Sharing, preserving and reusing language data in research infrastructures

Koenraad De Smedt

University of Bergen

Trondheim, September 13, 2012

Language data diversity

- Nearly all human communication is in language form: books, newspapers and other printed materials, e-books, websites, blogs, tweets, radio, tv and film, dialog, lectures, etc.
- 130 million books on Google Books (in 2010, i.e. 4% of all books)
- There is an enormous variety of data in/about language: source texts, translated/edited texts, annotated texts, speech and video recordings, transcriptions, concordances, parallel and comparable corpora, dictionaries, word nets, termbanks, grammars, eye tracking data, dialect maps, etc.

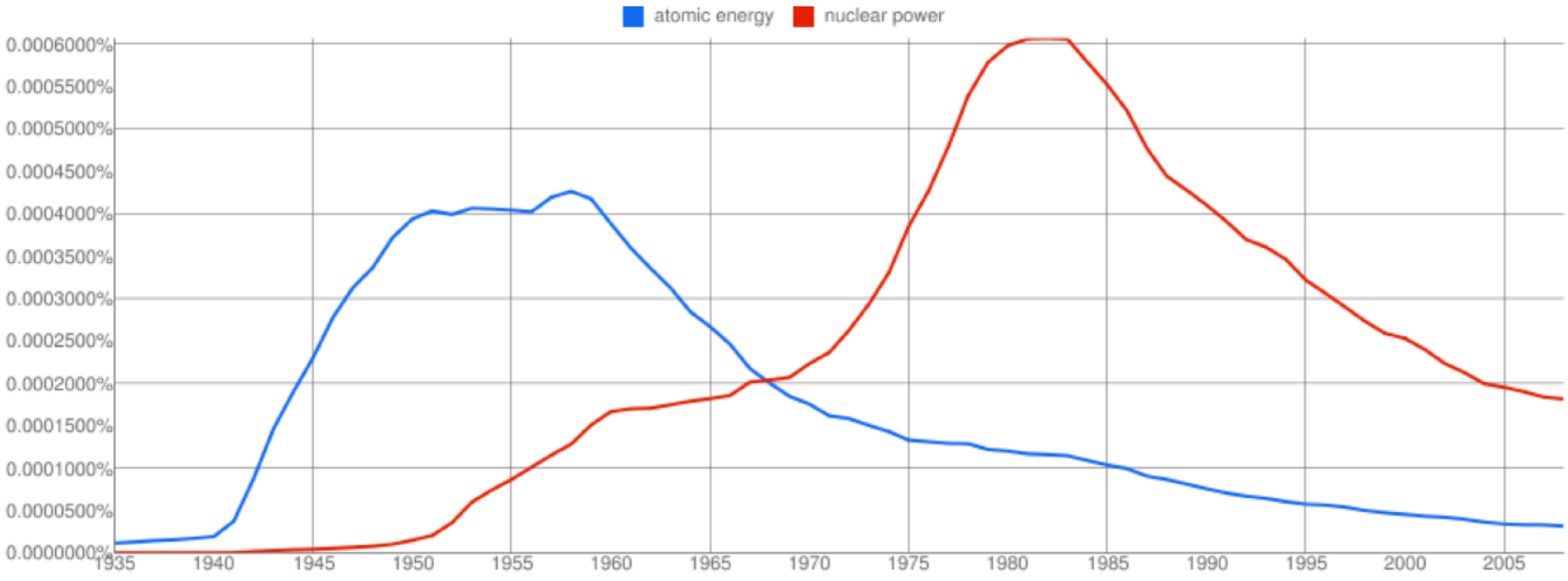
Things you can do with large language resources

Graph these **case-sensitive** comma-separated phrases:

atomic energy,nuclear power

between 1935 and 2008 from the corpus English with smoothing of 4 .

Search lots of books



Search in Google Books:

1935 - 1947	1948 - 1954	1955 - 1956	1957 - 1985	1986 - 2008	atomic energy (English)
1935 - 1964	1965 - 1980	1981 - 1982	1983 - 1999	2000 - 2008	nuclear power (English)

Run your own experiment! Raw data is available for download here.

EC survey on scientific information in the digital age

As for the question of access to research data, the vast majority of respondents (87 %) disagreed or disagreed strongly with the statement that there is no access problem for research data in Europe. The barriers to access research data considered very important or important by respondents were: lack of funding to develop and maintain the necessary infrastructures (80 %); and insufficient national/regional strategies/policies (79 %). There was strong support (90 % of responses) for research data that is publicly available and results from public funding to be, as a matter of principle, available for reuse and free of charge on the Internet.

Data management challenges

- Primary data are often unmanaged, messy, unstructured, and ambiguous
- Annotation is vital (open-ended, many-layered, task dependent, not always reusable)
- Data collections and tools are very scattered in the community, invisible
- There is a multitude of obsolete formats and decaying information bearers
- Good analysis tools are missing, unreliable or not easily adaptable
- Few humanities data collections have a permanent source of funding
- Many original language data are restricted by copyright and privacy concerns

CLARIN mandate

“CLARIN is committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities.”

Usage scenario

- A researcher authenticates at her own organization and creates a “virtual” collection of resources from different repositories
- She finds materials on the basis of browsing a catalogue, searching through metadata, or searching in resource content
- To be granted access to this distributed dataset she signs the appropriate licenses
- She is then able to use a workflow specification tool and process this virtual collection using LT tools in the form of reliable distributed web services which she is authorized to use.
- The data can be added to a repository and can be cited and reused through permanent identifiers (PIDs).

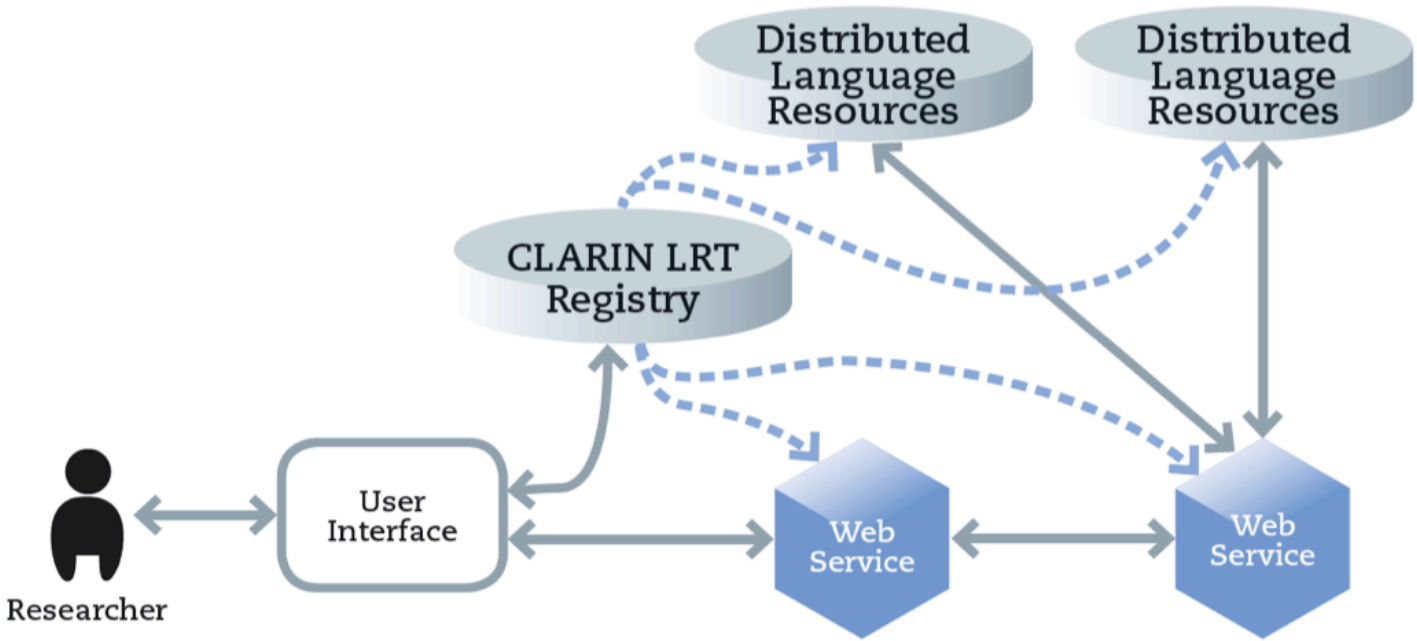
Some eHumanities needs

- Cooperation between stakeholders must lead to increased sharing, visibility and accessibility
- Workflows with tools from different sources, tailoring to data formats and user needs
- Updates and versioning must be supported (extensions to corpora, reparsing, new annotations etc.)
- Rights, licences and privacy must be respected
- Provenance information and persistency must be guaranteed
- Visualizations and user-adapted representations
- Some data is irreplaceable and must be archived and curated securely and indefinitely
- Data must be citable and deserves credit

Infrastructure components

- Trusted repositories for data and services e.g. CLARIN centers
- Metadata catalog for browsing and searching
- Virtual collection registries to store user specified collections and share them
- AAI infrastructure for technical, organizational, legal issues
- Distributed workspaces and workflows
- Persistent identification of resources to make references last

Web services



CLARIN

Common Language Resources and Technologies Infrastructure

Cooperative infrastructure project aimed at managing language data, primarily serving Language Sciences and Humanities

- Most prominent project on ESFRI roadmap within Humanities
- Preparatory phase Jan. 2008 – June 2011 with 24 countries
- Legal entity: European Research Infrastructure Consortium (ERIC) founded on Feb. 29, 2012
- National funding (CLARINO in Norway is funded by RCN)

CLARIN layers

1. Coordination and governance layer (ERIC)
2. Infrastructure layer (long-term national responsibility)
3. Content creation layer (short-term projects by countries, institutions)

Other infrastructure projects may directly or indirectly contribute to CLARIN goals.

CLARIN network formation

- Type A, *Infrastructure* centers
- Type B, *Service* centers
- Type C, *Data and metadata* centers

CLARINO project plan (2012 – 2017)

WP1 Centres Setup

WP2 National Registry and Long-Term Archiving

WP3 Trusted AAI

WP4 Electronic Editions Platform with IDP

WP5 Glossa Integration

WP6 Corpuscle Integration

WP7 Terminology Integration

WP8 Language Analysis Portal

WP9 Tool Adaptation

WP10 Data and Metadata Adaptation

WP11 Management and Dissemination

WP12 Use Cases, Evaluation and Delivery

Links

- CLARIN: clarin.eu
- META-SHARE, sharing of resources: meta-net.eu
- Exploration of syntax and semantics (RCN): iness.uib.no