Chapter 5

A rule-based machine translation system from Serbo-Croatian to Macedonian

Hrvoje Peradin, Francis Tyers
University of Zagreb, Universitat d'Alacant

Abstract

This paper describes the development of a one-way machine translation system from Serbo-Croatian to Macedonian on the Apertium platform. Details of resources and development methods are given, as well as an evaluation, and general directives for future work.

5.1 Introduction

The modern Macedonian language was standardised in 1944. and is the official language of the Republic of Macedonia.

Serbo-Croatian is a term that encompasses four standard languages (Bosnian, Croatian, Montenegrin and Serbian) based on the *neoshtokavian* dialect. The standardisation of the language started in the 19th century, as an attempt to unify the literary and linguistic traditions of the south Slavic area. The standard remained pluricentric until the dissolution of Yugoslavia. Due to the large similarities between the standards we have decided to group them into one module, with a common mode for analysis. Having in mind future work we have added separate modes for generation of Bosnian, Croatian and Serbian.[1]

Serbo-Croatian and Macedonian are largely mutually intelligible, however despite their close relation the differences in morphology create difficulties in translation. For this reason the system is currently mono-directional (sh→mk). The direction was chosen since it is

[1] The standardisation process of Montenegrin is under way, and we are awaiting the outcome to implement a separate mode.

easier to construct a system translating from a more detailed morphology (e.g. Macedonian "во куќа" can be translated both as "u kući" [in the house.LOC] and "u kuću"[into the house.ACC]).

Other systems currently supporting the languages are notably Google Translate[2] and Systran.[3]

The language pair `apertium-sh-mk`[4] is available under GNU GPL.

5.2 Design

The Apertium platform

The Apertium[5] platform follows a modular machine translation model. Morphological analysis of the source text is performed by a letter transducer compiled from a morphological lexicon,[6] and cohorts[7] obtained in this manner go through a disambiguation process. Disambiguated readings proceed to a bilingual dictionary also performed by a letter transducer and through a two-level syntactic transfer, which performs word reordering, deletions, insertions, and basic syntactic chunking. The final module is a letter transducer that generates surface forms in the target language from the bilingual transfer output.

Constraint Grammar

The disambiguation in this language pair is performed by a Constraint Grammar (CG) module[8]. CG is a paradigm that uses hand-written rules to reduce the problem of linguistic ambiguity. A series of context-dependent rules are applied to a stream of tokens and readings for a given surface form are excluded, selected or assigned additional tags.

5.3 Development

Resources

Although some resources for morphological analysis of Serbian and Croatian exist (Vitas and Krstev, 2004, Vitas et al., 2003, Agić et al., 2008, Šnajder et al., 2008), to our knowledge there are none freely available for either Serbian, Bosnian or Croatian. Thus the monolingual dictionary for Serbo-Croatian has been developed almost from scratch, with the aid of a Croatian grammar (Barić et al., 1997), and on-line resources such as *Hrvatski jezični portal*,[9]

[2] Supports Croatian, Serbian and Macedonian.

[3] Language pairs Serbian→English, Croatian→English.

[4] `http://wiki.apertium.org/wiki/Serbo-Croatian_and_Macedonian`

[5] `http://wiki.apertium.org/`

[6] A morphological lexicon contains ordered pairs of word surface forms and their lemmatised analyses.

[7] A cohort consists of a surface form and one or more readings containing the lemma of the word and the morphological analysis.

[8] Implemented in the CG3 formalism, using the `vislcg3` compiler, available under GNU GPL. For a detailed reference see: `http://beta.visl.sdu.dk/cg3.html`

[9] `http://hjp.srce.hr`

wiktionaries and Wikipedia, as well as an SETimes corpus[10] (Tyers and Alperen, 2010) and a corpus composed from the Serbian, Bosnian, Croatian and Serbo-Croatian Wikipedias.

Bilingual resources available were also scarce. We used a parallel corpus obtained from SETimes, and a Serbian–Macedonian dictionary.[11]

The morphological analyser/generator for Macedonian was taken from `apertium-mk-bg` (Rangelov, 2011), which is freely available under GNU GPL. For reference on the Macedonian language we used the SEELRC reference grammar[12] and Дигитален речник на македонскиот јазик.[13]

Analysis and generation

The morphological analyser for Serbo-Croatian was written in the XML formalism of *lttoolbox*[14] (Ortiz-Rojas et al., 2005), almost entirely from scratch, with the aim to match the lexicon of the analyser from `apertium-mk-bg`. Since we intended to create a resource for all three standards, a paradigm was assigned to the reflex of the vowel yat[15] to enable analysis of both ekavian and ijekavian dialects (for a more detailed reference on Serbo-Croatian dialects see Brown and Alt, 2004), and the extended metadix format was used to enable separating different standards by analysis and generation modes.

The basic inflectional paradigms were taken from the Croatian grammar (Barić et al., 1997), and further refined according to new entries (e.g. with voice changes not covered by basic declension patterns).

The entries were made mostly manually, with some proper nouns obtained semi-automatically from the Macedonian dictionary.

Disambiguation

As there was no reliable, free training corpus, and the target-language based training of Sánchez-Martínez et al. (2008) only supports 1-stage transfer, we elected to do the disambiguation solely by a Constraint Grammar module, and omit the statistical tagger component standardly used in Apertium language pairs. In case of remaining ambiguity, the system picks the first analysis from the output of the disambiguation module.

The following are examples of disambiguation rules:

- Preposition-based case disambiguation:

  (9) ...u mojoj kući...
  [in.PR.GEN/ACC/LOC] [my.PRN.DAT/LOC] [house.DAT/LOC]
  (in my house)

[10] `http://opus.lingfil.uu.se/SETIMES.php`

[11] `http://rechnik.on.net.mk/`

[12] `http://slaviccenters.duke.edu/projects/grammars`

[13] A digital dictionary of the Macedonian language, `http://www.makedonski.info/`

[14] `http://wiki.apertium.org/wiki/Lttoolbox`

[15] Typically in ekavian it is either a long or short "e", while in ijekavian the long variant is reflected as "ije", and the short as "je".

```
REMOVE Prep + $$Case IF (1 Nominal - $$Case)
REMOVE Nominal + $$Case IF (NOT -1 Prep + $$Case) (NOT -1 Modifier + $$Case)
```

The first rule cleans a grammatical case from a preposition[16] if it is not followed by a noun, pronoun or adjective in the same case. The second rule, similarly, cleans a case from a noun, pronoun or adjective if it is not preceded directly either by a preposition which governs the case or a modifier (e.g. adjective or demonstrative pronoun) in the same case.[17] Thus the whole phrase is correctly disambiguated as locative.

- Noun phrases:

(10) ...lijepa žena...
[pretty.ADJ.(NT.PL)/(F.SG)] [woman.N.(F.SG)/(F.PL)]
(a pretty woman)

```
REMOVE Modifier + $$GenNum IF (1 Nominal - $$GenNum)
REMOVE Nominal + $$GenNum IF (-1 Modifier - $$GenNum)
```

These rules operate on noun phrases, and use the gender and number agreement to eliminate grammatically impossible readings. In this example the first rule removes the neuter reading from the adjective, since the noun it agrees with does not have the neuter gender. The adjective is then left only with the singular reading and the second rule proceeds to remove the plural reading from the noun.

- Adverb / adjective ambiguity:

(11) On puno radi.
[he] [full.(ADJ.NT.SG)/(ADV)] [works.VB]
(He works a lot.)

```
SELECT Adverb IF (0 Adverb OR Adjective) (1 Verb)
```

This simple rule resolves a common ambiguity by selecting the adverb reading if the word is followed by a verb.

- Dative / locative ambiguity:

(12) Brod prilazi luci.
[ship] [approaches] [harbour.DAT/LOC]
(The ship is approaching the harbour.)

```
SELECT Dative IF (0 Dative OR Locative) (NOT -1 Prep) (NOT -1 Modifier +
Locative)
```

The cases are orthographically identical, however locative is purely prepositional, so in most cases the ambiguity is easily resolved by selecting dative if the phrase is not preceded by a locative preposition.

[16] The cases the prepositions govern are marked on the analyses of the prepositions.

[17] The $$ prefix signifies unification, i.e. iteration over the set of all grammatical cases.

Lexical transfer

The bilingual lexicon was written using the *lttoolbox* format, and composed mostly manually, with paradigms added to compensate the tag set differences. Translation entries were added according to the lexicon from the Macedonian analyser. Having in mind future work, translations specific solely to Bosnian, Croatian or Serbian standard were grouped in respective sections.

Syntactic transfer

Despite the close relation of the two languages, there are substantial differences in morphology, and structures with analogous functionality are not necessarily morphologically cognate. Therefore we have used a two level syntactic transfer.

The first level performs tag mappings, normalisation (e.g. case to nominative, infinitive to present), rudimentary transformations, and packing of phrases in syntactic chunks.

Examples of transfer rules:

- The future tense:

  (13) Ja ću gledati[18] → Јаз ќе гледам
  [I] [will.CLT.P1.SG] [watch.INF] → [I] [will.CLT] [watch.PRES.P1.SG]
  (I will watch.)

  Serbo-Croatian uses a clitic + infinitive form with a declinable clitic, while Macedonian uses a frozen clitic form, and the person/number is marked on the verb. Thus several rules were written to match occurrences of future tense and transfer the information in translation.

- Clitic reordering:

  (14) Okrenut ću se → Ќе се обрнам
  [turn.INF] [will.CLT.P1.SG] [myself.CLT] →
  [will.CLT] [myself.CLT] [turn.PRES.P1.SG]
  (I will turn myself around)

  The order of clitics in both languages is different, so a series of rules was written to rearrange them.

- Cases as prepositional phrases:

  (15) Let avionom[19] → Летање со авион
  [flight] [by aeroplane.INS] → [flight] [with] [aeroplane]
  (Flying by an aeroplane.)

[18] The encliticised future tense forms (gledat ću / gledaću) are handled equally.

[19] The Croatian normative 'zrakoplov' is also accepted and translated equally.

Table 5.1: Status of `apertium-sh-mk` as of April 11 2011.

Module	Entries / Rules
Serbo-Croatian dictionary	7564
Macedonian dictionary	8672
Bilingual dictionary	9985 (unique)
Transfer rules (1 and 2)	51 + 11
Serbo-Croatian CG	170

While Serbo-Croatian has seven morphological cases, Macedonian has completely replaced its declension system with analytic, prepositional and clitic constructions. The second level of transfer replaces simple noun and adjective phrases with prepositional constructions.

- Inference of definiteness:

  (16) U sastavu Vojske Srbije → Во составот на Српската војска[20]
  [in] [composition] [of Serbian Army] → [in] [composition.DEF] [of Serbian Army]
  (In the composition of the Serbian Army)

  The definite article in Macedonian has no analogy in Serbo-Croatian (except to some extent the definiteness of adjectives). This transfer rule infers definiteness for a common noun preceding a proper noun in genitive.

- A clear definiteness transfer:

  (17) Lijep dan → Ував ден
  [lovely.IND] [day] → [lovely.IND] [day]
  (A lovely day)
  Lijepi dan → Убавиот ден
  [lovely.DEF] [day] → [lovely.DEF] [day]
  (The lovely day)

  For a class of adjectives in Serbo-Croatian definiteness can be distinctly marked. In such cases it can be directly used in translation.

Status

The current status of the language pair is given in Table 5.1.

5.4 Evaluation

This section presents an evaluation of the system performance, with coverage measured on two corpora, and a quantitative analysis.

[20] The article in Macedonian attaches to the first constituent of the noun phrase.

Table 5.2: Coverage

Corpus	Coverage	Std. dev.
Wikipedia (sh+bs+sr+hr)	73.12%	0.36
SETimes (sr + hr)	82.64%	0.38

Coverage

The data for coverage of the Serbo-Croatian analyser is given in Table 5.2. Coverage is naive, it means that for any given form in the source language at least one analysis has been given. The analyser has been tested on a combined Wikipedia corpus, and on a corpus of Serbian and Croatian SETimes articles. The corpora was divided in four parts and average coverage calculated.

Quantitative evaluation

Quantitative evaluation has been performed on four articles from SETimes. The articles were translated by Apertium, and post-edited by a human translator.

The first two articles were selected with nearly full coverage to get an idea of how disambiguation and transfer rules work in ideal circumstances, while the remaining two provide an assessment of the system's practical quality.

The word error rate (WER) and the position-independent error rate (PER) were calculated by the number of changes the human editor needed to make. Results are given in Table 5.3.

Table 5.3: Quantitative evaluation

Article	OOV[1]	Words	WER	PER	Translit.[2]
`setimes.pilots.txt`	0.4%	454	29.9%	20.5%	97.5%
`setimes.tablice.txt`	0.4%	470	48.1%	34.6%	85.2%
`setimes.klupa.txt`	18.1%	480	60.4%	46.8%	82.7%
`setimes.povijest.txt`	14.2%	529	53.4%	40.5%	84.8%

[1] Out of vocabulary words

[2] Baseline WER, obtained by transliteration of the source text

Common problems

Although CG rules successfully rule out quite a lot of grammatically impossible analyses, the number of rules for this language pair is quite low, so disambiguation is not always correct.

Another obvious source of errors are unknown words, which typically disrupt the flow of disambiguation, especially when they occur inside noun phrases.

The definite article is quite difficult to infer. Though in limited cases it can be transferred from definite adjectives, or guessed from specific context, there is e.g. no straightforward way to mark a subject previously introduced in discourse as definite.

Serbo-Croatian cases do not translate consistently to prepositional constructions. A notable example is the partitive vs. possessive genitive. The phrase "čaša vode" can be translated as "чаша вода" ("a glass of water") or "чашата на вода" (the water's glass).

Both languages have a very free word order of the main constituents. For instance, an adjective can agree with a noun arbitrarily far to the left:

(18) *Vožnja*.N.FEM zrakoplovom ... bila je *odlučujuća*.ADJ.FEM → *Возенје*.N.NEUT со авионот ... беше *решавачка*.ADJ.FEM

[The airplane *ride* ... was *decisive*]

If the noun changes gender in translation, the adjective is not matched to it, and retains the source language gender.

5.5 Discussion

This paper presented the design and an evaluation of a language pair for the Apertium platform. It is the first rule-based MT system between Serbo-Croatian[21] and Macedonian, and the morphological analyser and CG module are currently only such open-source resources for the languages.

The system was dubbed by a native speaker as overall fine, there are obvious errors, but the output is legible and easily post-editable.

A significant part of the problems is typical for a system in such an early phase of development. The morphological lexicons for both languages are small, and the same remark can be made for the number of disambiguation rules.

Some ambiguities that arise in analysis of the source language are difficult or impossible to resolve in a simple rule-based manner, which suggests that the system should be combined with machine learning and statistical methods.

In terms of future work the essential task is to increase coverage, to enable working with larger corpora, and to improve the disambiguation rules, which make a significant contribution to translation quality.

Acknowledgements

The development of this language pair was funded as a part of the Google Summer of Code.[22] Many thanks to Jimmy O'Regan, Kevin Unhammer and other Apertium contributors for their help on Apertium, to Tino Didriksen for his help on CG, to Tihomir Rangelov for his advice in the initial phase of this work, and to Dime Mitrevski for his input on the Macedonian language.

[21] It is to our knowledge the first MT system supporting Bosnian.

[22] `http://code.google.com/soc/`

Bibliography

Agić, Ž., M. Tadić, and Z. Dovedan. 2008. Improving part-of-speech tagging accuracy for Croatian by morphological analysis. *Informatica* 32(4):445–451.

Barić, E., M. Lončarić, D. Malić, S. Pavešić, M. Peti, V. Zečević, M. Znika, et al. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.

Brown, W. and T. Alt. 2004. *A handbook of Bosnian, Serbian, and Croatian*. SEELRC.

Forcada, M.L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J.A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F.M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation* pages 1–18.

Ortiz-Rojas, S.O., M.L. Forcada, and G.R. Sánchez. 2005. Construcción y minimización eficiente de transductores de letras a partir de diccionarios con paradigmas. *Procesamiento de Lenguaje Natural* 35:51–57.

Rangelov, T. 2011. Rule-based machine translation between Bulgarian and Macedonian. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation (2011: Barcelona)*.

Sánchez-Martínez, F., J.A. Pérez-Ortiz, and M.L. Forcada. 2008. Using target-language information to train part-of-speech taggers for machine translation. *Machine Translation* 22(1):29–66.

Tyers, F. and M.S. Alperen. 2010. South-East European times: A parallel corpus of Balkan languages. In *Forthcoming in the proceedings of the LREC workshop on "Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*.

Vitas, D. and C. Krstev. 2004. Intex and Slavonic morphology. *INTEX pour la linguistique et le traitement automatique des langues, Presses Universitaires de Franche-Comté* pages 19–33.

Vitas, D., G. Pavlović-Lažetić, C. Krstev, L. Popović, and I. Obradović. 2003. Processing Serbian written texts: An overview of resources and basic tools. In *Workshop on Balkan Language Resources and Tools*, vol. 21, pages 97–104.

Šnajder, J., Bojana B. Dalbelo Bašić, and M. Tadić. 2008. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing and Management* 44(5):1720–1731.