Services of LINDAT/CLARIN Centre

Pavel Straňák, Jozef Mišutka, Eva Hajičová, Jan Hajič
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranske nám. 25, 118 00 Praha 1, Czechia
{hajic, hajicova, misutka, stranak}@ufal.mff.cuni.cz

Keywords: web applications, NLP services for general scientific public, web services

1. Introduction

We will introduce the services and online applications developed so far in LINDAT/CLARIN and available to general scientific public for non-commercial use. Several services require authentication and they use Clarin-recommended mechanisms for that, as we shall explain in Section 4., but all services that can be run freely, without any authentication, are run that way.

LINDAT/CLARIN provides services from creation and many types of annotation of data, through tier visualisation and sophisticated methods of search in both lexical and textual resources. We discuss how our services integrate with Clarin central infrastructural services like Virtual Language Observatory (VLO), Federated Content Search (FCS), or web-service chaining tool Weblicht. We also describe how our data repository and other services that require user authentication for some functionality employ Clarin Service Provider Federation (SPF) and Edugain integration to be available to as many users as possible.

2. LINDAT/CLARIN Portal

We provide a central portal of our activities at `http://lindat.mff.cuni.cz`. The services for our users (which includes data depositors) are organised into three main sections:

- Repository for safely storing, accessing and referencing data and software
- Corpus manager providing search, advanced linguistic metrics, and FCS integration for all corpora stored in our data repository
- Web application and services. All our web applications provide a graphical web fronted and also at least a basic REST API that allows the application to be used programatically as a web service.

  Except for the repository and the corpus manager Kontext, LINDAT/CLARIN provides two other big and complex web applications: Treex::Web engine for orchestration and efficient parallel execution of NLP scenarios, and PML::TQ search engine for searching any treebanks regardless of their type and annotation schema. Both of these applications include SVG visualisation of results (if they are trees).

  The remaining ten web applications we currently provide are either interfaces to lexicon and knowledge databases, or simpler tools that provide one task and this APIs allow them to be easily linked to ad hoc chains.

2.1. Standards for LINDAT/CLARIN Applications and Services

All applications that wish to qualify as LINDAT/CLARIN official services must have a formal project including source code management and issue tracking and this project must be Open Access. These projects are managed either at our departmental server `redmine.ms.mff.cuni.cz` using Redmine system, or at Github (`github.com/ufal`), Bitbucket, SourceForge and other popular source code management services.

All such projects must have a stable maintainer that is responsible for code quality and also for the fact that build scripts and possible installers of the software work.

All our web applications are installed in a stable, secure and scalable way: a small cluster dedicated to running web applications runs only these. Each application runs in its own virtual machine. All applications include REST API that allows them to be run e.g. from a simple script for batch processing and chaining of services.[1]

3. Projects, Apps, Services

In this section we present in detail all our official LINDAT/CLARIN web applications and services. In this abstract we give just three examples. Some services are still in development but will be ready before the conference.

3.1. Repository

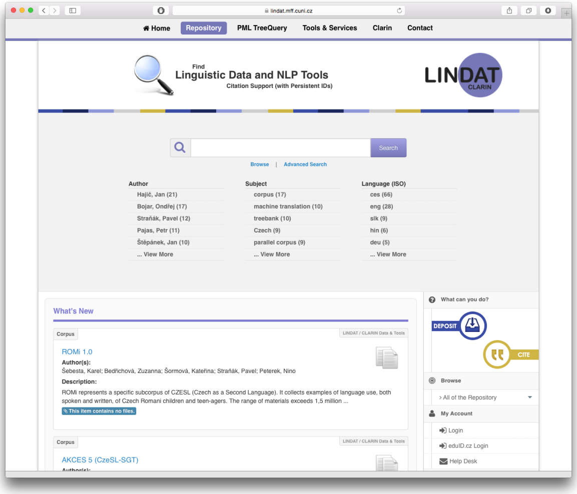
Our repository (Pajas et al., 2014) for linguistic data and tools is built on the most popular free repository software DSpace. Our development includes some heavy customisations of DSpace, especially licensing modules, new user interface and various administration improvements. The repository runs at `lindat.mff.cuni.cz/repository` and its software project is managed at GitHub.[2]

The whole of the repository is accessible without any restrictions, only uploading data and downloads of datasets

[1] except applications where it doesn't really make sense, e.g. Czech Language Guide

[2] To be done before the conference. Currently it is managed at `http://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications`.

with specific licensing restrictions require lures to log in as we describe in Section 4.



Author	Subject	Language (ISO)
Hajič, Jan (21)	corpus (17)	ces (66)
Bojar, Ondřej (17)	machine translation (10)	eng (28)
Straňák, Pavel (12)	treebank (10)	slk (9)
Pajas, Petr (11)	Czech (9)	hin (6)
Štěpánek, Jan (10)	parallel corpus (9)	deu (5)
... View More	... View More	... View More

What's New

Corpus — LINDAT / CLARIN Data & Tools

ROMi 1.0

Author(s):
Šebesta, Karel; Bedřichová, Zuzanna; Šormová, Kateřina; Straňák, Pavel; Peterek, Nino

Description:
ROMi represents a specific subcorpus of CZESL (Czech as a Second Language). It collects examples of language use, both spoken and written, of Czech Romani children and teen-agers. The range of materials exceeds 1,5 million ...

This item contains no files.

Corpus — LINDAT / CLARIN Data & Tools

AKCES 5 (CzeSL-SGT)

Author(s):

Figure 1: Repository homepage

3.2. Treex::Web

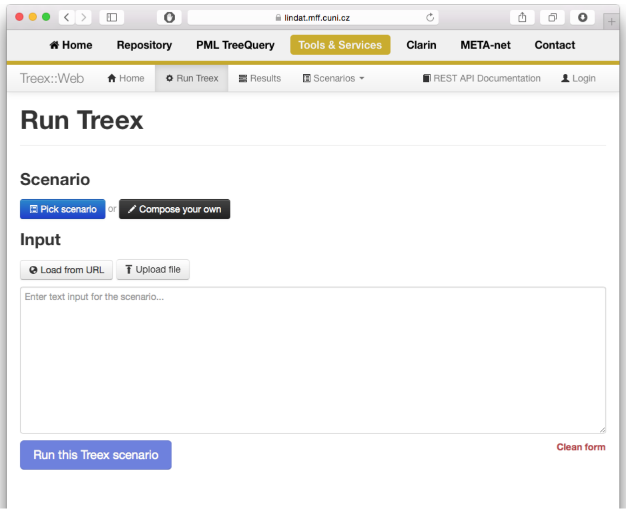
Treex::Web (Sedlák, 2014) is a web user interface for Treex. Treex is a highly modular NLP software system implemented in Perl programming language under Linux. It is primarily aimed at Machine Translation, making use of the ideas and technology created during the Prague Dependency Treebank project. At the same time, it can facilitate and accelerate development of software solutions of many other NLP tasks, especially due to re-usability of the numerous integrated processing modules (called blocks), which are equipped with uniform object-oriented interfaces.

The web interface of Treex::Web allows to pick a ready-made scenario, e.g. Czech-English translation, and run it, create your own scenario, save the results to disk or visualise them directly (especially relevant for results that include syntactic trees), etc.

3.3. MorphoDiTa

MorphoDiTa stands for "Morphological Dictionary and Tagger". It is an open-source tool for morphological analysis of natural language texts. It performs morphological analysis, morphological generation, tagging and tokenization and is distributed as a standalone tool or a library, along with trained linguistic models. For Czech MorphoDiTa achieves state-of-the-art results with a throughput around 10-200K words per second. MorphoDiTa 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.

The web interface provides a simple user interface that allows casual users to paste and analyse or lemmatise a piece of text and tweak the formatting of results to best fit their workflow. The REST API provides all the options available



Run Treex

Scenario

Input

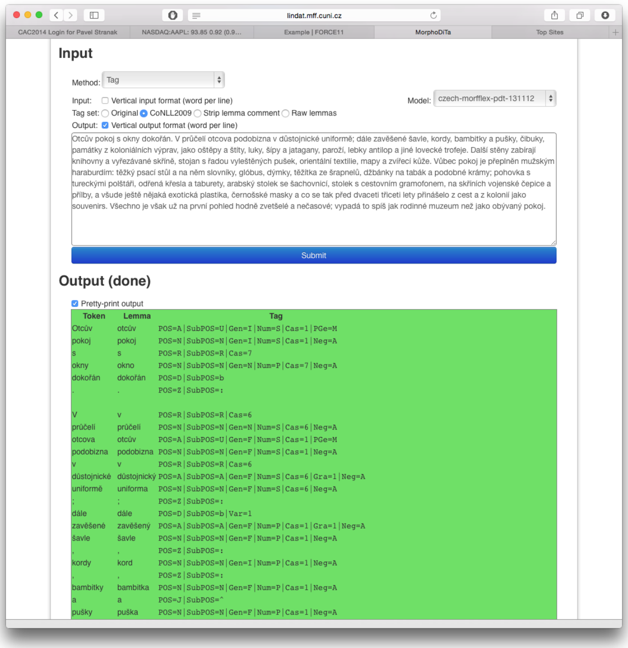
Enter text input for the scenario...

Figure 2: Treex::Web

in the web interface and can serve to efficiently process larger amounts of data and provide output in the desired format.

MorphoDiTa is an open-source project and the LINDAT/CLARIN service based on it is freely available for non-commercial purposes. The library is distributed under LGPL and the currently available associated models and data under CC BY-NC-SA, although for some models the original data used to create the model may impose additional licensing conditions.

Non-commercial use of the service doesn't require any logins or verifications, it if freely available.


lindat.mff.cuni.cz
CAC2014 Login for Pavel Stranak
NASDAQ:AAPL: 93.85 0.92 (0.9...
Example | FORCE11
MorphoDiTa
Top Sites
Submit

Input

Method: Tag

Input: □ Vertical input format (word per line)

Model: czech-morfflex-pdt-131112

Tag set: ○ Original ● CoNLL2009 ○ Strip lemma comment ○ Raw lemmas

Output: ☑ Vertical output format (word per line)

Otcův pokoj s okny dokořán. V průčelí otcova podobizna v důstojnické uniformě; dále zavěšené šavle, kordy, bambitky a pušky, čibuky, památky z koloniálních výprav, jako oštěpy a štíty, luky, šípy a jatagany, paroží, lebky antilop a jiné lovecké trofeje. Další stěny zabírají knihovny a vyřezávané skříně, stojan s řadou vyleštěných pušek, orientální textilie, mapy a zvířecí kůže. Vůbec pokoj je přeplněn mužským haraburdím: těžký psací stůl a na něm slovníky, glóbus, dýmky, těžítka ze šrapnelů, džbánky na tabák a podobné krámy; pohovka s tureckými polštáři, odřená křesla a taburety, arabský stolek se šachovnicí, stolek s cestovním gramofonem, na skříních vojenské čepice a přílby, a všude ještě nějaká exotická plastika, černošské masky a co se tak před dvaceti třiceti lety přinášelo z cest a z kolonií jako souvenirs. Všechno je však už na první pohled hodně zvetšelé a nečasové; vypadá to spíš jak rodinné muzeum než jako obývaný pokoj.

Output (done)

☑ Pretty-print output

```
  Token       Lemma                              Tag
Otcův       otcův       POS=A|SubPOS=U|Gen=I|Num=S|Cas=1|PGe=M
pokoj       pokoj       POS=N|SubPOS=N|Gen=I|Num=S|Cas=1|Neg=A
s           s           POS=R|SubPOS=R|Cas=7
okny        okno        POS=N|SubPOS=N|Gen=N|Num=P|Cas=7|Neg=A
dokořán     dokořán     POS=D|SubPOS=b
.           .           POS=Z|SubPOS=:

V           v           POS=R|SubPOS=R|Cas=6
průčelí     průčelí     POS=N|SubPOS=N|Gen=N|Num=S|Cas=6|Neg=A
otcova      otcův       POS=A|SubPOS=U|Gen=F|Num=S|Cas=1|PGe=M
podobizna   podobizna   POS=N|SubPOS=N|Gen=F|Num=S|Cas=1|Neg=A
v           v           POS=R|SubPOS=R|Cas=6
důstojnické důstojnický POS=A|SubPOS=A|Gen=F|Num=S|Cas=6|Gra=1|Neg=A
uniformě    uniforma    POS=N|SubPOS=N|Gen=F|Num=S|Cas=6|Neg=A
;           ;           POS=Z|SubPOS=:
dále        dále        POS=D|SubPOS=b|Var=1
zavěšené    zavěšený    POS=A|SubPOS=A|Gen=F|Num=P|Cas=1|Gra=1|Neg=A
šavle       šavle       POS=N|SubPOS=N|Gen=F|Num=P|Cas=1|Neg=A
,           ,           POS=Z|SubPOS=:
kordy       kord        POS=N|SubPOS=N|Gen=I|Num=P|Cas=1|Neg=A
,           ,           POS=Z|SubPOS=:
bambitky    bambitka    POS=N|SubPOS=N|Gen=F|Num=P|Cas=1|Neg=A
a           a           POS=J|SubPOS=^
pušky       puška       POS=N|SubPOS=N|Gen=F|Num=P|Cas=1|Neg=A
```

Figure 3: MoirphoDiTa web interface with output in the popular “CoNLL-2009” tabular format

4. Integration with Central Clarin infrastructural Services

We will demonstrate which applications are integrated with which Clarin technologies and services and for what result.
CMDI, **ISOCat** and **OAI-PMH** in the repository alows to provide highly accurate metadata, export them to VLO and also automatically integrate web services into Weblicht by providing correct CMDI metadata.
FCS is integrated in the corpus manager Kontext and thus allows all our available corpora to be accessible in Clarin FCS Agregator for the first search and then in the second step an corpus can be searched locally using a more expressive search language.
Shibboleth is currently integrated in Česílko, PML-TQ, Treex::Web, multimodal corpus search system Dialogy.org (corpus ROMi) – for accessing restricted datasets – and in the repository **to allow interactive submissions and license signing** for academic (ACA) and restricted (RES) datasets.
We have also started to make our services or complete scenarios available as **tools in WebLicht**. The first tool that is already available is Czech-Slovak machine translation system Česílko (Hajič et al., 2012) and more will be available by the time of the full paper and the Clarin conference.

5. Conclusion

Czech Clarin centre LINDAT/CLARIN is a consortium of four leading NLP centres in the country. They run a single unified portal that strives to present useful services for all scientist that work with language data.
We work hard to make our services as accessible and simple to use as possible, from an option to immediately and interactively deposit and publish linguistic data or tools , through various web applications and services that provide access to structured linguistic data (corpora, lexicons) or allow to detect linguistic structures in data, to visualisation of the results.
All our applications and services are also managed as open source projects and correctly licensed under popular open licenses.
Most of our datasets are available freely at least for academic research. After consultations within CLARIN Legal Issues Committee (CLIC) we allow our corpora including treebanks to be searched completely freely.
All our services that need to authenticate users do so via Shibboleth module. LINDAT/CLARIN is a member of Czech national federation EduID.cz through which it became member of Clarin Service Provider Federation and EduGAIN, thus allowing maximal number of users to securely and comfortably use our services.

6. Acknowledgements

Place all acknowledgements (including those concerning research grants and funding) in a separate section at the end of the article.

7. References

Hajič, J., Kuboň, V., and Homola, P. (2012). Česílko. In *LINDAT/Clarin*. http://hdl.handle.net/11858/00-097C-0000-0006-AAFE-A. 4.

Pajas, P., Vandas, K., Mišutka, J., Kamran, A., Jawaid, B., Košarko, O., Sedlák, M., Josífko, M., Straňák, P., and Hajič, J. (2014). Linguistic digital repository based on DSpace. In *LINDAT/Clarin*. http://hdl.handle.net/11858/00-097C-0000-0023-4087-6. 3.1.

Sedlák, M. (2014). Treex::web. In *LINDAT/Clarin*. http://hdl.handle.net/11858/00-097C-0000-0023-44AF-C. 3.2.