



FACULTY  
OF ARTS

Masaryk University

# Towards the New Czech Grammar-checker

RASLAN 2018

Vojtěch Mrkývka

mrkyvka@phil.muni.cz

December 7, 2018



# Goal

- New grammar-checker of Czech
- Web-based application
- Using new and existing tools developed at MU



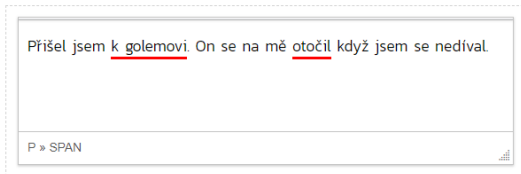
# Motivation

- There are tools existing / in development at MU
- Current best Czech GC is part of proprietary system
- Create an alternative to applications like Grammarly but for Czech

# The current interface



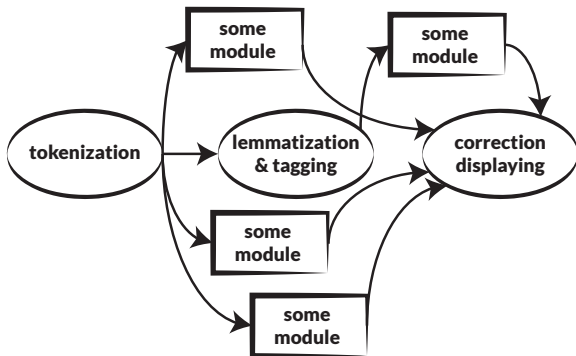
## Rozhraní korektoru



## The current interface

- Based on on-line text processor tinyMCE
- Mostly in JavaScript as tinyMCE modules
- Asynchronous process
- Communication with backend via AJAX

## Processing diagram



## Correction displaying

0	1	2	3	4	5	6	7
The		dog		is		runing.	

**Tokens to display mistake at: 6**

**Correction: 6/runing/running**



## Implemented modules

<b>Correction</b>	<b>TP</b>	<b>FP</b>	<b>TN</b>	<b>FN</b>	<b>pre</b>	<b>rec</b>
<b>Misspellings (excl. proper nouns)</b>	24	0	487	16	1,000	0,600
<b>Misspellings (incl. proper nouns)</b>	7	17	497	6	0,292	0,538
<b>Vocalisation of prepositions</b>	4	0	8	0	1,000	1,000
<b>Multiple whitespaces</b>	4	0	515	0	1,000	1,000
<b>Whitespace in the interpunction proximity</b>	7	0	119	0	1,000	1,000
<b>Conditionals</b>	2	0	1	0	1,000	1,000
<b>Commas in a sentence</b>	3	0	0	4	1,000	0,429





## A problem with testing

- Testing data were too small
- Mistakes were artificial
- ⇒ Need for API & collection of correctly annotated genuine texts

## Implemented modules

<b>Correction</b>	<b>TP</b>	<b>FP</b>	<b>TN</b>	<b>FN</b>	<b>pre</b>	<b>rec</b>
<b>Misspellings (excl. proper nouns)</b>	24	0	487	16	1,000	0,600
<b>Misspellings (incl. proper nouns)</b>	7	17	497	6	0,292	0,538
<b>Vocalisation of prepositions</b>	4	0	8	0	1,000	1,000
<b>Multiple whitespaces</b>	4	0	515	0	1,000	1,000
<b>Whitespace in the interpunction proximity</b>	7	0	119	0	1,000	1,000
<b>Conditionals</b>	2	0	1	0	1,000	1,000
<b>Commas in a sentence</b>	3	0	0	4	1,000	0,429

## A problem with spell-checking

- Precision is low
- Not often updated dictionary
- ⇒ Method of adding new words, using different lexicon...

## A problem with error reporting

- Allow users to flag miscorrections
- How to not display miscorrection afterwards?
- ⇒ Probably module-dependent



Thank you for your attention!

This work was supported by the project of specific research *Čeština v jednotě synchronie a diachronie* (Czech language in unity of synchrony and diachrony; project no. MUNI/A/0862/2017).