LINDAT CLARIAH-CZ podzimní tutoriály

Národní galerie Praha - Veletržní palác
Dukelských hrdinů 47, Praha 7
Auditorium, 6. patro

tutoriál
17.10.2019
9-13 hod.

Praktická část

09:00-11:00 Nástroje pro zpracování textů
- Webové služby https://lindat.cz/services
- Lehký úvod do strojového učení
- Demonstrace spuštění vybraných nástrojů

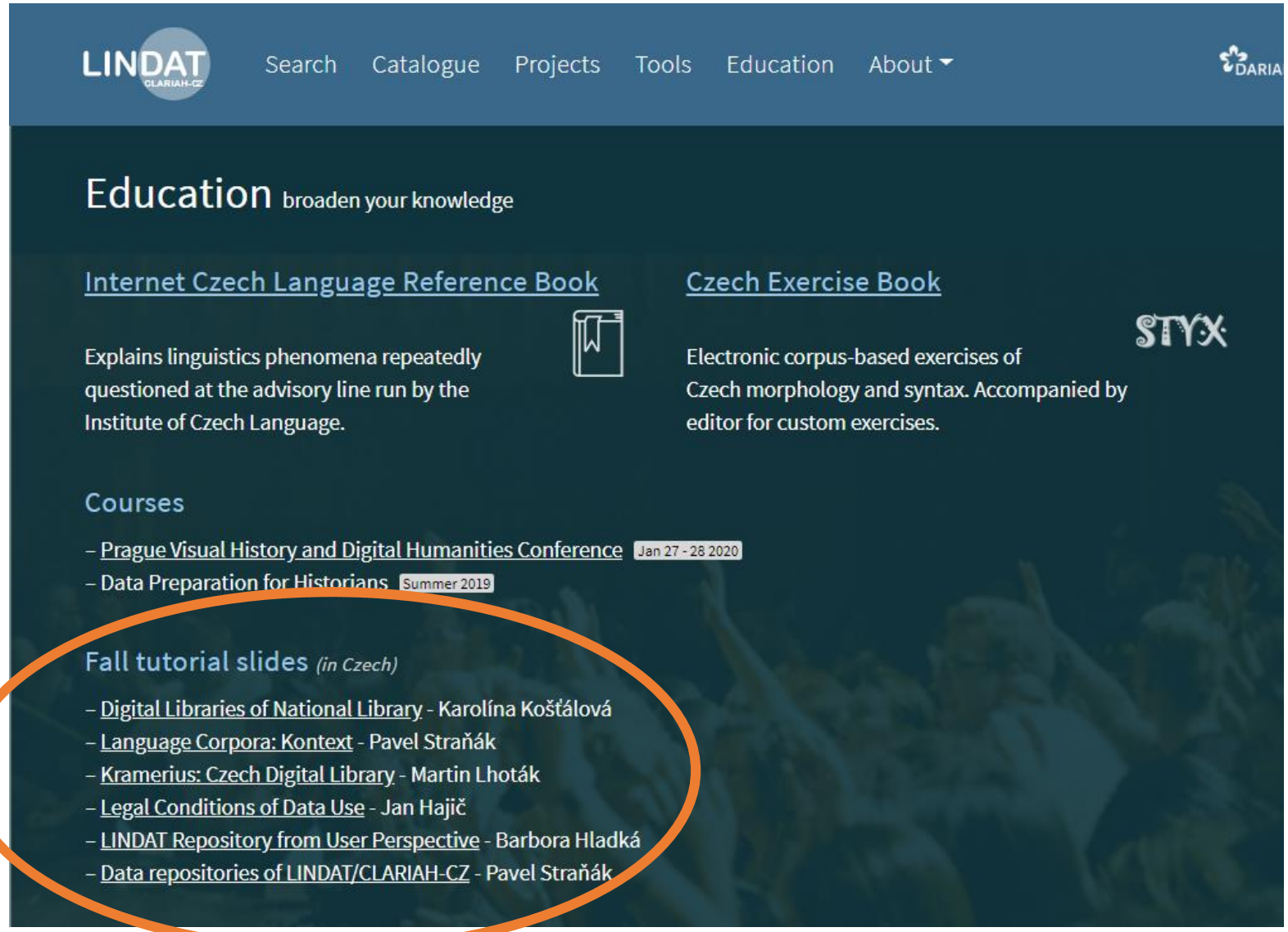
Prezentační část

11:30-12:00 Bibliografie dějin českých zemí
12:00-13:00 Diskuse

- Otevřeno všem zájemcům
- Registrace není nutná a účast je bezplatná
- Každý účastník tutoriálu obdrží dárek

Materiály z prvního podzimního tutoriálu 26/9/19

https://clariah.lindat.cz/#education



Nástroje pro zpracování textů

Barbora Hladká
ÚFAL MFF UK

LINDAT
CLARIAH-CZ

Nástroje pro zpracování textů

jsou automatické procedury, které provádí lingvistickou analýzu textu,

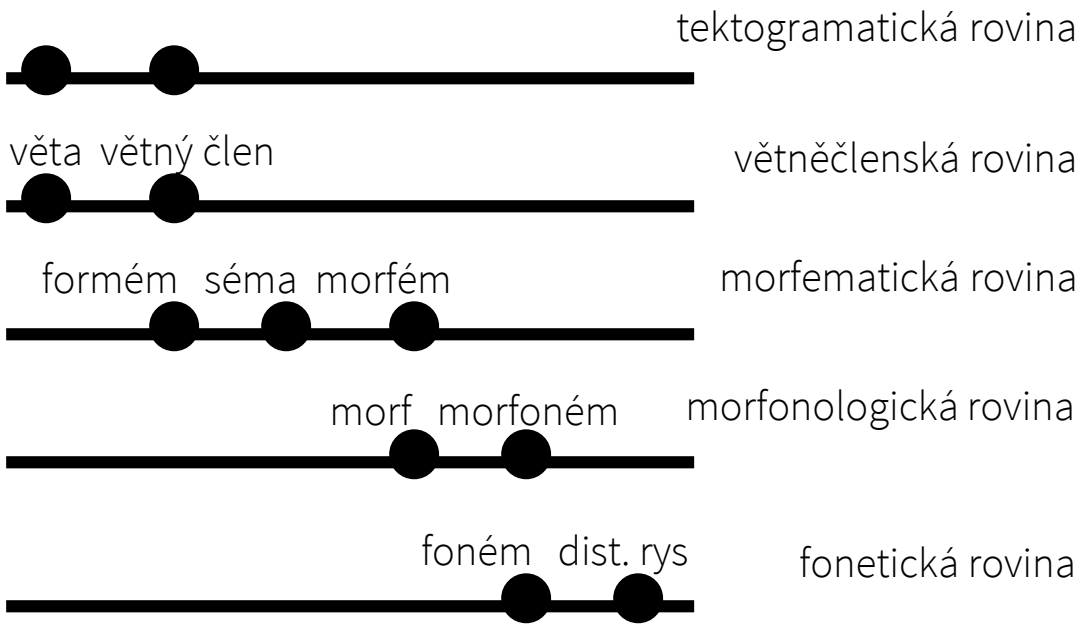
tj. obohacují text o jazykovědné informace.

Označujeme je též jako lingvistické nástroje.

Hierarchie nástrojů navázaná na formální popis jazyka

Zaměříme se na nástroje v prostředí http://lindat.cz, které navazují na Funkční generativní popis jazyka (Sgall, 1967)

- jazyk jako systém
- roviny systému
- jednotky rovin



Nástroje z pohledu uživatele

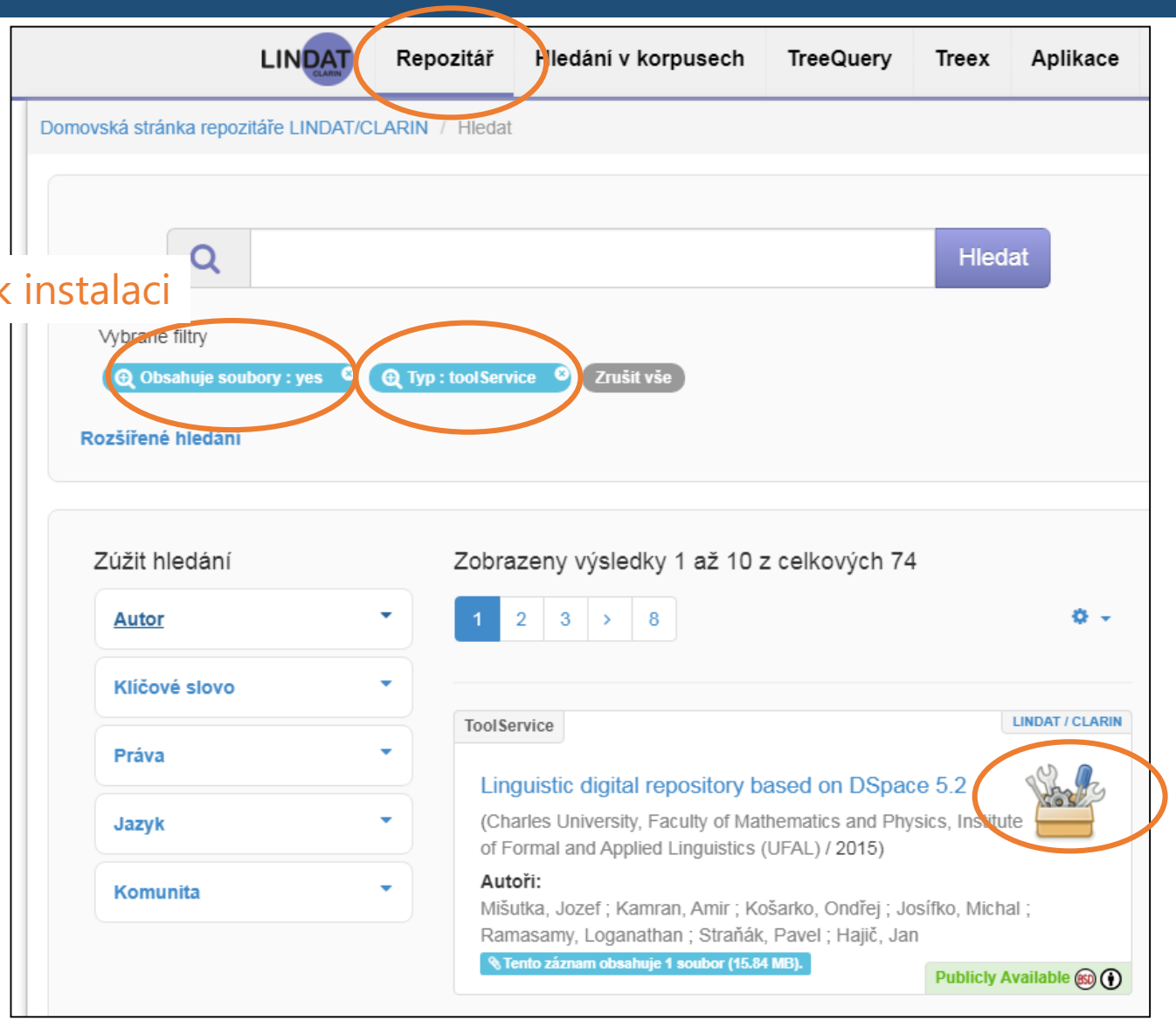
Co dělají a jak se používají

- vlastní instalace i v druhé části tutoriálu s Janou
- demo v prohlížeči
- webová služba

https://lindat.cz/repository
Nástroje k vlastní instalaci

datový repozitář,
i nástroje jsou data

k instalaci



https://lindat.cz/repository
Nástroje k vlastní instalaci

Ve výsledcích hledání jsou nejenom lingvistické nástroje, ale i aplikace, např.

- FairyTale Child (pid)

  chatbot simulující zvídavé dítě

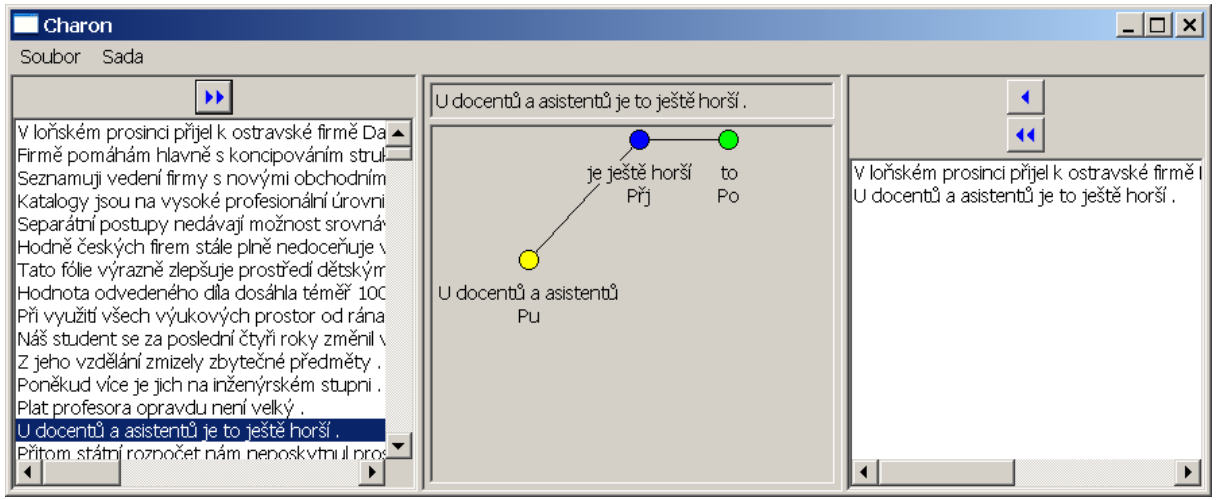
- STYX (pid)

  elektronická cvičebnice

  české morfologie a syntaxe

  s 11 tis. větami

Dítě	Povídej mi pohádku!
Uživatel	O červené karkulce
Dítě	Jaká to byla karkulka?
Uživatel	červená :)
Dítě	Byla hodně červená?
Uživatel	Jako jablíčko



https://lindat.cz/services

Dema a webové služby lingvistických nástrojů

LINDAT | Repozitář | Hledání v korpusech | TreeQuery | Treex | Aplikace | O n

Treex::Web - vysoce modulární NLP framework online

Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

Sedlák, Michal, 2014, *Treex::Web*, 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-0000-0023-44AF-C.

Share:

Treex::Web

Autoři:
Martin Popel, Zdeněk Žabokrtský, Jan Ptáček, Petr Pajas

Popis:
Treex (dříve TectoMT) je vysoce modulární NLP softwarový systém implementovaný pod Linuxem v programovacím jazyce Perl. Systém je zaměřený především na strojový překlad a využívá myšlenek a technologií vytvořených v rámci projektu Prague Dependency Treebank. Zároveň se předpokládá, že může významně usnadnit a urychlit vývoj softwarových řešení mnoha dalších úkolů NLP, a to zejména s ohledem na znovuvyužitelnost mnoha integrovaných procesních modulů (tzv. bloků), které jsou vybaveny jednotným objektově orientovaným rozhraním.

Stránka Projektu | Spustit

- Treex::Web
- Česílko
- PML-Tree Query
- PDT-Vallex
- EngVallex
- CzEngVallex
- MorphoDiTa
- NameTag
- ILRB
- ElixirFM
- Dialogy.Org
- Korektor
- KonText
- Parsito
- KER
- UDPipe
- EVALD 1.0
- EVALD 1.0 for Foreigners
- Machine translation
- UWebASR

Dema a webové služby
Záznamy

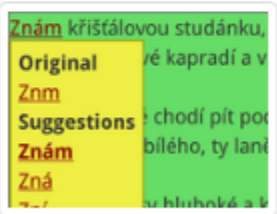
Korektor

! citace

Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

Straka, Milan and Richter, Michal, 2015, *Korektor 2*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-1469.

Share:



Autoři:
Michal Richter, Pavel Straňák, Milan Straka

Popis:
Korektor je statistický nástroj pro kontrolu pravopisu a (částečně) gramatiky. Tento nástroj vznikl jako součást diplomové práce Michala Richtera Pokročilý korektor češtiny, ale je dále rozvíjen. Existují dvě verze: unixová command-line utilita (testováno na Debian, Ubuntu a OS X) a OS X SpellServer se systémovou službou, která je integrována do nativních OS X GUI aplikací.

Stránka Projektu | Spustit

- Treex::Web
- Česílko
- PML-Tree Query
- PDT-Vallex
- EngVallex
- CzEngVallex
- MorphoDiTa
- NameTag
- ILRB
- ElixirFM
- Dialogy.Org
- Korektor
- KonText
- Parsito
- KER
- UDPipe
- EVALD 1.0
- EVALD 1.0 for Foreigners
- Machine translation
- UWebASR

Dema a webové služby
Záznamy

LINDAT/CLARIN / Services / Korektor

Korektor

About | Run | REST API Documentation

Korektor is a statistical spellchecker and (occasional) grammar checker released under 2-Clause BSD license and versioned using Semantic Versioning.

Korektor started with Michal Richter's diploma thesis Advanced Czech Spellchecker, but it is being developed further. There are two versions: a *command line utility* (tested on Linux, Windows and OS X) and a *REST service* with publicly available API and HTML front end.

The original *OS X SpellServer* providing System Service integrating Korektor with native OS X GUI applications is no longer developed, but do not hesitate to contact us if you are interested in it.

Copyright 2015 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.

Basic info

Authors	Michal Richter, Milan Straka, Pavel Straňák
Homepage	http://ufal.mff.cuni.cz/korektor
Development repository	https://github.com/ufal/korektor.git
Status	Stable, version 2.0.0
OS	Linux, Windows, OS X
License of the library	BSD 2-Clause
License of the models	CC BY-NC-SA
Contact	straka@ufal.mff.cuni.cz

Dema a webové služby
Záznamy

LINDAT/CLARIN / Services / Korektor

uživatelsky velmi pohodlné

Korektor

About | Run | REST API Documentation

Korektor is a statistical spellchecker and (occasional) grammar checker released under 2-Clause BSD license and versioned using Semantic Versioning.

Korektor started with Michal Richter's diploma thesis Advanced Czech Spellchecker, but it is being developed further. There are two versions: a *command line utility* (tested on Linux, Windows and OS X) and a *REST service* with publicly available API and HTML front end.

The original *OS X SpellServer* providing System Service integrating Korektor with native OS X GUI applications is no longer developed, but do not hesitate to contact us if you are interested in it.

Copyright 2015 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.

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. If you perform corrections to the output (either by choosing other suggestions or by manually correcting the text), please use the **Submit corrected text** button to send the corrected text to us. All comments and reactions are welcome.

Model: Czech

czech-130202

Task: Spellcheck | Generate Diacritics | Strip Diacritics

Přílyš žluťoučky kůň úpěl ďábelské ódi.

uživatel vloží text k opravě

Dema a webové služby
Záznamy

LINDAT/CLARIN / Services / Korektor

Korektor

s Janou

About | Run | REST API Documentation

Korektor web service is available on `http(s)://lindat.mff.cuni.cz/services/korektor/api/`.

Table of Contents

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.

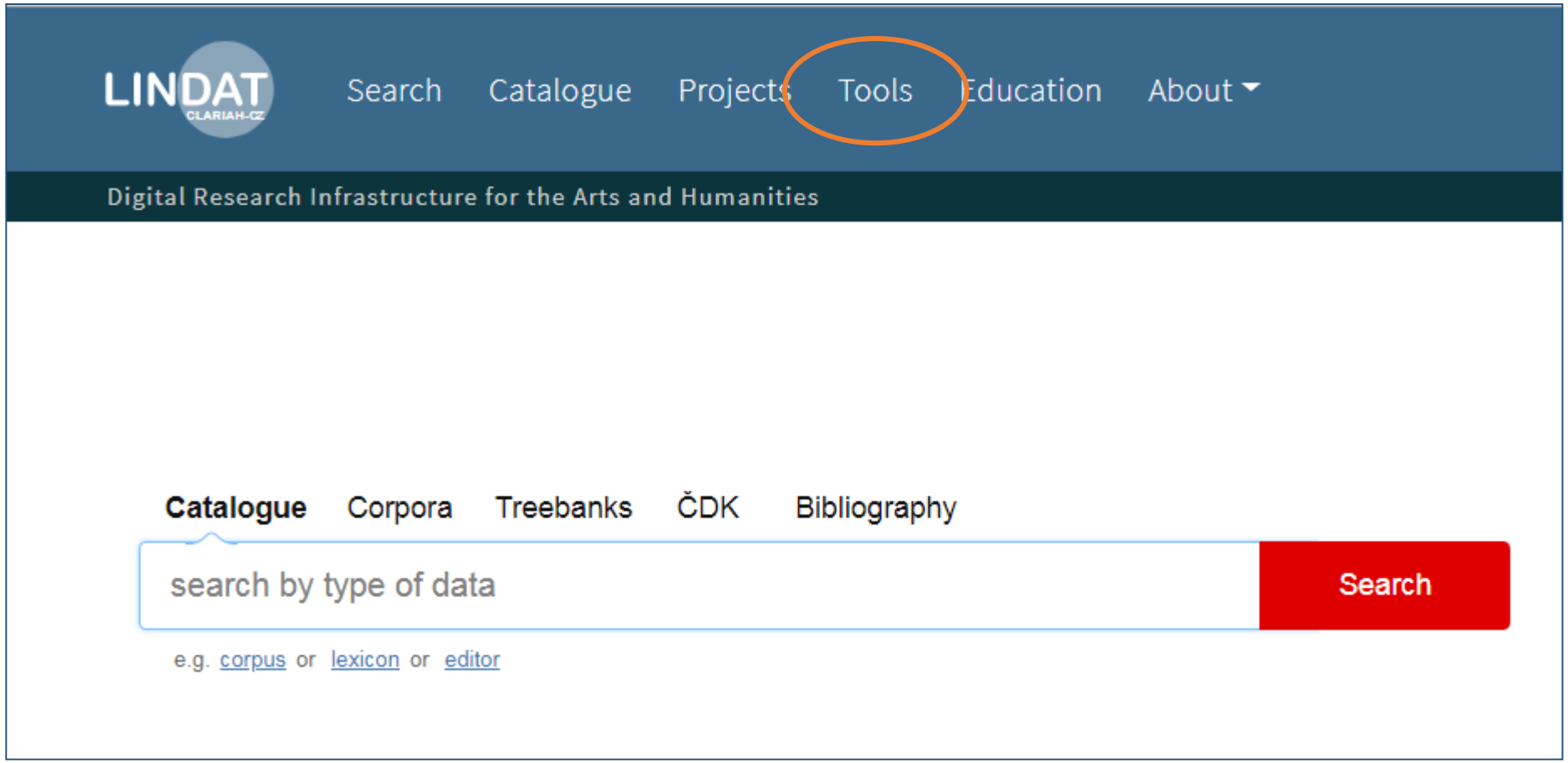
API Reference

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

Zpět k hierarchii nástrojů

https://clariah.lindat.cz



Zpět k hierarchii nástrojů https://lindat.cz/services https://clariah.lindat.cz

LINDAT CLARIAH-CZ

Tools for Natural Language Processing

Terms of use
Quality monitoring

Enrich your texts
Make them more searchable

Phonetics, Phonology

- UWebASR (audio transcription, Czech)

Morphology and tagging

- MorphoDiTa (Czech)
- ElixirFM (Arabic)
- UDPipe

Syntactic parsing

- Parsito
- UDPipe
- Korektor (spellchecker, Czech)

Discourse, Pragmatics

- NameTag (named-entity recognition, Czech, English)
- KER (keyword extraction, Czech, English)
- Evald (coherence evaluation, native speakers of Czech)
- Evald (coherence evaluation, non-native speakers of Czech)

Sentiment Analysis

Machine Translation

- Transformer (Czech, English, French)
- Česílko (Czech, Slovak)

Natural language processing

- Treex (Czech, English, Latin)

Search

- Kontext (concordances, collocations, word frequencies)
- PML Tree Query search (treebank search)
- PDT-Vallex (lexicon, Czech)
- EngVallex (lexicon, English)
- CzEngVallex (lexicon, Czech, English)
- Internet Language Reference Book (Czech)
- Dialogy.org (audio-visual corpora search)

Treex::Web
Česílko
PML-Tree Query
PDT-Vallex
EngVallex
CzEngVallex
MorphoDiTa
NameTag
ILRB
ElixirFM
Dialogy.Org
Korektor
KonText
Parsito
KER
UDPipe
EVALD 1.0
EVALD 1.0 for Foreigners
Machine translation
UWebASR

Ukázky dem

Tools for Natural Language Processing

Terms of use
Quality monitoring

Enrich your texts
Make them more searchable

Phonetics, Phonology

- UWebASR (audio transcription, Czech)

Morphology and tagging

- MorphoDiTa (Czech)
- ElixirFM (Arabic)
- UDPipe

Syntactic parsing

- Parsito
- UDPipe
- Korektor (spellchecker, Czech)

Discourse, Pragmatics

s Janou

- NameTag (named-entity recognition, Czech, English)
- KER (keyword extraction, Czech, English)
- Evald (coherence evaluation, native speakers of Czech)
- Evald (coherence evaluation, non-native speakers of Czech)

Sentiment Analysis

Machine Translation

- Transformer (Czech, English, French)
- Česílko (Czech, Slovak)

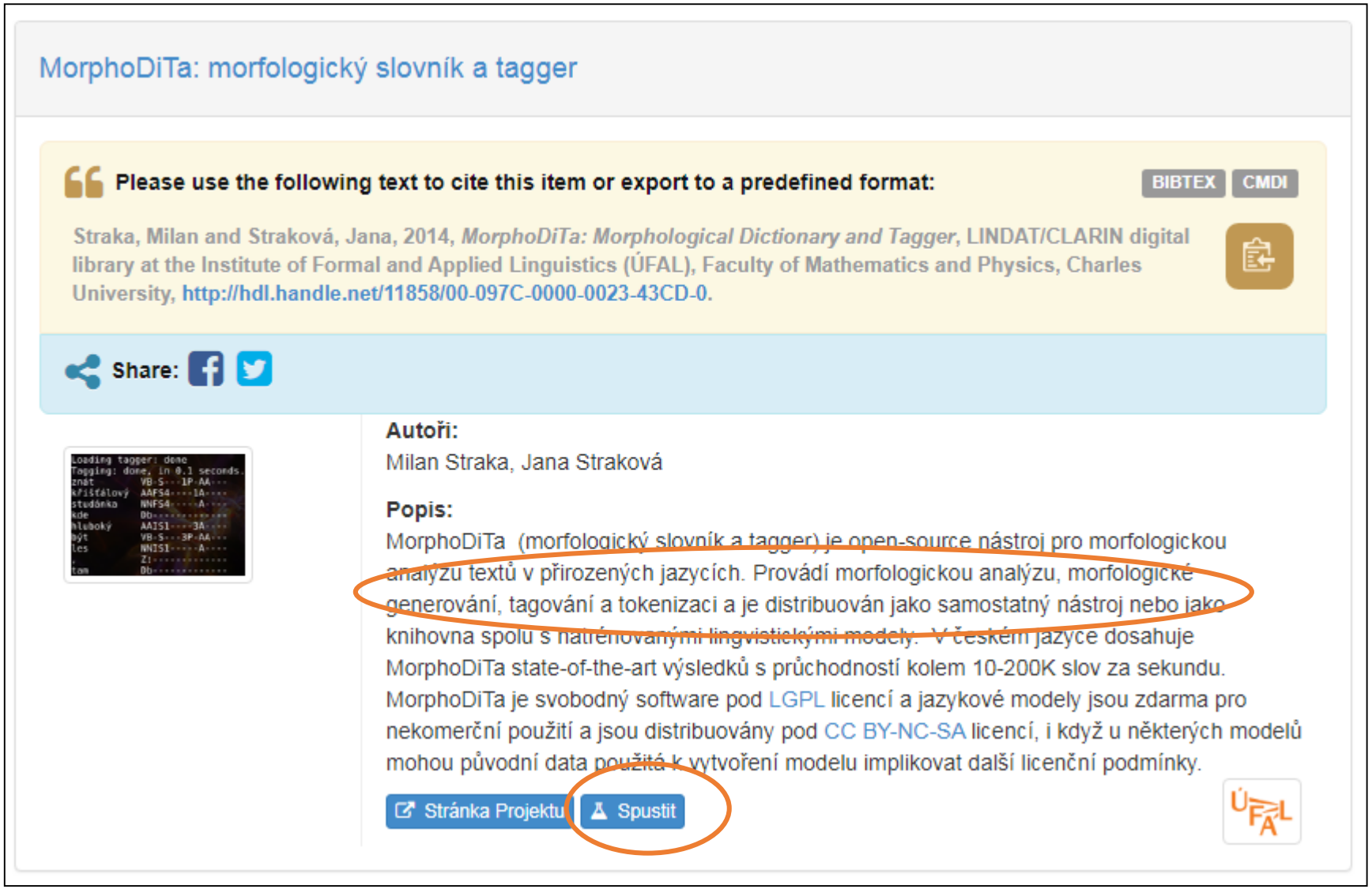
Natural language processing

- Treex (Czech, English, Latin)

Search

- Kontext (concordances, collocations, word frequencies)
- PML Tree Query search (treebank search)
- PDT-Vallex (lexicon, Czech)
- EngVallex (lexicon, English)
- CzEngVallex (lexicon, Czech, English)
- Internet Language Reference Book (Czech)
- Dialogy.org (audio-visual corpora search)

MorphoDita (pid) -- demo



- tokenizace
  je segmentace textu na slovní jednotky
- lematizace
  je převedení slova do základního tvaru
- tagování
  je zjednoznačnění morfologické analýzy slov v textu

MorphoDiTa (pid) -- demo

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.

viz dále lehký úvod do strojového učení

Model: Czech English Slovak

czech-morfflex-pdt-161115

Task: Tag Lemmatize Analyze Generate Tokenize

Use morphological guesser for unknown words

Input: Plain text Vertical (word per line)

Tag set: Original CoNLL2009 Strip lemma comment Raw lemmas

Derivation: No morphological derivation Replace lemma by root Append path to root Append whole derivation tree

Output: Formatted XML (format description) Vertical (format description)

A Input Text | Input File

1 Satoranský řádil v přípravě v Torontu. Chicago zvítězilo na palubovce úřadujícího mistra NBA

uživatel vloží text ke zpracování

MorphoDiTa(pid) -- demo

2 Process Input

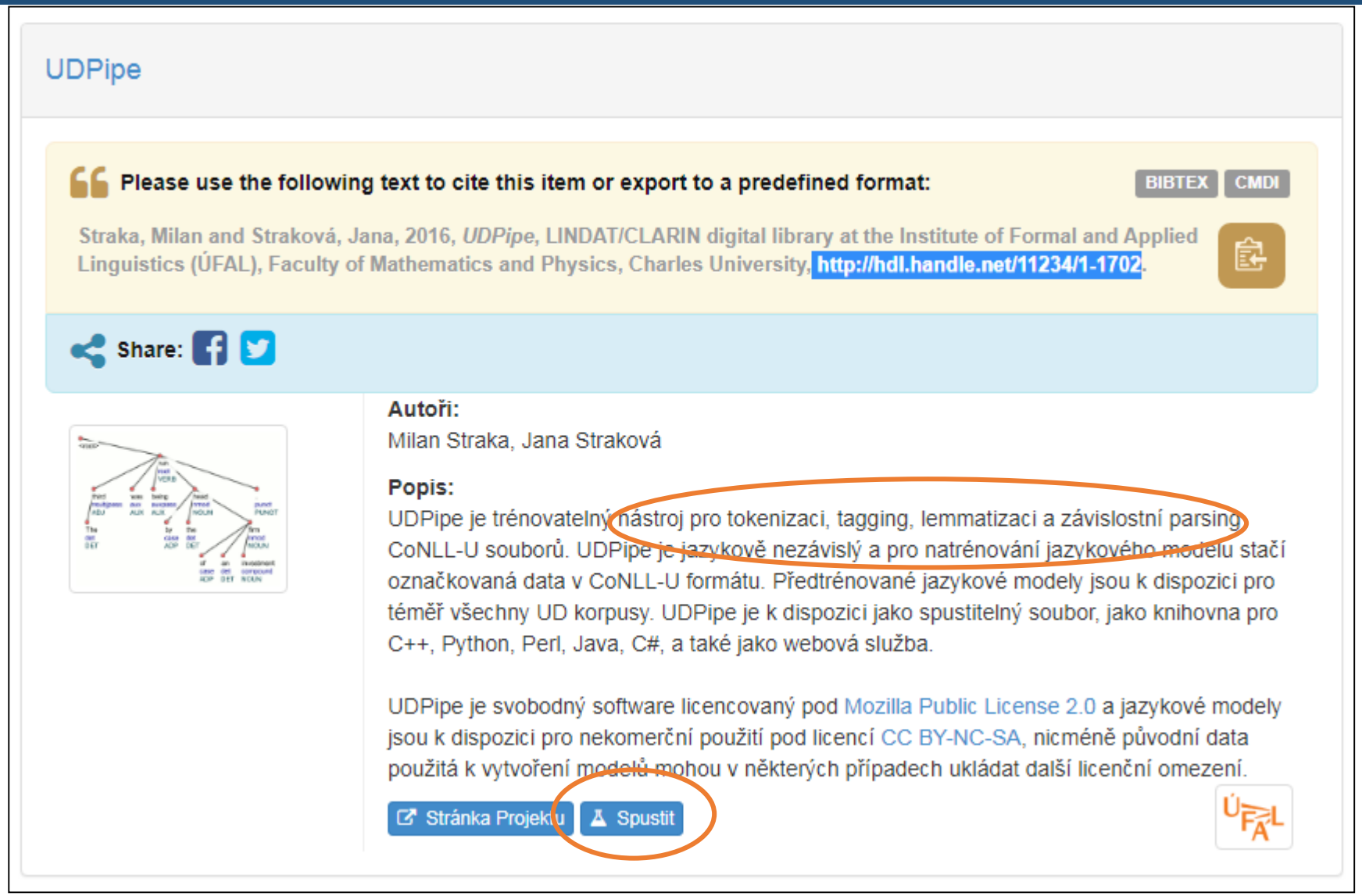
Output 3

Save Output File

Token	Lemma	Tag
Satoranský	Satoranský_;S	NNMS1-----A----
řádil	řádit_:T	VpYS---XR-AA---
v	v-1	RR--6----------
přípravě	příprava	NNFS6-----A----
v	v-1	RR--6----------
Torontu	Toronto_;G	NNNS6-----A----
.	.	Z:----------
Chicago	Chicago_;G	NNNS1-----A----
zvítězilo	zvítězit_:W	VpNS---XR-AA---
na	na-1	RR--6----------
palubovce	palubovka	NNFS6-----A----
úřadujícího	úřadující_^(*5ovat)	AGMS2-----A----
mistra	mistr	NNMS2-----A----
NBA	NBA_:B_;K	NNFXX-----A---8

15-ti poziční morfologické značky

UDPipe (pid) -- demo



UDPipe

Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

Straka, Milan and Straková, Jana, 2016, *UDPipe*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-1702.

Share:

Autoři:
Milan Straka, Jana Straková

Popis:
UDPipe je trénovatelný nástroj pro tokenizaci, tagging, lemmatizaci a závislostní parsing CoNLL-U souborů. UDPipe je jazykově nezávislý a pro natrénování jazykového modelu stačí označkovaná data v CoNLL-U formátu. Předtrénované jazykové modely jsou k dispozici pro téměř všechny UD korpusy. UDPipe je k dispozici jako spustitelný soubor, jako knihovna pro C++, Python, Perl, Java, C#, a také jako webová služba.

UDPipe je svobodný software licencovaný pod Mozilla Public License 2.0 a jazykové modely jsou k dispozici pro nekomerční použití pod licencí CC BY-NC-SA, nicméně původní data použitá k vytvoření modelů mohou v některých případech ukládat další licenční omezení.

Stránka Projektu | Spustit

- tokenizace
  je segmentace textu na slovní jednotky
- lematizace
  je převedení slova do základního tvaru
- tagování
  je zjednoznačnění morfologické analýzy slov v textu
- parsování je syntaktická analýza vět
- Universal Dependencies je způsob reprezentace syntaktické analýzy vět

UDPipe(pid) -- demo

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: UD 2.4 (description) UD 2.3 (description) UD 2.0 (description) UD 1.2 (description)

czech-pdt-ud-2.4-190531

Actions: Tag and Lemmatize Parse

Advanced Options

Input Text | Input File

1 Satoranský řádil v přípravě v Torontu. Chicago zvítězilo na palubovce úřadujícího mistra NBA.

uživatel vloží text ke zpracování

UDPipe(pid) -- demo

2 Process Input

3 Output Text | Show Table | Show Trees

Save Output File

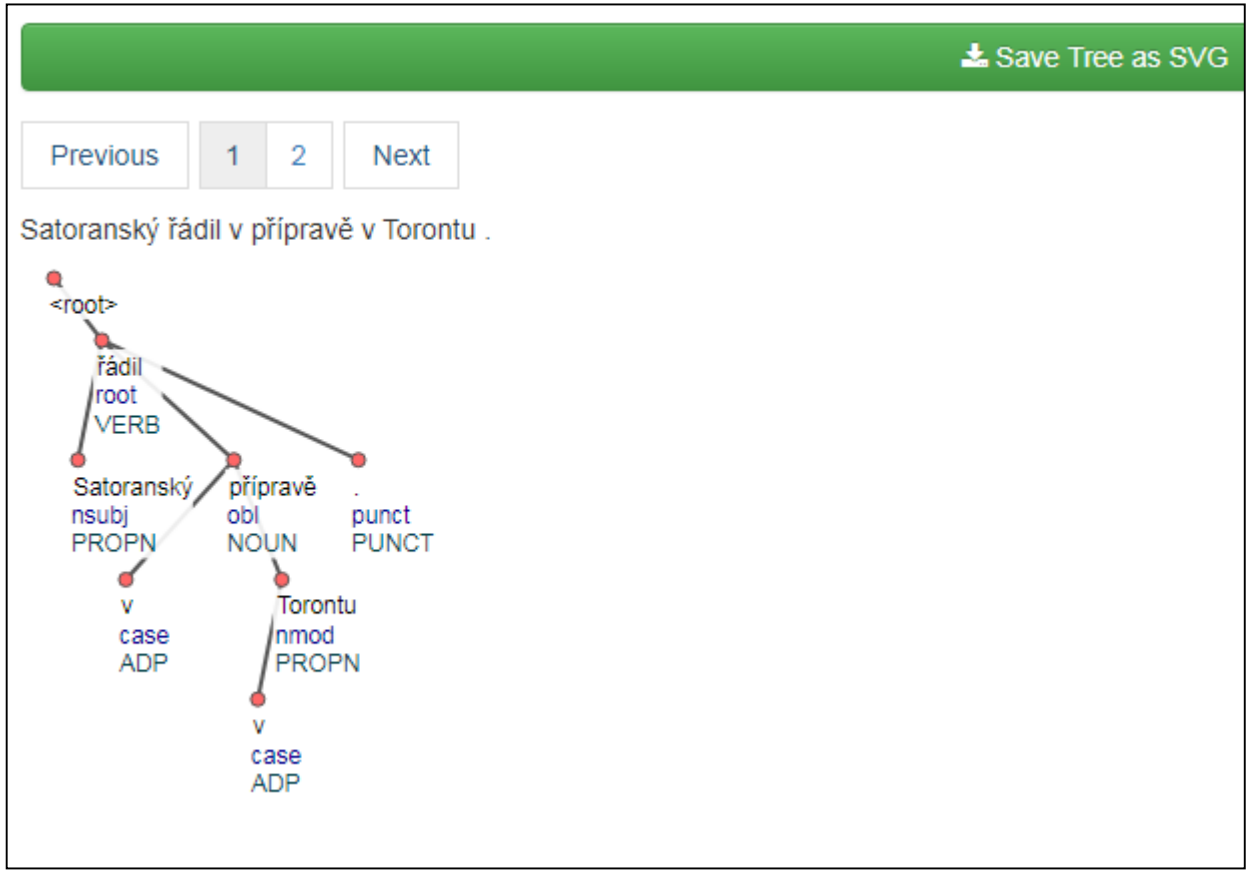
textový výstup (tzv. CoNLL-U format)

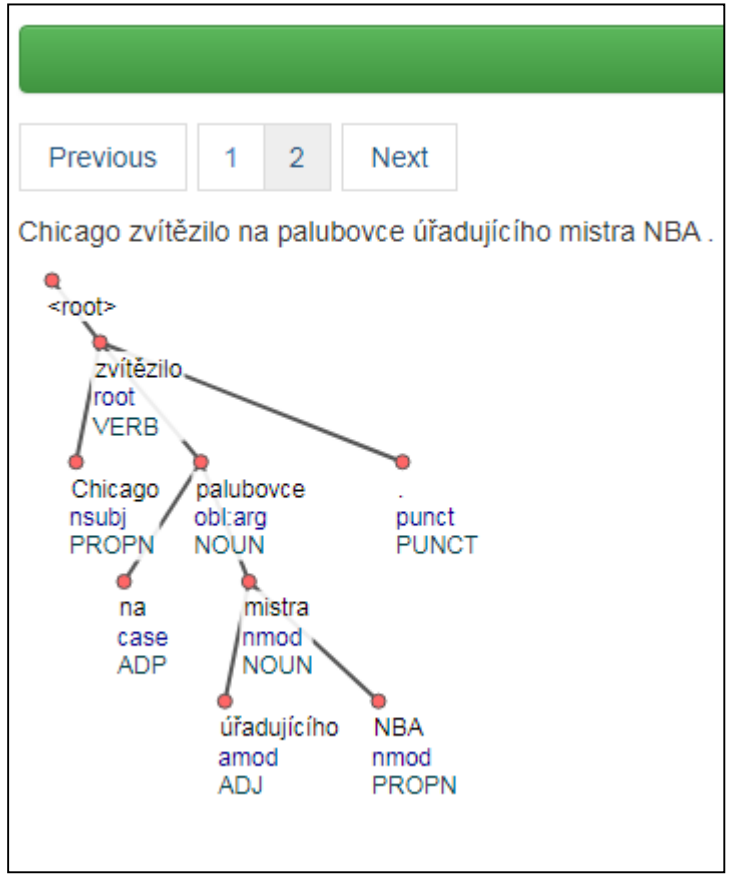
```
# newdoc
# newpar
# sent_id = 1
# text = Satoranský řádil v přípravě v Torontu.
1	Satoranský	Satoranský	PROPN	NNMS1-----A----	Animacy=Anim|Case=Nom|Gender=Masc|NameType=Sur|Number=Sing|Polarity=Pos	2	nsubj	_	_
2	řádil	řádit	VERB	VpYS---XR-AA---	Aspect=Imp|Gender=Masc|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act	0	root	_	_
3	v	v	ADP	RR--6----------	AdpType=Prep|Case=Loc	4	case	_	_
4	přípravě	příprava	NOUN	NNFS6-----A----	Case=Loc|Gender=Fem|Number=Sing|Polarity=Pos	2	obl	_	_
5	v	v	ADP	RR--6----------	AdpType=Prep|Case=Loc	6	case	_	_
6	Torontu	Toronto	PROPN	NNNS6-----A----	Case=Loc|Gender=Neut|NameType=Geo|Number=Sing|Polarity=Pos	4	nmod	_	SpaceAfter=No
7	.	.	PUNCT	Z:----------	_	2	punct	_	_

# sent_id = 2
# text = Chicago zvítězilo na palubovce úřadujícího mistra NBA.
1	Chicago	Chicago	PROPN	NNNS1-----A----	Case=Nom|Gender=Neut|NameType=Geo|Number=Sing|Polarity=Pos	2	nsubj	_	_
2	zvítězilo	zvítězit	VERB	VpNS---XR-AA---	Aspect=Perf|Gender=Neut|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act	0	root	_	_
3	na	na	ADP	RR--4----------	AdpType=Prep|Case=Acc	4	case	_	_
4	palubovce	palubovka	NOUN	NNMS4-----A----	Animacy=Anim|Case=Acc|Gender=Masc|Number=Sing|Polarity=Pos	2	obl:arg	_	_
5	úřadujícího	úřadující	ADJ	AGMS4-----A----	Animacy=Anim|Aspect=Imp|Case=Acc|Gender=Masc|Number=Sing|Polarity=Pos|Tense=Pres|VerbForm=Part|Voice=Act	6	amod	_	_
6	mistra	mistr	NOUN	NNMS4-----A----	Animacy=Anim|Case=Acc|Gender=Masc|Number=Sing|Polarity=Pos	4	nmod	_	_
7	NBA	NBA	PROPN	NNFXX-----A---8	Abbr=Yes|Foreign=Yes|Gender=Fem|NameType=Com|Polarity=Pos	6	nmod	_	SpaceAfter=No
8	.	.	PUNCT	Z:----------	_	2	punct	_	SpaceAfter=No
```

UDPipe(pid) -- demo

grafický výstup





https://lindat.mff.cuni.cz/services/translation/

LINDAT CLARIAH-CZ

LINDAT Translation

Success

Translate | Docs

You'll be shown the status (started/queued) of your translation task above this panel.

Models

cs-en (Czech->English (CUBBITT))

Source

cs

Target

en

advanced

Input sentences

Tutoriály jsou primárně určeny výzkumníkům humanitních, kulturních a kreativních oborů, odborníkům v jazykových technologiích a studentům napříč těmito obory. Tutoriály budou přehlídkou, inspirací a motivací zároveň, a proto mohou upoutat pozornost odborníků i z jiných oblastí.

Translation

Tutorials are primarily for researchers in the humanities, cultural and creative fields, professionals in language technology and students across these disciplines. Tutorials will be a show, inspiration and motivation at the same time, and can therefore attract the attention of experts from other areas.

Translate | Choose file

Credits: The service runs systems trained by Martin Popel (en-cs; cs-en; en-fr; fr-en), Shantipriya Parida (en-hi) and Dušan Variš (en-ru, ru-en, en-de, de-en).

Introducing Martin Popel

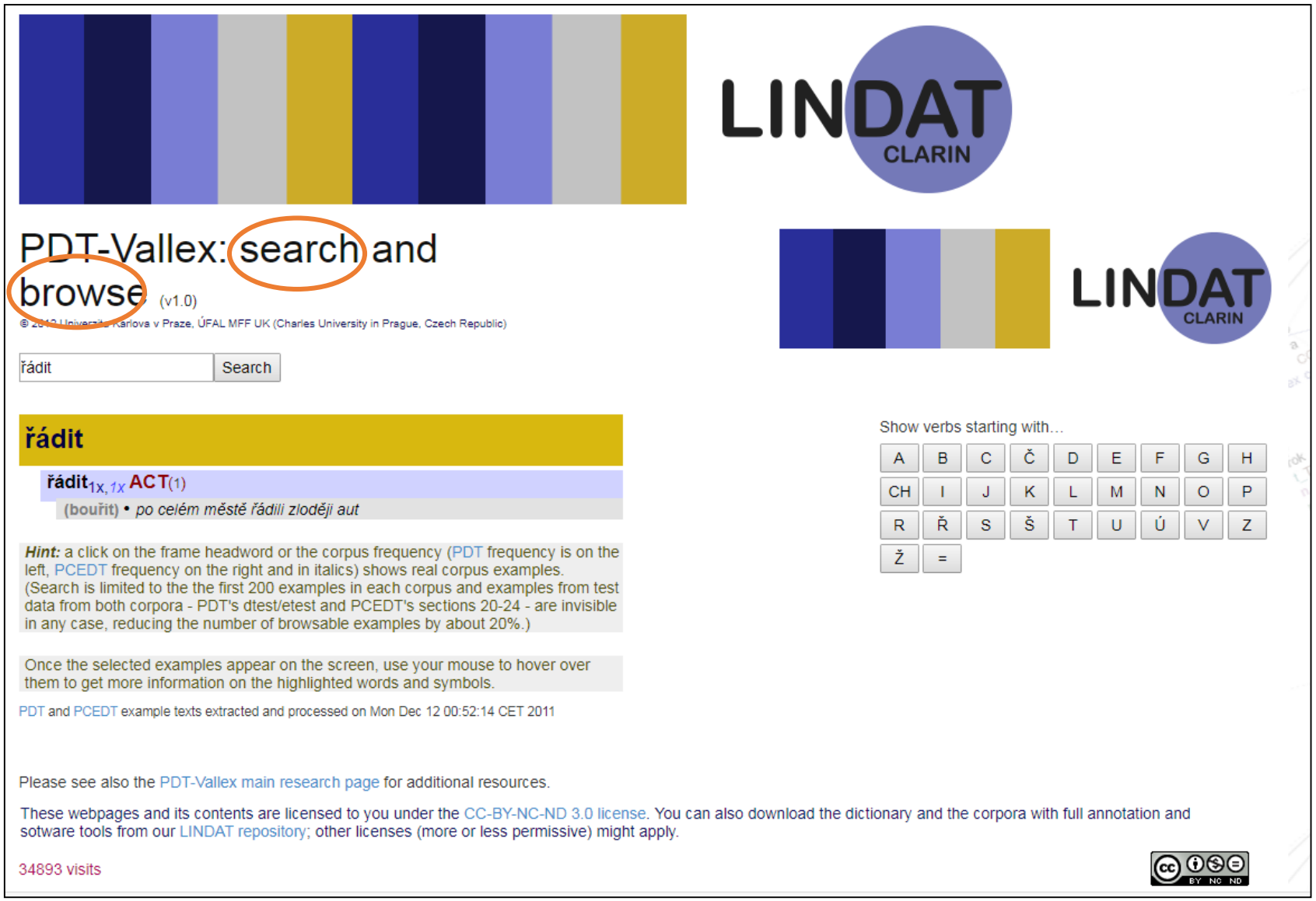
I developed a neural MT system CUNI-Transformer, which won in the English-Czech WMT2018 shared task. I also work on Universal Dependencies using Udapi.

I teach Language Data Resources, Modern Methods in Computational Linguistics and Natural language processing on computational cluster.

I am the technical editor of the PBML journal.

PDTVallex (pid): demo

valenční slovník s webovým rozhraním



https://lindat.cz/services

Tools for Natural Language Processing

Terms of use
Quality monitoring

Enrich your texts
Make them more searchable

Phonetics, Phonology

- UWebASR (audio transcription, Czech)

Morphology and tagging

- MorphoDiTa (Czech)
- ElixirFM (Arabic)
- UDPipe

Syntactic parsing

- Parsito
- UDPipe
- Korektor (spellchecker, Czech)

Discourse, Pragmatics

- NameTag (named-entity recognition, Czech, English)
- KER (keyword extraction, Czech, English)
- Evald (coherence evaluation, native speakers of Czech)
- Evald (coherence evaluation, non-native speakers of Czech)

Sentiment Analysis

Machine Translation

- Transformer (Czech, English, French)
- Česílko (Czech, Slovak)

Natural language processing

- Treex (Czech, English, Latin)

Search

- Kontext (concordances, collocations, word frequencies)
- PML Tree Query search (treebank search)
- PDT-Vallex (lexicon, Czech)
- EngVallex (lexicon, English)
- CzEngVallex (lexicon, Czech, English)
- Internet Language Reference Book (Czech)
- Dialogy.org (audio-visual corpora search)

LINDAT CLARIAH-CZ podzimní tutoriály

Národní galerie Praha - Veletržní palác
Dukelských hrdinů 47, Praha 7
Auditorium, 6. patro

tutoriál
17.10.2019
9-13 hod.

Praktická část

09:00-11:00 Nástroje pro zpracování textů
- Webové služby https://lindat.cz/services
- Lehký úvod do strojového učení
- Demonstrace spuštění vybraných nástrojů

Prezentační část

11:30-12:00 Bibliografie dějin českých zemí
12:00-13:00 Diskuse

- Otevřeno všem zájemcům
- Registrace není nutná a účast je bezplatná
- Každý účastník tutoriálu obdrží dárek

Výzkumná infrastruktura LINDAT/CLARIAH-CZ

(LM2018101 a CZ.02.1.01/0.0/0.0/16_013/0001781; původně LM2010013, LM2015071)

je financována Ministerstvem školství, mládeže a tělovýchovy České republiky

v programu LM „Velké infrastruktury".