The MULTEXT-East Slovene Lexicon*

Tomaž Erjavec
Dept. for Intelligent Systems, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

This paper describes the lexicon of the Slovene language created in the scope of the EU Copernicus project MULTEXT-East. The project developed a multi-lingual corpus of text and speech data, covering six languages, including Slovene, and lexical resources covering the corpus data. The Slovene lexicon contains the full inflectional paradigms for over 15,000 lemmas; it has over half a million entries, where each entry gives the word-form, its lemma and morphosyntactic description. These descriptions are constructed according to the MULTEXT-East grammar, which follows international recommendations and is harmonised for seven languages. The MULTEXT-East Slovene lexicon is freely available for research purposes.

1 Introduction

While standardised, large-scale language resources exist or are under development for most western languages there have, so far, been few comparable efforts for Central and Eastern European (CEE) languages. The MULTEXT-East (Multilingual Text Tools and Corpora for Eastern and Central European Languages, 1995–97) project [3], a spin-off of the EU MULTEXT project [7], was intended to fill this gap by developing language resources for six CEE languages (Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene) and adapting existing tools and standards to them. MULTEXT-East produced the following results:

1. test and adaptation of language standards;
2. an annotated multilingual corpus;
3. morpho-lexical resources;
4. adaptation of the MULTEXT corpus tools.

In this paper we concentrate on (3), and, partially, on (1), i.e. on the specification of the morpho-lexical encoding and on the lexica of the project [6]; the focus is on the Slovene language portion of these resources.

The development of the morpho-lexical resources proceeded in three stages. First, harmonised *morphosyntactic descriptions* (MSDs) were developed for the languages of the project. These are used to describe a word as being, say, a proper inanimate masculine noun in singular accusative. The second stage was in building the actual lexica which cover lexical stock of the corpus collected in the project. Finally, the developed lexica were used to MSD annotate a portion of the MULTEXT-East corpus. This paper describes the results of the first two of the above stages, as applied to the Slovene data.

The rest of the paper is structured as follows: Section 2. describes the structure of the (Slovene) MULTEXT-East lexicon; Section 3. gives the grammar of the MSDs; Section 4. provides some data on the scope and coverage of the Slovene lexicon; and Section 5. concludes with the utility and availability of the developed lexicon.

2 The structure of the lexicon

A MULTEXT-East lexicon contains lexical entries, one entry per line. A lexical entry has three fields, separated by the tabulator character:

word-form ⟨TAB⟩ **lemma** ⟨TAB⟩ **MSD**

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. '*čemer koli*', '*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 '='.

*Proceedings of the 7th Electrotechnical Conference ERK '98, Portorož, Slovenia, Volume B, pp. 189-192.

The *MSD* is the morphosyntactic description of the word-form. 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, in Figure 1, 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

Figure 1: Sample lexical entries

3 The Morphosyntactic Specifications

The syntax and semantics of the MULTEXT-East MSDs are given in the *morphosyntactic specifications* of the project [4]. These specifications have been developed in the formalism and on the basis of specifications for six Western European languages of the EU MULTEXT project [1]; the MULTEXT project produced its specifications in cooperation with EAGLES (Expert Advisory Group on Language Engineering Standards) [2].

The MULTEXT-East morphosyntactic specifications contain, along with introductory matter, also:

1. the list of defined categories (parts-of-speech)

2. common tables of attribute-values

3. language particular tables

Of the MULTEXT-East *categories*, Slovene uses Noun (N), Verb (V), Adjective (A), Pronoun (P), Adverb (R), Adposition (S),[1] Conjunction (C), Numeral (M), Interjection (I), Residual (X),[2] Abbreviation (Y), and Particle (Q).

The *common tables* 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. The format of the common tables is exemplified by the start of the Noun table, given in Figure 2.

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

...

Figure 2: Example of Common MSD Table: Nouns

The common tables have a strictly defined format, which enables the automatic expansion and validation of MSDs.[3]

[1] Adpositions include prepositions and postpositions; Slovene uses only prepositions.

[2] Residual is a category encompassing unknown (unanalysable) lexical items. It appears only once in the Slovene lexicon, for `2+2=5`.

[3] But note that the table extracts given in this paper have been edited to conform with the formatting requirements of the paper.

Finally, the *language particular tables* are, again, organised by category, and provide commentary and examples on the attributes and values of the language. They also specify additional constrains on the allowed MSDs. The Slovene section gives, for each attribute value, its Slovene equivalent together with an example, and possible notes. Furthermore, it gives feature coocurrence restrictions on the allowed combinations of values.

To exemplify, we give in Figure 3 the valid combinations of qualificative adjectives (e.g. *lep, lepi, lepega, . . . , najlepšim*). The table constrains the marking for definiteness only to the nominative and non-animate accusative masculine singular and animacy only to the accusative masculine singular, both in the positive degree.

```
*** **** **** **** **** **** ---- ---- ----
PoS Type Degr Gend Numb Case Def  (Cltc)Anim
*** **** **** **** **** **** ---- ---- ----
 A   f     p    m    s     n  [ny]    -   -
 A   f     p    m    s     a  [ny]    -   n
 A   f     p    m    s     a     -    -   y
 A   f     p    m    s   [gdli]  -    -   -
 A   f     p    m  [pd] [ngdali] -    -   -
 A   f     p  [fs] [spd] [ngdali] -   -   -
 A   f   [cs] [mfs] [spd] [ngdali] -  -   -
```

Figure 3: Example of Language Specific MSD Constraints: Slovene qualificative adjectives

The morphosyntactic specifications thus provide the grammars for the MSDs of the MULTEXT-East languages. The greatest worth of these specifications is that they provide an attempt at a morphosyntactic encoding standardised across languages. In addition to already encompassing seven typologically very different languages, the structure of the specifications and of the MSDs is readily extensible to new languages.

As far as the Slovene specifications go, it should be noted that another morphosyntactic encoding has been recently proposed [8]; in contrast to the multilingual, and, on the meta-level, English MULTEXT-East specification, it uses Slovene terms, and codes derived from them. For example, the equivalent of the MSD `Vcps-sma` is the code `GLBme` (*GLagol, opisni deležnik na* -l biti, *moški spol, ednina*). This makes it more acceptable to speakers of Slovene; as it accounts only for Slovene, it can also follow more closely the grammar of Slovene, and the codes can be shorter.

To give an impression of the information content of the Slovene MSDs and their distribution, Table 1 gives, for each category, the number of *attributes* in the category, the total number of *values* for all attributes in the category; the number of different MSDs in the *lexicon*, and, finally, in the MSD and lemma annotated MULTEXT-East Slovene *corpus*, comprising the Slovene translation of Orwell's "1984".

PoS	Att	Val	Lex	Cor
Noun	5	16	99	74
Verb	8	26	128	93
Adjective	7	22	279	169
Pronoun	11	36	1,335	594
Adverb	2	4	3	3
Adposition	3	8	6	6
Conjunction	2	4	3	2
Numeral	7	23	226	80
Interjection	0	0	1	1
Residual	0	0	1	1
Abbreviation	0	0	1	1
Particle	0	0	1	1
All	45	139	2,083	1,025

Table 1: Slovene morphosyntactic distribution

4 Lexical Coverage

Which word-forms should go into the lexicon? MULTEXT-East adopted a combination of an external, and an internal approach. The externally given criterion was the corpus of the MULTEXT-East project; for the Slovene part, it encompassed cca. 300.000 words, comprising translated and native fiction and newspaper articles. The words encountered in this corpus were assigned lemmas, and these were included in the lexicon, until the number of lemmas reached 15.000, the size required by the project plan. This gives a lexicon of medium size, which displays good coverage of high and medium frequency words, but gives low-frequency words only for the Slovene corpus of the project.

Rather than including only the word-forms encountered in the corpus, the more informative option of giving the complete inflectional paradigms of the lemmas was chosen. As this generative approach depends on an internally given morphological model, the paradigms are not fully validated in practice, and, in some cases, tend to be overgenerous. Spelling out full paradigms for a language as inflectionally rich as Slovene also leads to a large lexicon, cca. 17 MB. However, this approach (rather than using some sort of morphological compression) has the advantage of keeping the underlying model as simple and explicit as possible. This makes it better for hand-corrections, various experiments and interchange.

Table 2 gives some quantitative data on the Slovene lexicon. For each category, we give the number of lexical entries, the number of different word-forms, and different lemmas (=). Dividing the number of entries by lemmas gives us the average paradigm size of the category in question, while dividing by word-forms gives the average morphosyntactic ambiguity of words in a corpus.

PoS	Entries	WForms	Lemmas
Noun	127,811	61,525	7,465
Verb	110,949	78,001	3,699
Adjective	310,754	64,604	4,621
Pronoun	3,654	732	105
Adverb	7,415	7,395	442
Adposition	123	109	79
Conjunction	39	38	39
Numeral	4,401	832	181
Interjection	10	10	10
Residual	1	1	1
Abbreviation	48	48	48
Particle	76	76	76
All	565,281	201,011	16,766

Table 2: Slovene lexical distribution

5 Conclusions

The article has presented the Slovene lexical resources, i.e. the morphosyntactic specifications and the lexicon, that have been developed in the MULTEXT-East project. These resources present the first widely available standardised encoding for Slovene lexical data. A lexicon is one of the basic language resources and a core lexicon, such a the one presented here, can serve as the basis for the construction of more extensive or targeted lexica, for annotating Slovene corpora, or as a dataset for machine learning of morphological relations.

One of the objectives of the MULTEXT-East project has been to make its resources freely available for research purposes. In the scope of the TELRI (Trans European Language Resources Infrastructure) concerted action, the complete results of TELRI and MULTEXT-East have been released on a CD-ROM [5]. Extensive information and the complete documentation of the project are also available via WWW, on *http://nl.ijs.si/ME/*. It should be noted, however, that the results of the project, including the lexical resources, are being maintained; the Slovene resources presented in this article have already been updated since the finish of MULTEXT-East. To get the latest version, interested parties should contact the author.

6 Acknowledgements

Many people have, directly or indirectly contributed to the results described in this paper. Peter Holozan and Vladimír Petkevič gave essential input for the morphosyntactic specifications for Slovene. Velimir Gjurin, France Žagar, David Stermole, and Lydia Sinapova gave much appreciated comments and suggestions. All errors, of course, remain the responsibility of the author. The involvement of the company Amebis d.o.o. in the MULTEXT-East project was essential for the production of the Slovene resources, especially the lexicon described in this paper; it was derived by them via their spelling checker Besana. Aleksandra Bizjak and Primož Jakopin, from the Institute for Slovene Language "Fran Ramovš", ZRC SAZU performed the MSD tagging of the MULTEXT-East Slovene corpus. The work reported here was funded by the EU project Cop 106, MULTEXT-East and by the supporting grant MS-3/95, U3-BR-273/95 of the Ministry of Science and Technology of Slovenia.

References

[1] N. Bel, N. Calzolari & M. Monachini (eds.). Common specifications and notation for lexicon encoding and preliminary proposal for the tagsets. MULTEXT Deliverable D1.6.1B, ILC, Pisa, 1995.

[2] N. Calzolari & J. McNaught (eds.). Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages. EAGLES Document EAG—CLWG—MORPHSYN/R, ILC, Pisa, 1996. http://www.ilc.pi.cnr.it/EAGLES/home.html.

[3] L. Dimitrova, T. Erjavec, N. Ide, H-J. Kaalep, V. Petkevič & D. Tufiş. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL '98*, Montréal, 1998.

[4] T. Erjavec & Monica Monachini (eds.). Specifications and notation for lexicon encoding. MULTEXT-East Final Report D1.1F, Jožef Stefan Institute, Ljubljana, 1997.

[5] T. Erjavec, A. Lawson & L. Romary (eds.). East meets West: A Compendium of Multilingual Resources. CD-ROM, 1998. TELRI Association e.V. ISBN: 3-922641-46-6.

[6] N. Ide, D. Tufiş & T. Erjavec. Development and Assessment of Common Lexical Specifications for Six Central and Eastern European Languages. In *Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98*, pp. 233–240, Granada, 1998. ELRA.

[7] N. Ide and J. Véronis. Multext (multilingual tools and corpora). In *Proceedings of the 15th Interna-*

tional Conference on Computational Linguistics, pp. 90–96, Kyoto, 1994.

[8] P. Jakopin & A. Bezjak. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična Revija*, 45(3-4):513–532, 1997.