FACULTY
OF ARTS
Masaryk University

Comparing majka and MorphoDiTa for Automatic Grammar Checking

Jakub Machura, Helena Geržová,
Markéta Masopustová and Marie Valíčková
{415795,400133,428801,415295}@mail.muni.cz

December 6, 2019

Table of Contents

FACULTY
OF ARTS
Masaryk University

Backround

- Grammaticon (Lingea)
  - support ended in 2014
- Kontrola české gramatiky (V. Petkevič)
  - limited functionality since its launch
- “PLINkor”
  - a new project of the automatic online language checker is in development, using the SET parser

FACULTY
OF ARTS
Masaryk University

Motivation

- majka+DESAMB make mistakes
- MorphoDiTa is maintained and has good results

FACULTY
OF ARTS
Masaryk University

Obstacles I. Punctuation

S1: Muž, který je_1 pravděpodobně unesl, je_2 běloch.

<s>		
Muž	muž	k1gMnSc1
<g>		
,	,	kIx,
který	který	k3yRgMnSc1
je_1	on	k3xPp3gNnSc4
pravděpodobně	pravděpodobně	k6eAd1
unesl	unést	k5eAaPmAgMnS
<g/>		
,	,	kIx,
je_2	být	k5eAaImIp3nS
běloch	běloch	k1gMnSc1
<g/>		
.	.	kIx.
</s>		

FACULTY
OF ARTS
Masaryk University

Obstacles I. Punctuation

S2: Muž, který je_1 pravděpodobně unesl je_2 běloch.

<s>		
Muž	muž	k1gMnSc1
<g/>		
,	,	kIx,
který	který	k3yRgMnSc1
je_1	on	k3xPp3gNnSc4
pravděpodobně	pravděpodobně	k6eAd1
unesl	unést	k5eAaPmAgMnS
<g/>		
je_2	on	k3xPp3gNnSc4
běloch	běloch	k1gMnSc1
<g/>		
.		kIx.
</s>		

FACULTY
OF ARTS
Masaryk University

Obstacles II. Homonymy

S3: V druhém šuplíku najdeš správné **koření**.

<s>		
V	v	k7c6
druhém	druhý	k4xOgInSc6
šuplíku	šuplík	k1gInSc6wH
najdeš	najít	k5eAaPmIp2nS
správné	správný	k2eAgNnSc1d1
koření	koření	k1gNnSc1
<g/>		
.	.	kIx.
</s>		

FACULTY
OF ARTS
Masaryk University

Obstacles II. Homonymy

S4: Běž a rychle kup **koření** v supermarketu.

<s>		
Běž	běžet	k5eAaImRp2nS
a	a	k8xC
rychle	rychle	k6eAd1
kup	kup	k1gInSc1
koření	kořenit	k5eAaImIp3nS
v	v	k7c6
supermarketu	supermarket	k1gInSc6
<g/>		
.	.	kIx.
</s>		

FACULTY
OF ARTS
Masaryk University

MorphoDiTa

- a complex tool (dictionary and tagger), maintained
- available at CLARIN/LINDAT
- however, uses positional tag set:
  - the SET parser (and we) works with attributive tag set
  - the posttags switch

FACULTY
OF ARTS
Masaryk University

The posttags switch example

S1: Muž, který je pravděpodobně unesl, je běloch.

			--posttags
Muž	muž	NNMS1-----A----	k1gMnSc1eA;cap
,	,	Z:----------	kI
který	který	P4YS1----------	k3yRgMgInSc1
je	on	PPXP4--3-------	k3xPg.nPc4p3
pravděpodobně	pravděpodobně	Dg-------1A----	k6d1eA
unesl	unést	VpYS---XR-AA---	k5mAgMgInSp.mReA
,	,	Z:----------	kI
je	být	VB-S---3P-AA---	k5mInSp3mIeA
běloch	běloch	NNMS1-----A----	k1gMnSc1eA
.	.	Z:----------	kI

FACULTY
OF ARTS
Masaryk University

Punctuation I

- testing on DESAM corpus (contains 61 098 commas)
- MorhopoDiTa did not bring better results, however it seems it is better in case of homonymy

FACULTY OF ARTS
Masaryk University

Punctuation II

Total of commas: 61 098	majka + DESAMB					MorhoDiTa				
Rules	**TP**	**FP**	**FN**	**P (%)**	**R (%)**	**TP**	**FP**	**FN**	**P (%)**	**R (%)**
All rules	33 833	2 457	27 265	**93,23**	**55,37**	33 808	2 741	27 290	92,50	55,33
1. Connector	32 806	2 256	28 292	**93,57**	53,69	32 805	2 609	28 293	92,63	53,69
2. Coordination	1 025	224	60 073	82,07	**1,68**	1 005	145	60 093	**87,39**	1,64
3. Coordination	1 034	94	60 064	91,67	**1,69**	804	56	60 294	**93,49**	1,32

FACULTY
OF ARTS
Masaryk University

Zeugma I

- what is zeugma:
  - one expression is in semantic or syntactic relation with two other paratactically connected expressions (e.g. two verbs), but the whole structure is grammatically defective
- for testing purpose only 20 verb & their rules for detection of zeugma
- two datasets
  - “test_set_with_errors_2” – set with errors only
  - “test_set_mixed_1” – coordinating structures consisting of a tested verb and another verb
- both tested analyzers had more or less similar results

FACULTY OF ARTS
Masaryk University

Zeugma II

	test_set_mixed_1			test_set_with_errors_2		
	TP	FP	Precision (%)	TP	FN + TP	Recall (%)
majka + DESAMB	314	57	84,64	227	483	47,00
MorphoDiTa	359	50	87,78	225	483	46,58

FACULTY
OF ARTS
Masaryk University

Subject-predicate agreement I

- for testing purpose only subject-predicate agreement with a simple subject
- data set: 124 sentences; 34 correct, 90 with errors
- first observation
  - majka+DESAMB behave cautiously (rather to avoid of making false report)
  - MorphoDiTa reports more mistakes, but often they are false positives

Subject-predicate agreement II

	TP	FP	FN	Precision (%)	Recall (%)
majka	29	15	65	65,9	30,9
MorphoDiTa	40	48	54	45,5	42,6

FACULTY OF ARTS
Masaryk University

Subject-predicate agreement III

- detailed inspection showed mistake in the posttags switch
  - attributive tags specify plural feminine and masculine inanimate gender
  - positional tags have ambiguous position, which means plural feminine and also masculine inanimate gender
  - these types of words get tag `k5mAnPgIgFnPp.mReA`, but the SET cannot process the whole tag and works only with masculine inanimate (`gI`)
- second observation
  - after repair, number of false positives dropped greatly

Subject-predicate agreement IV

	TP	FP	FN	Precision (%)	Recall (%)
majka	29	15	65	65,9	30,9
MorphoDiTa	40	48	54	45,5	42,6
MorphoDiTa (after repair)	40	12	54	76,9	42,6

FACULTY
OF ARTS
Masaryk University

“Other” rules

- no difference between majka+DESAMB and MorphoDiTa
- the posttag switch does not convert tag for colloquial expressions

FACULTY
OF ARTS
Masaryk University

Conclusion

- testing did not prove that MorhpoDiTa system would arrange a big difference (except for the homonymy)
- the posttags switch needs to be tuned up
- disambiguation needs to be tuned up (by improving DESAMB or by developing/finding better tagger)
- majka dictionary should be updated

Thank you for your attention!

Question about...
... punctuation? → ask Jakub: 415795@mail.muni.cz
... zeugma? → ask Helena 400133@mail.muni.cz
... agreement? → ask Marie 428801@mail.muni.cz
... colloquial expressions? → ask Markéta 415295@mail.muni.cz