

CLARIN

Common Language Resources and Technology Infrastructure



CLARIN: a Pan-European Research Infrastructure for Language Resources and Technologies

Martin Wynne

OUCS

OeRC

Linguistics

martin.wynne@oucs.ox.ac.uk

Language Resources and Technologies



- Linguistic corpus (a principled collection of texts sampled to be representative of a particular language variety for the purposes of empirical linguistic research)
- Audio and video corpora
- Lexical resources (wordlists, dictionaries, morphological tables, semantic resources, ontologies)
- Language documentation
- Language processing tools (for annotation, analysis, linking, editing, speech recognition and synthesis, translation, summarisation, text mining, internet search etc)
- Processing environments and workflow management tools
- Other language resources...

LRTs in Oxford



- Linguistic corpora in the Oxford Text Archive (OTA), e.g.:
 - Old English Corpus
 - British Academic Writing in English
 - Parsed Corpus of Early English Correspondence
 - COMIC – Commerical Italian
 - Lancaster Corpus of Chinese
 - The Electronic Text Corpus of Sumerian Literature
- British National Corpus
- Audio resources in the Phonetics Laboratory (IViE)
- Other scholarly research (in ModLang, Comlab, Linguistics, Psychiatry, Oriental Studies, English, History, etc.)
- Other?



- Founded 1976 by Lou Burnard
- Collecting and distributing then-rare electronic texts
- Developing guidelines and good practice (e.g. TEI)
- Central to humanities computing
- 1996-2008 Arts and Humanities Data Service - AHDS Literature, Languages and Linguistics
- c. 1400 collections; mostly text, but some images, audio, video, websites
- Literary, linguistic and related humanities disciplines
- Some outdated legacy data
- Many text encoding formats
- TEI XML metadata
- Several levels of access restrictions

The problems with LRTs



- Many archives known only to certain communities
- Archives are mostly unconnected, and data difficult to find
- Every archive has its own standards for storage and access
 - usually only simple retrieval of files (text, audio or video documents)
- Not sufficient incentives to share resources
- Resources are in different formats, follow different standards, are described in differing ways
- Tools are hard to use for non-specialist
- Tools and data are not available for online processing
- Many researchers are not aware of the potential benefits of using language and speech technology tools

The CLARIN vision



A researcher in Oxford from his desktop computer can:

- single sign-on with local authentication
- search for, find and obtain authorization to use corpora in Oxford, Prague and Bergen
- select the precise dataset to work on, and save that selection
- run semantic analysis tools from Budapest and statistical tools from Tübingen over the dataset
- use computational power from the local or national computing centre where necessary
- save the workflow and results of the analysis, and share those results with collaborators in Paris, Vienna and Zagreb
- Discuss, annotate and iteratively tweak and re-run the analyses with collaborators

The CLARIN Mission



- what?
 - create a research infrastructure that makes language resources and technologies (LRT) available to scholars of all disciplines, especially humanities and social sciences
- how?
 - unite existing digital archives into a federation of connected archives with unified web access
 - provide language and speech technology tools as web services operating on (language) data in archives

This represents the first coordinated and comprehensive attempt to address the technical, legal, administrative and financial barriers to the effective use of LRTs in academic research.

Why now?



- Maturity of language resources and technology field
 - well-established data centres
 - large number of widely used resources and tools
 - established networks of collaboration
 - wide agreement that there are current technical, legal and administrative barriers that need to be addressed
- Lots of restructuring and rebuilding going on
- Growing number of digitisation projects producing potentially useful language resources, but outside of the linguistics communities
- Existing and emerging international initiatives for generic computing and research infrastructure (access and authorisation, digital libraries, Grid)
- ESFRI Roadmap and FP7 infrastructure funding schemes

What is CLARIN?



- CLARIN stands for
 - Common Language Resources and Technology Infrastructure (for the Humanities and Social Sciences)
- European Seventh Framework Program (FP7) Research Infrastructure project
 - started 1st January 2008
- 3 phases
 - preparatory phase
 - 2008-2010: planning, building a prototype
 - budget: 4.1 M€ from EC, ??? M€ from participating countries
 - construction phase
 - 2011-2015: build and populate with resources and tools
 - exploitation phase
 - 2016-: CLARIN in full service
- Overall budget (2008-2020): ca 200 M€

Who we are?



- CLARIN consortium
 - 32 partners from 22 EU and associated countries
- CLARIN community
 - 140-odd members in 32 countries
- leading partners include:
 - Utrecht University (*Steven Krauwer, coordinator*)
 - Max Planck Institute Nijmegen (*Peter Wittenburg*)
 - Hungarian Academy of Sciences (*Tamás Váradi*)
 - Oxford University (*Martin Wynne*)
 - Tübingen University (*Erhard Hinrichs*)
 - Helsinki University (*Kimmo Koskiennemi*)
 - *plus many more*

CLARIN technical work



Promoting collaboration and interoperability between European language resource repositories, particularly in relation to:

- Persistent identifiers
- Component metadata
- Trust domains
- Long-term Preservation and Access
- Service centres
- Virtual collections
- Standards and best practices
- Concept registry services

See the CLARIN Short Guides at <http://www.clarin.eu/>

Oxford's role in CLARIN



- Oxford Text Archive (originally as part of the AHDS) has been one of the main architects of the initiative, and is a member of the project consortium, receiving c. €300k in the Preparatory Phase. Martin Wynne is UK national representative, sits on the Executive Board, and is responsible for liaison with DARIAH.
- OTA participate in work packages relating to:
 - Mapping the environment and building links with relevant bodies and initiatives in the Humanities;
 - Designing the technical infrastructure;
 - Developing OTA as a centre in the infrastructure;
 - Obtaining national support;
 - Co-ordinating the project and the network.

OTA in CLARIN: things to do



1. XSLT script for TEI -> DC and TEI -> OLAC (and future CLARIN common metadata) crosswalks
2. Enhance metadata with new categories and corrections
3. Implement Shibboleth access layer
4. Register PIDs
5. Review and revise EULA
6. Review and revise deposit licence
7. Evaluate, prioritize and enhance collections

CLARIN in the Oxford landscape



- OTA as a node in a federated research data management landscape in Oxford (with focus on specific types of language resources)
- Trailblazer in the development of infrastructure services (e.g. implementing access via Shibboleth, PIDs, metadata mappings)
- Gateway to language resource collections and technology services in other institutions
- Assisting in promoting interoperability of other digital collections in Oxford with international infrastructure initiatives (e.g. collections in OULS, OUP, etc.)

CLARIN

Common Language Resources and Technology Infrastructure



Thank you for your attention

CLARIN has received funding from
the European Community's Seventh Framework Programme
under grant agreement n° 212230

DARIAH



- Another project for the Humanities, in the ESFRI roadmap, and funded by the EC as a preparatory phase infrastructure project
- Supporting digitisation, preservation, sharing and re-use of research data across the Arts and Humanities
- Led by DANS, a national data centre in the Netherlands
- UK participants:
 - CeRch, King's College, London
 - Archaeology Data Service, University of York
 - Oxford University Computing Services
- Started 1st September 2008, running for 2 years