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

Hrvoje Peradin, Filip Petkovski and Francis Tyers

May 27, 2014

The Balkan peninsula

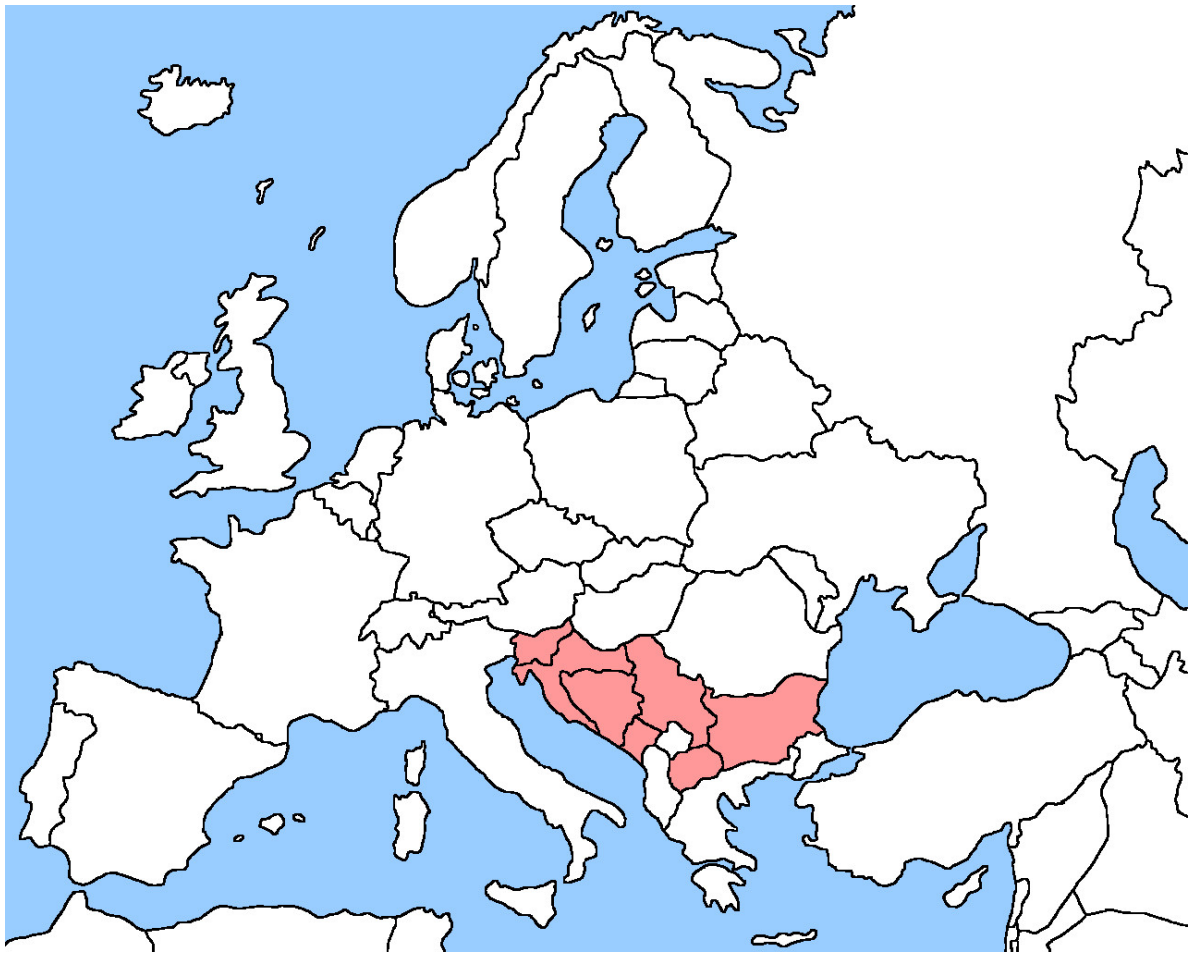

Figure : The Balkan peninsula

The Balkan peninsula

Figure : The Balkan peninsula

Introduction

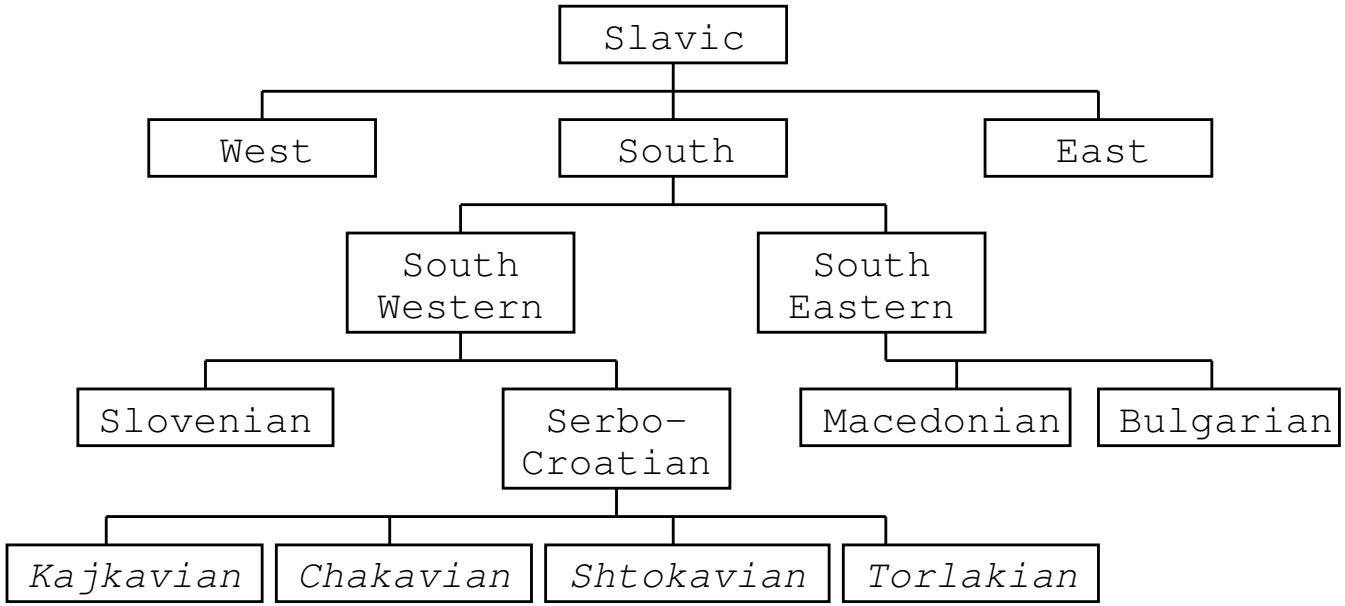

Figure : A traditional division of the South-Slavic languages. All four standard varieties of Serbo-Croatian (Bosnian, Croatian, Montenegrin, and Serbian) are based on the shtokavian dialect.

The Apertium platform

The Apertium platform is a modular machine translation system. The core setup consists of:

- A letter transducer morphological lexicon
- Morphological disambiguation module
- Lexical selection module
- Syntactic transfer module
- A letter transducer generator

Resources

Several resources were used extensively throughout the development process:

- On the Serbo-Croatian side we used the Croatian Language Portal (Hrvatski jezični portal)
- The Amebis Besana flective lexicon and other resources were used for the Slovenian side
- EUDict was used for the bilingual dictionary as the main resource

Morphological analysis and generation

- The basis for this language pair are the morphological lexicons for Serbo-Croatian and Slovenian developed during Google Summer of Code 2011
- The Serbo-Croatian dictionary was developed as part of a Serbo-Croatian–Macedonian pair
- The Slovenian dictionary was developed within a Slovenian–Spanish project
- Both lexicons had to be synchronized and trimmed down since they were developed with different tagsets and frequency lists

Disambiguation

- No freely available tools for disambiguation existed at the time when the project was being developed
- Satisfactory results were not obtained with the cannonical statistical tagger used in Apertium language pairs
- We relied solely on Constraint Grammar, which chooses the first output analysis in the case of remaining ambiguity
- Many of the rules developed earlier for Serbo–Croatian were reused for Slovenian

Lexical transfer

- Lexical transfer was done using an *lttoolbox* letter transducer and a bilingual dictionary.
- Additional paradigms were added for simple tagset mismatches which could be easily resolved in this stage
  - One example are cases when the adjective on one side is synthetic, and on the other analytic (*zdravije* vs *bolj zdravo*)

Lexical selection

- Due to a lack of a parallel corpus, Lexical selection was done based on hand-written rules
- For cases not covered by our hand-written rules, the default translation from the bilingual dictionary was chosen
- The lexical selection module was used mainly for:
  - Phonetics-based lexical selection: words that differe in a single phoneme, like *točno* and *tačno* (accurate)
  - Croatian months have Slavic-derived names and differ from the original Latin names (*siječanj* – January)

Syntactic transfer

- Most transfer rules were written to:
  - Bridge tagset differences
  - Cover different word orders