

Vybrané nástroje na příkazové řádce

Jana Straková
ÚFAL MFF UK

Výběr nástrojů pro analýzu textu

- **UDPipe**: nástroj pro analýzu textu
 - 60 jazyků
 - rozdělí text na věty, slova
 - provede **morfologickou analýzu** - slovní druhy, pády, rody, vidy, ...
 - provede **lemmatizaci** - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
 - provede **závislostní analýzu** - gramatika (podmět, přísudek, přívlastky, ...)
 - <http://ufal.mff.cuni.cz/udpipe>

Výběr nástrojů pro analýzu textu

- **UDPipe**: nástroj pro analýzu textu
 - 60 jazyků
 - rozdělí text na věty, slova
 - provede **morfologickou analýzu** - slovní druhy, pády, rody, vidy, ...
 - provede **lemmatizaci** - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
 - provede **závislostní analýzu** - gramatika (podmět, přísudek, přívlastky, ...)
 - <http://ufal.mff.cuni.cz/udpipe>
- **NameTag**: hledá vlastní jména v textu (“Praha”, “Jan Novák”, ...)
 - <http://ufal.mff.cuni.cz/nametag>

Výběr nástrojů pro analýzu textu

- **UDPipe**: nástroj pro analýzu textu
 - 60 jazyků
 - rozdělí text na věty, slova
 - provede **morfologickou analýzu** - slovní druhy, pády, rody, vidy, ...
 - provede **lemmatizaci** - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
 - provede **závislostní analýzu** - gramatika (podmět, přísudek, přívlastky, ...)
 - <http://ufal.mff.cuni.cz/udpipe>
- **NameTag**: hledá vlastní jména v textu (“Praha”, “Jan Novák”, ...)
 - <http://ufal.mff.cuni.cz/nametag>
- **Korektor**: spellchecker a oprava gramatiky
 - <http://ufal.mff.cuni.cz/korektor>

Výběr nástrojů pro analýzu textu

- **UDPipe**: nástroj pro analýzu textu
 - 60 jazyků
 - rozdělí text na věty, slova
 - provede **morfologickou analýzu** - slovní druhy, pády, rody, vidy, ...
 - provede **lemmatizaci** - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
 - provede **závislostní analýzu** - gramatika (podmět, přísudek, přívlastky, ...)
 - <http://ufal.mff.cuni.cz/udpipe>
- **NameTag**: hledá vlastní jména v textu (“Praha”, “Jan Novák”, ...)
 - <http://ufal.mff.cuni.cz/nametag>
- **Korektor**: spellchecker a oprava gramatiky
 - <http://ufal.mff.cuni.cz/korektor>

vzorový nástroj
pro dnešní tutoriál

Výběr nástrojů pro analýzu textu

- **UDPipe**: nástroj pro analýzu textu
 - 60 jazyků
 - rozdělí text na věty, slova
 - provede **morfologickou analýzu** - slovní druhy, pády, rody, vidy, ...
 - provede **lemmatizaci** - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
 - provede **závislostní analýzu** - gramatika (podmět, přísudek, přívlastky, ...)
 - <http://ufal.mff.cuni.cz/udpipe>
- **NameTag**: hledá vlastní jména v textu (“Praha”, “Jan Novák”, ...)
 - <http://ufal.mff.cuni.cz/nametag>
- **Korektor**: spellchecker a oprava gramatiky
 - <http://ufal.mff.cuni.cz/korektor>

příklad na doma
nebo když zbyde čas

Co budeme potřebovat

- osobní počítač s operačním systémem Unix / Windows
- webový prohlížeč
- příkazová řádka
 - Linux: Start -> Terminál / Emulátor terminálu
 - Windows:
 - [Cygwin](#) - program napodobující chování Unixových systémů ve Windows
 - nebo [WSL - Windows Subsystem for Linux](#)

Data

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK

Korpus: Czech Sociological Review 1993-2016

- Czech Sociological Review 1993-2016

- Hladik, Radim, 2018, *Czech Sociological Review 1993-2016*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11372/LRT-2703>
- Korpus **stáhneme a rozbalíme** (návod: [Předchozí tutoriál](#))

The screenshot displays the LINDAT/CLARIN digital library interface. At the top, there is a navigation bar with the LINDAT logo and menu items: Repozitář, Hledání v korpusech, TreeQuery, Treex, Aplikace, O nás, and CLARIN. Below the navigation bar, the page title is 'Czech Sociological Review 1993-2016'. A search bar is visible in the top right corner. The main content area features a citation recommendation: 'Pro citaci použijte následující text, nebo export do připraveného formátu:' followed by a citation for Hladik, Radim (2018) and a 'BIBTEX' button. Below this is a 'Sdílet:' section with social media icons for Facebook and Twitter. At the bottom left, the 'Autori' (Hladik, Radim) and 'Identifikátor' (<http://hdl.handle.net/11372/LRT-2703>) are listed. On the right side, there is a sidebar with the LINDAT and CLARIN logos, a search bar, and buttons for 'DEPOSIT' and 'CITE'. The sidebar also includes a 'Kde začít?' section with a question mark icon and a 'Procházet' section with a dropdown menu.

Prohlížení korpusu

- Přípona “tsv” = tab separated value = obsahuje sloupce.

- Prohlížení v textovém editoru: Poznámkový blok, PSPad, vim, Emacs, ...
- Prohlížení z příkazové řádky: `less soc_review_corpus_1993-2016.tsv`

původní text rozdělený na slova

```

Terminal
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|Stranou stranou Db-----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|je být VB-S---3P-AA---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|ponechávána ponechávat :T_^(*4at) VsQW---XX-AP---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|centralizace centralizace NNFS1-----A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|jako jako-1 J,-----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|nutná nutný AAFS1----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|procesní procesní AAFS1----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|forma forma NNFS1-----A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|každého každý AAIS2----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|státu stát-1_^ (státní útvar) NNIS2----A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|a a-1 J^-----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|specifické specifický AAIP1----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|důvody důvod NNIP1-----A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|československé československý AAIP1----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|rovněž rovněž Db-----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|. Z:-----
  
```

Prohlížení korpusu

- Přípona “tsv” = tab separated value = obsahuje sloupce.
 - Prohlížení v textovém editoru: Poznámkový blok, PSPad, vim, Emacs, ... **slova v základním tvaru (lemmata)**
 - Prohlížení z příkazové řádky: `less soc_review_corpus_1993-2016.tsv`

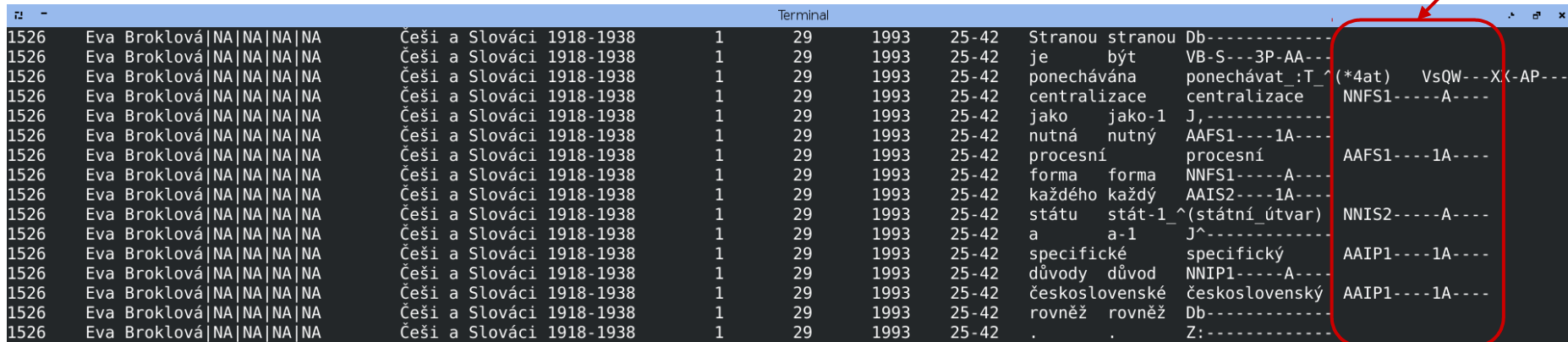
```

Terminal
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|Stranou|stranou|Db-----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|je|být|VB-S---3P-AA---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|ponechávána|ponechávat:T_^(*4at)|VsQW---XX-AP---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|centralizace|centralizace|NNFS1----A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|jako|jako-1|)-----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|nutná|nutný|AAFS1----1A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|procesní|procesní|AAFS1----1A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|forma|forma|NNFS1----A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|každého|každý|AAIS2----1A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|státu|stát-1_^(státní_ú tvar)|NNIS2----A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|a|a-1|)-----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|specifické|specifický|AAIP1----1A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|důvody|důvod|NNIP1----A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|československé|československý|AAIP1----1A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|rovněž|rovněž|Db-----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|.|.|Z:-----
  
```

Prohlížení korpusu

- Přípona “tsv” = tab separated value = obsahuje sloupce.
 - Prohlížení v textovém editoru: Poznámkový blok, PSPad, vim, Emacs, ...
 - Prohlížení z příkazové řádky: `less soc_review_corpus_1993-2016.tsv`

morfologické značky



```

1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|Stranou|stranou|Db-
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|je|být|VB-S---3P-AA--
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|ponechávána|ponechávat_
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|centralizace|centralizace|
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|jako|jako-1|J,
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|nutná|nutný|AAFS1----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|procesní|procesní|AAFS1----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|forma|forma|NNFS1----A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|každého|každý|AAIS2----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|státu|stát-1_^(státní_útv
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|a|a-1|J^
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|specifické|specifický|AAIP1----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|důvody|důvod|NNIP1----A----
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|československé|československý|AAIP1----1A---
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|rovněž|rovněž|Db-
1526 Eva Broklová|NA|NA|NA|NA|NA|NA|NA|NA|Češi a Slováci|1918-1938|1|29|1993|25-42|.|. |Z:
  
```

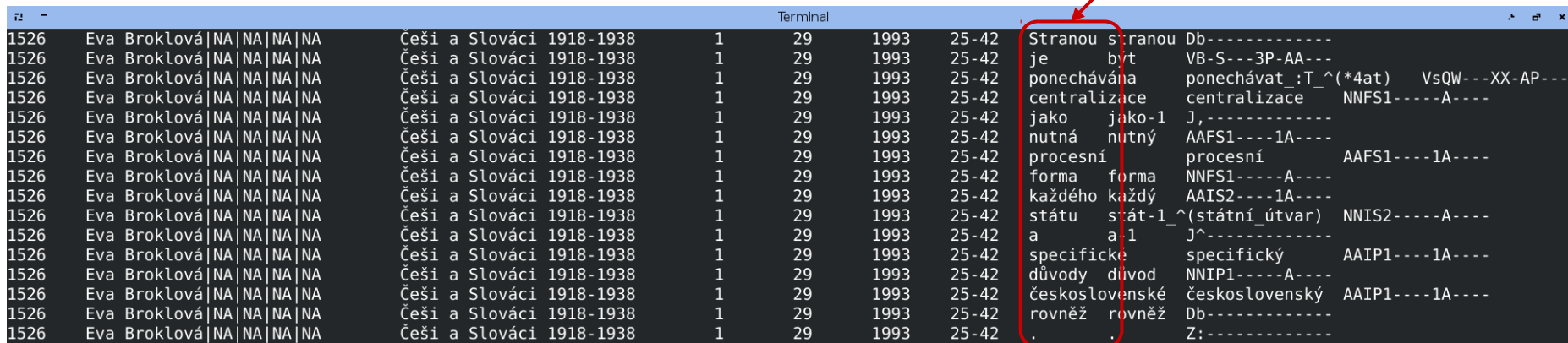

Motivace: O kom se psalo?

- Chceme analýzu pro **vlastní jména** -> hledáme **pojmenované entity**
- **NameTag**: nástroj pro rozpoznávání pojmenovaných entit

Vyjmeme vstupní text rozdělený na slova (8. sloupec) a uložíme zvlášť:

```
cut -f8 soc_review_corpus_1993-2016.tsv > slova.txt
```

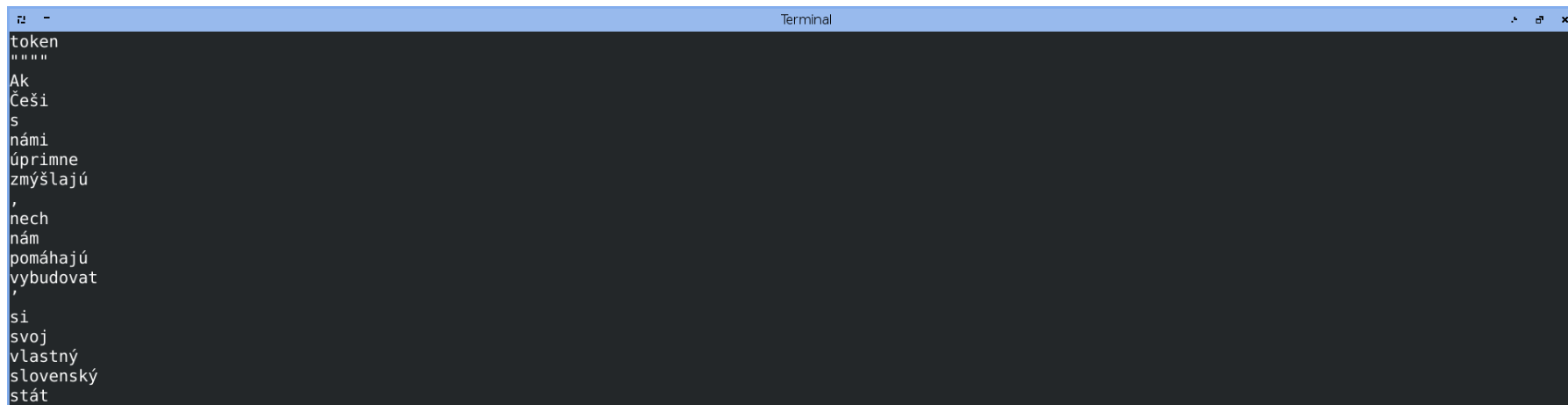
původní text rozdělený na slova



	1	2	3	4	5	6	7	8	9	10	11	12
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	Stranou stranou Db-----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	je být VB-S---3P-AA---
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	ponechávána ponechávat :T_^(*4at) VsQW---XX-AP---
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	centralizace centralizace NNFS1-----A----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	jako jako-1 J,-----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	nutná nutný AAFS1----1A----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	procesní procesní AAFS1----1A----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	forma forma NNFS1-----A----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	každého každý AAIS2----1A----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	státu stát-1_^(státní útvar) NNIS2-----A----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	a a-1 J^-----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	specifické specifický AAIP1----1A----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	důvody důvod NNIP1-----A----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	československé československý AAIP1----1A----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	rovněž rovněž Db-----
1526	Eva Broklová	NA	NA	NA	NA	Češi a Slováci	1918-1938	1	29	1993	25-42	. Z:-----

- Prohlédneme si obsah souboru slova.txt:

```
less slova.txt
```

A terminal window titled "Terminal" with a dark background and light text. The output of the 'less' command is displayed as follows:

```
token  
" " " "  
Ak  
Češi  
s  
námi  
úprimne  
zmýšľajú  
,  
nech  
nám  
pomáhajú  
vybudovat  
,  
si  
svoj  
vlastný  
slovenský  
stát
```

Způsoby využití nástrojů z LINDAT

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK

Tři způsoby použití nástrojů

- Demo v prohlížeči
 - pouze pro jednoduché příklady
 - nelze zpracovat větší data

Tři způsoby použití nástrojů

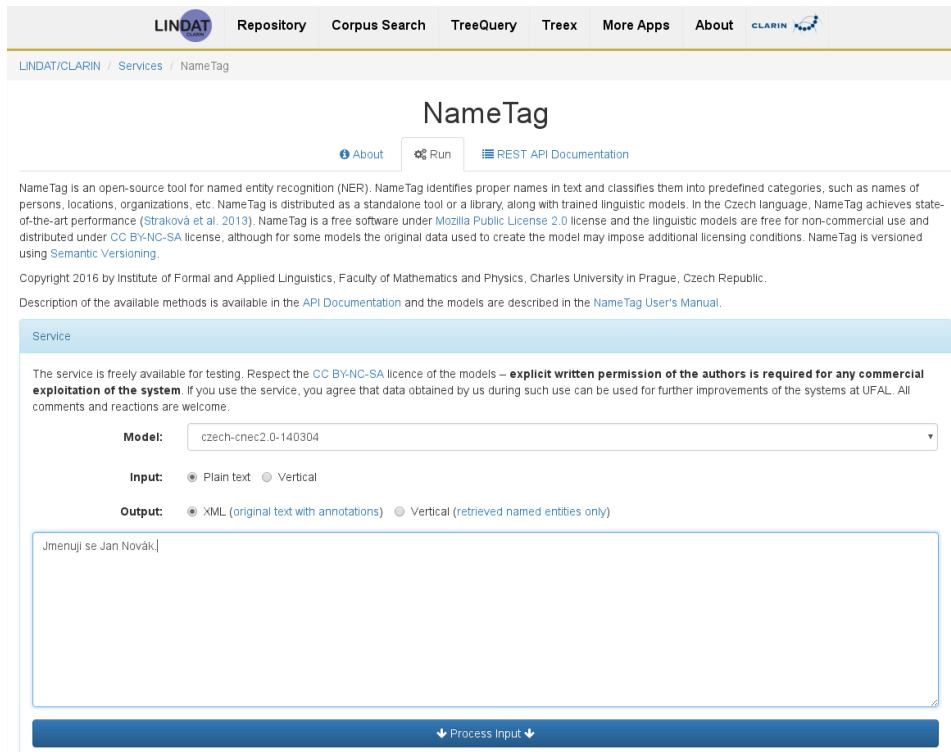
- **Demo** v prohlížeči
 - pouze pro jednoduché příklady
 - nelze zpracovat větší data
- **Instalace nástroje (a modelů)** do vlastního počítače
 - Lze zpracovávat velká data
 - Mohu trénovat vlastní modely
 - Nástroj lze zapojit do většího frameworku

Tři způsoby použití nástrojů

- **Demo** v prohlížeči
 - pouze pro jednoduché příklady
 - nelze zpracovat větší data
- **Instalace nástroje (a modelů)** do vlastního počítače
 - Lze zpracovávat velká data
 - Mohu trénovat vlastní modely
 - Nástroj lze zapojit do většího frameworku
- **REST API** je způsob, jak snadno získávat informace z **webové služby**
 - Nástroj “žije” na serveru jako **webová služba**, my (klient) posíláme dotazy a dostáváme zpět odpovědi od serveru
 - Odpadá instalace nástroje a správa modelů

Demo

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK



The screenshot shows the NameTag web interface. At the top, there is a navigation bar with the LINDAT logo and links for Repository, Corpus Search, TreeQuery, Treex, More Apps, and About. Below this is a breadcrumb trail: LINDAT/CLARIN / Services / NameTag. The main heading is "NameTag". There are three tabs: "About", "Run", and "REST API Documentation". The "Run" tab is active. The main content area contains a paragraph describing NameTag as an open-source tool for named entity recognition (NER). It mentions that NameTag identifies proper names in text and classifies them into predefined categories. It also states that NameTag is distributed as a standalone tool or a library, along with trained linguistic models. The text is licensed under CC BY-NC-SA, and the linguistic models are free for non-commercial use. A copyright notice for 2016 by the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic, is provided. A link to the API Documentation and the NameTag User's Manual is included. Below this is a "Service" section with a text area for input and a "Process Input" button. The input field contains the text "Jmenuji se Jan Novák".

LINDAT
Repository Corpus Search TreeQuery Treex More Apps About CLARIN

LINDAT/CLARIN / Services / NameTag

NameTag

About Run REST API Documentation

NameTag is an open-source tool for named entity recognition (NER). NameTag identifies proper names in text and classifies them into predefined categories, such as names of persons, locations, organizations, etc. NameTag is distributed as a standalone tool or a library, along with trained linguistic models. In the Czech language, NameTag achieves state-of-the-art performance (Straková et al. 2013). NameTag is a free software under Mozilla Public License 2.0 license and the linguistic models are free for non-commercial use and distributed under CC BY-NC-SA license, although for some models the original data used to create the model may impose additional licensing conditions. NameTag is versioned using Semantic Versioning.

Copyright 2016 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.

Description of the available methods is available in the [API Documentation](#) and the models are described in the [NameTag User's Manual](#).

Service

The service is freely available for testing. Respect the [CC BY-NC-SA](#) licence of the models – **explicit written permission of the authors is required for any commercial exploitation of the system**. If you use the service, you agree that data obtained by us during such use can be used for further improvements of the systems at UFAL. All comments and reactions are welcome.

Model: czech-cnec2.0-140304

Input: Plain text Vertical

Output: XML (original text with annotations) Vertical (retrieved named entities only)

Jmenuji se Jan Novák

↓ Process Input ↓

Instalace nástroje

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK

Instalace

Instalace probíhá stažením a rozbalením do adresáře.

Vytvoříme si nový adresář např. `nametag`:

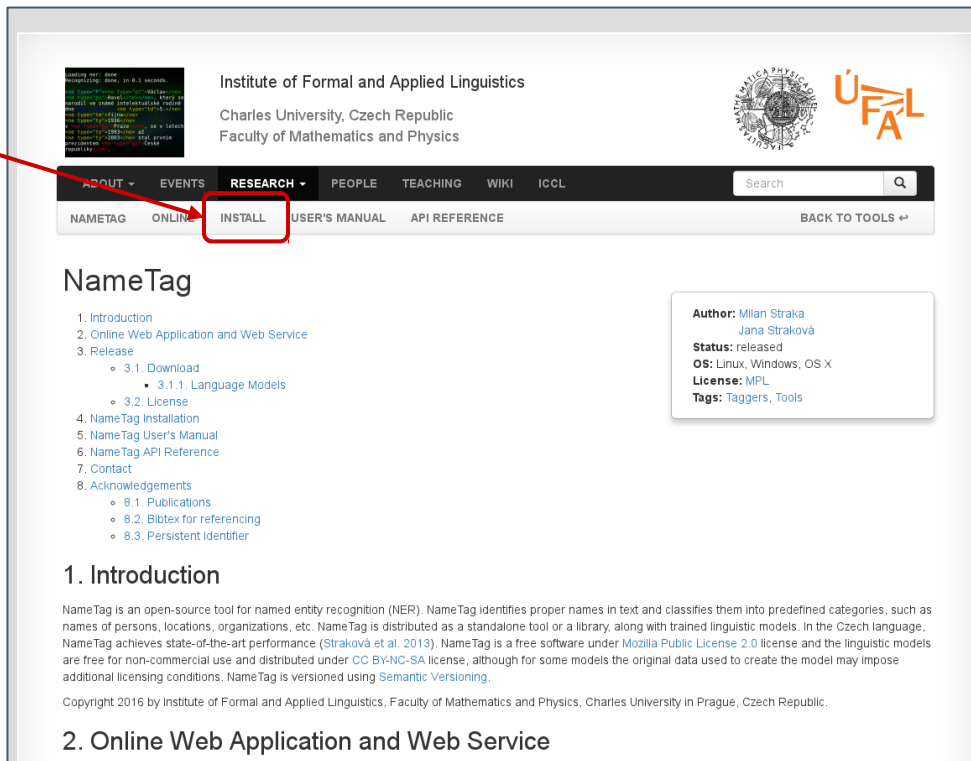
- z příkazové řádky Unix: `mkdir nametag`
- případně jak jsme zvyklí ve správci

Stačí umístit do osobních složek uživatele, např. `/home/<uživatel>/nametag`.

Budeme potřebovat:

1. program `NameTag`
2. předtrénovaný model k programu

Instalace



Institute of Formal and Applied Linguistics
Charles University, Czech Republic
Faculty of Mathematics and Physics

ABOUT ▾ EVENTS RESEARCH ▾ PEOPLE TEACHING WIKI ICCL

SEARCH

NAMETAG ONLINE **INSTALL** USER'S MANUAL API REFERENCE BACK TO TOOLS ↵

NameTag

- 1. Introduction
- 2. Online Web Application and Web Service
- 3. Release
 - 3.1. Download
 - 3.1.1. Language Models
 - 3.2. License
- 4. NameTag Installation
- 5. NameTag User's Manual
- 6. NameTag API Reference
- 7. Contact
- 8. Acknowledgements
 - 8.1. Publications
 - 8.2. Bibtex for referencing
 - 8.3. Persistent Identifier

Author: Milan Straka
Jana Straková
Status: released
OS: Linux, Windows, OS X
License: MPL
Tags: Taggers, Tools

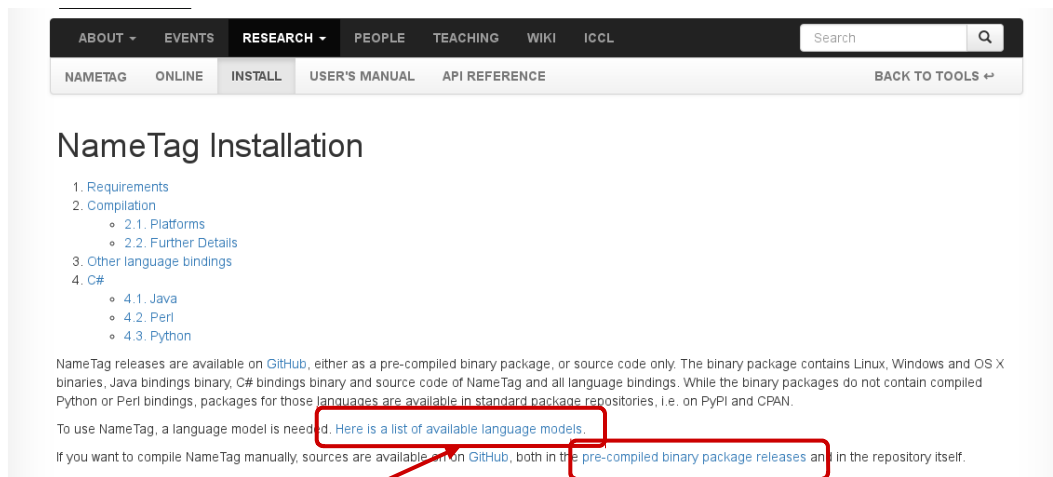
1. Introduction

NameTag is an open-source tool for named entity recognition (NER). NameTag identifies proper names in text and classifies them into predefined categories, such as names of persons, locations, organizations, etc. NameTag is distributed as a standalone tool or a library, along with trained linguistic models. In the Czech language, NameTag achieves state-of-the-art performance (Straková et al. 2013). NameTag is a free software under Mozilla Public License 2.0 license and the linguistic models are free for non-commercial use and distributed under CC BY-NC-SA license, although for some models the original data used to create the model may impose additional licensing conditions. NameTag is versioned using Semantic Versioning.

Copyright 2016 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.

2. Online Web Application and Web Service

<http://ufal.mff.cuni.cz/nametag>



The screenshot shows the 'NameTag Installation' page. The navigation bar includes 'ABOUT', 'EVENTS', 'RESEARCH', 'PEOPLE', 'TEACHING', 'WIKI', 'ICCL', a search box, and a 'BACK TO TOOLS' link. Below the navigation bar, there are links for 'NAMETAG', 'ONLINE', 'INSTALL', 'USER'S MANUAL', and 'API REFERENCE'. The main content area has the title 'NameTag Installation' and a list of sections: 1. Requirements, 2. Compilation (with sub-sections 2.1. Platforms and 2.2. Further Details), 3. Other language bindings, and 4. C# (with sub-sections 4.1. Java, 4.2. Perl, and 4.3. Python). A paragraph of text follows, mentioning that NameTag releases are available on GitHub and that binary packages are available in standard package repositories like PyPI and CPAN. Two red boxes highlight specific links: one around 'Here is a list of available language models.' and another around 'pre-compiled binary package releases'. Two red arrows point from these boxes to the explanatory text below the screenshot.

1. odkaz na předtrénovaný model

2. odkaz na program v repozitáři GitHub

<http://ufal.mff.cuni.cz/nametag/install>

Instalace - stažení modelu

LINDAT CLARIN | Repozitář | Hledání v korpusech | TreeQuery | Treex | Aplikace | O nás | CLARIN

Domovská stránka repozitáře LINDAT/CLARIN / Zobrazit záznam

Czech Models (CNEC) for NameTag

Pro citaci použijte následující text, nebo export do připraveného formátu: BIBTEX CMDI

Straka, Milan and Straková, Jana, 2014, Czech Models (CNEC) for NameTag, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11858/00-097C-0000-0023-7D42-8>.

Sdílet: f t

LINDAT / CLARIN

Autoři	Straka, Milan ; Straková, Jana
Identifikátor	http://hdl.handle.net/11858/00-097C-0000-0023-7D42-8
URL projektu	http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec
Datum vydání	2014-03-04
Typ	languageDescription
Velikost	31 mb
Jazyky	Czech
Popis	Czech models for NameTag, providing recognition of named entities. The models are trained on Czech Named Entity Corpus 2.0 and 1.1.
Nakladatel	Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Licenční kategorie: **Publicly Available**
Licence: Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)

CC BY-NC-SA

Název	czech-cnec-140304.zip
Velikost	30.81 MB
Formát	application/zip
Popis	Czech Models (CNEC) for NameTag
MD5	d1ec1ff8aa3b22a015d5000073bec7f

Stáhnout soubor | Náhled

Instalace - stažení programu

ufal / nametag

Watch 13 Star 26 Fork 5

Code Issues 2 Pull requests 2 Projects 0 Security Insights

Be notified of new releases [Dismiss](#)

Create your free GitHub account today to subscribe to this repository for new releases and build software alongside 40 million developers.

[Sign up](#)

Releases Tags

Latest release

v1.1.2
5c65d87

NameTag 1.1.2

foxiik released this on Jul 1, 2017 · 32 commits to master since this release

Changes since NameTag 1.1.1:

- Allow specifying custom path to C++ library in Java.
- Fix bug causing a memory leak on g++.
- Add --log option to the REST server.

Assets 3

nametag-1.1.2-bin.zip	9.47 MB
Source code (zip)	
Source code (tar.gz)	

Instalace - rozbalení

Program i model rozbalíme a máme nainstalováno.

- z příkazové řádky Unix:
 - `unzip czech-cnec-140304.zip`
 - `unzip nametag-1.1.2-bin.zip`
- případně jak jsme zvyklí rozbalovat zip soubory ve správci

První spuštění

Najdeme odpovídající binárku programu pro náš počítač v adresáři `nametag-1.1.2-bin`:

- 32-bit linux: `bin-linux32`, 64-bit linux: `bin-linux64`
- 32-bit Windows: `bin-win32`, 64-bit Windows: `bin-win64`

Spuštění programu: `nametag-1.1.2-bin/bin-linux64/run_ner`

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spuštění rozpoznávače - Uživatelský manuál

Institute of Formal and Applied Linguistics
Charles University, Czech Republic
Faculty of Mathematics and Physics

ABOUT ▾ EVENTS RESEARCH ▾ PEOPLE TEACHING WIKI ICCL

NAME TAG ONLINE INSTALL **USER'S MANUAL** API REFERENCE

Search

BACK TO TOOLS ←

NameTag User's Manual

1. Czech NameTag Models
 - 1.1. Download
 - 1.2. Acknowledgements
 - 1.2.1. Publications
 - 1.3. Czech Named Entity Corpus 2.0 Model
 - 1.4. Czech Named Entity Corpus 1.1 Model
- 2. Running the Recognizer**
 - 2.1. Input Formats
 - 2.2. Output Formats
3. Running the Tokenizer
 - 3.1. Output Formats
4. Running REST Server
5. Training of Custom Models
 - 5.1. Training data
 - 5.2. Tagger
 - 5.2.1. Lemma Structure
 - 5.3. Feature Templates
 - 5.4. Running train_ner

<http://ufal.mff.cuni.cz/nametag/users/manual>

Spuštění rozpoznávače - Uživatelský manuál

2. Running the Recognizer

The NameTag Recognizer can be executed using the following command:

```
run_ner recognizer_model
```

The input is assumed to be in UTF-8 encoding and can be either already tokenized and segmented, or it can be a plain text which is tokenized and segmented automatically.

Any number of files can be specified after the `recognizer_model`. If an argument `input_file:output_file` is used, the given `input_file` is processed and the result is saved to `output_file`. If only `input_file` is used, the result is saved to standard output. If no argument is given, input is read from standard input and written to standard output.

The full command syntax of `run_ner` is

```
Usage: run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
        --output=vertical|xml
```

2.1. Input Formats

The input format is specified using the `--input` option. Currently supported input formats are:

- `untokenized` (default): the input is tokenized and segmented using a tokenizer defined by the model.
- `vertical`: the input is in vertical format, every line is considered a word, with empty line denoting end of sentence.

2.2. Output Formats

The output format is specified using the `--output` option. Currently supported output formats are:

- `xml` (default): Simple XML format without a root element, using `<sentence>` element to mark sentences and `<token>` element to mark tokens. The recognized named entities are encoded using `<ne type="...">` element.

Example input:

```
Václav Havel byl český dramatik, esejista, kritik komunistického režimu a později politik.
```

A NameTag identifies a first name (`pf`), a surname (`ps`) and a person name container (`P`) in the input (line breaks added):

```
<sentence><ne type="P"><ne type="pf"><token>Václav</token></ne> <ne type="ps"><token>Havel</token></ne></ne>
<token>byl</token> <token>český</token> <token>dramatik</token><token>,</token> <token>esejista</token><token>,</token>
<token>kritik</token> <token>komunistického</token> <token>režimu</token> <token>a</token> <token>později</token>
<token>politik</token></token></sentence>
```

- `vertical`: Every found named entity is on a separate line. Each line contains three tab-separated fields: `entity_range`, `entity_type` and `entity_text`. The `entity_range` is composed of token identifiers (counting from 1 and including end-of-sentence, if the input is also `vertical`, token identifiers correspond exactly to line numbers) of tokens forming the named entity and `entity_type` represents its type. The `entity_text` is not strictly necessary and contains space separated words of this named entity.

Example input:

```
Václav Havel byl český dramatik, esejista, kritik komunistického režimu a později politik.
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
Loading ner: done
<sentence><token>"</token><ne type="P"><ne type="pf"><token>Václav</token></ne> <ne type="ps"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním</token>
<token><token>prezidentem</token> <ne type="gc"><token>České</token> <token>republiky</token></ne><token>.</token><token>"</token></sentence>
Recognizing done, in 0.000 seconds.
gris@ametyst:~/nametag_tutorial$
```


Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spuštění rozpoznávače - parametr

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spustíme rozpoznávač s parametrem:

```
nametag-1.1.2-bin/bin-linux64/run_ner --output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - parametr

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spustíme rozpoznávač s parametrem:

```
nametag-1.1.2-bin/bin-linux64/run_ner --output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner --output=vertical czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
Loading ner: done
2,3   P      Václav Havel
2     pf     Václav
3     ps     Havel
7,8   gc     České republiky
Recognizing done, in 0.000 seconds.
```

Spuštění rozpoznávače na našich datech

```
token
" " " "
Ak
Češi
s
námi
úprimne
zmýšľajú
,
nech
nám
pomáhajú
vybudovať
,
si
svoj
vlastný
slovenský
stát
```

Hladik, Radim, 2018, *Czech Sociological Review 1993-2016*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University,
<http://hdl.handle.net/11372/LRT-2703>

Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```


Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Spuštění rozpoznávače na našich datech

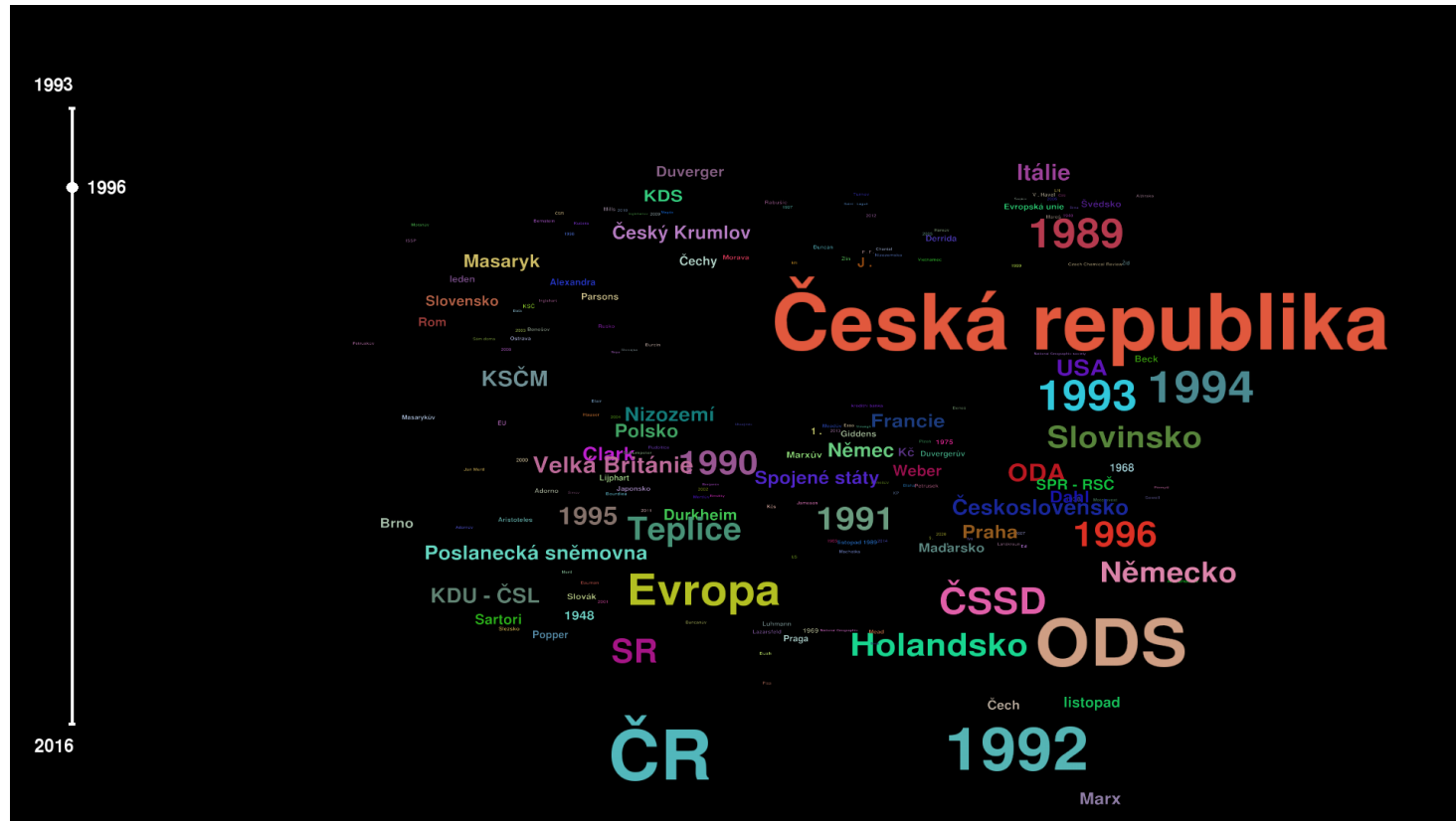
```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Výstup rozpoznávače

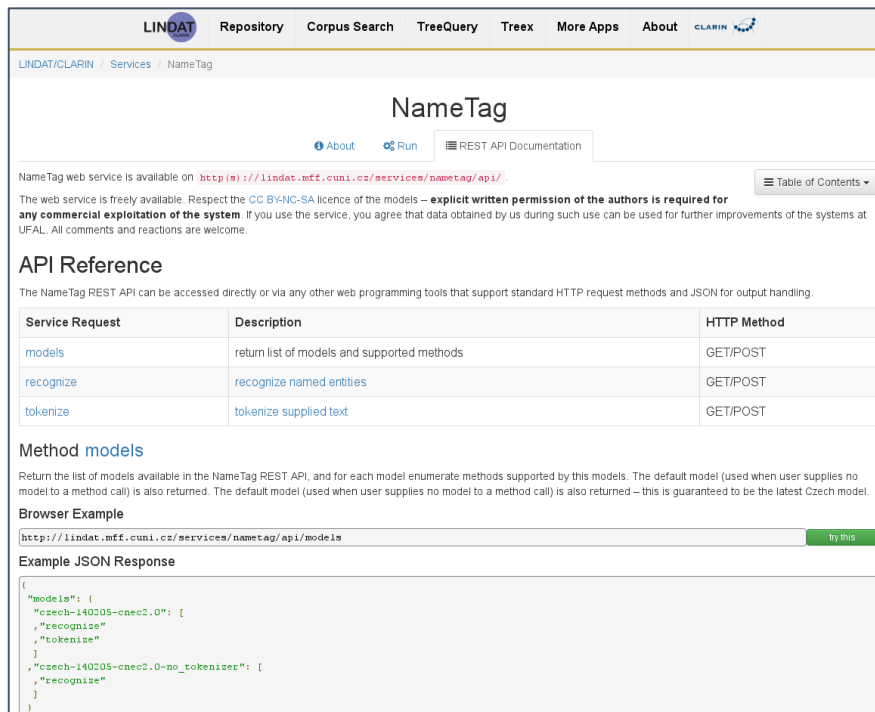
id	typ	text
3	gc	Ak
4	pc	Češi
42,43	gc	slovenskej republiky
122	pc	Čechů
124	pc	Slovákům
197	gc	Slovenska
271	gc	ČSR
338	ps	Henlein
345	gc	Slovensku
469,470	gr	Podkarpatské Rusi
520	gc	Slovenska
537	gr	Slezsku
539	ps	Masaryk
618	ty	1918
642	gc	Rakouska

Vizualizace nejčastějších vlastních jmen



Webová služba

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK



The screenshot shows the LINDAT/CLARIN NameTag REST API reference page. The page has a navigation bar with links for Repository, Corpus Search, TreeQuery, Treex, More Apps, and About. The main content area is titled "NameTag" and includes a "REST API Documentation" tab. Below the title, there is a "Table of Contents" dropdown menu. The page contains a "NameTag web service is available on" link, a "The web service is freely available. Respect the CC BY-NC-SA licence of the models – explicit written permission of the authors is required for any commercial exploitation of the system." notice, and an "API Reference" section. The API Reference section includes a table with columns for Service Request, Description, and HTTP Method. Below the table, there is a "Method models" section with a description and a "Browser Example" section with a URL input field and a "try this" button. Finally, there is an "Example JSON Response" section with a code block containing JSON data.

LINDAT Repository Corpus Search TreeQuery Treex More Apps About CLARIN

LINDAT/CLARIN / Services / NameTag

NameTag

About Run REST API Documentation

NameTag web service is available on [http\(s\)://lindat.mff.cuni.cz/services/nametag/api/](http(s)://lindat.mff.cuni.cz/services/nametag/api/).

The web service is freely available. Respect the [CC BY-NC-SA](#) licence of the models – **explicit written permission of the authors is required for any commercial exploitation of the system**. If you use the service, you agree that data obtained by us during such use can be used for further improvements of the systems at UFAL. All comments and reactions are welcome.

Table of Contents

API Reference

The NameTag REST API can be accessed directly or via any other web programming tools that support standard HTTP request methods and JSON for output handling.

Service Request	Description	HTTP Method
models	return list of models and supported methods	GET/POST
recognize	recognize named entities	GET/POST
tokenize	tokenize supplied text	GET/POST

Method [models](#)

Return the list of models available in the NameTag REST API, and for each model enumerate methods supported by this models. The default model (used when user supplies no model to a method call) is also returned. The default model (used when user supplies no model to a method call) is also returned – this is guaranteed to be the latest Czech model.

Browser Example

<http://lindat.mff.cuni.cz/services/nametag/api/models> [try this](#)

Example JSON Response

```
{
  "models": {
    "czech-140205-cnec2.0": [
      "recognize",
      "tokenize"
    ],
    "czech-140205-cnec2.0-no_tokenizer": [
      "recognize"
    ]
  }
}
```

<http://lindat.mff.cuni.cz/services/nametag/api-reference.php>

REST API v prohlížeči

Do adresy v prohlížeči zadáme

`http://lindat.mff.cuni.cz/services/nametag/api/recognize?data=Václav Havel byl prvním prezidentem České republiky.`

REST API v prohlížeči

Do adresy v prohlížeči zadáme

`http://lindat.mff.cuni.cz/services/nametag/api/recognize?data=Václav Havel byl prvním prezidentem České republiky.`

Vrátí se nám

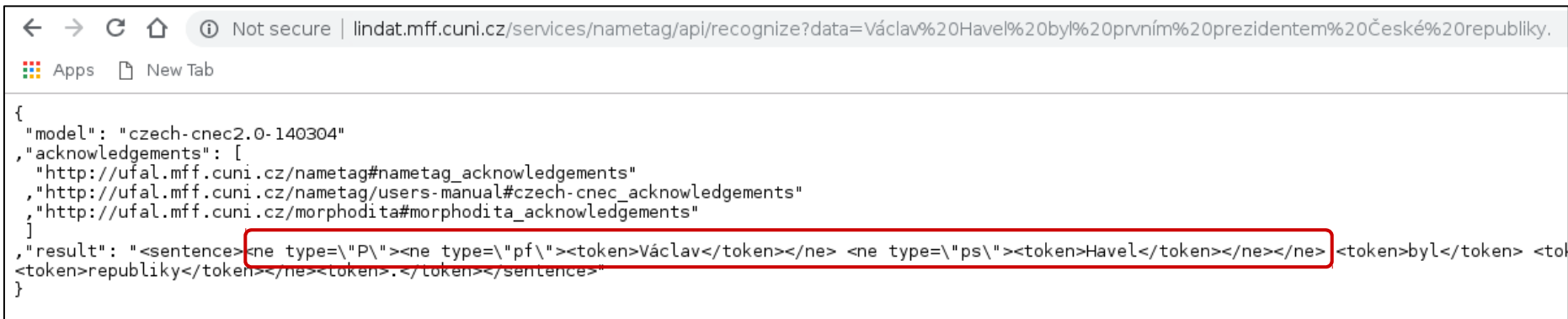
```
← → ↻ 🏠 ⓘ Not secure | lindat.mff.cuni.cz/services/nametag/api/recognize?data=Václav%20Havel%20byl%20prvním%20prezidentem%20České%20republiky.  
📱 Apps 📄 New Tab  
{  
  "model": "czech-cnec2.0-140304"  
  , "acknowledgements": [  
    "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"  
    , "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"  
    , "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"  
  ]  
  , "result": "<sentence><ne type='P'><ne type='pf'><token>Václav</token></ne> <ne type='ps'><token>Havel</token></ne></ne> <token>byl</token> <to</token>republiky</token></ne><token>.</token></sentence>"  
}
```

REST API v prohlížeči

Do adresy v prohlížeči zadáme

`http://lindat.mff.cuni.cz/services/nametag/api/recognize?data=Václav Havel byl prvním prezidentem České republiky.`

Vrátí se nám



```
{
  "model": "czech-cnec2.0-140304"
, "acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
, "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
, "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
]
, "result": "<sentence><ne type=\"P\"><ne type=\"pf\"><token>Václav</token></ne> <ne type=\"ps\"><token>Havel</token></ne></ne> <token>byl</token> <token>republiky</token></ne></sentence>."
}
```

REST API v prohlížeči

Do adresy v prohlížeči zadáme

`http://lindat.mff.cuni.cz/services/nametag/api/recognize?data=Václav Havel byl prvním prezidentem České republiky.`

Vrátí se nám



```
{
  "model": "czech-cnec2.0-140304"
, "acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
, "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
, "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
]
, "result": "<sentence><ne type='P\\'><ne type='pf\\'><token>Václav</token></ne> <ne type='ps\\'><token>Havel</token></ne></ne> <token>byl</token> <token>republiky</token></ne><token>.</token></sentence>"
}
```

formát JSON

API pomocí CURL (příkazová řádka)

Na příkazové řádce zadáme

```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'  
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

API pomocí CURL (příkazová řádka)

Na příkazové řádce zadáme

```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'  
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

curl = Client for URLs

API pomocí CURL (příkazová řádka)

Na příkazové řádce zadáme

```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'  
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```



data

API pomocí CURL (příkazová řádka)

Na příkazové řádce zadáme

```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'
```

```
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```



webová služba

API pomocí CURL (příkazová řádka)

Na příkazové řádce zadáme

```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'  
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

```
gris@ametyst:~$ curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.' http://lindat.mff.cuni.cz/services/nametag/api/recognize  
{  
  "model": "czech-cnec2.0-140304"  
  , "acknowledgements": [  
    "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"  
    , "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"  
    , "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"  
  ]  
  , "result": "<sentence><ne type=\"P\"><ne type=\"pf\"><token>Václav</token></ne> <ne type=\"ps\"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním</token> <token>prezidentem</token> <ne type=\"gc\"><token>České</token> <token>republiky</token></ne><token>.</token></sentence>"  
}  
gris@ametyst:~$ █
```


CURL - přidání parametru

```
gris@ametyst:~$ curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.' http://lindat.mff.cuni.cz/services/nametag/api/recognize
{"model": "czech-cnec2.0-140304",
"acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements",
  "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements",
  "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
],
"result": "<sentence><ne type=\"P\"><ne type=\"pf\"><token>Václav</token></ne> <ne type=\"ps\"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním</token> <token>prezidentem</token> <ne type=\"gc\"><token>České</token> <token>republiky</token></ne><token>.</token></sentence>"
}
gris@ametyst:~$
```

CURL - přidání parametru

```
gris@ametyst:~$ curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.' http://lindat.mff.cuni.cz/services/nametag/api/recognize
{"model": "czech-cnec2.0-140304",
  "acknowledgements": [
    "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements",
    "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements",
    "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
  ],
  "result": "<sentence><ne type=\\\"P\\\"><ne type=\\\"pf\\\"><token>Václav</token></ne> <ne type=\\\"ps\\\"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním</token> <token>prezidentem</token> <ne type=\\\"gc\\\"><token>České</token> <token>republiky</token></ne><token>.</token></sentence>"
}
```

CURL s parametrem:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

CURL - přidání parametru

```
gris@ametyst:~$ curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.' http://lindat.mff.cuni.cz/services/nametag/api/recognize
{"model": "czech-cnec2.0-140304",
  "acknowledgements": [
    "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements",
    "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements",
    "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
  ],
  "result": "<sentence><ne type=\"P\"><ne type=\"pf\"><token>Václav</token></ne> <ne type=\"ps\"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním</token> <token>prezidentem</token> <ne type=\"gc\"><token>České</token> <token>republiky</token></ne><token>.</token></sentence>"
}
```

CURL s parametrem:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

```
gris@ametyst:~$ curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
{"model": "czech-cnec2.0-140304",
  "acknowledgements": [
    "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements",
    "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements",
    "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
  ],
  "result": "1,2\tP\tVáclav Havel\n1\tpf\tVáclav\n2\tps\tHavel\n6,7\tgc\tČeské republiky\n"
```

CURL - vstup ze souboru, výstup do souboru

Větu uložíme do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

CURL - vstup ze souboru, výstup do souboru

Větu uložíme do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

CURL se vstupem ze souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

CURL - vstup ze souboru, výstup do souboru

Větu uložíme do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

CURL se vstupem ze souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

Přesměrování výstupu do souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize  
> result_json.txt
```

CURL - vstup ze souboru, výstup do souboru

Větu uložíme do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

CURL se vstupem ze souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

Přesměrování výstupu do souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize  
> result_json.txt
```

```
gris@ametyst:~$ curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize > result_json.txt
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           % Done   0     363  100   346    602    574   ---:--:--  ---:--:--  601
gris@ametyst:~$ cat result_json.txt
{
  "model": "czech-cnec2.0-140304"
, "acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
, "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
, "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
]
, "result": "1,2\tp\tVáclav Havel\n1\tpf\tVáclav\n2\tps\tHavel\n6,7\tgc\tČeské republiky\n"
}
gris@ametyst:~$
```

CURL - konverze JSON výstupu do textu

```
{
  "model": "czech-cnec2.0-140304"
, "acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
, "http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
, "http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
]
, "result": "1,2\tP\tVáclav Havel\n1\tpf\tVáclav\n2\tps\tHavel\n6,7\tgc\tČeské republiky\n"
}
```

```
cat result_json.txt | PYTHONIOENCODING=utf-8 python3 -c "import json,sys; print(json.load(sys.stdin)
['result'], end='')" > result.txt
```

```
gris@ametyst:~$ cat result_json.txt | PYTHONIOENCODING=utf-8 python3 -c "import json,sys; print(json.load(sys.stdin)
['result'], end='')" > result.txt
gris@ametyst:~$ cat result.txt
1,2    P      Václav Havel
1      pf     Václav
2      ps     Havel
6,7    gc     České republiky
gris@ametyst:~$
```

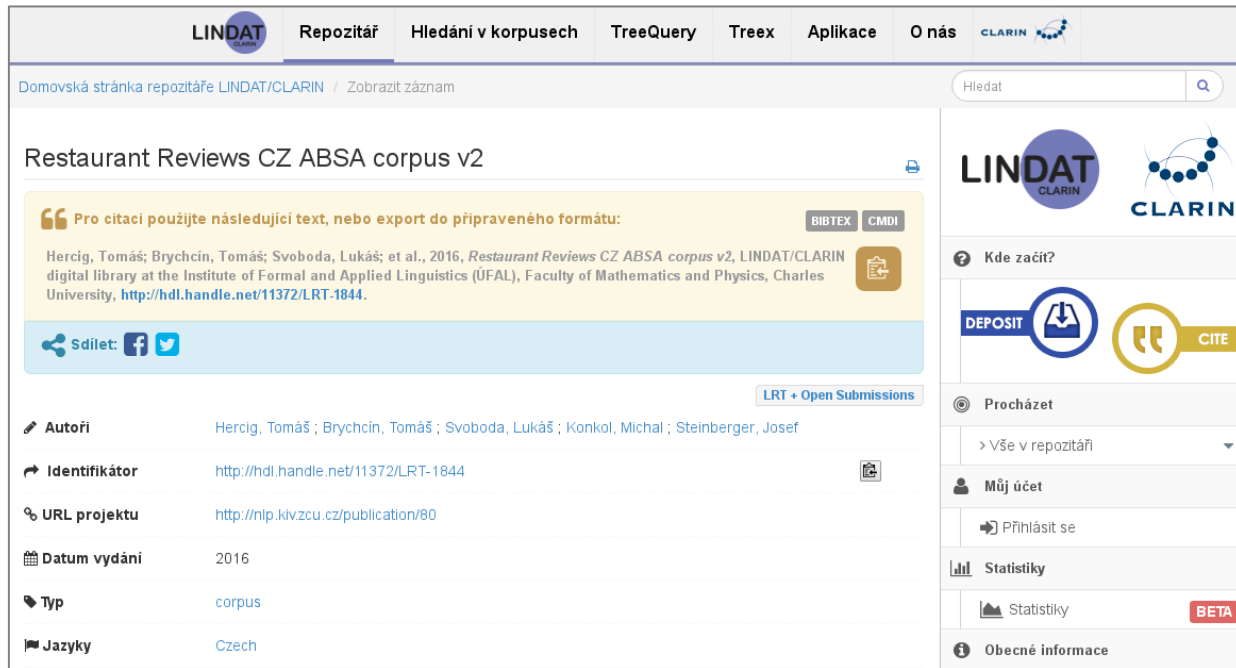

Další příklad: Korektor

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK

Korpus: Restaurant Reviews CZ ABSA corpus v2

Restaurant Reviews CZ ABSA corpus v2

Hercig, Tomáš; Brychcín, Tomáš;
Svoboda, Lukáš; et al., 2016, *Restaurant Reviews CZ ABSA corpus v2*,
LINDAT/CLARIN digital library at the
Institute of Formal and Applied
Linguistics (ÚFAL), Faculty of
Mathematics and Physics, Charles
University,
<http://hdl.handle.net/11372/LRT-1844>



The screenshot shows the LINDAT/CLARIN digital library interface. The main content area displays the record for 'Restaurant Reviews CZ ABSA corpus v2'. A yellow box contains citation information and export options (BIBTEX, CMDI). Below this is a metadata table with fields for authors, identifiers, project URL, publication date, type, and language. The right sidebar contains navigation and utility links like 'Kde začít?', 'DEPOSIT', 'CITE', 'Procházet', 'Můj účet', 'Statistiky', and 'Obecné informace'.

Pro citaci použijte následující text, nebo export do připraveného formátu:	BIBTEX	CMDI
Hercig, Tomáš; Brychcín, Tomáš; Svoboda, Lukáš; et al., 2016, <i>Restaurant Reviews CZ ABSA corpus v2</i> , LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11372/LRT-1844 .		

Autoři	Hercig, Tomáš ; Brychcín, Tomáš ; Svoboda, Lukáš ; Konkol, Michal ; Steinberger, Josef
Identifikátor	http://hdl.handle.net/11372/LRT-1844
URL projektu	http://mlp.kiv.zcu.cz/publication/80
Datum vydání	2016
Typ	corpus
Jazyky	Czech

Prohlížení korpusu

```
<?xml version="1.0" encoding="UTF-8"?>
<Reviews>
  <Review rid="ALL">
    <sentences>
      <sentence id="1411">
        <text>Jídlo taky nic moc</text>
        <Opinions>
          <Opinion category="null" from="0" polarity="negative" target="Jídlo" to="5"/>
          <Opinion category="food" from="0" polarity="negative" target="NULL" to="0"/>
        </Opinions>
      </sentence>
      <sentence id="1410">
        <text>Jídlo od Nás dostává 5 z 5 a pan Doksanský (spolumajitel) i jeho personál jsou profesí
        pití ke konkrétním jídlům</text>
        <Opinions>
          <Opinion category="null" from="44" polarity="positive" target="spolumajitel" to="56"/>
          <Opinion category="null" from="65" polarity="positive" target="personál" to="73"/>
          <Opinion category="null" from="137" polarity="positive" target="pití" to="141"/>
          <Opinion category="null" from="156" polarity="positive" target="jídlům" to="162"/>
          <Opinion category="null" from="122" polarity="positive" target="víno" to="126"/>
          <Opinion category="food" from="0" polarity="positive" target="NULL" to="0"/>
          <Opinion category="service" from="0" polarity="positive" target="NULL" to="0"/>
        </Opinions>
      </sentence>
      <sentence id="941">
        <text>Vloni trochu na kvalitě polevili,ale nadále bezva</text>
```

Motivace: doplnění chybějící diakritiky

```
<sentence id="942">
  <text>Vsichni vime, ze Bio maso je podstatne drazsí nez maso klasické, ale ceny zde jsou opravdu na ceny z Bio chovu, není něco špatné?</text>
  <Opinions>
    <Opinion category="null" from="50" polarity="neutral" target="maso" to="54"/>
    <Opinion category="null" from="69" polarity="negative" target="ceny" to="73"/>
    <Opinion category="null" from="17" polarity="neutral" target="Bio maso" to="25"/>
    <Opinion category="food" from="0" polarity="neutral" target="NULL" to="0"/>
    <Opinion category="price" from="0" polarity="negative" target="NULL" to="0"/>
  </Opinions>
</sentence>
```

```
<sentence id="10001">
  <text>V hamburgeru byl jakýsi malinky karbanátek a nejvýraznější bylo celkove rajce.</text>
  <Opinions>
    <Opinion category="null" from="72" polarity="positive" target="rajce" to="77"/>
    <Opinion category="null" from="32" polarity="negative" target="karbanátek" to="42"/>
    <Opinion category="null" from="2" polarity="negative" target="hamburgeru" to="12"/>
    <Opinion category="food" from="0" polarity="negative" target="NULL" to="0"/>
  </Opinions>
</sentence>
```

Motivace: doplnění chybějící diakritiky

```
<sentence id="942">
  <text>Vsichni víme, ze Bio maso je podstatne drazsí nez maso klasické, ale ceny zde jsou opravdu na ceny z Bio chovu, není něco špatné?</text>
  <Opinions>
    <Opinion category="null" from="50" polarity="neutral" target="maso" to="54"/>
    <Opinion category="null" from="69" polarity="negative" target="ceny" to="73"/>
    <Opinion category="null" from="17" polarity="neutral" target="Bio maso" to="25"/>
    <Opinion category="food" from="0" polarity="neutral" target="NULL" to="0"/>
    <Opinion category="price" from="0" polarity="negative" target="NULL" to="0"/>
  </Opinions>
</sentence>
```

```
<sentence id="10001">
  <text>V hamburgeru byl jakýsi malinky karbanátek a nejvýraznější bylo celkové rajce.</text>
  <Opinions>
    <Opinion category="null" from="72" polarity="positive" target="rajce" to="77"/>
    <Opinion category="null" from="32" polarity="negative" target="karbanátek" to="42"/>
    <Opinion category="null" from="2" polarity="negative" target="hamburgeru" to="12"/>
    <Opinion category="food" from="0" polarity="negative" target="NULL" to="0"/>
  </Opinions>
</sentence>
```

Korektor: spellchecker a oprava gramatiky

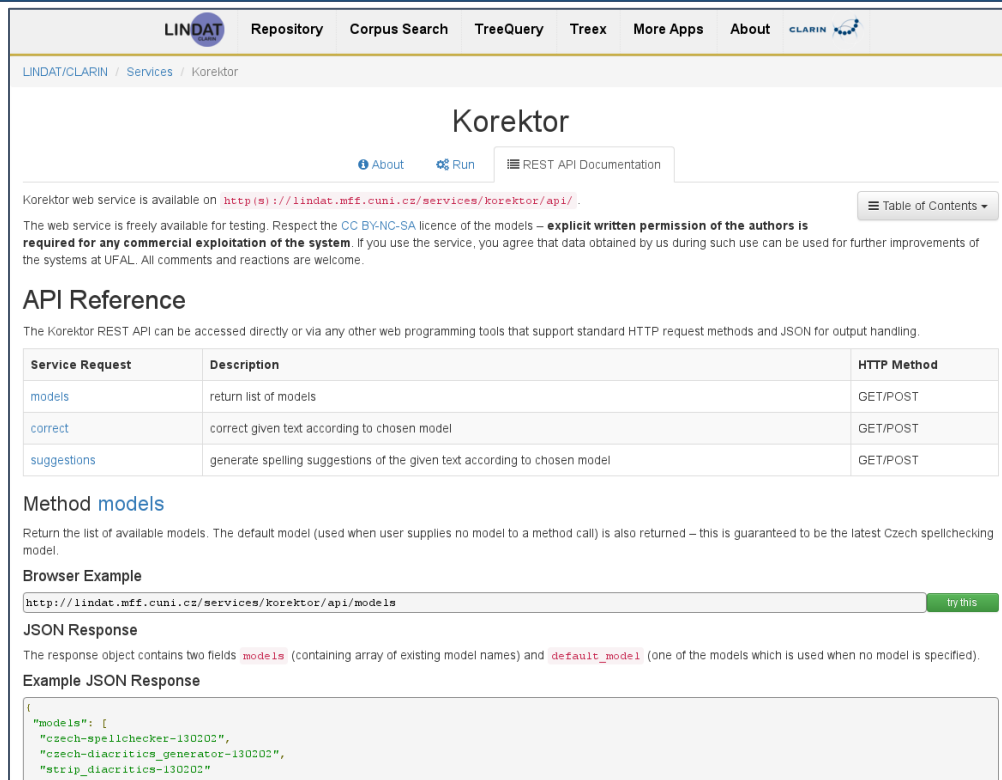
<http://ufal.mff.cuni.cz/korektor>

Získáme všechny věty (řádky s obsahem <text>), značku <text> odstraníme:

```
grep "<text>" CzechABSA-v2_format_SemEval2016.xml | sed "s/<text>//;s/<\/text>//" > texty.txt
```

```
Jídlo taky nic moc
Jídlo od Nás dostává 5 z 5 a pan Doksanský (spolumajitel) i jeho personál jsou profesionálové kteří Vám doporučí kvalitní víno nebo jiné pití k
e konkrétním jídlům
Vloni trochu na kvalitě polevili,ale nadále bezva
Kdo ví, třeba bych tam ještě po pozitivní konstruktivní kritice dostal přes ústa
Kdo Vám v dnešní době nabídne výběr z 5 skutečně fresh jídel, která mají i super chuť
Vsichni víme, že Bio maso je podstatně dražší než maso klasické, ale ceny zde jsou opravdu na úrovni z Bio chovu, není nic špatného?
Jistě se k vám brzy vrátím a všem vřele tuto restauraci doporučuji
Mají tu fantastické jídlo a příjemné prostředí!
Všichni byli sice milí, ale v pořadí vypadala hospůdka U Štěpána rozhodně lákavěji
Ku příkladu dnes 29.10.2013 nabízejí LOSOSO S GRILOVANOU ZELENINOU, RISSOTO a nesmím zapomenout na FRANFURSKOU POLÉVKU
Všechno bylo naprosto bez obtíží, obsluha super, jídlo bez vady
Když budu chtít zas někdy někoho pozvat, rozhodně ne sem
všechno je super ,tak se nenechte odradit
Vynikající obsluha pozorná vítající hned při příchodu
Rozsah jídelního lístku velký ,jídlo nám vždy chutnalo
RESTAURACI VŠEM DOPOPRUČUJI
Pro mě prostě vynikající steaky v příjemném prostředí, příště kupuji zase a moc se těším!
Pro změnu byl steak krvavý a nechutný
SKVĚLÁ KUČTIČKA, JÍDLO CHUTNÉ A PIVO??? DOBŘE CHLAZENÉ
S obsluhou jsme byli spokojeni
Poté přišel nejlepší vedoucí, jídlo odnesl a donesl nové
Prostředí nic moc
```

Korektor jako webová služba



The screenshot shows the web interface for the Korektor REST API. At the top, there is a navigation bar with the LINDAT logo and links for Repository, Corpus Search, TreeQuery, Treex, More Apps, and About. Below the navigation bar, the page title "Korektor" is displayed, along with tabs for "About", "Run", and "REST API Documentation". The main content area contains the following information:

Korektor web service is available on <https://lindat.mff.cuni.cz/services/korektor/api/>.

The web service is freely available for testing. Respect the [CC BY-NC-SA](#) licence of the models – **explicit written permission of the authors is required for any commercial exploitation of the system**. If you use the service, you agree that data obtained by us during such use can be used for further improvements of the systems at UFAL. All comments and reactions are welcome.

API Reference

The Korektor REST API can be accessed directly or via any other web programming tools that support standard HTTP request methods and JSON for output handling.

Service Request	Description	HTTP Method
models	return list of models	GET/POST
correct	correct given text according to chosen model	GET/POST
suggestions	generate spelling suggestions of the given text according to chosen model	GET/POST

Method [models](#)

Return the list of available models. The default model (used when user supplies no model to a method call) is also returned – this is guaranteed to be the latest Czech spellchecking model.

Browser Example

JSON Response

The response object contains two fields: `models` (containing array of existing model names) and `default_model` (one of the models which is used when no model is specified).

Example JSON Response

```
{
  "models": [
    "czech-spellchecker-130202",
    "czech-diacritics_generator-130202",
    "strip_diacritics-130202"
  ]
}
```

<https://lindat.mff.cuni.cz/services/korektor/api-reference.php>

Zjištění dostupných modelů

```
curl http://lindat.mff.cuni.cz/services/korektor/api/models
```


Zjištění dostupných modelů

```
curl http://lindat.mff.cuni.cz/services/korektor/api/models
```

```
gris@amethyst:~/ufal/work/prednasky_a_slajdy/2019_lindat_tutorial$ curl http://lindat.mff.cuni.cz/services/korektor/api/models
{"models": ["czech-spellchecker-130202", "czech-diacritics_generator-130202", "strip_diacritics-130202"], "default_model": "czech-spellchecker-130202"}
```

Zjištění dostupných modelů Korektoru

```
curl http://lindat.mff.cuni.cz/services/korektor/api/models
```

```
gris@ametyst:~/ufal/work/prednasky_a_slajdy/2019_lindat_tutorial$ curl http://lindat.mff.cuni.cz/services/korektor/api/models
{
  "models": [
    "czech-spellchecker-130202",
    "czech-diacritics_generator-130202",
    "strip_diacritics-130202"
  ],
  "default_model": "czech-spellchecker-130202"
}
```

Doplnění diakritiky pomocí Korektoru

```
curl -F 'data=@texty.txt' -F 'model=czech-diacritics_generator-130202'  
http://lindat.mff.cuni.cz/services/korektor/api/correct | PYTHONIOENCODING=utf-8 python -c  
"import sys,json; sys.stdout.write(json.load(sys.stdin)['result'])" > opravene_texty.txt
```

Doplnění diakritiky pomocí Korektoru

```
curl -F 'data=@texty.txt' -F 'model=czech-diacritics_generator-130202'  
http://lindat.mff.cuni.cz/services/korektor/api/correct | PYTHONIOENCODING=utf-8 python -c  
"import sys,json; sys.stdout.write(json.load(sys.stdin)['result'])" > opravene_texty.txt
```

Doplnění diakritiky pomocí Korektoru

```
curl -F 'data=@texty.txt' -F 'model=czech-diacritics_generator-130202'  
http://lindat.mff.cuni.cz/services/korektor/api/correct | PYTHONIOENCODING=utf-8 python -c  
"import sys,json; sys.stdout.write(json.load(sys.stdin)['result'])" > opravene_texty.txt
```

Jídlo od Nás dostává 5 z 5 a pan Doksanský (spolumajitel) i jeho personál jsou profesionálové kteří Vám doporučí kvalitní víno nedlům

Vloni trochu na kvalitě polevili,ale nadále bezva

Kdo ví, třeba bych tam ještě po pozitivní konstruktivní kritice dostal přes ústa

Kdo Vám v dnešní době nabídne výběr z 5 skutečně fresh jídel, která mají i super chuť

Všichni víme, že Bio maso je podstatně dražší než maso klasické, ale ceny zde jsou opravdu na úrovni Bio chovu, není něco špatně?

Jistě se k vám brzy vrátím a všem vřele tuto restauraci doporučuji

Mají tu fantastické jídlo a příjemné prostředí!

Všichni byli sice milí, ale v pořadí vypadala hospůdka U Štěpána rozhodně lákavěji

Ku příkladu dnes 29.10.2013 nabízejí LOSOSO S GRILOVANOU ZELENINOU, RISSOTO a nesmím zapomenout na FRANFURSKOU POLÉVKU

Všechno bylo naprosto bez obtíží, obsluha super, jídlo bez vady

Když budu chtít zas někdy někoho pozvat,rozhodně ne sem

všechno je super ,tak se nenechte odradit