



UiO : **Det humanistiske fakultet**

Janne Bondi johannessen, Anders Nøklestad, Joel Priestley and Kristin Hagen

WP5: Glossa Integration



WP5 Glossa integration

- The current Glossa corpus interface and analysis tool will be integrated in the CLARINO infrastructure through the following extensions, carried out by UiO (at Text Laboratory, ILN in cooperation with USIT):
- Needs analysis among national and international users
- Automation of corpus building by creating the necessary server software
- A framework for the use of *internal and external* text analysis and management tools, and integration of this framework into the corpus construction pipeline
- A framework for metadata-based user interface and a port for metadata harvesting by the national registry

- Integration of the user management and access control features into a federated user authentication framework.
- A fully self-service workflow for creating and managing the user's corpora.
- User communication features such as self-service access control and online documentation, comments, blogs etc.
- Glossa will be provided as an online web-based service hosted in cooperation with USIT at the University of Oslo.

- The Glosa corpus system is used for many different types of corpora:
 - Monolingual written language
 - Multilingual written language (= parallel corpora / translation corpora)
 - Monolingual speech corpora
 - Multilingual speech corpora

Lexicographic Bokmål Corpus

Lexicographic corpus for Norwegian
Bokmål

Glossa ([my results](#) | [my annotations](#) | [statistics](#) | [full query help](#))

valg »

Regular expressions: **Hits per page:** **Randomize**
Search within: **Max results :** **Skip tot. freq.** **Context:**
 sentence word
 left right

tittel **tittel-id**
samling **type**

issn/isbn **utgiver** **utgivelsessted** **utgivelsesår**

kategori **emne**

navn **fødested**
type **kjønn** **fødselsår**

[Velg subkorpus](#)

Display:
Search within:

Flash QT

Oslo Multilingual Corpus

The Oslo Multilingual Corpus

Glossa ([my results](#) | [my annotations](#) | [statistics](#) | [full query](#) | [help](#))

[translations](#) [drevet](#)

Norwegian ▾
+
valg »
-

[add phrase](#) [delete phrase](#)

Regular expressions: **Search within:**
Hits per page: **Max results :**
Randomize Skip tot. freq. Context:
 sentence word
 left right

classcode ⁺ **database** ⁺ translated ⁻
 n: original
 y: translation

publisher ⁺ **publication place** ⁺ **publication date** ⁺

title ⁺ **title-id** ⁺

author ⁺ **translator** ⁺

language variety ⁺

Display: **Search within:**

[Search corpus](#)

[Reset form](#)

[Show texts](#)


[Save subcorpus](#)

[Choose subcorpus](#)

Flash QT

The Oslo Multilingual Corpus

Developed by [The Sprik project](#),
in cooperation with The Text Laboratory

tektlab.



The RuN Corpus

Norwegian ▾
+
valg » -

[add phrase](#) [delete phrase](#)

Regular expressions: **Search within:**
Hits per page: **Max results :**
Randomize Skip total frequency Context:
 sentence word
 left right

[Search corpus](#)

[Reset form](#)

classcode ⁺ **database** ⁺ translated [▾]
 n: original
 y: translation

publisher ⁺ **publication place** ⁺ **publication date** ⁺

title ⁺ **title-id** ⁺

author ⁺ **translator** ⁺

language variety ⁺

[Show texts](#)

[Show external text list](#)

[Save subcorpus](#)

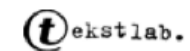
[Choose subcorpus](#)

Display: **Search within:**

Flash QT

The RuN Corpus

Developed by [The RuN project](#)
in cooperation with [The Text](#)



NoTa-Oslo speech corpus

æøå...»
+
-
criteria»

add phrase delete phrase

Regulære uttrykk: Treff per side: Tilfeldig utvalg
Søk innen: Maks resultater : Ingen totalfrekvens

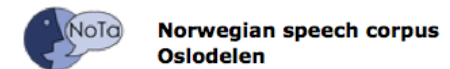
Søk i korpus
Slett skjema

informant + representativ utvalg + situasjon +
oppvokst + bosted + bodd lengst +
detaljert + detaljert + detaljert +
aldersgruppe + kjønn +
utdannelse + yrke +

Show informants
Lagre subkorpus
Velg subkorpus

Display: Search within:

Flash QT



Developed at tekstlab.



Nordic Dialect Corpus

Nordic Dialect Corpus

Glossa ([my results](#) | [my annotations](#) | [statistics](#) | [full query](#) | [help](#))

æøå...»

criteria»



- [Transcription guidelines, translation list etc](#)
- [Recording locations](#)
- [Transcriptions](#)

Regular expressions: **Search within:** **Hits per page:** **Max results :** Randomize Skip tot. freq. Orthographic Phonetic Both

informant
country region area place
agegroup sex rec (year) genre

[Choose subcorpus](#)

Display: **Search within:**
Flash QT



Different metadata (search categories)

Lexicographic Bokmål Corpus

tittel

2001 årsrapport
Arbeid, velferd og inkludering
Automatisering i maskinfaget
Avfall – avskaffelsen av kastesystemet
Bokvalitet og hverdagsliv for eldre
Bolignormer, helse og velferd
Cultural Studies

[>]
[<]

choose ▾

samling

issn/isbn utgiver utgivelsessted utgivelsesår

kategori emne

AV0%: Aviser og periodika
SA0%: Sakprosa
SK0%: Skjønnlitteratur
TV0%: TV
UP0%: Upublisert materiale / annet

[>]
[<]

choose ▾

kategori (detaljer)

navn fødested

type kjønn fødselsår

Oslo Multilingual Corpus

classcode database translated

n: original
 y: translation

publisher publication

..
A/S Hjemmet – Fagpresseforlaget
AA Publishing
Actes Sud
Ad Notam Gyldendal
Ad Notam Gyldendal AS
Addison-Wesley Publishing Company

[>]
[<]

choose ▾

title title-id

author translator

??
A.B.P. de Lemos
Adelina Antunes
Agnete Øye
Alain Gnaedig
Alken Bruns
American Institute for Contemporary Germ...

[>]
[<]

choose ▾

More different metadata

informant + representativ utvalg + situasjon +

oppvokst + bosted - bodd lengst +

rest vest [>] [<] velg

detaljert + detaljert + detaljert +

aldersgruppe + kjønn +

utdannelse + yrke +

NoTa-Oslo

NDC

Regular expressions: Hits per page: 20 Randomize Orthographic
Search within: s Max results : 2000 Skip tot. freq. Phonetic
Both

informant +

country + region + area + place +

agegroup + sex + rec (year) + genre +

Corpora that use Glossa

- The European Parliamentary Comparable and Parallel Corpora (ECPC) (under development):
<http://www.ecpc.uji.es/EN/home.php?language=en>
- Oslo Multilingual Corpus (Johansson and Hofland, 1994):
<http://www.hf.uio.no/ilos/OMC/>
- RUN Parallel Corpus:
<http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/run/corpus/>
- Lexicographical Bokmål Corpus (Fjeld 2008):
<http://www.hf.uio.no/iln/forskning/samlingene/bokmal/index.html#bokmalskorpus>
- Lule Sámi Corpus:
<http://giellatekno.uit.no/doc/lang/corpus/corpus-smj.html>
- Macedonian text corpus:
http://www.tekstlab.uio.no/glossa/html/index_dev.php?corpus=mak
- Mörkuð íslensk málheild (Icelandic Corpus):
<http://mim.hi.is/>
- North Sámi Corpus:
<http://giellatekno.uit.no/doc/lang/corpus/corpus-sme.html>
- Big Brother Corpus (Speech), Norwegian:
<http://www.tekstlab.uio.no/nota/bigbrother/>
- Nordic Dialect Corpus (Speech):
<http://www.tekstlab.uio.no/nota/scandiasyn/>
- Ruija Speech Corpus of Kven:
<http://www.hf.uio.no/iln/tjenester/kunnskap/sprak/korpus/talesprakskorpus/ruija/index.html>
- NoTa Oslo Speech Corpus
<http://www.tekstlab.uio.no/nota/oslo/>
- TAUS Speech Corpus of Norwegian:
<http://www.tekstlab.uio.no/nota/taus/index.html>
- UPUS Speech Corpus Multiethnic Norwegian:
<http://www.hf.uio.no/iln/forskning/prosjekter/opus/>
- Finlandssvensk: Svenska Litteratursällskapet I Finland
<http://www.sls.fi/>

Automation of corpus building (text):

pipeline

Files with text and metadata (e.g. in TEI format)

Text extraction

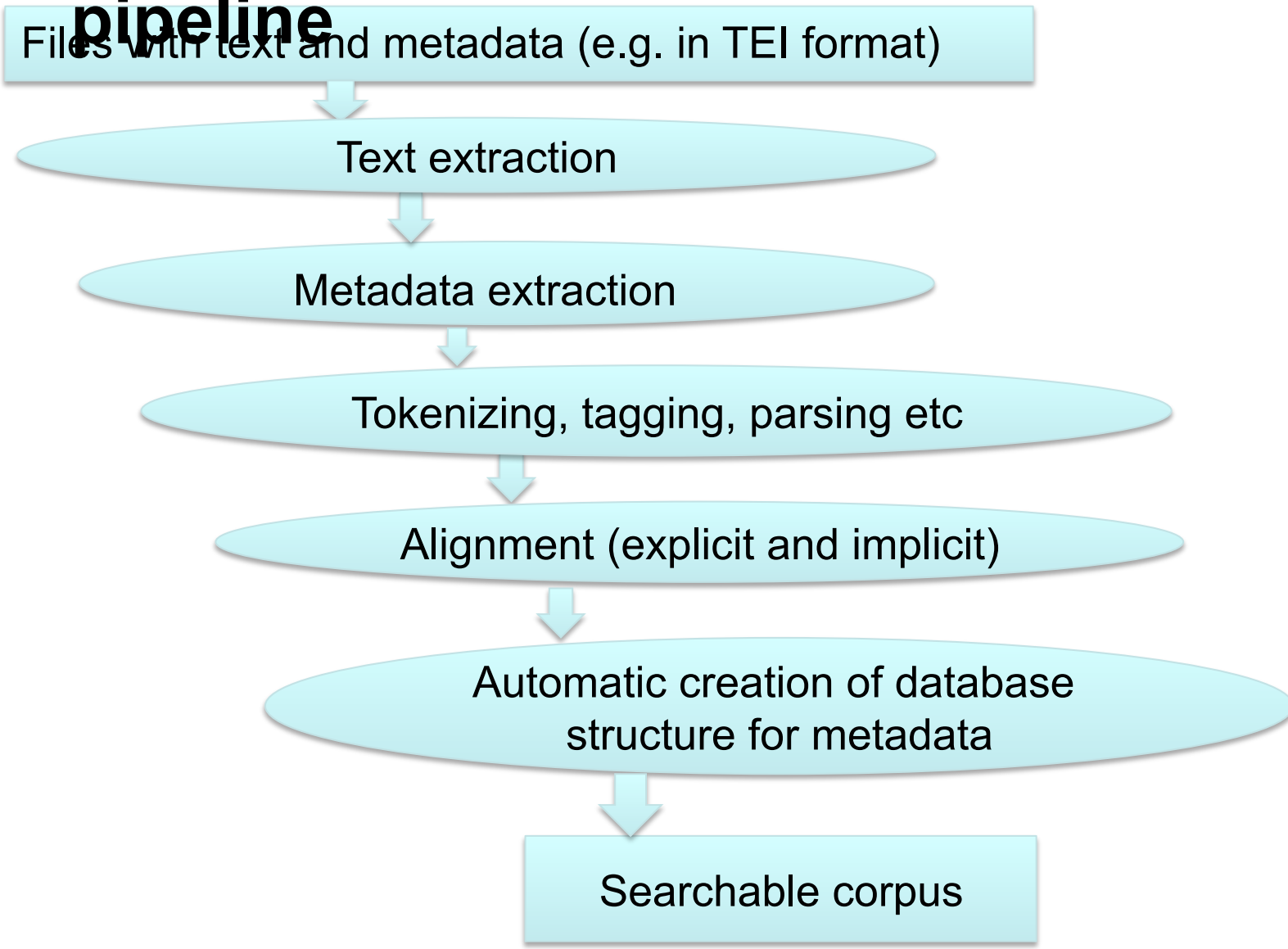
Metadata extraction

Tokenizing, tagging, parsing etc

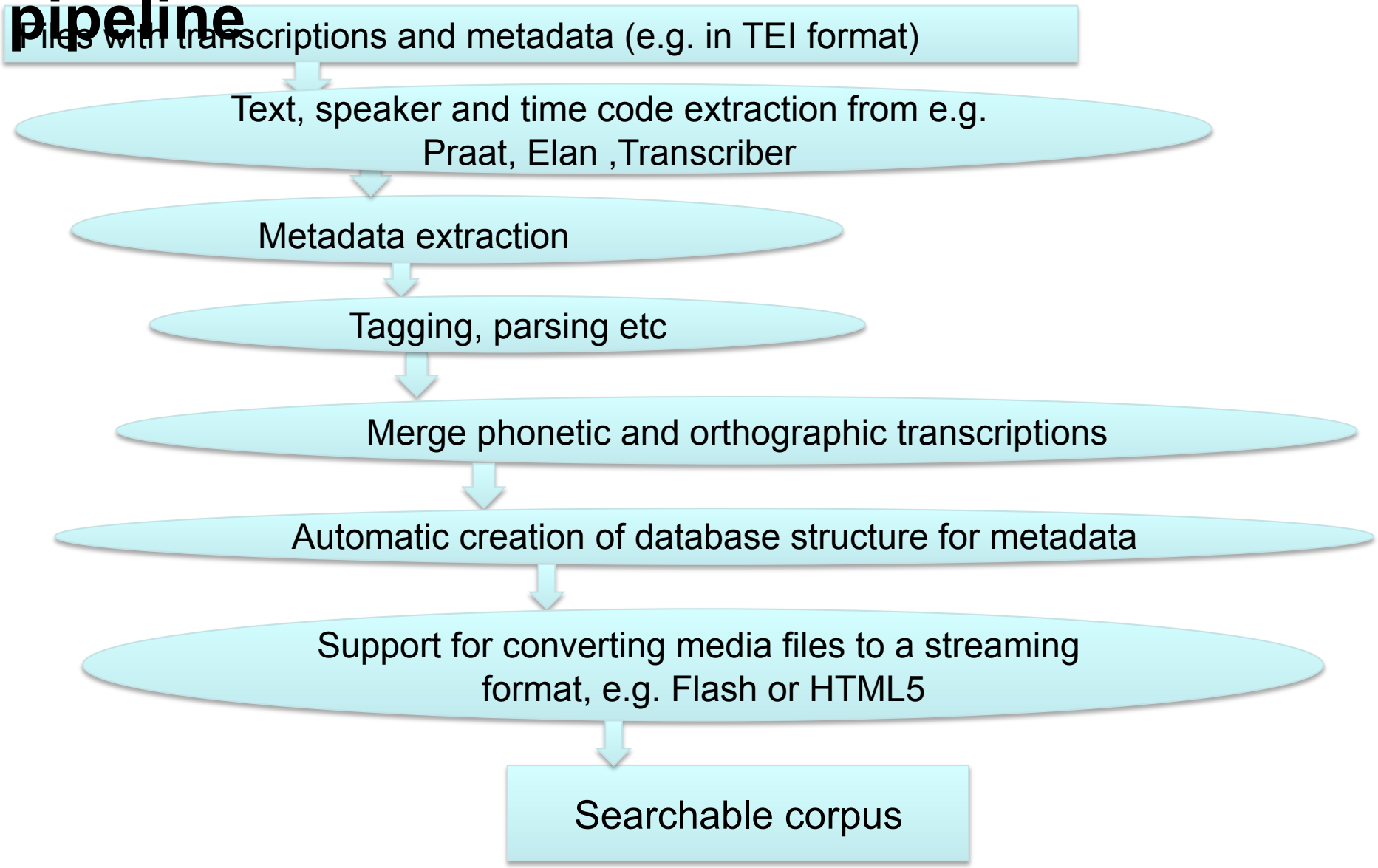
Alignment (explicit and implicit)

Automatic creation of database structure for metadata

Searchable corpus



Automation of corpus building (speech): pipeline



Use of internal and external tools

- Build in support for use of text analysis and results handling tools
- Text analysis: tokenizers, taggers, parsers etc.
- Results handling tools: visualization modes (e.g. syntactic trees), multi-modal display, statistics
- Internal tools: bundled with Glossa
- External tools: accessed through the CLARIN network



Metadata-based user interface

- Perhaps ideally: user interface generated directly from corpus data
- No, problem when corpus texts contain different data
 - Example: multi-lingual corpora with different tagsets for different languages, must be harmonized
- Better: interface generated from metadatabase

Interface generation

- Set of available metadata input fields generated from metadatabase
- Manual grouping and ordering of the fields for improved structure and final finish of the interface

Metadata harvesting

- Set up a web service for Glosa metadata to be harvested by the CLARIN thresher



Needs analysis among national and international users

- Already ongoing
- Findings so far:
 - Need to handle more alphabets and right-to-left writing systems
 - Stepwise interface from simple to advanced
 - Different academic fields have different needs
 - Virtually unlimited need wrt results handling options, e.g. statistics, visualizations etc.

الحق



Access control

- Integration of the user management and access control features into a federated user authentication framework.
- CLARINO plan: Uninett will through Feide offer authentication and authorization services to the CLARIN AAI federation.
 - What about user groups who are not registered in Feide?
- The Glossa access control will adapt the standards and procedures developed by Uninett

A fully *self-service* workflow for creating and managing the user's corpora

- Interface for
 - uploading files or selecting files from the CLARIN depot
 - Defining metadata if not included in the text files
 - grouping texts into corpora and subcorpora
 - Selecting language, taggers and other annotation tools to be run on the texts
- To be done in close collaboration with the LAP work package

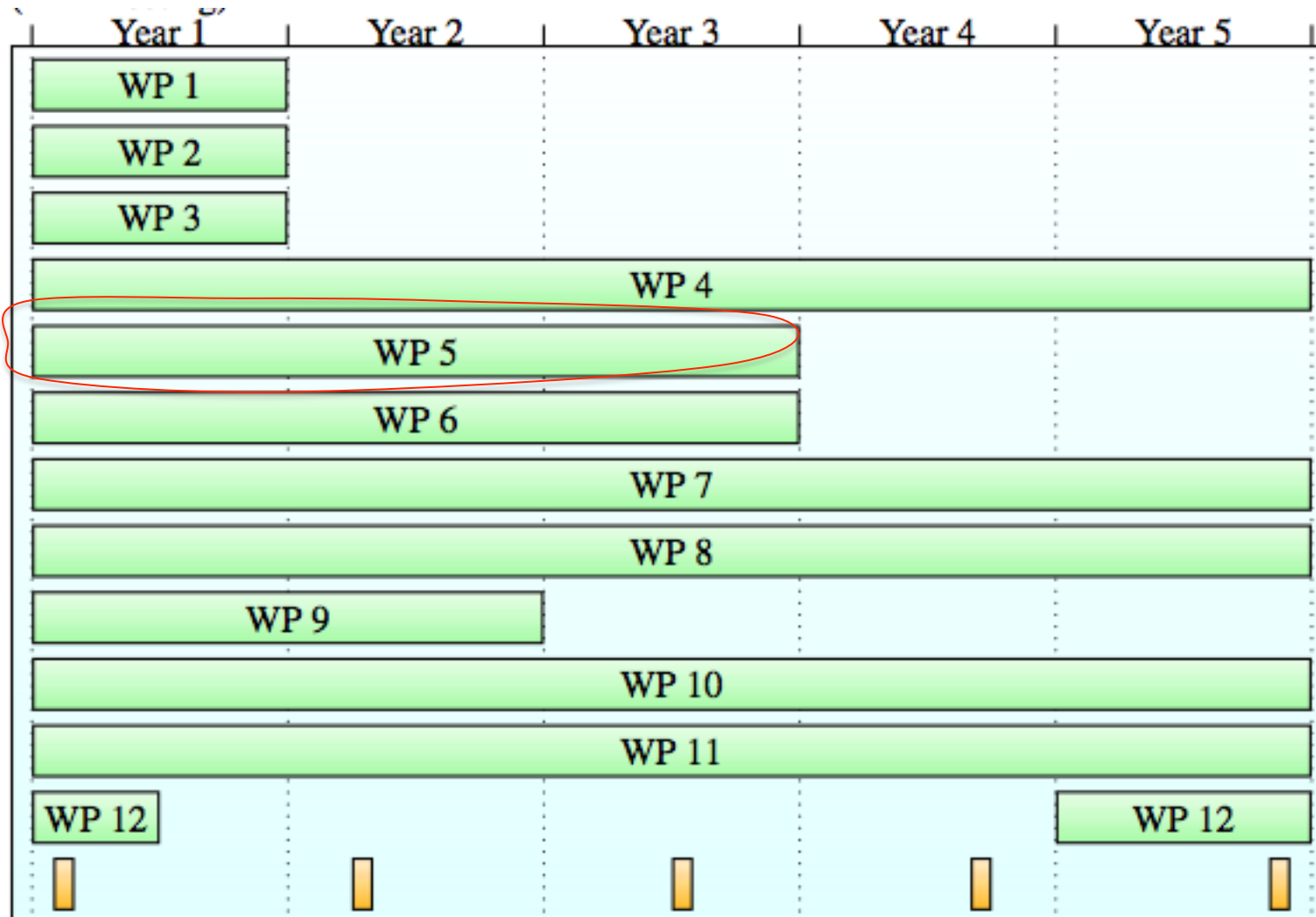


User communication features



- Self-service access control, inside and outside CLARIN
- Online documentation
- User-supplied comments
- User-supplied corrections and annotations

Time schedule



Year 1: 3 man-months

- Handling of different alphabets and right-to-left writing
- Initial design of corpus creation pipeline
 - based on a variety of procedures used for existing corpora
- Decoupling the front-end (web browser code) and back-end (server code), enabling
 - support for different user interfaces
 - JSON-based API for the back-end that can be used by the CLARIN infrastructure

- Thank you for your time!