Linguistic Services and Resources @ ILC

riccardo.delgratta@ilc.cnr.it

A pipeline for Italian

From plain text to complete linguistic annotations

- NER
- Morpho-syntactic analysis

Based on

1. Freeling
2. Self developed tools (C++)
   a. For managing geographical entities (namely from geonames)
   b. For performing linguistic analysis

A pipeline for Italian: I/O Formats (1)

Freeling

- Input: text
- Output: tabbed (up to pos)

GeoNerD

- Input: tabbed
- Output: augmented tabbed (with geographical informations)

A pipeline for Italian: I/O Formats (2)

Chunker

- Input: tabbed + rule files
- Output: a proprietary format

Ideal

- Input: chunker output + rules
- Output: a proprietary (tabbed) format or a JSON (for NER)

OPENER TOOL(s)

- Linguistic Tools available for different languages
- Up to constituency parser, NER, Opinion Mining, Sentiment Analysis
- Developed in various programming languages, Java, Perl, Python...
- Based on available tools, OpenNLP, Alpino, TreeTagger…
- Trained for Italian by ILC-staff: work on models, rules
- I/O specifics:
  - KAF Kyoto Annotation Format similar to Lexical Markup Language
  - JSON
- Converters from KAF to JSON available
- Checkout at https://github.com/opener-project

Latin Lemmatizer

- Based on LemLat
- Developed in Java
- Provides
  - A web application
  - Rest (GET) services
- Aims at
  - POST services
  - Integration in WebLicht
- Git
  - https://github.com/cnr-ilc/latmorphwebapp

Panacea toolset

- Different linguistic services (similar to opener’s)
- Use soaplab to get services from command line
- Interesting because they offer a SOAP end point through a web application
- Time to restify?
-
- In panacea we provided also lexicons for multiwords in Italian (LFM) in specific domain (e.g environment)