Linked Linguistic Corpus Data

Meritxell Gonzalez, Marta Lanau, Gerlinde Schneider, Christian Fäth

Motivation

Investigate how to model corpus data as linked data in order to link it to other resources

Improve corpus data quality

Enrich our annotations by linking to dictionaries, corpora, taxonomies, any open data, etc.

Datasets

- A subset of OUP’s Komodo English corpus annotated with Lemma and POS
  - +300 documents, CONLL-x format
  - News genre
  - Including some document metadata

- Spanish Reading corpus for Sintaxis Histórica de la Lengua Española (2014, Company Company)
  - +1.200 examples, 13th - 21st century

Tasks breakdown

1) Transform data into CoNLL Format
2) Build a CoNLL RDF Pipeline
3) Link to the OLiA Reference model
    a) Create an appropriate Linking Model to the OLiA Reference model for the AAIF data
    b) Find a Linking Model for the OUP data
4) Link to Babelnet by writing a SPARQL update
5) Do some basic disambiguation by checking the POS Tag, we use the lexinfo linking model and our linking model to do that
6) Extract chunks in other corpora and compare their annotation to improve and enrich our annotation
7) Model metadata with dc:terms (extend it if needed)