FACULTY
OF ARTS
Masaryk University

Using Syntax Analyser SET as a Grammar Checker for Czech

Marie Novotná, Markéta Masopustová
{428801,415295}@mail.muni.cz

December 7, 2018

FACULTY
OF ARTS
Masaryk University

Table of Contents

FACULTY
OF ARTS
Masaryk University

What is done?

- Grammaticon (Lingea)
  - support ended in 2014
- Kontrola české gramatiky (V. Petkevič)
  - limited functionality since its launch

FACULTY
OF ARTS
Masaryk University

The SET parser

- syntactic analyser @ FI MUNI
- accessible, flexible
- easy to work with

FACULTY
OF ARTS
Masaryk University

Attributive adjective-noun agreement

- not solved yet -> new category *modifier-bad*
- parts:
  - colloquial ending *-ej*, e.g. *Dobrej den.*
  - colloquial ending *-ý*, e.g. *hezký holky*
  - dual ending *-ma*, e.g. *s těma velkýma stolama*

FACULTY
OF ARTS
Masaryk University

Attributive adjective-noun agreement

Results

- test set: 230 sentences

TP	FN	FP	recall	precision
92	44	1	0.68	0.99

FACULTY
OF ARTS
Masaryk University

Multiple subject-predicate agreements

- limited to two subjects
- problems:
  - the SET does not work with coordination -> usage of extensive set of rules
  - many problems with past tense

TP	FN	FP	recall	precision
24	15	18	0.62	0.57

FACULTY
OF ARTS
Masaryk University

Colloquial expressions in written texts

Problem set I

- word order
  - enclitics (e.g. **Si pořídím nové kolo.*)
  - prepositions (e.g. *Dospěl k pro něj těžké otázce.*)
- excess of demonstrative nouns
- repeating the same expressions within one sentence (e.g. *Však to byla ona, však ji všichni poznali.*)
- pleonasms (e.g. *Představ si, že k nákupu dávali dárek zadarmo!*)
- absurd superlatives (e.g. *Až se překladatelé naučí nejzákladnější základy základní terminologie...*)

FACULTY
OF ARTS
Masaryk University

Colloquial expressions in written texts

Problem set II

- wrong use of the word *jakýkoli* (e.g. *Demisi mohl podat *jakýkoli z členů vlády.*)
- forgetting double conjunctions (like **jednak–druhak*, *buď–(a)nebo* and more)
- colloquial expressions in written texts (e.g. *Můžu s tebou jít, ale ten film musí být fakt dobrej.*)
- wrong usage of pronoun *který/jenž*
- other mistakes (like **vyjímka*, **pernamentka*, **dvěmi*, **pane Straka*, **reprezentanté*, **datumu*, misuse of conjunction *mimo*)

FACULTY OF ARTS
Masaryk University

Colloquial expressions in written texts

Results

rule	TP	FN	FP	precision	recall
enclitics correct sentences	136	0	0	1	1
enclitics bad senteces	41	0	309	0.117	1
prepositions	35	3	5	0.875	0.921
demonstrative nouns	59	1	1	0.983	0.983
pleonasms	131	17	0	1	0.885
double conjunctions	54	7	19	0.740	0.885
gender *který* correct sentences	151	0	1	0.993	1
gender *který* bad sentences	37	3	133	0.218	0.925
colloquial expressions	16	1	8	0.667	0.941

Other rules have 100 % success rate.

FACULTY
OF ARTS
Masaryk University

Conclusion

- the SET is suitable for usage as a grammar checker
- current tools have their limitations
- this work will be used as a starting point for a new grammar checker project

FACULTY
OF ARTS
Masaryk University

Thank you for your attention!