FACULTY
OF ARTS
Masaryk University

Recent Advancements on the New Online Proofreader of Czech

RASLAN 2019

Vojtěch Mrkývka
mrkyvka@phil.muni.cz

December 6, 2019

FACULTY
OF ARTS
Masaryk University

Let's state the facts

1. The best proofreader of Czech is part of proprietary system
2. There is no alternative (free of proprietary) providing similarly good results

FACULTY
OF ARTS
Masaryk University

The project

In 2018 a new project was submitted to and accepted by the Technology Agency of the Czech Republic (TA ČR).

- Goal: New web-based proofreading software
- Started in February 2019, ending in 2022

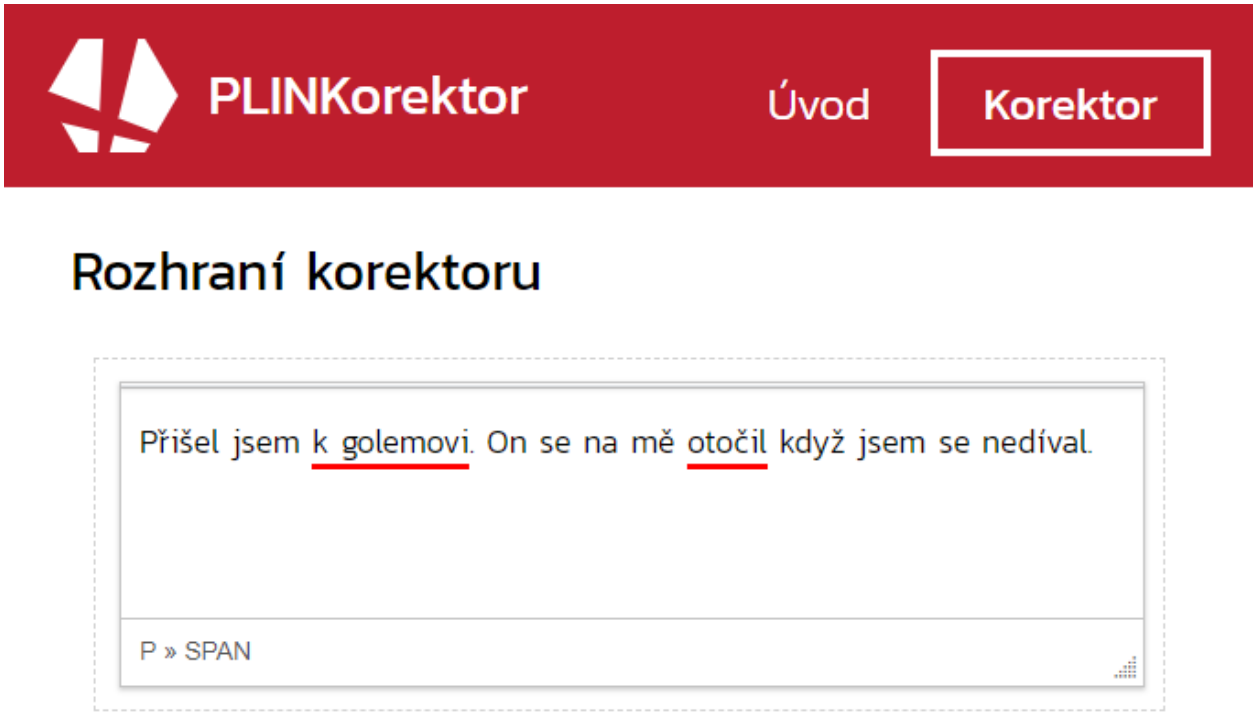
FACULTY
OF ARTS
Masaryk University

Used tools

- Tools developed in the NLP Center, FI MU
- Rulesets made by students
  (bachellor/master theses, doctoral studies)
- Additional external datasets or tools

FACULTY
OF ARTS
Masaryk University

The current interface



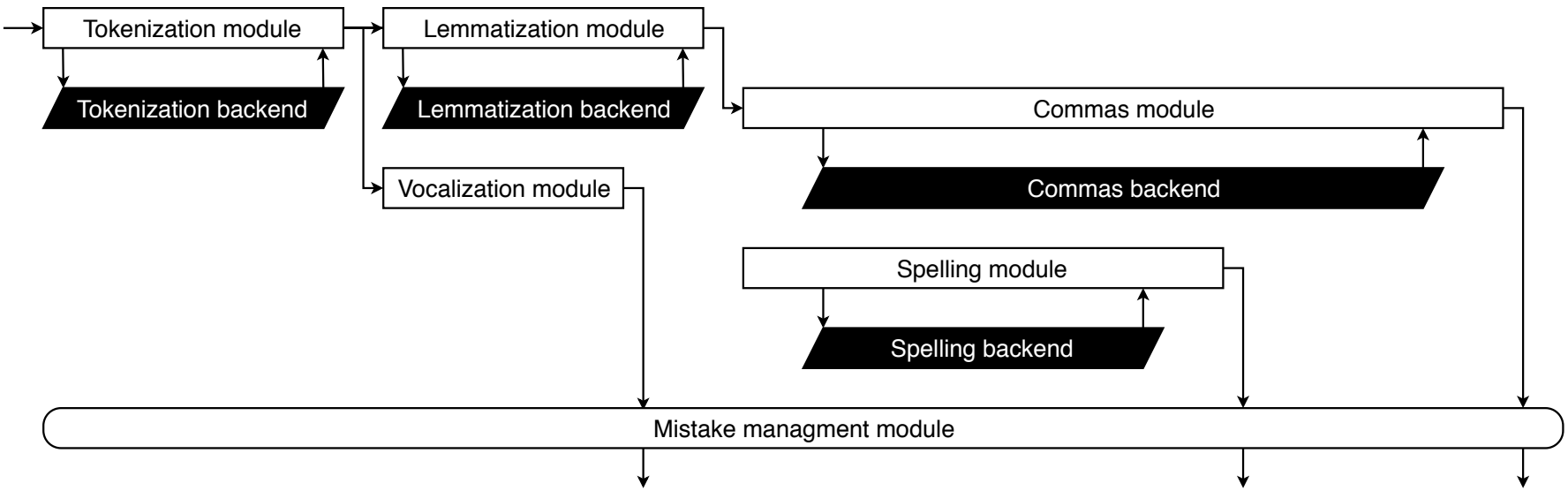
FACULTY
OF ARTS
Masaryk University

The first version

- Based on on-line text processor TinyMCE
- Series of separate modules
- Javascript & AJAX connection to backend (usually in Python) if necessary

FACULTY
OF ARTS
Masaryk University

The first version



FACULTY
OF ARTS
Masaryk University

The pros & cons of the first version

Pros

- Easy developement of separate modules
- Little to no bottlenecking due to asynchronous processing

Cons

- Too tightly bound to the single piece of software (TinyMCE)
- No control over potential extensions
- User-unfriendly installation process

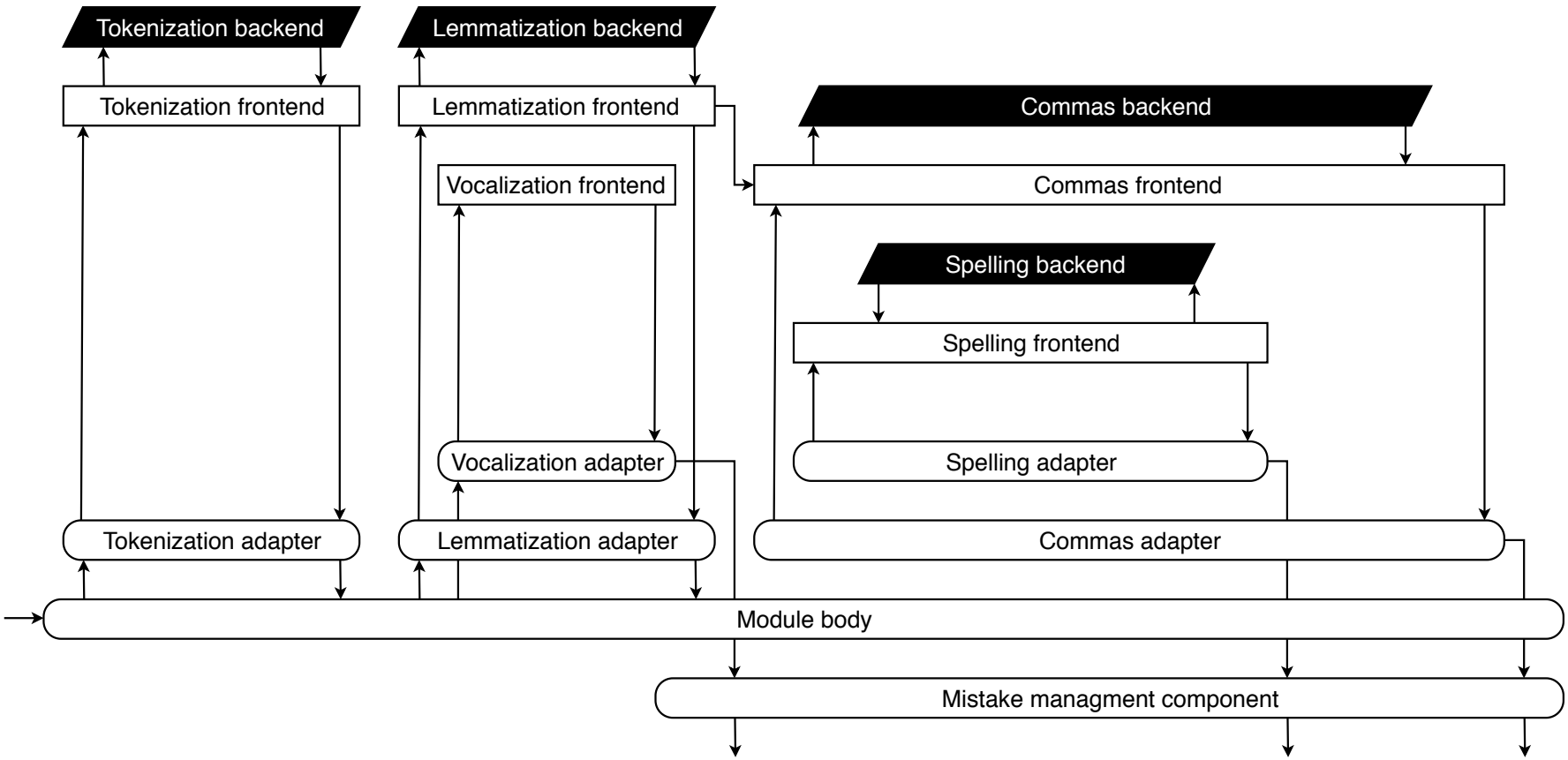
FACULTY
OF ARTS
Masaryk University

The second version

- Proofreader code independent on TinyMCE
- TinyMCE edition contained single module
- Javascript & AJAX connection to backend (usually in Python) if necessary
- Unfinished due to change of the assignment

FACULTY OF ARTS
Masaryk University

The second version



FACULTY OF ARTS
Masaryk University

The pros & cons of the second version

Pros

- Keeps most of the advantages of the first version
- Strict separation between proofreader and text processor code $\Rightarrow$ portability
- Improvement in user-friendliness as the proofreader (TinyMCE version) could have been distributed as a single package

Cons

- Not very suitable for non-web based applications

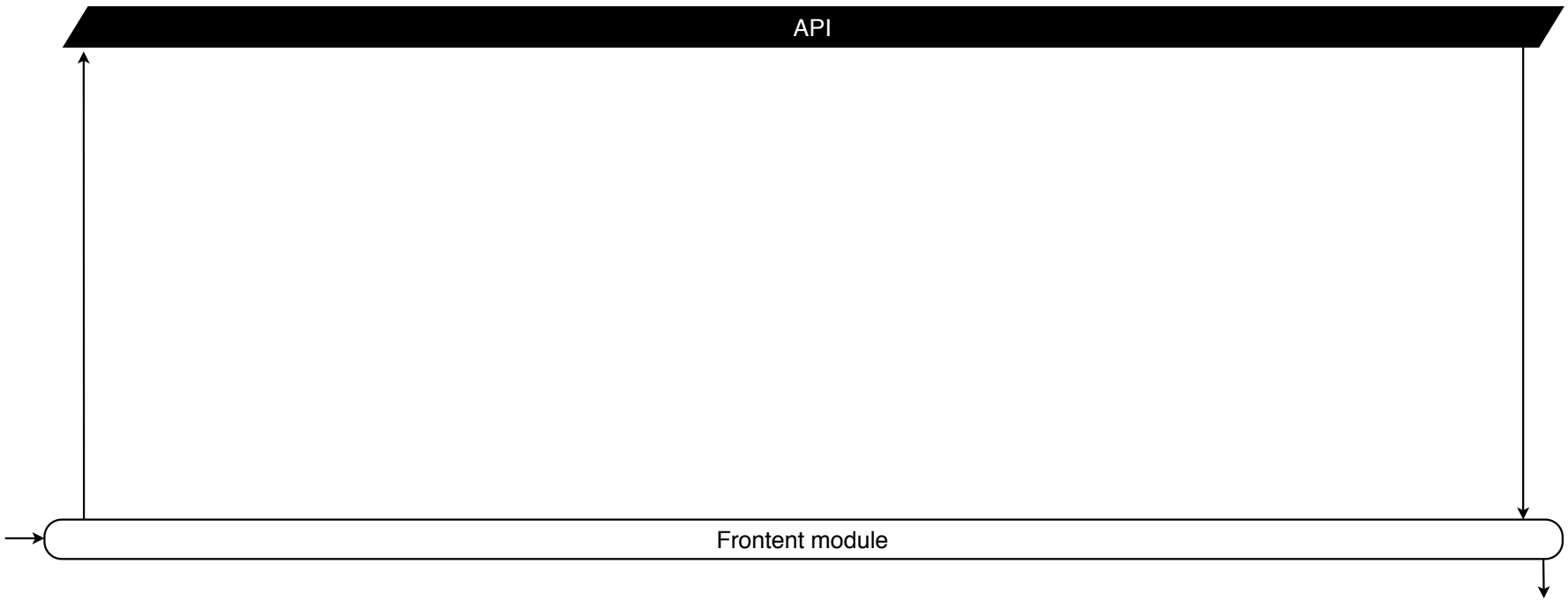
FACULTY
OF ARTS
Masaryk University

The third and current version

- Brand new $\Rightarrow$ first preview available from November
- We were addressed by potential customer
  $\Rightarrow$ slight alteration of the assignment
- More suitable for non-web based applications
- Single API written in Python

FACULTY
OF ARTS
Masaryk University

The third and current version



FACULTY OF ARTS
Masaryk University

The pros & cons of the third version

Pros

- Easy to use for web based apps and non-web based apps alike
- Portability improvements due to single input
- Better for collecting statistical data for future improvement
- Easy to be used for automatic testing

Cons

- Single output $\Rightarrow$ possibly lower performance

FACULTY
OF ARTS
Masaryk University

Currently included and worked on modules

Current modules

- *commas*, *non-grammatical constructions*, *"other"*
  usage of SET parser for proofreading

Worked on modules

- *spellchecker*
  data from Internet Language Manual
  (Internetová jazyková příručka, IJP)
- *interpunction*
  regular expressions designed by Zbyněk Michálek

FACULTY
OF ARTS
Masaryk University

Processing speed

- Optimatization of used data structures (named tuples instead of dictionaries)
- Asynchronous processing using `asyncio` module – already included
- Multiprocessing using `multiprocessing` module – to be included

FACULTY
OF ARTS
Masaryk University

Big data testing & alternative tagger

- We are getting testing data from the cooperating news company
- Currently the normalized input is being made for the big data testing to be started
- Optional use of different morphological analysis tools to achieve the best performance
- MorphoDiTa is currently being implemented
- We will surely try the new neural tagger presented earlier by Marie Stará

FACULTY
OF ARTS
Masaryk University

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/1061/2018) and by the Technology Agency of the Czech Republic under the project TL02000146.