

LANGUAGE RESEARCH INFRASTRUCTURE IN CZECHIA

LINDAT/CLARIN

PAVEL STRAŇÁK



MINISTRY OF EDUCATION
YOUTH AND SPORTS

LINDAT/CLARIN

LINDAT/CLARIN

LINDAT/CLARIN

LINgUistic

LINDAT/CLARIN

LINGuistic

DATa ... very broadly

LINDAT/CLARIN

LINgustic

DATa

/

CLARIN

LINDAT/CLARIN

LINguistic

DATa

/

Common

Language

Research and technology

INfrastructure

LINDAT/CLARIN

- Czech national project; node of CLARIN ERIC
- Operational since 2014
- Users:
 - Researchers in SSH and Computational Linguistics
- Technology:
 - Repository (resources), Services, Applications
- Knowledge, Support and Training

LANGUAGE TECHNOLOGY

- Natural Language Processing - NLP
 - Analysis, synthesis of spoken and written language
 - Machine Translation, Information Extraction, ...
 - Search in texts, audio, video, images
- State-of-the-art technology in NLP
 - “Statistical” methods:
 - Machine learning incl. neural networks
 - Need for (large) Language Resources – Texts, multimodal
 - Repositories, identification, replication of experiments, standards

USERS

- Everyone
 - communicates in and works with natural language!
- ... immediate users of the infrastructure:
 - Language Technology researchers
 - Universities, Research organisations
 - Need lots of data, easy to get, clean open licensing
 - “Content” users:
 - Linguists, historians, teachers, psychologists, sociologists, ...
 - Need identifiable data, preprocessed, searchable, easy-to-use services and applications
 - General public:
 - proper language use – IRLG; holocaust documentation: CVH Malach

Data Repository

PRESERVE AND FIND LANGUAGE DATA AND NLP TOOLS



Search

[Advanced Search](#)

Author	Subject	Language (ISO)
Hajič, Jan (47)	Germanistik (47)	English (222)
Žabokrtský, Zdeněk (32)	machine translation (39)	Czech (192)
Straka, Milan (29)	corpus (34)	German (159)
Zeman, Daniel (29)	treebank (30)	Dutch (92)
Bojar, Ondřej (28)	morphology (26)	Spanish (83)
... View More	... View More	... View More

What's New

ToolService

LINDAT / CLARIN

CorpusExplorer

Author(s):

Rüdiger, Jan Oliver

Description:

Software for corpus linguists and text/data mining enthusiasts. The CorpusExplorer combines over 45 interactive



? What can you do?

DEPOSIT



CITE



🎯 Browse

> All of the Repository

DATA REPOSITORY

- ~ 500 registered users
 - submitters & users signing licenses (not everything can be Open Access)
- 200+ Data Records
 - > 1000 Metadata Records
 - 80 languages
- 100 TB+ Data in Repository (+ 1PB of UCS Shoah Foundation Archive)

DATA REPOSITORY

- ~ 500 registered users
 - submitters & users signing licenses (not everything can be Open Access)
- 200+ Data Records
 - > 1000 Metadata Records
 - 80 languages
- 100 TB+ Data in Repository (+ 1PB of UCS Shoah Foundation Archive)

The screenshot displays the LINDAT/CLARIN Repository search interface. At the top, navigation links include "LINDAT/CLARIN", "Repository", "TreeQuery", "TreeX", and "More Apps". Below the navigation is a search bar with a magnifying glass icon and a "Search" button. A link for "Advanced Search" is located below the search bar.

The main content area is divided into two columns. The left column, titled "Limit your search", contains several filter dropdown menus: "Author", "Subject", "Rights", "Language (ISO)", "Type", "Contain Files", and "Community". The right column, titled "Showing 1 through 10 out of 1038 results", features a pagination control with buttons for "1", "2", "3", ">", and "104".

Below the filters, three search results are displayed, each with a category label in a grey box:

- Corpus**: "AKCES 2 ver. 2" (Charles University in Prague, ÚČJTK / 2013-12-18). Author(s): Šebesta, Karel ; Goláňová, Hana. This item contains 1 file (3.85 MB). Publicly Available (CC BY).
- LexicalConceptualResource**: "A Gold Standard Word Alignment for English-Swedish (2015-10-12)" (Linköping University / 2015-10-12). Author(s): Ahrenberg, Lars ; Holmqvist, Maria. This item contains 1 file (590 KB). Publicly Available (CC BY).
- ToolService**: "MorphoDiTa: Morphological Dictionary and Tagger" (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2014-02-14). Author(s): Straka, Milan ; Straková, Jana. This item contains no files.

DATA REPOSITORY

- Safe preservation (upload and don't worry)
- Discovery & Reuse
- Direct data citation (works in Google Scholar)
- Licensing (Open Access, but also more options)
- Versioning
- Language data and tools
- Worldwide (for everyone), easy to use

The screenshot shows the LINDAT/CLARIN Repository search results page. At the top, there are navigation links for "LINDAT/CLARIN Repository Home" and "Search". Below this is a search bar with a magnifying glass icon and a "Search" button. Underneath the search bar is a link for "Advanced Search".

The main content area is divided into two columns. The left column, titled "Limit your search", contains several dropdown menus for filtering results: "Author", "Subject", "Rights", "Language (ISO)", "Type", "Contain Files", and "Community".

The right column, titled "Showing 1 through 10 out of 1038 results", displays a list of search results. Each result is shown in a card format with a title, a brief description, the author(s), and a file size indicator. The first result is "AKCES 2 ver. 2" by Šebesta, Karel ; Goláňová, Hana, with a file size of 3.85 MB. The second result is "A Gold Standard Word Alignment for English-Swedish (2015-10-12)" by Ahrenberg, Lars ; Holmqvist, Maria, with a file size of 590 KB. The third result is "MorphoDiTa: Morphological Dictionary and Tagger" by Straka, Milan ; Straková, Jana, with no files. Each result card also includes a "Publicly Available" badge and a Creative Commons license icon.

UPLOAD AND DON'T WORRY

How to Deposit

Only authenticated users can deposit items. If you cannot find your home organisation in the Login dialog list of organisations then register at clarin.eu and authenticate using "clarin.eu website account". In case you cannot use any authentication method above or if you encounter a problem, do not hesitate to contact our [Help Desk](#) and we can create a local account for you.

Step 1: Login

To start a new submission you have to login first. Click Login under My Account in the right menu panel.

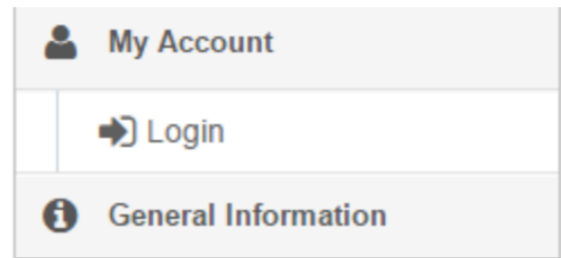
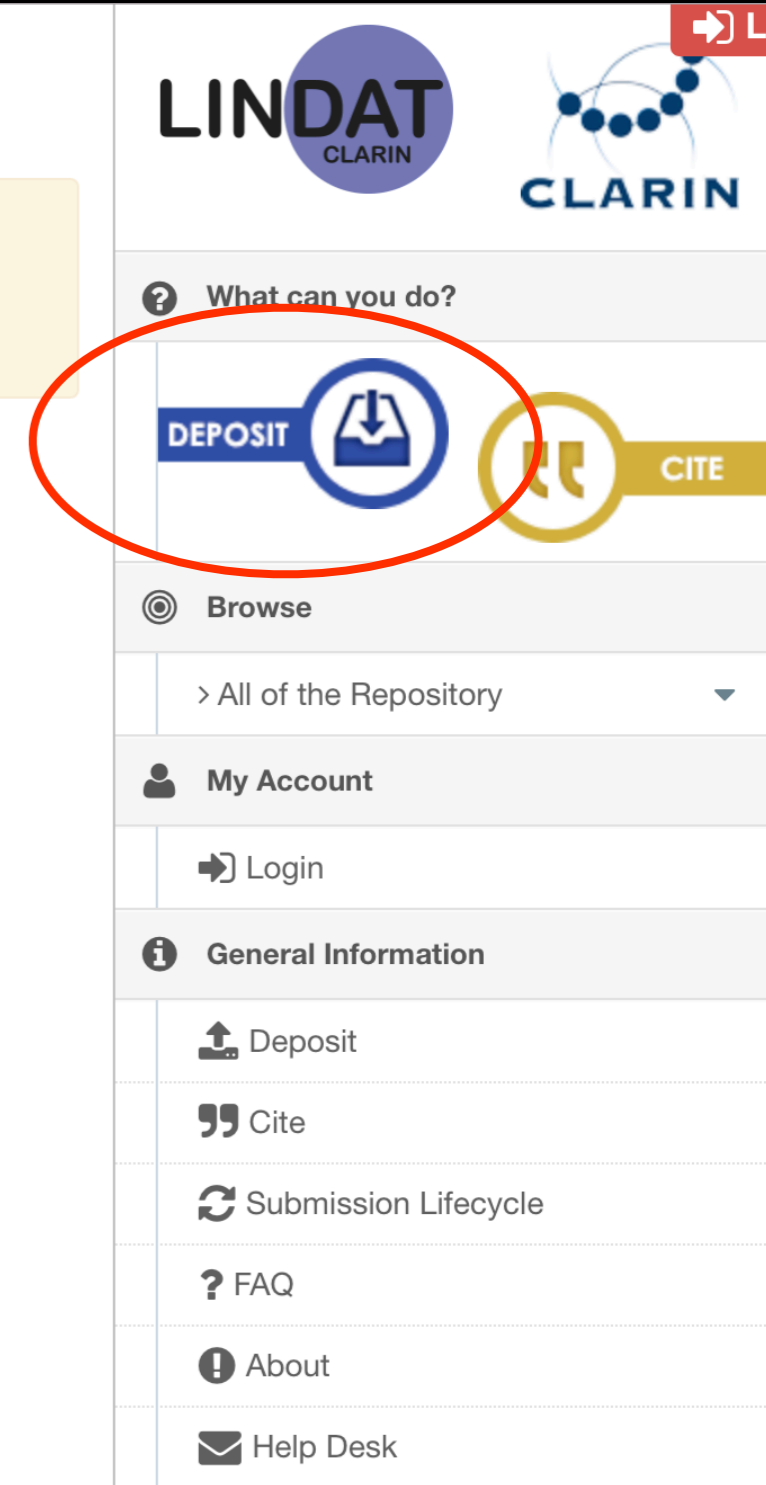
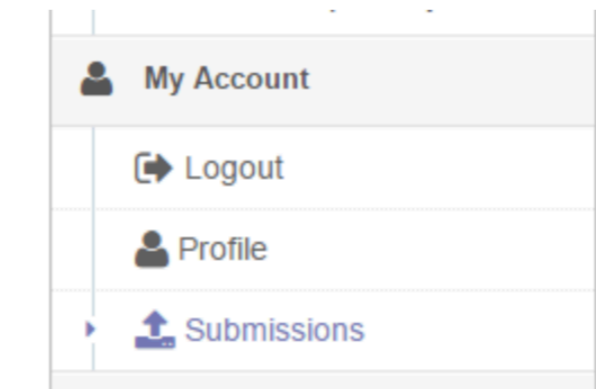



Fig1. Menu Login

Step 2: Starting a new submission

Now you have a new menu item 'Submissions' under My Account. Click on Submissions to go to the Submissions screen.



FACETED SEARCH



[Advanced Search](#)


Limit your search

- Author ▾
- Subject ▾
- Rights ▾
- Language (ISO) ▾
- Type ▾
- Contain Files ▾
- Community ▾

Showing 1 through 10 out of 1038 results

1 2 3 > 104 ⚙️ ▾


Corpus LINDAT / CLARIN

[AKCES 2 ver. 2](#) 


(Charles University in Prague, ÚČJTK / 2013-12-18)

Author(s):
Šebesta, Karel ; Goláňová, Hana

📎 This item contains 1 file (3.85 MB).

Publicly Available 

LexicalConceptualResource LRT + Open Submissions

[A Gold Standard Word Alignment for English-Swedish \(2015-10-12\)](#) 

(Linköping University / 2015-10-12)

Author(s):
Ahrenberg, Lars ; Holmqvist, Maria

📎 This item contains 1 file (500 KB).



What can you do?



Browse

> All of the Repository ▾

My Account

Login

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

Help Desk

FACETED SEARCH

Search

[Advanced Search](#)

Limit your search

- Author ▾
- Subject ▾
- Rights ▾
- Language (ISO) ▾
- Type ▾
- Contain Files ▾
- Community ▾

Showing 1 through 10 out of 1038 results

1
2
3
>
104
⚙️ ▾

Corpus
LINDAT / CLARIN

AKCES 2 ver. 2

(Charles University in Prague, ÚČJTK / 2013-12-18)

Author(s):
Šebesta, Karel ; Goláňová, Hana

📎 This item contains 1 file (3.85 MB).
Publicly Available
© ⓘ ⚙️

LexicalConceptualResource
LRT + Open Submissions

A Gold Standard Word Alignment for English-Swedish (2015-10-12)

(Linköping University / 2015-10-12)

Author(s):
Ahrenberg, Lars ; Holmqvist, Maria

📎 This item contains 1 file (500 KB).



🔍 What can you do?



🎯 Browse

> All of the Repository ▾

👤 My Account

➔ Login

📄 General Information

📄 Deposit

🗉 Cite

🔄 Submission Lifecycle

? FAQ

📄 About

✉ Help Desk

DISCOVERY

GOOGLE



prague dependency treebank 3.0



Vše

Obrázky

Nákupy

Mapy

Videa

Více

Nastavení

Nástroje

Přibližný počet výsledků: 13 900 (0,44 s)

Vědecké články o prague dependency treebank 3.0

Prague dependency treebank 3.0 - Bejček - Počet citací tohoto článku: 47

Prague Dependency Treebank - Hajič - Počet citací tohoto článku: 385

The Prague dependency treebank - Böhmová - Počet citací tohoto článku: 423

Prague Dependency Treebank 3.0 | ÚFAL

<https://ufal.mff.cuni.cz/pdt3.0> ▼ Přeložit tuto stránku

Introduction. The **Prague Dependency Treebank 3.0** (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed and improved in various aspects. Moreover ...

The Prague Dependency Treebank 2.0.

<https://ufal.mff.cuni.cz/pdt2.0/> ▼ Přeložit tuto stránku

The **Prague Dependency Treebank 2.0** (PDT 2.0) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation ... Please note that new versions of this corpus have been published: PDT 3.0 (2013), PDiT 1.0 (2012), PDT 2.5 (2012).

Prague Dependency Treebank 3.0 (PDT 3.0)

https://lindat.mff.cuni.cz/repository/xmlui/bitstream/.../PDT30_index_lindat.html?...

Prague Dependency Treebank 3.0 (PDT 3.0). Overview. The **Prague Dependency Treebank 3.0** (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed ...

CREDIT FOR DATA

enTenTen

“ Please use the following text to cite this item or export to a predefined format: ”

BIBTEX CMDI

Masaryk University, NLP Centre, 2011, *enTenTen*, 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-0000-0001-CCDF-8>.



This resource is also integrated in following services:

Share:

KonText

Item identifier	http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8
Date issued	2011-12-16
Type	corpus
Language(s)	English
Description	Very large English web corpus enTenTen, comprising 3,268,798,627 tokens.
Publisher	Masaryk University, NLP Centre
Acknowledgement	Lexical Computing Ltd.
Subject(s)	English large corpus
Collection(s)	LINDAT / CLARIN Data & Tools

[Show full item record](#)

What can you do?

DEPOSIT CITE

Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Context

> Edit this item

> Export Item

> Export Metadata

Administrative

Control Panel

Access Control

AS OPEN AS POSSIBLE

Choose a License

Answer the questions or use the search to find the license you want

↻ Start again



What do you want to deposit?

Software

Data

Search for a license...

Public Domain Mark (PD)

The work identified as being free of known restrictions under copyright law, including all related and neighboring rights.

Publicly Available



Public Domain Dedication (CC Zero)

CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

Publicly Available



OPEN DATA

AS OPEN AS POSSIBLE (NOT MORE)

Publisher Faculty of Arts, Institute of the Czech National Corpus, Charles University in Prague

Acknowledgement Ministerstvo školství, mládeže a tělovýchovy
 Project code: LM2011023
 Project name: Český národní korpus

Subject(s) representative corpus written language

Collection(s) LINDAT / CLARIN Data & Tools

[Show full item record](#)

Files in this item



Download instructions for command line

This item is Academic Use and licensed under:
 Czech National Corpus (Shuffled Corpus Data)



CLEAR RULES
 CUSTOM LICENSES
 LICENSE SIGNING

Name	syn2015.gz
Size	1.48 GB
Format	application/x-gzip
Description	corpus
MD5	e0242cc77e999794af6cfaf57f843c12



[Download file](#)

PREFER LATEST, PRESERVE ALL

Project name: Internet jako jazykový korpus

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: LN00A063

Project name: Centrum počítační lingvistiky

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: MSM 0021620838

Project name: Moderní metody, struktury a systémy informatiky

Subject(s)

MorphoDiTa

Czech

morphological analysis

morphological generation

PoS tagging

Collection(s)

LINDAT / CLARIN Data & Tools



This item is replaced by a newer submission:

<http://hdl.handle.net/11234/1-1836>

Please refer to the submission above for the latest available data. If you nevertheless need the original data, please click [here](#).

List all versions ▼

VERSIONING

PREFER LATEST, PRESERVE ALL

Collection(s)

LINDAT / CLARIN Data & Tools

Other versions

List all versions ▾

- ▶ Czech Models (Morfflex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
- Czech Models (Morfflex CZ 160310 + PDT 3.0) for MorphoDiTa 160310
- Czech Models (Morfflex CZ + PDT) for MorphoDiTa

[Show full item record](#)

Files in this item



Download instructions for command line

This item is **Publicly Available** and licensed under:

Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)



Name	czech-morfflex-pdt-161115.zip
Size	69.18 MB
Format	application/zip
Description	Czech Models (Morfflex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
MD5	adde38cd363219759e19165b06baa4ce



[Download file](#)

[Preview](#)

REPOSITORY SOFTWARE

- CLARIN-DSpace
 - DSpace + licensing, versioning and more
- LINDAT's project converting to community
- 13 deployments
10 countries

The screenshot shows the GitHub interface for the repository 'ufal / clarin-dspace', which is a fork of 'DSpace/DSpace'. The repository has 8,658 commits and 43 branches. The current branch is 'clarin', which is 1368 commits ahead and 2394 commits behind the 'DSpace:master' branch. A recent merge by user 'kosarko' is shown, merging the 'release-2018.01' branch into 'clarin'. Below the merge, a list of sub-projects is displayed, each with a folder icon and a description of the issue it resolves.

This repository Search Pull requests Issues

ufal / clarin-dspace
forked from DSpace/DSpace

Code Issues 90 Pull requests 0 Projects 0

clarin-dspace digital repository based on DSpace and LINDAT/CLARIN

repository dspace open-source pid

8,658 commits 43 branches

Branch: clarin New pull request

This branch is 1368 commits ahead, 2394 commits behind DSpace:master.

kosarko Merge branch 'release-2018.01' into clarin

dspace-api	Resolves #412 - Access other versions
dspace-jsui	Resolves #816 - Merge DSpace-5.8 (#
dspace-lni	Resolves #816 - Merge DSpace-5.8 (#
dspace-oai	Resolves #816 - Merge DSpace-5.8 (#
dspace-rdf	Resolves #816 - Merge DSpace-5.8 (#
dspace-rest	Resolves #412 - Access other versions
dspace-services	Resolves #816 - Merge DSpace-5.8 (#
dspace-solr	Resolves #816 - Merge DSpace-5.8 (#
dspace-sword	Resolves #816 - Merge DSpace-5.8 (#
dspace-swordv2	Resolves #816 - Merge DSpace-5.8 (#

SERVICES

- Web services
 - REST, stable services only
 - open source software, high-standards requirements
- Applications and user services
 - Wrappers around services (web-based applications)
 - Internet Language Reference Book (at Institute for Czech Language)
- “Physical Service”: Centre for Visual History Malach
 - Shoah collection access point
 - Reusable audio search technology

Web Services and Applications

20 SERVICE TYPES (WITH 100+ CORPORA/LEXICONS)

NameTag

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

BIBTEX

CMDI

Straka, Milan and Straková, Jana, 2014, *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-0000-0023-43CE-E>.



Share:

```

Loading ner: done
Recognizing: done, in 0.1 seconds.
<ne type="P">Václav</ne>
<ne type="ns">Havel</ne>/ne>, který se
narodil ve známé intelektuální rodině
<ne type="td">5.</ne>
<ne type="td">11</ne>/ne>
<ne type="ty">1936</ne>
v <ne type="q">Praze</ne>, se v letech
<ne type="ty">1993</ne> až
<ne type="ty">2003</ne> stal prvním
prezidentem <ne type="q">České
republiky</ne>.
    
```

Authors:

Milan Straka, Jana Straková

Description:

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 LGPL 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.

[PROJECT HOME](#)

[Run](#)



Treex::Web

Česílko

PML-Tree Query

PDT-Vallex

EngVallex

CzEngVallex

MorphoDiTa

NameTag

ILRB

ElixirFM

Dialogy.Org

Korektor

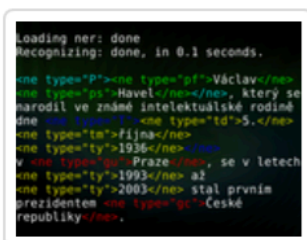
“ Please use the following text to cite this item or export to a predefined format:

BIBTEX CMDI

Straka, Milan and Straková, Jana, 2014, *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-0000-0023-43CE-E>.



Share:



Authors:

Milan Straka, Jana Straková

Description:

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 [LGPL](#) 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.

[PROJECT HOME](#) [Run](#)



Tree::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

Transformer-EN-CS

The Internet Language Reference Book

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

BIBTEX CMDI

Šmerk, Pavel; Pravdová, Markéta; Beneš, Martin; et al., 2009, *The Internet Language Reference Book*, 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-0000-0023-8BD2-2>.



Share:

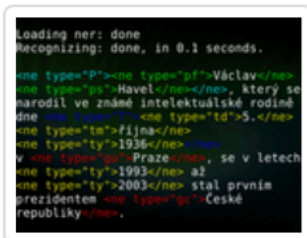
“ Please use the following text to cite this item or export to a predefined format:

BIBTEX CMDI

Straka, Milan and Straková, Jana, 2014, *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-0000-0023-43CE-E>.



Share:



Authors:

Milan Straka, Jana Straková

Description:

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 [LGPL](#) 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.

[PROJECT HOME](#) [Run](#)



Tree::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

Transformer-EN-CS

The Internet Language Reference Book

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

BIBTEX CMDI

Šmerk, Pavel; Pravdová, Markéta; Beneš, Martin; et al., 2009, *The Internet Language Reference Book*, 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-0000-0023-8BD2-2>.



Share:

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

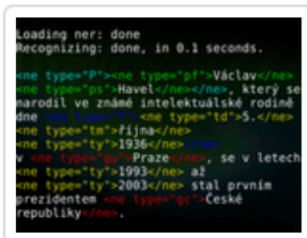
BIBTEX

CMDI

Straka, Milan and Straková, Jana, 2014, *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-0000-0023-43CE-E>.



Share:



Authors:

Milan Straka, Jana Straková

Description:

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 [LGPL](#) 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.

PROJECT HOME

Run



Tree::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

Transformer-EN-CS

The Internet Language Reference Book

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

BIBTEX

CMDI

Šmerk, Pavel; Pravdová, Markéta; Beneš, Martin; et al., 2009, *The Internet Language Reference Book*, 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-0000-0023-8BD2-2>.



Share:


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

BIBTEX

CMDI

Straka, Milan and Straková, Jana, 2014, *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-0000-0023-43CE-E>.



 Share:

```
loading ner: done
recognizing: done, in 0.1 seconds.
<ne type="P"><ne type="P1">Václav</ne>
<ne type="ns">Havel</ne></ne>, který se
narodil ve známé intelektuální rodině
<ne type="tm">fjina</ne>
<ne type="ty">1936</ne>
v <ne type="lo">Praze</ne>, se v letech
<ne type="ty">1993</ne> až
<ne type="ty">2003</ne> stal prvním
prezidentem <ne type="lo">České
republiky</ne>.
```

Authors:

Milan Straka, Jana Straková

Description:

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 [LGPL](#) 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.

 PROJECT HOME

 Run



Tree::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

Transformer-EN-CS

The Internet Language Reference Book


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

BIBTEX

CMDI

Šmerk, Pavel; Pravdová, Markéta; Beneš, Martin; et al., 2009, *The Internet Language Reference Book*, 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-0000-0023-8BD2-2>.



 Share:

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

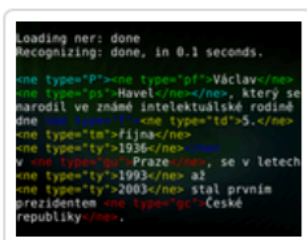
BIBTEX

CMDI

Straka, Milan and Straková, Jana, 2014, *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-0000-0023-43CE-E>.



Share:



Authors:

Milan Straka, Jana Straková

Description:

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 [LGPL](#) 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.

[PROJECT HOME](#)

[Run](#)



The Internet Language Reference Book

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

BIBTEX

CMDI

Šmerk, Pavel; Pravdová, Markéta; Beneš, Martin; et al., 2009, *The Internet Language Reference Book*, 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-0000-0023-8BD2-2>.



Share:

Tree::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

Transformer-EN-CS

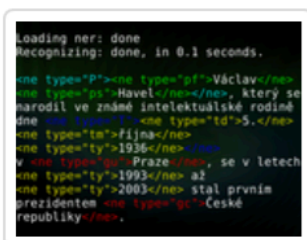
“ Please use the following text to cite this item or export to a predefined format:

BIBTEX CMDI

Straka, Milan and Straková, Jana, 2014, *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-0000-0023-43CE-E>.



Share:



Authors:

Milan Straka, Jana Straková

Description:

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 [LGPL](#) 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.

[PROJECT HOME](#)

[Run](#)



The Internet Language Reference Book

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

BIBTEX CMDI

Šmerk, Pavel; Pravdová, Markéta; Beneš, Martin; et al., 2009, *The Internet Language Reference Book*, 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-0000-0023-8BD2-2>.



Share:

Tree::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

Transformer-EN-CS

UDPipe

[About](#)[Run](#)[REST API Documentation](#)

UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given any data in [CoNLL-U format](#). Trained models are provided for nearly all [UD treebanks](#). UDPipe is available as a binary for Linux/Windows/OS X, as a library for C++, Python, Perl, Java, and as a web service. [Third-party R CRAN package](#) also exists.

UDPipe is a free software distributed under the [Mozilla Public License 2.0](#) and the linguistic models are free for non-commercial use and distributed under the [CC BY-NC-SA](#), although for some models the original data used to create the model may impose additional licensing conditions. UDPipe is versioned using [Semantic Versioning](#).

Copyright 2017 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic.

Description of the available methods is available in the [API Documentation](#) and the models are described in the [UDPipe 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: UD 2.0 ([description](#)) CoNLL17 Baseline UD 2.0 ([description](#)) UD 1.2 ([description](#))

 czech-ud-2.0-170801

Actions: Tag and Lemmatize Parse

▼ Advanced Options

 Input Text

 Input File

UDPipe

[About](#)[Run](#)[REST API Documentation](#)

UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given an in [CoNLL-U format](#). Trained models are provided for nearly all [UD treebanks](#). UDPipe is available as a binary for Linux/Windows/OS X, as a library for C++, Python, Perl, Java as a web service. [Third-party R CRAN package](#) also exists.

UDPipe is a free software distributed under the [Mozilla Public License 2.0](#) and the linguistic models are free for non-commercial use and distributed under the [CC BY-NC-SA](#) although for some models the original data used to create the model may impose additional licensing conditions. UDPipe is versioned using [Semantic Versioning](#).

Copyright 2017 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic.

Description of the available methods is available in the [API Documentation](#) and the models are described in the [UDPipe 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: UD 2.0 ([description](#)) CoNLL17 Baseline UD 2.0 ([description](#)) UD 1.2 ([description](#))

 czech-ud-2.0-170801

Actions: Tag and Lemmatize Parse

▼ Advanced Options

 Input Text

 Input File

UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given a dataset in [CoNLL-U format](#). Trained models are provided for nearly all [UD treebanks](#). UDPipe is available as a binary for Linux/Windows/OS X, as a library for C++, Python, Perl, Java, and JavaScript, as a web service. [Third-party R CRAN package](#) also exists.

UDPipe is a free software distributed under the [Mozilla Public License 2.0](#) and the linguistic models are free for non-commercial use and distributed under the [CC BY-NC-SA](#) license, although for some models the original data used to create the model may impose additional licensing conditions. UDPipe is versioned using [Semantic Versioning](#).

Copyright 2017 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic.

Description of the available methods is available in the [API Documentation](#) and the models are described in the [UDPipe 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. Any comments and reactions are welcome.

Model: UD 2.0 ([description](#)) CoNLL17 Baseline UD 2.0 ([description](#)) UD 1.2 ([description](#))

 english-ud-2.0-170801

Actions: Tag and Lemmatize Parse

▼ Advanced Options

 Input Text

 Input File

In the first place I hope you will live twenty-three years longer. Mr. Tom Lefroy's birthday was yesterday, so that you are very near of an age.

After this necessary preamble I shall proceed to inform you that we had an exceeding good ball last night, and that I was very much disappointed at not seeing Charles Fowle of the party, as I had previously heard of his being invited. In addition to our set at the Harwoods' ball, we had the Grants, St. Johns, Lady Rivers, her three daughters and a

In the first place I hope you will live twenty-three years longer. Mr. Tom Lefroy's birthday was yesterday, so that you are very near of an age.

After this necessary preamble I shall proceed to inform you that we had an exceeding good ball last night, and that I was very much disappointed at not seeing Charles Fowle of the party, as I had previously heard of his being invited. In addition to our set at the Harwoods' ball, we had the Grants, St. Johns, Lady Rivers, her three daughters and a

Process Input

A Output Text

Show Table

Show Trees

Save Output File

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# newdoc									
# newpar									
# sent_id = 1									
# text = In the first place I hope you will live twenty-three years longer.									
1	In	in	ADP	E	_	4	case	_	_
2	the	the	DET	RD	Definite=Def PronType=Art	4	det	_	_
3	first	first	ADJ	NO	Degree=Pos NumType=Ord	4	amod	_	_
4	place	place	NOUN	S	Number=Sing	6	obl	_	_
5	I	I	PRON	PE	Number=Sing Person=1 PronType=Prs	6	nsubj	_	_
6	hope	hope	VERB	V	Mood=Ind Number=Sing Person=1 Tense=Pres VerbForm=Fin	0	root	_	_

In the first place I hope you will live twenty-three years longer. Mr. Tom Lefroy's birthday was yesterday, so that you are very near of an age.

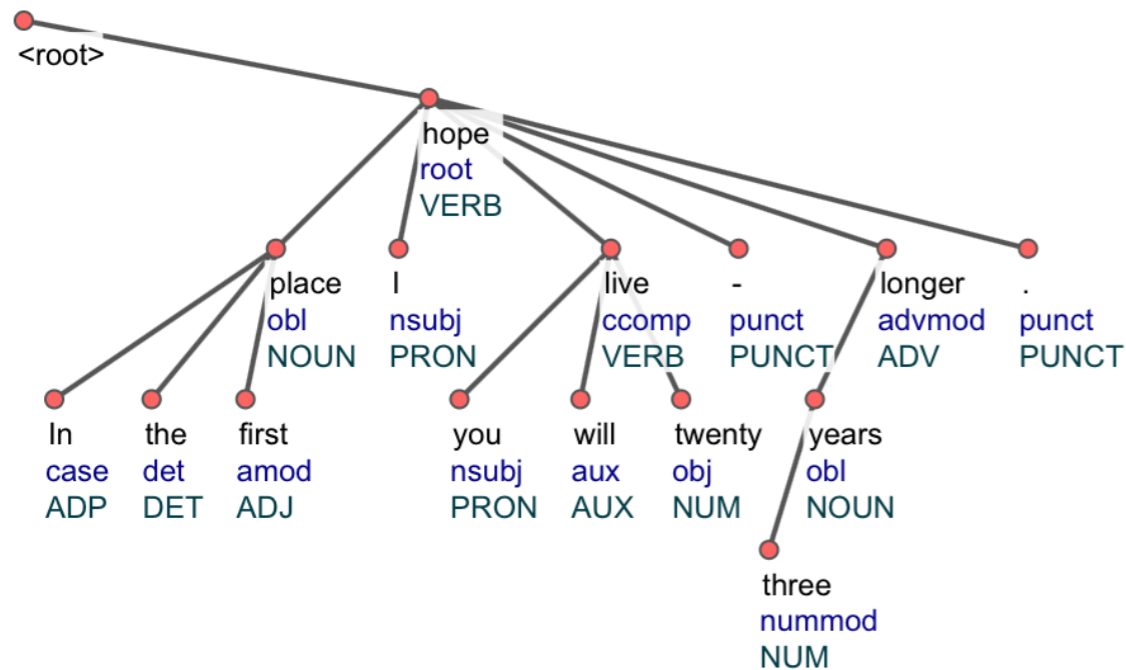
After this necessary preamble I shall proceed to inform you that we had an exceeding good ball last night, and that I was very much disappointed at not seeing Charles Fowle of the party, as I had previously heard of his being invited. In addition to our set at the Harwoods' ball, we had the Grants, St. Johns, Lady Rivers, her three daughters and a

Process Input

Save Tree as SVG

Previous 1 2 3 4 5 6 7 8 9 10 11 12 ... Next

In the first place I hope you will live twenty - three years longer .



UDPipe

[About](#)[Run](#)[REST API Documentation](#)

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

[Table of Contents](#)

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.

API Reference

The UDPipe 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
process	process supplied data	GET/POST

Method [models](#)

Return the list of models available in the UDPipe REST API, and for each model enumerate components supported by this models (model can contain a [tokenizer](#), [tagger](#) and a [parser](#)). 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

[try this](#)

Example JSON Response

```
{
  "models": {
    "czech-ud-1.2-160523": ["tokenizer", "tagger", "parser"],
    "english-ud-1.2-160523": ["tokenizer", "tagger", "parser"]
  },
  "default_model": "czech-ud-1.2-160523"
}
```

Method [process](#)

The response is in [JSON](#) format of the following structure:

```
{
  "model": "Model used",
  "acknowledgements": ["URL with acknowledgements", ...],
  "result": "processed_output"
}
```

The `processed_output` is the output of the UDPipe in the requested output format.

Browser Examples

```
http://lindat.mff.cuni.cz/services/udpipe/api/process?tokenizer&tagger&parser&data=Děti pojedou k babičce. Už se těší.
```

try this

Model Selection

There are several possibilities how to select required model using the `model` option:

- If `model` option is not specified, the default model (returned by `models` method) is used – this is guaranteed to be the latest Czech model.
- The `model` option can specify one of the models returned by the `models` method.
- The `model` option may be only several first words of model name. In this case, the latest most suitable model is used.
- The `model` can be ISO 639-1 or ISO 639-2 code of a language. If available, newest model for the requested language is used.

i Note that the last two possibilities allow using `czech`, `cs`, `ces`, `cze`, `english`, `en` or `eng` as models.

Accessing API using Curl

The described API can be comfortably used by `curl`. Several examples follow:

Passing Input on Command Line (if UTF-8 locale is being used)

```
curl --data 'tokenizer=&tagger=&parser=&data=Děti pojedou k babičce. Už se těší.' http://lindat.mff.cuni.cz/services/udpipe/api/process
```

Using Files as Input (files must be in UTF-8 encoding)

```
curl -F data=@input_file.txt -F tokenizer= -F tagger= -F parser= http://lindat.mff.cuni.cz/services/udpipe/api/process
```

Specifying Model Parameters

```
curl -F data=@input_file.txt -F model=english -F tokenizer= -F tagger= -F parser= http://lindat.mff.cuni.cz/services/udpipe/api/process
```

Converting JSON Result to Plain Text

```
curl -F data=@input_file.txt -F model=english -F tokenizer= -F tagger= -F parser= http://lindat.mff.cuni.cz/services/udpipe/api/process |
PYTHONIOENCODING=utf-8 python -c "import sys,json; sys.stdout.write(json.load(sys.stdin)['result'])"
```



- Preservation and dissemination of language data
- Creation of linguistic datasets
- Creation of language processing tools
- Language processing services
- Search interfaces for language datasets:
corpora, dictionaries, audio-visual data
- Support in utilising all of this for your research



- <http://lindat.cz> (<http://ufal.mff.cuni.cz/lindat>)
- <https://lindat.mff.cuni.cz/en/services>
- <http://hdl.handle.net/11858/00-097C-0000-0023-8BD2-2>
(<http://prirucka.ujc.cas.cz/>)
- <http://malach-centrum.cz>

Thank you!

<http://lindat.cz>

