Tagging Slavic Corpora

Tomaž Erjavec
Department for Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia

Talk given at
SFB441
Tübingen,
December 15th, 1999,
Revised and made avilable on
http://nl.ijs.si/et/talks/SFB441/
Ljubljana
Dec 19th 1999

Overview of the talk:

- What the talk is really about
- Tagging: tools and resources
- Bosnian, Croatian and Serbian resources
- Tagging the Slovene(-English) ELAN corpus with MULTEXT-East MSDs
- Conclusions

Topic of the Talk

Tagging Slavic corpora =
annotating Serbo-Croatian words in context with morphosyntactic information using a trainable PoS tagger

Above, in the interest of terminological simplicity, Serbo-Croatian is taken mean the union of Bosnian, Croatian and Serbian.
Needed:

- various pre-processing tools
  (tokeniser, segmenter, ...)
- various resources
  (tagset, lexicon(s), hand annotated corpus)
- PoS tagger

Another way:

- use integrated, interactive environment
- outsource

Tagging SC is similar to tagging other languages, but here and for Slavic in general:

- less prefabricated resources
- more tags in standard 'PoS' tagset (lower accuracy)

Example

Text:

- že omenjeni zavod za živinorejo, ki bo vsaj v prehodnem obdobju nadaljeval naloge na področju zootehnike;
- the above-mentioned Livestock-Breeding Centre, which will continue the tasks in zootechnics, at least for the transitional period;

Result:

```
<seg id="kmet.sl.87" corresp="kmet.en.87">
<c>-</c>
<w msd="Q" lemma="že">že</w>
<w msd="Vmp--pmp" lemma="omeniti">omenjeni</w>
<w msd="Ncmsa--n" lemma="zavod">zavod</w>
<w msd="Spsa" lemma="za">za</w>
<w msd="Ncfsa" lemma="živinoreja">živinorejo</w>
<c>,</c>
<w msd="Css" lemma="ki">ki</w>
<w msd="Vcif3s" lemma="biti">bo</w>
<w msd="Q" lemma="vsaj">vsaj</w>
<w msd="Spsl" lemma="v">v</w>
<w msd="Afpnsl" lemma="prehoden">prehodnem</w>
<w msd="Ncnsl" lemma="obdobje">obdobju</w>
<w msd="Vmps-sma" lemma="nadaljevati">nadaljeval</w>
<w msd="Ncfpa" lemma="naloga">naloge</w>
<w msd="Spsl" lemma="na">na</w>
<w msd="Ncnsl" lemma="področje">področju</w>
<w msd="Ncfsg">zootehnike</w>
<c>;</c>
</seg>
```


Steps in corpus preparation

Tool chain:

1. acquisition of digital source
2. conversion to standard format:
   character sets, structure, emphasis, quotes, headers
3. segmentation:
   sentences, tokens, capitalisation, abbreviations, compounds, numerals
4. morphological analysis:
   lexicon, (rules), tagset
5. tagging (parsing)
6. iterative improvement:
   interactive environment
7. . . . exploitation (eg concordancing)

Resources needed

Tools can be off-the-shelf, but resources are dependent on language:

1. Tagset:
   pragmatic: made to fit the task and tools
   canonic: reuse of existing resources, comparable
2. Lexicon:
   core: closed class words, irregular words
   reference: general lexical stock
   dynamic: unknown words
3. Hand-tagged corpus:
   ML approaches
   size/quality

SC: research and resources

- Danko Šipka's MIG:
  morphological analyser/parser, lexicon
- Computing Research Laboratory, NMSU:
  Corelli Project, Rapid Deployment Morphology
- Oslo Corpus of Bosnian Texts
- Marko Tadić et al.
  mtadic@ffzg.hr:
  Croatian 'old' lemmatised corpus, lexicon?, Croatian National Corpus
- Duško Vitas, Cvetana Krstev, Belgrade
  vitas@matf.bg.ac.yu, cvetana@matf.bg.ac.yu:
  YU Intex core lexicon, tagged samples of corpus
- Saša Kostić, *akostic@f.bg.ac.yu*:
  Serbian Corpus: cca 11M words, morphosyntactically tagged, original version 30 years old, now being computer coded.
- Moskovljević, Jasmina, Andjelković, Darinka
  jasmina@ubbg.etf.bg.ac.yu:
  Belgrade Corpus of Child Language, in progress.

Some related languages:

- Slovenian: MULTEXT-East, ELAN, FIDA
- Czech: CnC, Dependency annotated treebank, work on tagging by Hladka & Hajič
- Bulgarian: Proceedings 'Recent Advances in Natural Language Processing', Spoken corpus 1 & 2
- TELRI Tractor, ELRA, ...

IJS-ELAN

- EU MLIS project ELAN
  European Language Activity Network
- Parallel Slovene-English corpus: (1M words)
- sentence aligned, tokenised
- encoded in accordance with TEI
- free distribution, sampler, headers & sampler in HTML, WWW concordancing (CQP)

15 component texts:

Title	kB	kW
Constitution of the Republic of Slovenia	364	20
Speeches by the President of Slovenia, M. Kučan	1102	69
Functioning of the National Assembly	325	20
Slovenian Economic Mirror; 13 issues, 98/99	4056	239
National Environmental Protection Programme	1222	70
Europe Agreement	589	34
Europe Agreement - Annex II	483	25
Strategy for Integration into EU	1511	89
Programme for accession to EU - agriculture	543	29
Programme for accession to EU - economy	394	23
Vademecum by Lek	471	24
EC Council Regulation No 3290/94 - agriculture	1182	69
Linux Installation and Getting Started	3044	173
GNU PO localisation	353	13
G. Orwell: Nineteen Eighty-Four	6698	195

Tagging Slovene ELAN

1. Amebis lemmatisation
2. MULTEXT-East Slovene tagset, lexicon & tagged corpus:
   - Multilingual Text Tools and Corpora for Central and Eastern European Languages
   - Copernicus 106; '95 – '97
   - MULTEXT, EAGLES
   - *http://nl.ijs.si/ME/*, TELRI-CD
   - (English), Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene
     (Lithuanian, Latvian, Russian, Serbian)
3. Tagging with TnT

Input for Tagger

> It is expected that by 2008 a significant number of traditional air pollution problems will be eliminated...

```
<seg lang="sl">
<w lemma="do">Do</w>
<w lemma="let letati leto">leta</w>
<w type=dig>2008</w>
<w lemma="se">se</w>
<w lemma="pričakovati">pričakuje</w><c>,</c>
<w lemma="da dati">da</w>
<w lemma="biti">bo</w>
<w lemma="odpravljen odpraviti">odpravljen</w>
<w lemma="bistven">bistveni</w>
<w lemma="del delo deti">del</w>
<w lemma="tradicionalen">tradicionalnih</w>
<w lemma="problem">problemov</w>
<w lemma="onesnaženost">onesnaženosti</w>
...
```

The MULTEXT-East MSDs

Morphosyntactic specifications, HR by Marko Tadić.

```
Noun (N)

11 Positions

**** **** **** **** **** ---- ---- ---- ---- ---- ----
PoS  Type Gend Numb Case Def  Cltc Anim OwnN OwnP OwdN
**** **** **** **** **** ---- ---- ---- ---- ---- ----

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

```
% mtems-expand -brief Ncfsg
Ncfsg: Noun common feminine singular genitive
```


The Slovene Lexicon

```
avenij          avenija         Ncfdg
avenij          avenija         Ncfpg
avenija         =               Ncfsn
avenijah        avenija         Ncfdl
avenijah        avenija         Ncfpl
avenijam        avenija         Ncfpd
avenijama       avenija         Ncfdd
avenijama       avenija         Ncfdi
avenijami       avenija         Ncfpi
avenije         avenija         Ncfpa
avenije         avenija         Ncfpn
avenije         avenija         Ncfsg
aveniji         avenija         Ncfda
aveniji         avenija         Ncfdn
aveniji         avenija         Ncfsd
aveniji         avenija         Ncfsl
avenijo         avenija         Ncfsa
avenijo         avenija         Ncfsi
```

Cat	entries	wforms	lemmas	MSDs
Noun	127.811	61.525	7.465	99
Verb	110.949	78.001	3.699	128
Adjective	310.754	64.604	4.621	279
Pronoun	3.654	732	105	1.335
Adverb	7.415	7.395	442	3
Preposition	123	109	79	6
Conjunction	39	38	39	3
Numeral	4.401	832	181	226
Exclamation	10	10	10	1
Abbreviation	48	48	48	1
Particle	76	76	76	1
Σ	565.281	201.011	16.766	2.083

Intex (DELAS/DELAF) YU lexicon

```
gdekakav,ProA03.01       gdekakav,.ProA03.01:msn*:msa-
gdeko,ProN12             gdekakva,gdekakav.ProA03.01:msg*:nsg*:fsn*:np
gdekoji,ProA07           gdekakve,gdekakav.ProA03.01:fsg*:mpa*:fpn*:fp
gdetko,ProN12            gdekakvi,gdekakav.ProA03.01:mpn*
gdešto,AdvE,ProN13       gdekakvih,gdekakav.ProA03.01:mpg*:npg*:fpg*
gdjekakav,ProA03.01      gdekakvim,gdekakav.ProA03.01:msi*:nsi*:*pd*:*
...                      gdekakvima,gdekakav.ProA03.01:*pd*:*pi*:*pl*
ičiji,ProA05             gdekakvime,gdekakav.ProA03.01:msi*:nsi*
ja,ProN01                gdekakvo,gdekakav.ProA03.01:nsn*:nsa*:
kakav,ProA03.01          gdekakvog,gdekakav.ProA03.01:msg*:nsg*:msa+
kakavgod,ProA03.01       gdekakvoga,gdekakav.ProA03.01:msg*:nsg*:msa+
...                      gdekakvoj,gdekakav.ProA03.01:fsd*:fsl*
tvoj,ProA06              gdekakvom,gdekakav.ProA03.01:msd*:msl*:nsd*:n
vaš,ProA04               gdekakvome,gdekakav.ProA03.01:msd*:msl*:nsd*:
vi,ProN04                gdekakvomu,gdekakav.ProA03.01:msd*:msl*:nsd*:
šta,Adv*,Par*,ProN13     gdekakvu,gdekakav.ProA03.01:msd*:msl*:nsd*:ns
štagod,ProN13
```

PoS	DELAS		DELAF	
	KByte	entries	KByte	entries
nouns	50K	2720	627K	18241
adjectives	11K	630	565K	10956
verbs	42K	1884	1693K	49076
other	21K	1378	133K	3849
total	124K	6569	3009K	81152

The Slovene pre-tagged corpus

- Orwell's '1984'
- auto-tagged & checked by ZRC SAZU (Primož Jakopin, Aleksandra Bizjak)
- 90.792 words, 1.025 different MSDs (from 2.083 lexical)
- aligned with English, encoded in CES

The tagger

Qualities:

- availability
- accuracy
- robustness (large tagset!)
- speed

Evaluation on '1984':

- HMM, RBT, MET, MBT, TnT:
-

	TnT	MBT
Known	93.55%	93.58%
Unknown	60.77%	44.45%

TnT, 'Trigrams 'n Tags':

- research license
- Solaris, Linux
- fast, robust, well-designed
- unknown word guessing

TnT parameters

```
%% Statistically tagged file, Sun Dec  5 16:32:44 1999
%% lexicon     : mte.lex
%% ngrams      : mte.123
%% corpus      : elan-sl.t
%% model       : trigrams
%% sparse data : linear interpolation
%%   lambda1 = 1.292668e-01   lambda2 = 3.310223e-01   lambda3 = 5.3
%% unknown mode: lexicon entry @UNKNOWN
%% case of characters is significant
%% using suffix trie up to length 10
%% unknown words are marked with an asterisk (*)
%% Thorsten Brants, thorsten@coli.uni-sb.de

%% 177776 (30.04%) unknown tokens
%%      7141 recognized as cardinals/ordinals
%% 102761 tokens taken from the backup lexicon
%% avg. 10.87 tags/token, 1.71 tags/known token

izhajajoč               Afpmsnn *
iz                      Spsg
Temeljne                Afpfsg  *
ustavne                 Afpfsg  *
listine                 Ncfsg   *
o                       Spsl
samostojnosti           Ncfsl   *
```

Conclusions

- off the shelf tools are becoming available
- but automatic tagging requires substantial resources
- for good tagging, manual correction is necessary
- Question: what does a project require?