Adaptation of Czech syntactic analyzers for Slovak

Marek Medveď, Miloš Jakubíček, Vojtěch Kovář, Václav Němčík

Natural Language Processing Centre Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic

7.12.2012

Czech analyzers

- SYNT and SET syntactic analyzers have been developed in the NLP centre at Faculty of Informatics, Masaryk University.

Czech analyzers

- SYNT and SET syntactic analyzers have been developed in the NLP centre at Faculty of Informatics, Masaryk University.
- input for both systems is sentence might be vertical format, plaintext ...

Czech analyzers

- SYNT and SET syntactic analyzers have been developed in the NLP centre at Faculty of Informatics, Masaryk University.
- input for both systems is sentence might be vertical format, plaintext ...
- output for SYNT are a phrase-structure tree, a dependency graph and set of syntactic structures.

Czech analyzers

- SYNT and SET syntactic analyzers have been developed in the NLP centre at Faculty of Informatics, Masaryk University.
- input for both systems is sentence might be vertical format, plaintext ...
- output for SYNT are a phrase-structure tree, a dependency graph and set of syntactic structures.
- output for SET are a hybrid tree consisting of both dependency and constituent edges, a pure dependency tree, a pure constituent tree.

Slovak corpora

In our experiments, we used this three Slovak corpora:

Corpus	Number of tokens
r-mak 3.0	1.2M
skTenTen	876M
SDT	12 000

Table: Slovak corpora

Work structure

The project consists of three parts:

Work structure

The project consists of three parts:

- morphological tagging conversion

Work structure

The project consists of three parts:

- morphological tagging conversion
- lexical analysis adjustment in both parsers

Work structure

The project consists of three parts:

- morphological tagging conversion
- lexical analysis adjustment in both parsers
- grammar adaptation for both parsers

Morphological tag translation

Three ways of tag translation:

Morphological tag translation

Three ways of tag translation:

- rule tag translation

spevák (singer): SSfs1 ⟶ k1gFnSc1

Morphological tag translation

Three ways of tag translation:

- rule tag translation

  spevák (singer): SSfs1 ⟶ k1gFnSc1

- whitelist translation

  mnoho (lot of): NUns4 ⟶ k4xCgNnSc4
  ⟶ @k6eAdItQ

Morphological tag translation

Three ways of tag translation:

- rule tag translation

  spevák (singer): SSfs1 ⟶ k1gFnSc1

- whitelist translation

  mnoho (lot of): NUns4 ⟶ k4xCgNnSc4
  ⟶ @k6eAdItQ

- whitelist completion

  ktorý (who): PAms1 ⟶ k3gMnSc1
  ⟶ k3gMnSc1yR

Adaptation

- adaptation of lexical analysis

Adaptation

- adaptation of lexical analysis
- adaptation of context free grammar rules of SYNT

Adaptation

- adaptation of lexical analysis
- adaptation of context free grammar rules of SYNT
- adaptation of segmentation rules of SET

(„not + to be“) of SYNT

```
clause %> IS sth
clause %> ARE sth
clause %> VB12 sth
```

(„not + to be“) of SYNT

```
clause %> IS sth
clause %> ARE sth
clause %> VB12 sth
```

```
clause %> is sth
clause %> are sth
clause %> vb12 sth
is -> IS
is -> NOT IS
are -> ARE
are -> NOT ARE
vb12 -> VB12
vb12 -> NOT VB12
```

```
if (!strcmp(l_word[wi], "nie")) {
   lemma->preterm =
               __SYNT_NTERM_NOT;
 }
```

(„not + to be“) of SET

```
%TMPL: $PARTICIP $...* $NOT $BYBYT MARK 2 DEP 0    PROB 1000
%$BYBYT(word): som si sme ste
```

Evaluation

- We developed the program `sk2cs.py` which is a fast and easy to modify tool for tag translation.

Evaluation

- We developed the program `sk2cs.py` which is a fast and easy to modify tool for tag translation.
- Results for system `SYNT`:

Corpus	Number of sentences	Number of accepted
r-mak 3.0	74,127	77 %
SDT	12,762	76.9 %

Table: Evaluation of the coverage of `SYNT`

Number of sentences	77
Median number of trees	148
Average number of trees	71595.81
Average LAA of the first tree	87.13
Time per sentence	0.038 s

Table: Evaluation of the precision of `SYNT`

Evaluation

- Results for system SET:

Corpus	Number of sentences	Dependency precision
SDT	12,762	56.7 %

Table: Evaluation of the precision of SET

IAA of annotators

	Adv	Apos	Atr	AuxC	AuxO	AuxP	AuxR	AuxT	AuxV	AuxX	Coord	ExD	Obj	Pred	Oth
Adv	13581	2	1203	284	14	88	44	13	6	9	24	501	1827	763	1741
Apos	0	219	2	0	0	0	0	0	0	87	137	13	0	5	121
Atr	0	0	17226	39	32	12	4	6	8	5	4	281	1401	763	1717
AuxC	0	0	0	4064	6	51	0	1	2	18	123	25	192	2	892
AuxO	0	0	0	0	48	45	48	28	1	5	6	2	59	2	78
AuxP	0	0	0	0	0	10358	101	85	4	3	7	22	22	9	325
AuxR	0	0	0	0	0	0	700	2281	25	0	0	1	351	0	12
AuxT	0	0	0	0	0	0	0	949	100	0	1	3	222	2	28
AuxV	0	0	0	0	0	0	0	0	1056	1	0	4	36	81	95
AuxX	0	0	0	0	0	0	0	0	0	9383	953	261	3	4	279
Coord	0	0	0	0	0	0	0	0	0	0	4724	112	1	47	1246
ExD	0	0	0	0	0	0	0	0	0	0	0	2106	381	141	1452
Obj	0	0	0	0	0	0	0	0	0	0	0	0	12416	755	1714
Pred	0	0	0	0	0	0	0	0	0	0	0	0	0	11149	400
Oth	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28948

Table: annotators consistency

IAA of annotators

	Adv	Apos	Atr	AuxC	AuxO	AuxP	AuxR	AuxT	AuxV	AuxX	Coord	ExD	Obj	Pred	Oth	
Adv	13581	2	1203	284	14	88	44	13	6	9	24	501	1827	763	1741	0.32
Apos	2	219	2	0	0	0	0	0	0	87	137	13	0	5	121	0.63
Atr	1203	2	17226	39	32	12	4	6	8	5	4	281	1401	763	1717	0.24
AuxC	284	0	39	4064	6	51	0	1	2	18	123	25	192	2	892	0.29
AuxO	14	0	32	6	48	45	48	28	1	5	6	2	59	2	78	0.87
AuxP	88	0	12	51	45	10358	101	85	4	3	7	22	22	9	325	0.07
AuxR	44	0	4	0	48	101	700	2281	25	0	0	1	351	0	12	0.80
AuxT	13	0	6	1	28	85	2281	949	100	0	1	3	222	2	28	0.74
AuxV	6	0	8	2	1	4	25	100	1056	1	0	4	36	81	95	0.26
AuxX	9	87	5	18	5	3	0	0	1	9383	953	261	3	4	279	0.15
Coord	24	137	4	123	6	7	0	1	0	953	4724	112	1	47	1246	0.36
ExD	501	13	281	25	2	22	1	3	4	261	112	2106	381	141	1452	0.60
Obj	1827	0	1401	192	59	22	351	222	36	3	1	381	12416	755	1714	0.36
Pred	763	5	763	2	2	9	0	2	81	4	47	141	755	11149	400	0.21
Oth	1741	121	1717	892	78	325	12	28	95	279	1246	1452	1714	400	28948	0.26
																0.41

Table: annotators consistency

Conclusion

- We created first complete syntactic analyzers for Slovak language.

Conclusion

- We created first complete syntactic analyzers for Slovak language.
- We created easily modifiable program for tag set translation.

Conclusion

- We created first complete syntactic analyzers for Slovak language.
- We created easily modifiable program for tag set translation.
- The accuracy of adapted analyzers is about 77%.

Thank you for your attention.