Panel #4: Multilingual Corpus transformation into Linked Data

Roberto Navigli

Sapienza University of Rome

Providing semantically-annotated (possibly multilingual) corpora in RDF

MASC

- The Manually Annotated Sub-Corpus (MASC) is a project aimed at the construction of a large-scale corpora of English across different genres
- Sub-corpus drawn from the American National Corpus
- Intended to be useful for machine learning, genre-based analyses, etc.

MASC corpus statistics

- 500K words of written texts and transcribed speech
- 19 genres, among which:
  - Email, spam, blog
  - Letters
  - Newspaper, Journal
  - Movie scripts, travel guides
  - …

MASC into RDF format

- We converted the MASC corpus into RDF format and included additional semantic annotations
- We adopted the **NLP Interchange Format** (NIF) format, a RDF/OWL-based format that aims to achieve interoperability between NLP tools, language resources and annotations

NIF format

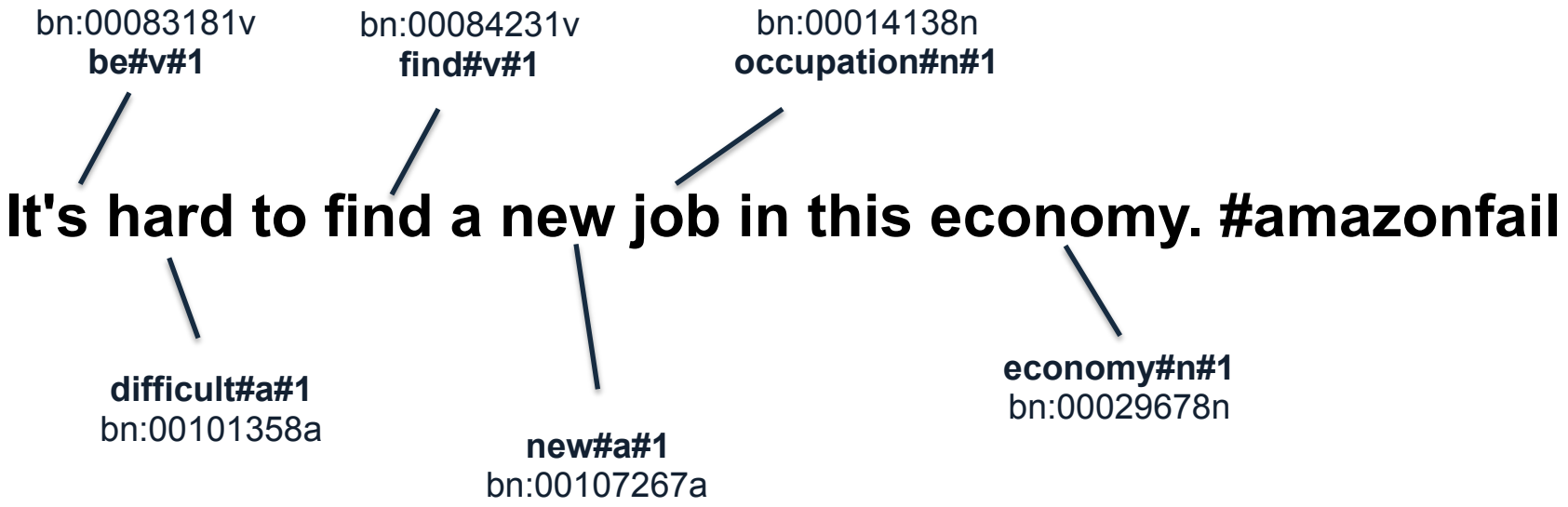
- Reuses existing standards (RDF, OWL 2, PROV…)
- NIF identifiers are used in the Internationalization Tag Set (ITS)
- Provides stable identifiers, persistent hosting, an open license and a community approved

Many corpora have already been converted into this format
(Wiki-Link, Brown, News-100 NER, Reuters-128 NER, KORE 50…)

MASC + BabelNet

[Moro et al., Annotating the MASC Corpus with BabelNet, LREC 2014]

- Automatic annotation, with both named entities and word senses



An excerpt from the RDF file

<**http://lcl.uniroma1.it/MASC-NEWS/written/journal/Article247_328-penn.xml#char=732,737**>

nif:anchorOf "**makes**"^^<http://www.w3.org/2001/XMLSchema#string> ;

nif:**beginIndex** "732"^^<http://www.w3.org/2001/XMLSchema#int> ;

nif:**endIndex** "737"^^<http://www.w3.org/2001/XMLSchema#int> ;

nif:nextWord <http://masc2rdf.com/masc_corpus/written/journal/Article247_328-penn.xml#char=738,741> ;

nif:previousWord <http://masc2rdf.com/masc_corpus/written/journal/Article247_328-penn.xml#char=723,731> ;

nif:referenceContext <http://masc2rdf.com/masc_corpus/written/journal/Article247_328-penn.xml> ;

itsrdf:taIdentRef <**http://babelnet.org/2.0/s00087106v**>

The start and end indices

BabelNet annotation (make#v#1)

The token («makes»)

The MASC document + reference to the token

Some statistics

- **392** files annotated
- **286,416** semantic annotations
- **730** average annotations per file
- **150,227** nouns disambiguated
  82,489 verbs disambiguated
  30,015 adjectives disambiguated
  23,685 adverbs disambiguated