Shallow-transfer rule-based machine translation for the Western group of South Slavic languages

Hrvoje Peradin
University of Zagreb
Faculty of Science
Department of Mathematics
hperadin@gmail.com

Filip Petkovski
University of Zagreb
Faculty of Electrical Engineering
and Computer Science
filip.petkovski@fer.hr

Francis M. Tyers
Dept. Lleng. i Sist. Inform.
Universitat d'Alacant
E-03899 Alacant
ftyers@dlsi.ua.es

Abstract

The South Slavic languages, spoken mostly in the Balkans, make up one of the three Slavic branches. The South Slavic branch is in turn comprised of two subgroups, the Eastern subgroup containing Macedonian and Bulgarian, and the western subgroup containing Serbo-Croatian and Slovenian. This paper describes the development of a bidirectional machine translation system for the western branch of South-Slavic languages — Serbo-Croatian and Slovenian. Both languages have a free word order, are highly inflected, and share a great degree of mutual inteligibility. We give details on resources and development methods used, as well as an evaluation, and general directions for future work.

1 Introduction

The South Slavic language branch, which is spoken mostly in the Balkans, makes up one of the three Slavic branches. The South Slavic branch itself is in turn comprised of two subgroups, the Eastern subgroup containing Macedonian and Bulgarian, and the western subgroup containing Serbo-Croatian and Slovenian.

The Serbo-Croatian (hbs)[1] dialects are the native language of most people in Serbia, Croatia, Montenegro and Bosnia and Herzegovina. They were formed on the basis of the *štokavian* dialects which got their name from the form *što* (or *šta*), which is used for the interrogative pronoun 'what?'. A second group of dialects from the Serbo-Croatian language group is the Čakavian group spoken in western Croatia, Istria, the coast of Dalmatia, and some islands in the Adriatic. Like the štokavian dialects, the *čakavian* dialects got their name from the form *ča* used for the same interrogative pronoun. Finally, the third main group of Serbo-Croatian dialects, spoken in northwestern Croatia, uses *kaj* instead of *što*, and is called *kajkavian*. An intermediate dialect between Serbo-Croatian, Bulgarian and Macedonian is the Torlakian dialect. The three or four standardised varieties of Serbo-Croatian are all based on the štokavian dialect.

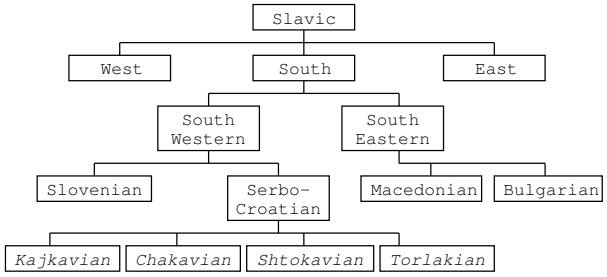

Figure 1: A traditional division of the South-Slavic languages. All four standard varieties of Serbo-Croatian (Bosnian, Croatian, Montenegrin, and Serbian) are based on the štokavian dialect.

Slovenian (slv) is the native language of Slovenia, and is also spoken in the neighbouring areas in Italy and Austria. While Slovenian has many different dialects, it shares some features with the Kajkavian dialects spoken in Croatia. Although the speakers of the different Serbo-Croatian dialects can understand each other without any serious difficulties, a Serbo-Croatian speaker can have a difficult time understanding a speaker of a Slovenian dialect.

[1] We use the term 'Serbo-Croatian' as an abbreviation for Bosnian-Croatian-Montenegrin-Serbian.

	Bosnian	Croatian	Montenegrin	Serbian
Čakavian	-	-i-,-e-,-je-	-	-
Kajkavian	-	-e-,-ie-,-ei,-i-	-	-
Štokavian	-ije-,-je-	-ije-,-je-,-i-	-ije-,-je-	-e-,-ije-,-je-
Torlakian	-	-	-	-e-

Table 1: Intersection of Serbo-Croatian languages and dialects. All four standard variants are based on the štokavian dialect, but other dialects are considered to *belong* to a standard. The -ije-, -e-, and -i- correspond to the *yat* reflex.

2 Design

2.1 The Apertium platform

The Apertium[2] platform is a modular machine translation system. The typical core layout consists of a letter transducer morphological lexicon[3]. The transducer produces cohorts[4] which are then subjected to a morphological disambiguation process. Disambiguated readings are then looked up in the bilingual dictionary, which gives the possible translations for each reading. These are then passed through a lexical-selection module (Tyers et al., 2012), which applies rules which select the most appropriate translation for a given source-language context. After lexical selection, the readings, which are now pairs of source and target language lexical forms are passed through a syntactic transfer module that performs word reordering, deletions, insertions, and basic syntactic chunking. The final module is another letter transducer which generates surface forms in the target language from the bilingual transfer output cohorts.

2.2 Constraint Grammar

This language pair uses a Constraint Grammar (CG) module[5] for disambiguation. The CG formalism consists of hand-written rules that are applied to a stream of tokens. Depending on the morphosyntactic context of a given token the rules select or exclude readings of a given surface form, or assign additional tags.

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

[3] A list of ordered pairs of word surface forms and their lemmatised analyses.

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

[5] 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`

3 Development

3.1 Resources

This language pair was developed with the aid of on-line resources containing word definitions and flective paradigms, such as *Hrvatski jezični portal*[6] for the Serbo-Croatian side. For the Slovenian side we used a similar online resource *Slovar slovenskega knjižnega jezika*,[7], and the *Amebis Besana* flective lexicon.[8]

The bilingual dictionary for the language pair was developed from scratch, using the *EUDict*[9] online dictionary and other online resources.

3.2 Morphological analysis and generation

The basis for this language pair are the morphological lexicons for Serbo-Croatian (from the language pair Serbo-Croatian–Macedonian, `apertium-hbs-mak`) and Slovenian (from the language pair Slovenian–Spanish, `apertium-slv-spa`). Both lexicons are written in the XML formalism of *lttoolbox*[10] (Ortiz-Rojas et al., 2005), and were developed as parts of their respective language pairs, during the Google Summer of Code 2011.[11] Since the lexicons had been developed using different frequency lists, and slightly different tagsets, they have been further trimmed and updated to synchronise their coverage.

3.3 Disambiguation

Though for both languages there exists a number of tools for morphological tagging and disambiguation (Vitas and Krstev, 2004, Agić et al., 2008, Šnajder et al., 2008), there are none freely available. Likewise, the adequately tagged corpora are mostly non-free (Erjavec, 2004,Tadić, 2002). Since both Serbo-Croatian and Slovenian

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

[7] `http://bos.zrc-sazu.si/sskj.html`

[8] `http://besana.amebis.si/pregibanje/`

[9] `http://eudict.com/`

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

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

are highly inflected languages, the automatically trained statistical tagger canonically used in Apertium language pairs would not give satisfactory results. For this reason we chose to use solely Constraint Grammar (CG) for disambiguation. The CG module does not provide complete disambiguation, so in the case of any remaining ambiguity the system picks the first output analysis.

Due to the similarities between the languages, we were able to reuse much of the rules developed earlier for Serbo-Croatian. Following are a few disambiguation rules examples:

- Simple adverb vs. adjective rule:

  (1) Ja često jedem ribu. ↔ Jaz pogosto jedem ribo.
  (I eat fish often.)

  For this phrase the morphological analyser gives:

```
"<Ja>"
    "free" prn pers p1 mfn sg nom
"<često>"
    "često" adv
;   "čest" adj pst nt sg nom ind
;   "čest" adj pst nt sg nom def
;   "čest" adj pst nt sg voc ind
;   "čest" adj pst nt sg voc def
;   "čest" adj pst nt sg acc ind
;   "čest" adj pst nt sg acc def
"<jedem>"
    "jesti" vb imperf tv pres p1 sg
;   "jesti" vb imperf ref pres p1 sg
;   "jesti" vb imperf iv pres p1 sg
"<ribu>"
     "riba" n f sg acc
```

  The rule used to disambiguate the adverb *često* in this phrase

```
SELECT Adv IF
    (0 Adv OR A)
    (1C V)
```

  selects the adverb reading in an adverb/adjective ambiguity if the word following is unambiguously a verb.

- Preposition based case disambiguation:

  (2) Za našo ljubo staro mater. ↔ Za našu dragu staru majku.
  (For our dear old mother.)

  Noun phrases in both languages typically generate a great number of ambiguities.

```
"<Za>"
     "za" pr acc
;    "za" pr gen
;    "za" pr ins
"<našo>"
     "naš" prn pos p1 f sg acc
;    "naš" prn pos p1 f sg ins
"<ljubo>"
     "ljub" adj f sg acc ind
;    "ljubo" adv sint
;    "ljub" adj nt sg nom ind
;    "ljub" adj nt sg acc ind
;    "ljub" adj f sg ins ind
"<staro>"
     "star" adj f sg acc ind
;    "staro" adv sint
;    "star" adj f sg ins ind
;    "star" adj nt sg nom ind
;    "star" adj nt sg acc ind
"<mater>"
     "mati" n f sg acc
;    "mati" n f du gen
;    "mati" n f pl gen
```

  First the rule

```
REMOVE Pr + $$Case IF
    (1 Nominal - $$Case)
```

  removes a case reading from a preposition if it is not followed by an adjective, noun or a pronoun in the same case, and then the rule

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

  in several passes removes the incorrect cases from the nouns and modifiers in the phrase.

- Noun heuristics:

  (3) ...izvršavanje dužnosti predstavnice...
  (...doing the duty of a representative...)

  Frequently in both languages if there is a sequence of nouns, the second noun is in genitive.

```
"<izvršavanje>"
    "izvršavanje" n nt sg nom
;   "izvršavanje" n nt sg voc
;   "izvršavanje" n nt sg acc
"<dužnosti>"
    "dužnost" n f sg gen
;   "dužnost" n f sg voc
;   "dužnost" n f pl voc
;   "dužnost" n f sg loc
;   "dužnost" n f sg dat
;   "dužnost" n f pl acc
;   "dužnost" n f sg ins
;   "dužnost" n f pl nom
;   "dužnost" n f pl gen
"<predstavnice>"
    "predstavnica" n f sg gen
;   "predstavnica" n f sg voc
;   "predstavnica" n f pl voc
;   "predstavnica" n f pl nom
;   "predstavnica" n f pl acc
```

  This simple heuristic selects a genitive reading of a noun after a noun:

```
SELECT: N + Gen IF (-1 N)
```

3.4 Lexical transfer

The lexical transfer was done with an *lttoolbox* letter transducer composed of bilingual dictionary entries. Additional paradigms were added to the transducer to compensate for the tagset notational differences.

3.5 Lexical selection

Since there was no adequate and free Slovenian – Serbo-Croatian parallel corpus, we chose to do the lexical selection relying only on hand-written rules in Apertium's lexical selection module (Tyers et al., 2012). For cases not covered by our hand-written rules, the system would choose the default translation from the bilingual dictionary. We provide examples of such lexical selection rules.

Phonetics-based lexical selection: many words from the Croatian and Serbian dialects differ in a single phoneme. An example are the words *točno* in Croatian and *tačno* in Serbian (engl. *accurate*). Such differences were solved through the lexical selection module using rules like:

```
<rule>
  <match lemma="točno" tags="adv.*">
    <select lemma="točno" tags="adv.*"/>
  </match>
</rule>
```

for Croatian, and

```
<rule>
  <match lemma="točno" tags="adv.*">
    <select lemma="tačno" tags="adv.*"/>
  </match>
</rule>
```

for Serbian and Bosnian.

Similarly, the Croatian language has the form *burza* (meaning stock exchange in English), while Serbian and Bosnian have *berza*. For those forms the following rules were written:

```
<rule>
  <match lemma="borza" tags="n.*">
    <select lemma="burza" tags="n.*"/>
  </match>
</rule>
```

for Croatian, and

```
<rule>
  <match lemma="borza" tags="n.*">
    <select lemma="berza" tags="n.*"/>
  </match>
</rule>
```

for Serbian and Bosnian.

Another example of a phonetical difference are words which have h in Croatian and Bosnian, but v in Serbian. Such words include *kuha* and *duhan* in Croatian and Bosnian, but *kuva* and *duvan* in Serbian. Similar rules were written for the forms for *porcelain* (procelan in Serbian and porculan in Croatian), *salt* (so and sol) and so on.

While the Serbian dialect accepts the Ekavian and Ikavian reflexes, the Croatian dialect uses only the Ijekavian reflex. Since the selection for the different reflexes of the *yat* vowel is done in the generation process, no rules were needed in the lexical selection module.

Internationalisms have been introduced to Croatian and Bosnian mainly through the Italian and German language whereas they have entered Serbian through French and Russian. As a result, the three dialects have developed different phonetic patterns for international words.

Examples of rules for covering such varieties include:

```
<rule>
  <match lemma="Betlehem" tags="np.*">
    <select lemma="Betlehem" tags="np.*"/>
  </match>
</rule>
```

for Croatian and Bosnian, and

```
<rule>
  <match lemma="Betlehem" tags="np.*">
    <select lemma="Vitlejem" tags="np.*"/>
  </match>
</rule>
```

for Serbian.

Finally, the Croatian months used for the Gregorian calendar have Slavic-derived names and differ from the original Latin names. For example, the Croatian language has the word *siječanj* for *January*, and the Serbian language has the word *januar*. These differences were also covered by the lexical selection module.

Besides the *yat* reflex, several other cases were not covered in the lexical selection module. These include the pronoun 'what' which has the form *što* in Croatian and the forms *što* and *šta* in Serbian and Bosnian, depending on whether the context is interrogative or relative. This ambiguity was left out of the lexical-selection module and was dealt with during generation.

3.6 Transfer

Serbo-Croatian and Slovenian are very closely related, and their morphologies are very similar. Most of the transfer rules written are thus either technical, or written to cover different word orders in the languages.

Following are examples of transfer rules, which also illustrate some contrastive characteristics of the languages:

- the future tense:

  (4) Gledal bom ↔ Gledat ću[12]
  [watch.LP.M.SG][be.CLT.P1.SG] ↔
  [watch.INF][will.CLT.P1.SG]
  (I will watch.)

  Both languages form the future tense in an analytic manner. While Slovenian uses a perfective form of the verb *to be* combined with the l-participle (analogous to Serbo-Croatian future II), Serbo-Croatian uses a cliticised form of the verb *to want* combined with the infinitive. Unlike the infinitive, the l-participle carries the information on the gender and number. Since in this simplest form we have no way of inferring the gender of the subject in the direction Serbo-Croatian → Slovenian the translation defaults to masculine.

- *lahko* and *moći*:

  (5) Bolezni lahko povzročijo virusi ↔ Bolesti mogu prouzročiti virusi
  [Diseases.ACC] [easily.ADV] [cause.P3.SG] [viruses.NOM] → [Diseases.ACC] [can.P3.SG] [cause.INF] [viruses.NOM]
  (Viruses can cause diseases.)

  Unlike its Serbo-Croatian cognate *lako* the adverb *lahko* in Slovenian, when combined with a verb has an additional meaning of *can be done*, expressed in Serbo-Croatian with the modal verb *moći*. Rules that cover these type of phrases normalise the target verb to infinitive, and transfer grammatical markers for number and person to the verb *moći*.

- *lahko* and conditional:

  (6) Lahko bi napravili ↔ Mogli bi napraviti
  [easily.ADV] [would.CLT.CND] [do.LP.PL] → [Can.LP.PL] [would.CLT.CND.P3.SG] [do.INF]
  (We/they could do)

  Another morphological difference is found in the conditional mood. The conditional marker in Serbo-Croatian is the aorist form of the verb *to be*, and carries the information on person and number[13]. Slovenian, and the majority of colloquial Serbo-Croatian varieties, use a frozen clitic form of the same verb.[14]

[12]The Serbo-Croatian analyser covers the encliticised future tense forms (gledat ću / gledaću) as well.

[13]*bih*, *bismo*, *biste*, or *bi*

[14]*bi* regardless of person and number

Dictionary	Paradigms	Entries	Forms
Serbo-Croatian	1,033	13,206	233,878
Slovenian	1,909	13,383	147,580
Bilingual	69	16,434	70,787

Table 2: Statistics on number of lexicon entries for each of the dictionaries in the system.

Thus in cases like this example, when it is impossible to exactly infer the person and number the system defaults to the colloquial form.

- *lahko* and conditional more complicated:

  (7) Mi bi lahko napravili ↔ Mi bismo mogli napraviti
  [We.P1.PL] [would.CLT.CND] [easily.ADV] [do.LP.PL] → [We.P1.PL] [would.CLT.CND.P3.PL] [can.LP.PL] [do.INF]
  (We could do)

  In this example the information on person and number is available on the pronoun *mi*, and can be copied in translation to the conditional verb.

- *treba* adverb to verb

  (8) je treba narediti → treba učiniti
  [is] [needed.ADV] [to be done.INF] → [needs.VB.P3.SG] [to be done.INF]
  (It needs to be done)

  Phrases with Slovenian adverb *treba* translate to Serbo-Croatian with the verb *trebati*. In its simplest form the phrase just translates as 3rd person singular.

  For the opposite direction *trebati* translates as the analogous verb *potrebovati*, so that no loss of morphological information occurs.

  (9) trebaju našu solidarnost → potrebujejo našu solidarnost
  (They need our solidarity)

  More complicated examples with different tenses and verb phrases involve word reodering:

  (10) narediti je bilo treba ↔ trebalo je napraviti
  [do.INF] [is.CLT.P3.SG] [was.LP.NT] [need.ADV] → [needed.LP.NT] [is.CLT.P3.SG] [do.INF]
  (It needed to be done.)

4 Evaluation

This sections covers the evaluation of the developed system. The system was tested by measuring the lexical coverage, and by performing a qualitative and a quantitative evaluation.

Type	hbs→slv	slv→hbs
Disambiguation	194	28
Lexical selection	–	42
Transfer	47	98

Table 3: Statistics on the number of rules in each direction. For the lexical selection rules, the number indicates that there are 42 rules for each of the three standard varieties currently supported.

Language	SETimes	Europarl
Serbo-Croatian	85.41%	–
Slovenian	–	95.50%

Table 4: Naïve coverage

Lexical coverage tested using existing free corpora, while the quantitative evaluation was performed on 100 postedited sentences (with 1,055 words in total) from the Slovenian news portal Delo [15].

4.1 Lexical coverage

Coverage for the Serbo-Croatian–Slovenian language pair was measured using both the SETimes (Tyers and Alperen, 2010) and Europarl (Koehn, 2005) corpora. We measured coverage naively, meaning that we assume a word is in our dictionaries if at least one of its surface forms is found in the corpus. We are aware of the shortcomings of such an evaluation framework, however we decided to use it because of its simplicity.

The Serbo-Croatian → Slovenian side was evaluated using the SETimes corpus. As SETimes does not cover Slovenian the Slovenian → Serbo-Croatian side was evaluated only on the EuroParl corpus. The results are shown in table 4.

4.2 Quantitative

The quantitative evaluation was performed by 5 articles from the Slovenian news portal Delo. The articles were translated from Slovenian using Apertium, and were later corrected by a human post-editor in order to get a correct translation. The Word Error Rate (WER) was calculated by counting the number of insertions, substitutions and deletions between the post-edited articles and the original translation. We used the freely available `apertium-eval-translator` for calculating the WER and for bootstrap resampling Koehn (2004). We also reported the percentage of out of vocabulary words (OOV), and the total number of words per article. The results are given in table 5.

We also calculated both metrics for the output of Google Translate[16] and the results are presented in the same tables.

Given all the assumptions the WER and PER metrics make, the results show that our system is comparable to Google Translate with regards to translation quality. The Slovenian → Serbo-Croatian translation seems to be better than the Serbo-Croatian → Slovenian one which is due to the fact that more effort was put into developing the former direction.

4.3 Qualitative

The biggest problems are currently caused by the incompleteness of our dictionaries. The issues caused by OOV words are twofold. The less important issue is the fact that the system is unable to provide a translation for the unknown words — although in many cases, such as with proper names, these may result in *free rides*, that is word is the same in both languages. However, the more important issue is that OOV words cause problems with disambiguation and transfer, since they break long chains of words into smaller ones and drastically reduce context information.

Next, we have seen that the number of disambiguation rules for Slovenian is not sufficient for high-quality disambiguation. The constraint grammar for the Slovenian side was written based on the constraint grammar for the Serbo-Croatian side, and it needs further work.

We have also noticed difficulties in the transfer because of the loose grammar of both sides. Adding additional rules does not significantly improve the performance of the system and OOV words make long transfer rules irrelevant.

Finally, because of the short timeframe, we were not able to work much on lexical selection. Our lexical-selection module is the least developed part of our system. We have not done any work on the Slovenian side and the number of rules for the Serbo-Croatian side is small. This is due to the fact that no reliable parallel corpus exists for this language pair.

[15] `http://www.delo.si/`

[16] `http://translate.google.com/`

Article	Words	% OOV		WER	
		Apertium	Google	Apertium	Google
maraton	243	16.8	–	**[42.85, 47.92]**	[64.39, 74.56]
sonce	169	17.7	–	**[32.65, 45.33]**	[47.27, 58.62]
merkator	414	16.9	–	**[38.78, 48.14]**	[56.13, 70.30]
volitve	229	13.9	–	[37.81, 53.36]	[46.66, 62.67]
maraton	245	37.7	–	[52.78, 56.25]	[45.58, 63.87]
sonce	171	17.5	–	[47.50, 62.79]	[32.10, 58.49]
merkator	424	12.9	–	[45.78, 56.56]	[48.46, 64.15]
volitve	226	16.8	–	[47.00, 58.44]	[38.09, 58.10]

Table 5: Results for Word Error Rate (WER) in the Slovenian→Serbo-Croatian direction (top) and Serbo-Croatian→Slovenian (bottom). Scores in bold show a statistically significant improvement over the other system according to bootstrap resampling at $p = 0.95$.

5 Future work

The greatest difficulties for our system are caused by the long phrases present and the loose and free word order in the South Slavic languages. Because of that, in future we plan to put more effort into dealing with those problems. We are aware of the fact that it is difficult to write transfer rules between the two sides, and we intend to address that issue by first improving the coverage of our dictionaries.

After expanding the dictionaries, we intend to put more time into developing the Slovenian constraint grammar, and improve transfer by taking into account wider context.

We intend to work on more Slavic language pairs, including Serbo-Croatian–Russian, and improve our existing ones, including Serbo-Croatian–Macedonian (Peradin and Tyers, 2012) using the resources and knowledge obtained by developing this language pair.

Finally, we will keep the resources updated based on the latest politico-linguistic developments, and we will add the Montenegrin language once the standard is completely agreed on.

6 Conclusions

This language pair was an encouraging take on a pair of closely related South-Slavic languages, and represents a satisfying conclusion to an MT chain of neighbouring languages (the pairs Serbo-Croatian–Macedonian and Macedonian–Bulgarian are also available in Apertium). While we are aware that it is still in its infancy, and has many flaws, it is a valuable free/open-source resource, and will serve as another solid ground for NLP in this language group.

Acknowledgements

The development of this language pair was funded as a part of the Google Summer of Code.[17] Many thanks to the language pair co-author Aleš Horvat and his mentor Jernej Vičič, and other Apertium contributors for their invaluable help and support.

References

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

Erjavec, T. (2004). MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Fourth Int. Conference on Language Resources and Evaluation, LREC*, volume 4, pages 1535–1538.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th MT Summit*, pages 79–86.

Ortiz-Rojas, S., Forcada, M., and Sánchez, G. (2005). Construcción y minimización eficiente

[17] http://code.google.com/soc/

de transductores de letras a partir de diccionarios con paradigmas. *Procesamiento de Lenguaje Natural*, 35:51–57.

Peradin, H. and Tyers, F. M. (2012). A rule-based machine translation system from Serbo-Croatian to Macedonian. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FREERBMT12)*, pages 55–65.

Tadić, M. (2002). Building the croatian national corpus. In *LREC2002 Proceedings, Las Palmas, ELRA, Pariz-Las Palmas*, volume 2, pages 441–446.

Tyers, F. and Alperen, M. (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*.

Tyers, F. M., Sánchez-Martínez, F., and Forcada, M. L. (2012). Flexible finite-state lexical selection for rule-based machine translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 213–220, Trento, Italy.

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

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