FCS @ ILC4CLARIN

…and something else...

Utrecht, May 21, 2019

Riccardo Del Gratta

Outline

- Current implementation
- Changes made to the SoftWare
- Dockerization
- What we offer
- What we (hopefully) are going to offer

Current implementation

- Korp Backend;
- KORP FCS End Point;
- Korp Frontend not yet implemented.

Korp Backend:	https://github.com/spraakbanken/korp-backend
Korp FCS End Point:	https://github.com/clarin-eric/fcs-korp-endpoint
Period:	January to March 2019
Ratio:	Same strategy we followed for CLARIN-DSPACE. Picking “something” which is used already, so that we can be helped when needed.

Changes to the software (1/2)

1. Possibility to manage different POS tagsets in the same End Point;
2. Added different query types from the aggregator.

1. The EP is configurable through a property file. One of this property says the POS tagset used to annotate the corpus:
   MYCORPUS1=UD,MYCORPUS2=EAGLES…
   The correct POS translator to and from UD is automatically loaded from a factory.

Ratio: ILC4CLARIN is planning to make various corpora available through FCS. Some of them are from several years ago, other more recent. So there are many POS tagsets. We decided to register the corpora in KORP as they are and to implement mappings to and from their tagsets and UD.

Changes to the software (2/2)

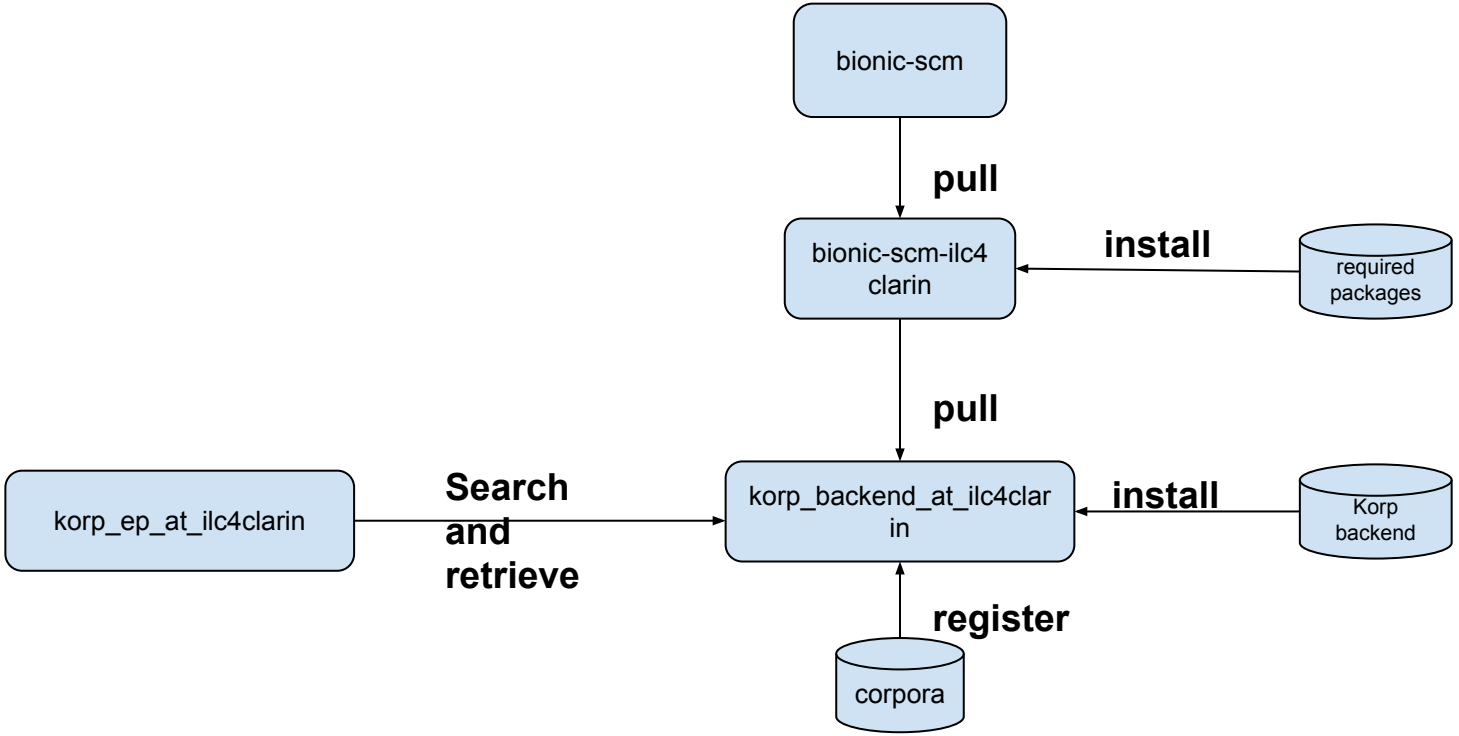
1. Possibility to manage different POS tagsets in the same End Point;

2. Added different query types from the aggregator.

2. Queries such as
[word = "..."],
[word = "..."] [pos = "..."],
[word = "..." & pos = "..."]
[word = "..." | pos = "..."]
[(word = "..." | pos = "..." & lemma = "...")]
[(word = "..." | pos = "..." | lemma = "...")]
[(word = "..." | pos = "...") | lemma = "..."]
….
Have been tested

ILC4CLARIN Korp FCS End Point: https://github.com/cnr-ilc/ilc4clarin-fcs-korp-endpoint

Dockerization (1/2)



Dockerization (2/2)

bionic-scm:	A minimal ubuntu image
bionic-scm-ilc4clarin:	All required software for KORP is installed on top of bionic-scm
korp_backend_at_ilc4clarin:	KORP is installed on top of bionic-scm-ilc4clarin. Corpora are registered.
korp_ep_at_ilc4clarin:	Minimal tomcat 9 with java 8. The FCS End Point is deployed here

What we Offer

- The Italian Treebanks
  - Contained in Universal Dependencies 2.3. (from LINDAT)
  - Licensed under different CC-*

UD_Italian-ISDT	An Italian corpus annotated according to the UD annotation scheme that was obtained by conversion from ISDT (Italian Stanford Dependency Treebank). Released for the dependency parsing shared task of Evalita 2014
UD_Italian-ParTUT	A conversion of a multilingual parallel treebank developed at the University of Turin and consisting of a variety of text genres, including talks, legal texts and Wikipedia articles, among others
UD_Italian-PoSTWITA	A collection of Italian tweets annotated in Universal Dependencies that can be exploited for the training of NLP systems to enhance their performance on social media texts
UD_Italian-PUD	A part of the Parallel Universal Dependencies (PUD) treebanks created for the CoNLL 2017 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies

What we are (hopefully) going to offer (1/2)

- PAROLE 3M Corpus
  - Balanced:
    - Newspaper
    - Miscellanea
    - Periodicals
    - Books
  - Different timespans
    - Newspaper (1992 - 1996)
    - Miscellanea (1987 - 1997)
    - Periodicals (1985 - 1988)
    - Books (1970 - 1995)

- PAROLE 20M
  - Discussions on licenses;
  - Balanced as well.

- Greek and Latin corpora
  - From Perseus project;
  - From students (Greek tragedies manually annotated).

What we (hopefully) are going to offer (2/2)

- PAROLE 3M Corpus
  - Word level (basic search):
  - Initially
    - Lemma and POS (advanced search)
      - Automatically annotated
      - POS tagset EAGLES (so the mapping to and from UD comes out again)
  - Later on
    - Lemma and POS (advanced search)
      - From SGML

- PAROLE 3M Organization
  - We will organize the corpora according to their original structure