

Creating Slovenian Language Resources for Development of Speech-to-Speech Translation Components

Darinka Verdonik, Matej Rojc, Zdravko Kačič

University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova ul. 17, Maribor, Slovenia
{darinka.verdonik,matej.rojc,kacic}@uni-mb.si

Abstract

Article brings detailed information about procedures of building Slovenian lexica within the LC-STAR project, and also detailed information about the size of that lexica. University of Maribor joined the LC-STAR project in order to provide appropriate language resources for developing speech-to-speech translation technology for Slovenian language. Lexica exists from three parts: 65.000 common words, 45.000 proper names and 6.000 special application domain words. All lexica will be morpho-syntactically tagged and phonetically transcribed. Quality of produced language resources is ensured by independent validation.

1. Introduction

In order to provide appropriate language resources for Slovenian language for developing speech-to-speech translation components (speech recognition, speech centered translation, text-to-speech synthesis) University of Maribor joined the LC-STAR project (www.lc-star.com; Hartikainen et al., 2003; Moreno, 2003) as external partner. Existing data for lexical language resources for Slovenian are only few (Multext-East (Erjavec, 1998), SImflex (Verdonik et al., 2002), Onomastica¹) and have many drawbacks like similar language resources for other languages: lack of coverage with respect to wide range of application domains, lack of suitability either for speech synthesis or speech recognition, lack of quality control, lack of standards, they are mostly suitable for research purposes. Since "the main objective of the LC-STAR project is to make large lexica available ... that cover a wide range of domains along with the development of standards relating to content and quality" (Hartikainen et al., 2003: 1529), we expect to obtain high quality data compatible with data for many other languages included in the project.

2. Word lists collection and size of word lists

The lexicon consist of three parts: 1) about 65.000 inflected common words entries, 2) about 45.000 proper names, 3) about 6.000 entries for special voice-driven application.

2.1. Common words

Common words entries are extracted from 12 mio. words clean corpora covering six major domains: sport, news, finance, culture and entertainment, consumer information, and personal communication. Major resources for collecting corpora were *Vecer* and *Delo*, two biggest Slovenian daily newspapers. For consumer communications domain texts from some supplements to those newspapers, texts from popular science magazine *Gea*, popular science portal *Svarog* and some manuals from Internet were added. Domains coverage is shown in table 1.

¹ Onomastica is electronic lexicon of Slovenian proper names, made within the framework of the Onomastica project COP57 from 1995.

Domains	Clean Corpora	
	Size	Distinct words
Sports/Games	1.888.753	42.029
News	2.178.834	63.912
Finance	3.411.268	62.837
Culture/Entertainment	2.716.028	82.305
Consumer Information	1.146.009	48.326
Personal communications	950.179	40.500
Total	12.291.071	139.645

Table 1: Size of clean corpora (i.e. without digits, punctuation marks, most common typos, proper names, abbreviations and singletons) for Slovenian language and distribution into domains.

From this corpora word lists were extracted. The size of word lists is shown in table 2. The column Number presents number of entries per domain and in total, and column Coverage presents coverage of input corpora in % for each domain and in total.

Domains	Different Words	
	Number	Coverage
Sports/Games	17.256	96%
News	31.479	96%
Finance	24.469	96%
Culture/Entertainment	40.847	96%
Consumer Information	28.418	96%
Personal communications	23.923	96%
Total	65.096	98,11%

Table 2: Word list: number of different words and their corresponding corpora coverage per domain and total.

2.1.1. Word list selection procedure

Procedure for extracting word list from corpora was the following:

1. The corpora was cleaned and tokenised for all domains. This involves removal of digits, punctuation marks and most common typos. Verification on size requirements on domains was performed.

2. Proper names and abbreviations were removed automatically as much as possible.
3. The number of occurrences of all distinct tokens (words) in the corpus was counted and words were sorted by the observed frequencies.
4. Since the words are repeated among word lists from different domains, word lists were first merged, and remaining proper names and abbreviations were semi-automatically detected (using the spellchecker for Slovenian) and manually verified. Then new word lists were constructed from word lists obtained in the 3rd step, eliminating manually found proper names and abbreviations.
5. A coverage target $t = 96\%$ was chosen to include some safety margin (according to the specifications at least 95% coverage was demanded). The relative frequencies were calculated and the running sum of relative frequencies, called the rank coverage, was calculated for each domain. We truncate the final word lists at the point where the target $t = 96\%$ was exceeded and output the j words into word list files for each domain.
6. Word lists for all domains were merged to obtain the final word list. The word list was longer than 50.000 entries according to the specifications.

2.1.2. Tokenisation procedure

Tokenisation procedure was the following:

1. White spaces, tabulator, new lines characters, punctuation marks “.,:;!#><’ and brackets []{}() separate tokens. Exceptions are: some abbreviations (if token is shorter than 4 characters, it is followed by dot (.) and next word isn't capitalized, it is considered for abbreviation), e-mails (include @ symbol), web addresses.
2. Words with hyphen are split into parts, if these parts exist as separate words.
3. Capitalized words occurring after dot (.), exclamation mark (!), question mark (?) or colon (:), are decapitalized. Remove digits and punctuation marks “.,:;!#><’[]{}()\$%&”
4. Tokens shorter than 4 characters and written with all letters capitalized are considered for abbreviations.
5. Tokens longer than 20 characters are removed.
6. Use capitalization as indicator for proper names.
7. Use corpus of Slovenian proper names constructed in Onomastica project to remove proper names left.
8. Remove singletons.
9. Use the spellchecker and manually remove abbreviations and proper names left.

2.1.3. Closed sets

In addition to extract common word lists from corpora a list of closed set (function) word classes was included. Function words are frequently used in various domains, however not all were covered by corpora collection. Therefore closed sets were added manually. For Slovenian language they are defined in 13 groups:

- adpositions
- conjunctions
- question particles, particles of agreement and denying
- demonstrative pronouns

- indefinite and negative pronouns + more pronoun types from Slovenian grammar: *poljubnostni, mnogostni, totalni, drugostni, istostni, celostni*
- interrogative pronouns
- personal pronouns
- personal possessive pronouns
- personal and possessive reflexive pronouns
- relative pronouns; includes two pronoun types from Slovenian grammar: *oziralni, oziralno poljubnostni*
- auxiliary verb *biti*
- modal verbs
- inflectional endings

All together more than 1000 word forms were collected this way, more than 700 of them were already collected within the corpora.

2.2. Proper names

The second part of lexica are proper names. The size of Slovenian proper names lexica is 45.027 different entries. This entries are divided in 3 major domains: first and last names, place names and organization names. Table 3 shows distribution of collected proper names in Slovenian LC-STAR lexicon per particular domain in number of entries and in %.

Domains	Size per domain	%
First and last names	22.469	49,2
Place names	11.828	25,9
Organizations	11.357	24,9
Total entries	45.654	100
Total different entries	45.027	

Table 3: Proper names. Domains and size per domain in number of words and in %.

Compound proper names, like *Stari_trg, Novo_mesto, Blejsko_jezero...*, count as one entry. In total 45.027 different entries were collected. Some of this entries can be in two or three different domains at the same time (because, for example, the same entry can be person name or organization name).

Most of resources were electronically available, however geographic names, major capitals, important and well known cities, cultural and historic places and some brand names had to be added manually. All organization and place names were manually checked (for correct capitalization etc.).

First and last names were taken from Onomastica, which contains names of all inhabitants in Slovenia from the year 1995. Names that include foreign characters (like y, q ...) or foreign sequence of characters (zz, eau...) were automatically eliminated. Final list was checked for including the most common person names.

All Slovenian city names and all Slovenian street names were included (sources Lexicon of Slovenian place names and Slovenian phone book), all countries and all capitals. Names of the biggest cities and names of most known,

popular or biggest geographic places were added manually.

From Slovenian phone book all names of companies and organizations at yellow pages were included. Some names were repeating only slightly changed (like primary schools, named by the name of the settlement (*Osnovna šola Dragonja, Osnovna šola Volčja vas...*)) – in this cases most of repeating entries were manually eliminated. Slovenian brand names were added manually. Most common international brand names were added partly manually and partly from the Internet.

2.3. Special application domain

The special application word list consists of numbers, letters, abbreviations and six major semantic domains related to voice-driven applications. For voice driven application a reference word list of 5.700 entries in US-English was provided by LC-STAR consortium and translated into other languages, including Slovenian. Table 4 shows domains and number of Slovenian entries per domain and in total.

Domains	No. of lexicon entries
Numbers, letters and digits	137
Abbreviations	373
US-English reference word list	
<i>Global domains</i>	
Measures	161
Abbreviations	491
Special signs	61
Domestic equipment	156
Health	108
Greetings	37
<i>(Information) Services/Retrieval</i>	
Financial/Commerce	667
Billing	666
<i>Travelling</i>	
Travelling	839
<i>Information Services</i>	
Services (general scenarios)	1212
<i>Retrieval/Controll</i>	
Retrieval	930
Control	534
<i>Telecommunications</i>	
Telecom	644
Web/Internet	543
Total	6040

Table 4: Domains and number of Slovenian entries per domain and in total.

3510 of total 6040 entries were already collected within common word list, selected from the 12. mio. corpora, therefore 2530 new entries were added to common words lexica. All together (common words, closed sets, proper names and special application words) 113.000 entries were collected.

3. Morpho-syntactic tags

Collected word lists will be morpho-syntactically tagged according to LC-STAR specifications (Hartikainen et al., 2003; www.lc-star.com) and phonetically represented in SAMPA symbols for Slovenian (Zemljak et al., 2002). Here we represent detailed POS scheme for Slovenian language, adjusted to formally specified grammar for LC-STAR lexica (Hartikainen et al., 2003).

NOM (Common and proper nouns)

Class: *common, PER* (person), *GEO* (geographic name), *COU* (country), *CIT* (city), *STR* (street), *COM* (organization), *BRA* (brand), *TOU* (cultural/historic place).
Number: *singular, plural, dual, invariant*.

Gender: *masculine, feminine, neuter, invariant*.

Case: *nominative, genitive, dative, accusative, locative, instrumentative, indeclinable*.

Type: *animated, not_animated* (applies only to *masculine, singular, accusative*).

ADJ (Adjective)

Number: *singular, plural, dual, invariant*.

Gender: *masculine, feminine, neuter, invariant*.

Case: *nominative, genitive, dative, accusative, locative, instrumentative, invariant*.

Degree: *positive, comparative, superlative*.

NUM (Numerals)

Number: *singular, plural, dual*.

Gender: *masculine, feminine, neuter, invariant*.

Case: *nominative, genitive, dative, accusative, locative, instrumentative, indeclinable*.

Type: *cardinal, ordinal, multiplicative, distributive*.

VER (Verb)

Number: *singular, plural, dual*.

Gender: *masculine, feminine, neuter* (applies only to *participle*).

Person: *1, 2, 3* (applies only to *indicative* and *imperative*).

Mood: *indicative, conditional, imperative, infinitive* (includes two types in Slovenian grammar, infinitive and supine), *participle* (The past participle is used for making all the compound active tenses (future, past, pluperfect) and is encoded *participle, NS* (for *tense*), *active*. The passive participles are encoded as *participle, NS* (for *tense*), *passive*, when they can be used in predicative position (e.g. *on je bil tepen* / he was beaten). When they can be used only in attributive position, they are classified as adjectives. The adjectival (e.g. *stokajoč* / moaning) and adverbial (e.g. *leže* / lying down) participles are classified as adjectives and adverbs respectively.)

Tense: *present, past, future*.

Voice: *active, passive*.

Polarity: *positive, negative* (applies only to present tense of verbs *biti* (to be), *hoteti* (to want), *imeti* (to have)).

Aspect: *imperfect, perfect*.

AUX (Auxiliary verb) (only verb *biti* (to be) in all it's forms)

Number: *singular, plural, dual*.

Gender: *masculine, feminine, neuter*.

Person: 1, 2, 3.

Mood: *indicative, conditional, imperative, infinitive, participle.*

Tense: *present, past, future.*

Voice: *active.*

Aspect: *imperfect.*

PRO (Pronoun)

Number: *singular, plural, dual, invariant.*

Gender: *masculine, feminine, neuter, invariant.*

Person: 1, 2, 3, *invariant* (applies only to personal and possessive pronouns).

Case: *nominative, genitive, dative, accusative, locative, instrumentative, indeclinable.*

Type: *personal, demonstrative, reflexive, indefinite* (includes many of pronoun types from Slovenian grammars: *nedoločni, poljubnostni, mnogostni, istostni, drugostni, celostni, nikalni*), *interrogative, relative* (includes two of pronoun types from Slovenian grammars: *oziralni, oziralno-poljubnostni*), *possessive.*

ADV (Adverb)

Degree: *positive, comparative, superlative.*

Type: *time, place, manner.*

CON (Conjunction)

ADP (Adposition)

INT (Interjection)

PAR (Particles)

4. Validation

One of advantages of LC-STAR project is assuring high quality of language resources. In order to achieve this all language resources, made within the project, will be validated by ELRA's validation centre SPEX (Speech Processing EXPertise centre), which is the main validation centre for the LC-STAR. The validation warrants that each partner's contribution meets the same high quality standards.

Validation is partly automatic and partly manual. SPEX has developed tools for formal, automatic checks. The tools were distributed to partners so they are able to pre-check their own lexicons. The formal check is re-done by SPEX to confirm and report the results. Manual checks on linguistic aspects of the lexicons are done by CST (Center for Sprogteknologi), Denmark. Manual validation is performed by independent native speaker experts and includes the validation of closed class word sets and of POS-tags.

5. Conclusion

The article presents Slovenian lexica for developing speech-to-speech translation systems, which is being built within the LC-STAR project. For Slovenian, like for many other languages, the lack of quality, lack of wide coverage, lack of suitability for speech synthesis or speech recognition and lack of standards at existing written language resources for speech-to-speech translation components can be observed. The developed Slovenian language resources as other developed resources in the framework of the LC-STAR project will overcome most of these drawbacks. The high quality lexica and corpora

will represent solid foundation for development of speech-to-speech translation components for Slovenian language.

6. References

- Erjavec, T., Ide, N. (1998). The Multext-East Corpus. In the Proceedings of First International Conference on Language Resources & Evaluation (pp. 971--974). Granda, Spain.
- Hartikainen, E., Maltese, G., Moreno, A. Shamma, S., Tiegenghain, U. (2003). Large Lexica for Speech-to-Speech Translation: From Specification to Creation. In Proceedings of the Eurospeech (pp. 1529--1532). Geneva.
- Moreno, A. (2003). Project presentation. In SEPLN2003. Madrid.
- Project homepage: <http://www.lc-star.com/>.
- Zemljak, M., Kačič, Z., Dobrišek, S., Gros, J., Weiss, P. (2002). Računalniški simbolni fonetični zapis slovenskega govora. Slavistična revija, 50(2), 159--169.
- Verdonik, D., Rojc, M., Kačič, Z., Horvat, B. (2002). Zasnova in izgradnja oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik. In Proceedings of Jezikovne tehnologije/Information Society Multi-Conference (pp. 44--48). Ljubljana, Slovenia.