East languages; the tabular structure facilitates the addition of new languages, and the new release adds Croatian (HR) to the specifications. The common tables have a strictly defined format, which enables the automatic expansion and validation of MSDs. The format of these tables is exemplified by the start of the Noun table, given below:

				EN	RO	SL	CS	BG	ET	HU	HR
P	ATT	VAL	C	x	x	x	x	x	x	x	x
1	Type	common	c	x	x	x	x	x	x	x	x
		proper	p	x	x	x	x	x	x	x	x
2	Gender	masculine	m	x	x	x	x	x			x
		feminine	f	x	x	x	x	x			x
		neuter	n	x	x	x	x	x			x
3	Number	singular	s	x	x	x	x	x	x	x	x
		plural	p	x	x	x	x	x	x	x	x
		dual	d			x	x				
	l.s.	count	t					x			
4	Case	nominative	n			x	x	x	x	x	x
		genitive	g			x	x		x	x	x
		dative	d			x	x			x	x
		accusative	a			x	x			x	x
		vocative	v		x		x	x			x
		locative	l			x	x				x
		instrumental	i			x	x			x	x
	l.s.	direct	r		x						
	l.s.	oblique	o		x						
	l.s.	partitive	1						x		
		illative	x						x	x	
		inessive	2						x	x	
		elative	e						x	x	

...

Example of Common MSD Table: Nouns

In addition to the common tables the specifications include also language particular tables, which are, again, organised by category and provide commentary and examples on the attributes and values of the language. The Slovene section additionally gives each attribute and value its Slovene equivalent thus enabling the localisation of the MSDs. Furthermore, the language particular sections give feature coocurrence restrictions on the allowed combinations of values as well as an exhaustive list of MSDs allowed for each category

2.2 The lexicons

The project delivered medium sized morphological lexica for the seven languages [8]. The Slovene lexicon contains the full inflectional paradigms for about 15,000 lemmas, giving a lexicon of over half a million entries. A MULTEXT(-East) lexicon contains lexical entries, one entry per line. Each entry has three fields: the word-form, its lemma and the morphosyntactic description.

The word-form is the word as it appears in the running text, modulo sentence initial capitalisation, e.g. *diskreditirajmo, Moloha, dvomišljenja.* A special case arises when more than one running word is taken as one word-form, e.g. *New York*. Here the underscore is used to join the words into the word-form. In the word-forms, as in the lemmas, SGML entities are used for the representation of non-ASCII characters, e.g. *čemer_koli*. The lemma is the unmarked form of the word, i.e. what would correspond to the headword in a dictionary, e.g. *diskreditirati, Moloh, dvomišljenje*. In cases where the word-form is the lemma itself, the lemma is entered as the equals sign, "=".

The MSD is the morphosyntactic description of the word-form, as explained above. The MSDs are provided as strings, using a linear encoding. In this notation, the first character denotes the part-of-speech, and, for the other characters in the string, the position corresponds to the part-of-speech determined attribute, and specific characters in

each position indicate the value for that attribute. So, for example, the MSD *Vmmp1p* expands to *PoS:Verb, Type:main, VForm:imperative, Tense:present, Person:first, Number:plural*. If a certain attribute does not apply, either to a language, to a combinations of attribute-values, or the the specific lexical item, then the value of that attribute is a hyphen. So, for example, the *Person* attribute of *Verb* is not relevant for *Type:participle*, hence *Vmps-sma* for *Verb main participle past (no Person) singular masculine active*. By convention, trailing hyphens are not included in the lexical MSDs.

To illustrate these points we give below a part of the lexical paradigm for the verb *to be*, i.e. *biti*.

```
bi      biti    Vcc
bil     biti    Vcps-sma
bila    biti    Vcps-dma
bila    biti    Vcps-pna
bila    biti    Vcps-sfa
bile    biti    Vcps-pfa
bili    biti    Vcps-dfa
bili    biti    Vcps-dna
bili    biti    Vcps-pma
bilo    biti    Vcps-sna
biti    =       Vcn
bo      biti    Vcif3s
```

Sample lexical entries

2.3 The corpus

The MULTEXT-East annotated multilingual corpus [9] is divided into a text corpus and a speech corpus. The text corpus consists of the parallel part and two comparable parts, where each of the three parts contains approximately 100.000 words per language. We here concentrate on the parallel part of the corpus, which has recently been corrected and re-released.

The multilingual parallel corpus consists of the novel *1984* by George Orwell in the English original and translations into the six languages of the project. The choice of this data was motivated by the availability of the translations in all six languages and the availability of the English original and the Slovene translation in digital form from the Oxford Text Archive via the European Corpus Initiative. The *1984* is the central component of the MULTEXT-East corpus; while the whole corpus has been bibliographically and structurally marked-up, the parallel part has been also sentence segmented, aligned, and, crucially, as annotated with word-level linguistic information.

In the second release, the multilingual *1984* is encoded as an parameterisation of the TEI, where each translation is a separate document and comprises a TEI header giving the bibliographic and other information about the file, and the body of the novel. The body itself contains structural markup for divisions (parts and chapters of the novel) and paragraphs. The value of the corpus comes from its linguistic markup; apart from sentence segmentation and tokenisation, each word is also marked up for context disambiguated hand-validated lemma and MSD. We illustrate the structure by giving the first sentence of the Slovene translation of the novel:

```
<text id="Osl." lang="sl">
<body>
<div type="part" id="Osl.1">
<div type="chapter" id="Osl.1.2">
<p id="Osl.1.2.2">
<s id="Osl.1.2.2.1">
<w lemma="biti" ana="Vcps-sma">Bil</w>
<w lemma="biti" ana="Vcip3s--n">je</w>
<w lemma="jasen" ana="Afpmsnn">jasen</w>
<c>,</c>
<w lemma="mrzel" ana="Afpmsnn">mrzel</w>
<w lemma="aprilski" ana="Aopmsn">aprilski</w>
<w lemma="dan" ana="Ncmsn">dan</w>
<w lemma="in" ana="Ccs">in</w>
<w lemma="ura" ana="Ncfpn">ure</w>
<w lemma="biti" ana="Vcip3p--n">so</w>
<w lemma="biti" ana="Vmps-pfa">bile</w>
<w lemma="trinajst" ana="Mcnpnl">trinajst</w>
<c>.</c>
</s>
```

Sample sentence for the *1984* MULTEXT-East corpus

The translations of the novel are sentence aligned with the English original; the alignment is stored in a separate file, by specifying ID links between the sentences. For easier processing, a "knitted" version of the corpus is also made available; it contains for each English segment that has translations into all the languages, a translation unit consisting of the seven translation segments.

While the annotated MULTEXT-East *1984* corpus is quite small, it had been the first morphosyntactically annotated corpus for Slovene and most of the other languages of the project. It has therefore served as a valuable "gold standard" dataset for studying word-class syntactic tagging and similar applications.

3. The IJS-ELAN corpus

While the MULTEXT-East parallel corpus is heavily annotated, it nevertheless consists of only one novel, with the Slovene part being the translation; both factors severely limit its usability. The EU MLIS project ELAN (European Language Activity Network) provided an opportunity to somewhat remedy this lack. Our contribution to ELAN was, in part, to collect and annotate a 1 million word Slovene-English parallel and sentence aligned corpus.

The IJS-ELAN corpus [10] contains fifteen recent texts, from interesting areas of text production. The texts and corpus encoding have been chosen so as to have minimal restriction on further use and could thus be made widely available as a standardised dataset for bilingual language engineering research.
The rest of this section briefly outlines the composition of the corpus and the manner of its encoding. Further information about the IJS-ELAN, including the TEI headers and a sampler in HTML, the complete corpus packed for downloading, and on-line concordancing can be found at the http://nl.ijs.si/elan/.

3.1. Corpus composition

The small scale of the project prohibited any attempt towards making an English-Slovene reference-type corpus except maybe at the level of encoding. The composition of the IJS-Elan corpus was motivated in part by considerations of usability, and in part by ease of acquisition. For usability, the corpus contains recent (90's) texts rich in terminology and from active topic areas. Ease of acquisition also played a decisive role in choosing the particular texts; we only considered texts where the original and translation were already available electronically, in one of a few formats: HTML, RTF, and SGML (QUERTZ DTD). A factor in selecting the component texts was the

willingness of the copyright holders to allow the inclusion of their texts in the corpus, with minimal restrictions on further distribution.

The corpus has fifteen components, which are mostly complete bi-texts, but with omissions of predominately non-textual data (numerical charts etc). In the corpus each bi-text is given its ID and constitutes, along with its header, one element of the corpus. The texts are usefully divided into those that have a Slovene original and an English translation, and those whose original is English, and the translation is into Slovene. Apart from there being linguistic differences due to the opposition original/translation, the two parts also have a quite different composition. The Slovene - English half has been, for the most part, acquired from various branches of the Slovene government and deal with economy and law. This part consists of eleven texts, containing somewhat more than half of the corpus material. The English-Slovene part of the corpus, on the other hand, is composed of only four elements, with two of these being full-length books. It also has different text types from the Slovene-English part: two components deal with computers, one with pharmaceuticals, while one is the *1984* novel taken over from the MULTEXT-East corpus.

3.2 Corpus encoding

The current version of the corpus is encoded as a translation-memory like parameterisation of the TEI, where each corpus component has its header and body, and the body is composed of translation units, each with two segments: one from the original and the other of the translation. The corpus is also tokenised into words and punctuation. To illustrate, we give below some short translation units from the corpus:

```
<tu lang="sl-en" id="ecmr.2">
<seg lang="sl"><w>AKTUALNA</w> <w>GIBANJA</w></seg>
<seg lang="en"><w>IN</w> <w>THE</w> <w>SPOTLIGHT</w></seg>
</tu>

<tu lang="sl-en" id="stra.2">
<seg lang="sl"><w>Ljubljana</w><c>,</c> <w>september</w>
<w type=dig>1997</w></seg>
<seg lang="en"><w>Ljubljana</w><c>,</c> <w>September</w>
<w type=dig>1997</w></seg>
</tu>

<tu lang="en-sl" id="gnpo.2">
<seg lang="en"><w>%s</w><c>:</c> <w>option</w>
<c type=open>'</c><c>--</c><w>%s</w><c type=close>'</c>
<w>doesn't</w> <w>allow</w> <w>an</w> <w>argument</w></seg>
<seg lang="sl"><w>%s</w><c>:</c> <w>izbira</w>
<c type=open>'</c><c>--</c><w>%s</w><c type=close>'</c>
<w>ne</w> <w>dovoljuje</w> <w>argumenta</w></seg>
</tu>
```

Translation units of the IJS-ELAN corpus

While the corpus is tokenised, it has not yet been tagged with MSDs or lemmatised. Because this would significantly increase the utility of the corpus we have started work on this issue, and the automatically tagged corpus should be available in 2001.

4. The Concede lexical database

The value of language resources is greatly enhanced if they share a common markup with an explicit minimal semantics. Achieving this goal for lexical databases is difficult, as large-scale resources can realistically only be obtained by up-translation from pre-existing dictionaries, each with its own proprietary structure. The EU project Concede (Consortium for Central European Dictionary Encoding, 1998-2000) built structured lexical databases (LDBs) derived from existing machine-readable dictionaries for the same set of languages as MULTEXT-East,

namely Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene [11]. One of the goals of the project was to deliver these databases as an integrated resource, complementing the *1984* parallel corpus. To achieve this the databases had to, as far as possible, share a common markup scheme, using the same tags and giving them the same interpretations.

While all the other languages took as their starting point monolingual dictionaries, the Slovene case was special in that we investigated a bi-lingual dictionary, namely a sample of the English-Slovene dictionary by the publishing house DZS. This dictionary is based on the Oxford University Press English(-French) dictionary, and is still in the process of being produced. While the Slovene-English sample LDB is to be made available, one of the requirements of the DZS publishing house was that this will only happen after the publication of the paper dictionary.

The original dictionaries came in a variety of legacy formats, from Word to SGML. To give us a common ground for comparison all the dictionaries were first converted to similar representations, which were various parameterisations of the TEI.dictionaries base tagset. At this stage, the guiding principle was to preserve or further detail the information found in the original digital format. The dictionaries samples used together 58 different elements. Most are taken directly from the TEI.dictionaries tag set, with some modifications and additions.

4.1 The Concede DTD

With the information in a standard format, we were in a position to develop a single DTD to cover all the dictionaries. We have used XML (Extensible Markup Language), due to its emergence as the de facto standard for data representation, and in order to take advantage of facilities developed within the XML framework, e.g. the Extensible Style Language (XSL). Our guiding principle was to provide a DTD with as few elements as possible, each with an unambiguous, clearly-defined interpretation. This task breaks naturally into two parts: content and structural elements.

For content, we identified an inventory of 23 TEI elements capable of representing all the content elements in the source dictionaries (not necessarily 1-1), and fixed their TEI interpretations. For structural elements, we followed the observations in [12] that certain underlying regularities exist in all print dictionaries (in particular, the use of a hierarchical organisation that enables the factoring of information over nested levels) and that all levels in dictionary hierarchies potentially contain the same elements. Therefore, we adopt a simple general scheme involving three structural elements:

- <struc> represents a node in the tree. These elements may be recursively nested at any level to reflect the structure of the corresponding tree. This is also the only element in the encoding scheme that corresponds to the tree structure; all other elements provide information associated with a specific node (i.e., the node corresponding to the immediately enclosing <struc> element).
- <alt> alternatives may appear within any <struc>. The use of this element corresponds to the shorthand often used in dictionary entries, where two equally applicable sets of information apply to the entire sub-tree, as where there are two possible spellings and two or more meanings, and either spelling can be coupled with any meaning.
- <brack> is a general-purpose bracketing element to group associated features.

To exemplify the structure of the Concede LDB, we give below and entry from the English-Slovene dictionary.

In comparison with the TEI dictionary encoding, the Concede LDB format is somewhat less informative and, in the case of the English-Slovene sample, also contains a percentage of conversion errors. However, it has a much simpler content model and a defined inheritance structure, making it easier for applications to exploit the dictionary information.

```
<entry id="ensl.2">
  <hw>Adam</hw>
  <pos>n</pos>
  <pron>"&amp;d.m</pron>
  <trans>
    <orth>Adam</orth>
  </trans>
  <struc type="eg-compound">
    <eg>
      <q>Adams ale/wine</q>
    </eg>
    <trans>
      <orth>voda</orth>
    </trans>
  </struc>
  <struc type="hom-idioms" id="ensl.2.5">
    <alt type="xr">
      <xr>know</xr>
      <xr>old</xr>
    </alt>
    <struc type="eg-idiom">
      <eg>
        <q>I don't know him from Adam</q>
      </eg>
      <trans>
        <orth>sploh ga ne poznam</orth>
      </trans>
    </struc>
  </struc>
</entry>
```

Example of a Concede LDB lexical entry

4.2 Sample application

The Concede project also addressed the integration of machine-readable dictionaries and lexical databases with corpora. As a demonstration of using the TEI and Concede LDB formats of the English-Slovene sample, we have converted some TEI dictionary entries into HTML, which are hyperlinked, via the LDB, to an on-line concordancing system [11].

The corpus used in the experiment is the English-Slovene parallel part of the MULTEXT-East corpus. The on-line query system has as its corpus processing backend the CQP system [13], which incorporates a powerful query language that allows querying for all of the corpus annotation. For the experiment we identified 29 dictionary entries, whose headwords do in fact appear in the Slovene *1984*. We then produced a Web rendering of these entries, where it is possible to click on elements, causing the retrieval of associated bi-lingual concordances. The queries are automatically constructed according to the (inherited) information available for the element in question.

The envisaged application of such an integration of a dictionary representation with corpus evidence is to either use it in the process of making the dictionary, or to give to the end-user of the dictionary a means to further supplement dictionary examples with concordances culled from the corpus.

The process of converting to the HTML dictionary/concordance representation first involved choosing suitable anchor elements for querying, e.g. the orthography, part-of-speech, translation etc. Queries are then constructed for each anchor element, taking into account the information in the anchor as well as that inherited from superordinate anchors. The type of query is dependent on the anchor element. For example, headwords are translated into the query *[lemma="*hw*"]*, meaning "find the lemma string *hw* of a token in the (default) English part of the corpus". After such

a query is constructed for each potential anchor, the query is run off-line to ensure hits in the corpus. If concordances are found, the query URL is hyperlinked to the HTML rendering of the TEI element bearing the appropriate ID. In our 29 dictionary entries there were 2870 potential anchor elements, of which 337 produced matches in the corpus.

Such an HTML encoding, in spite of the small size of the corpus, demonstrates how to exploit the LDB data and presents a method of visually combining corpus searches with information encoded in dictionaries.

5. Conclusions

The paper presented three Slovene-English language resources that were developed at our institute as datasets for exploring various aspects of computational linguistics and language technologies, and are suitable for MT applications.

As has been mentioned, the annotated MULTEXT-East corpus has been, in spite of its small size, very valuable, as it has been the first morphosyntactically annotated corpus for most of the languages of the project. The main application has been in using it as a training and testing set for learning and evaluating models and methods of word-class syntactic tagging, as discussed in e.g. [14] and [15]. Another application which we have investigated was learning rules for automatic lemmatisation of unknown (i.e. not contained in the lexicon) Slovene words, by using a combination of stochastic methods and inductive logic programming [16]. The new release of the corpus and associated resources should offer an even better test-bed for such experiments, which could also be expanded to a multilingual setting.

The IJS-ELAN aligned corpus is closer to being exploited for MT purposes: it has already been used in experiments involving bi-lingual terminological extraction [1], and similar experiments are planned in the future. As has been mentioned, we also plan to perform word-class syntactic tagging on the corpus and also extend it with new bi-texts, mostly from the 'Acquis Communautaire', i.e. the texts being prepared and translated for Slovenia's membership in the EU.

Finally, the Concede dictionary fragment and lexical database has seen some research in exploitation, as explained in the previous section. Further experiments will also focus on the integration of the lexical database with extracted translation equivalents from bi-lingual corpora.

References

[1] Dias, G.; Vintar, Š.; Pereira Lopes, G.; Guillore, S. (2000).
Normalising the IJS-ELAN Slovene-English Parallel Corpus for the Extraction of Multilingual Terminology.
In: Monachesi, P. (ed.) Proceedings of the CLIN '99 (Computational Linguistics in the Netherlands).

[2] Armstrong, S.; Kempen, M.; McKelvie, D.; Petitpierre, D.; Rapp, R.; Thompson, H. (1998).
Multilingual Corpora for Cooperation.
In the Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98, Granada.
pp. 579--980, 1998.

[3] McEnery, T.; Wilson, A.; Sanchez-Leon, F.; Nieto-Serrano, A. (1997).
Multilingual Resources in European Languages: Contributions of the CRATER Project.
Literary and Linguistic Computing 12/4, 1997.

[4] Johansson, S.; Ebeling, J.; Hofland, K. (1996).
Coding and aligning the English-Norwegian Parallel Corpus.
In K. Aijmer and B. Altenberg and M. Johansson (eds.) Languages in Contrast.
Lund University Press, pp. 87-112, 1996.

[5] Sperberg-McQueen, C. M.; Burnard, L., (1994).
Guidelines for Electronic Text Encoding and Interchange.
Chicago and Oxford, 1994.

[6] Dimitrova, L.; Erjavec, T.; Ide, N.; Kaalep, H.J.; Petkevič, V.; Tufis, D. (1998).
Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages.
COLING-ACL '98, pp. 315-319, 1998.

[7] Erjavec, T.; Lawson, A.; Romary, L. (eds.) (1998).
CD-ROM East meets West: A Compendium of Multilingual Resources.
TELRI Association e.V, 1998.

[8] Ide, N.; Tufis, D.; Erjavec, T. (1998).
Development and Assessment of Common Lexical Specifications for Six Central and Eastern European Languages.
In the Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98, Granada, pp. 233--240, 1998.

[9] Erjavec, T.; Ide, N. (1998).
The MULTEXT-East Corpus.
In the Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98, Granada, pp. 971--974, 1998.

[10] Erjavec, T. (1999).
The ELAN Slovene-English Aligned Corpus.
In the Proceedings of the Machine Translation Summit VII, Singapore, pp. 349-357, 1999.

[11] Erjavec, T.; Evans, R.; Ide, N.; Kilgarriff, A. (2000).
The Concede Model for Lexical Databases.
In the Proceedings of the Second International Conference on Language Resources and Evaluation, LREC'00", Athens, pp. 355--362, 2000.

[12] Ide, N.; Veronis, J. (1995).
Encoding Dictionaries.
In The Text Encoding Initiative: Background and Context.
Kluwer Academic Publishers, Dordrecht, pp. 167--180, 1995.

[13] Christ, O. (1994).
A Modular and Flexible Architecture for an Integrated Corpus Query System.
In the Proceedings of COMPLEX '94: 3rd conference on Computational Lexicography and Text Research.
1994.

[14] Džeroski, S.; Erjavec, T.; Zavrel, J. (2000).
Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets.
In the Proceedings of the Second International Conference on Language Resources and Evaluation, LREC'00", Athens, pp. 1099--1104, 2000.

[15] Hajič, J. (2000).
Morphological Tagging: Data vs. Dictionaries.
In the Proceedings of ANLP/NAACL 2000, Seatle.

[16] Džeroski, S.; Erjavec, T. (2000).
Learning to lemmatise Slovene Words.
In J. Cussens, S. Džeroski (eds.) Learning Language in Logic.
Springer, Lecture notes in artificial intelligence 1925., pp. 69-88, 2000.