CLARIN Workshop
Multilingual corpus annotation tools: development and integration

INTRODUCTION

Tomaž Erjavec, Darja Fišer

Dept. of Knowledge Technologies,
Jožef Stefan Institute, Ljubljana Dept. of Translation Studies,

Faculty of Arts, University of Ljubljana

CLARIN workshop, Jožef Stefan Institute
Ljubljana, 10th–11th November 2016

Overview

Introduction

Reasons for the proposal

- CLARIN.SI: working repository & concordancer
- Logical next step:
  corpus annotation for linguists → on-line workflows
  - we developed open & language independent annotation tools (ReLDI & Janes)
  - CLARIN-D WebLicht
  - JSI ClowdFlows and TextFlows
- So, why not have a workshop?
- Proposers: CLARIN.SI + CLARIN-DE, CLARIN Latvia

CLARIN Type II workshop

- The WS should result in a development plan
  max. 2 months after the WS: January 2017!
- The proposed development should address humanities & social sciences users
- Development:
  - max. 6 months duration
  - 3 PM (secondment) + travel costs
- Not necessary for all the WS participants to participate in the development

The time is right

- 60's meets EU: Open Source and Open Access
- Maturity of tools and resources:
  "little" effort needed to open & operationalise them
- Issues:
  - Maturity of platforms:
    ease of use, bugs & crashes
  - Scalability of solutions:
    data transfer overhead & hardware base
  - How many tools and resources are in fact openly available
  - Disparate standards in formats and linguistics formalisms

Standards, Resources, Tools

Standards

- Current situation:
  - Resource encoding: TEI, TCF, FoLiA, ...
  - Metadata encoding: DC, CMDI, TEI, ...
  - Tool outputs: tabular, TCF, Python objects, ...
  - Formalisms:
    - PoS tagsets: MULTEXT-East, STTS, UD, ...
    - Similarly for tokenisation, NER, syntactic annotation, ...
- Ideal outcomes (for me):
  - TEI + TCF (prior work in WebLicht)
  - UD + MULTEXT-East (prior work with STTS)

Resources

- Problems:
  - Different standards
  - Problems with finding and accessing
  - Copyright, Privacy protection, Terms-of-Use
- Training corpora (& lexicons):
  - Slovene OK
  - Availability of others?
- Testing and use corpora:
  - EU DGT
  - UD treebank
  - Historical data
  - On-demand creation (TweetCat + TweetGeo, Web crawl)

Tools

- Standards (formats & formalisms)
- Availability (licence, open source?)
- Exposing as web service
- Language independence
- Scalability (speed, max size of resource)

Possible use case: Twitter

- Provide sample datasets for Twitter data for a selection of languages of workshop participants (could use TweetCat / TweetGeo)
- Integrate a preprocessing toolchain for Twitter data for these languages
  - Non-standard tokenisation and sentence splitting
  - Normalisation
  - PoS-tagging and lemmatisation
  - Syntactic parsing
- Enable mounting the preprocessed data to a concordancer
- Undertake a common study and present at CLARIN 2017

User Perspective: WebLicht Tutorial

- Introduction on how to use the developed services
- Aimed at humanities scholars
- Localized into the national languages of workshop participants
  - `https://dhregensburg.files.wordpress.com/2015/07/weblicht-tutorial.pdf`
  - `https://www.youtube.com/watch?v=3RgRCEa6Smo`

Overview of the Workshop

Participants

- Jožef Stefan Institute:
  - CLARIN.SI: Tomaž Erjavec, Darja Fišer
  - ReLDI+Janes tools: Nikola Ljubešić
  - TextFlows: Senja Pollak, Matej Martinc
  - With special thanks to Tina Anžič for organisation
- CLARIN-DE / WebLicht:
  Erhard Hinrichs, Marie Hinrichs, Wei Qiu
- CLARIN Latvia:
  Inguna Skadina, Roberts Dargis, Lauma Pretkalnina
- CLARIN Estonia: Krista Liin
- LINDAT/Lindat: Pavel Stranak (Friday only)
- CLARIN-IT: Riccardo Del Gratta (Friday only)

WS Schedule

- On the Web
- Intro to WebLicht & TextFlows
- Overview of tools and resources (today and tomorrow)
- Today: Discussion
- Tomorrow: Hackaton & drafting the workplan

Practical matters

- OK to put presentations on the Web?
- Lunches at the JSI cafeteria
- Dinner at Špajza

CLARIN Workshop
Multilingual corpus annotation tools: development and integration

INTRODUCTION

Tomaž Erjavec, Darja Fišer

Dept. of Knowledge Technologies,
Jožef Stefan Institute, Ljubljana Dept. of Translation Studies,

Faculty of Arts, University of Ljubljana

CLARIN workshop, Jožef Stefan Institute
Ljubljana, 10th–11th November 2016