CLARIN: a pan-European research infrastructure for language resources

Martin Wynne
Martin.wynne@it.ox.ac.uk

Oxford e-Research Centre &
IT Services (formerly OUCS) &
Faculty of Linguistics, Philology and Phonetics,
University of Oxford

UCREL Seminar
Lancaster
Monday 23rd June 2014
Summary

• Many areas of linguistics (and related disciplines) have been thoroughly transformed by digital data, tools and methods
• Many areas of the humanities and social sciences are currently experiencing the digital turn (with great differences in the pace and impact of changes)
• One aspect of the digital turn across is the introduction of digital language technologies and resources across a number of disciplines
• There are numerous barriers to the successful deployment of these resources in real-life research scenarios
• CLARIN is an attempt to address these problems and more fully realize the potential for the use of language resources (across the humanities and social sciences)
• CLARIN is building the infrastructure to underpin, support and sustain research
• Is there a way forward for CLARIN in the UK?
Corpus Linguistics
Interoperability and sustainability for digital textual scholarship

Well-known problems with digital resources in the humanities of:
- fragmentation of communities, resources, tools;
- lack of connectness and interoperability;
- sustainability of online services;
- lack of deployment of tools as reliable and available services

There is a potential solution in distributed, federated infrastructure services.
CLARIN in a nutshell

WHAT IS CLARIN?
• European Research Infrastructure Consortium (ERIC)
• Node in the ESFRI roadmap
• Set of national infrastructures
• Network of technical centres, data providers, researchers, developers
• Community of researchers using language resources and tools to do new forms of digital research in the humanities and social sciences

WHO?
• Austria, Bulgaria, Czech Republic, Denmark, Estonia, Germany, The Netherlands, Norway, Poland, Dutch Language Union.
• Croatia, Finland, France & Sweden in the process of joining
• Contacts and activities in all other European countries

WHEN?
• Preparation: 2008-2011
• Construction: 2011-2015
• Operation: 2013-
• Long-term perspective

WHAT DOES CLARIN DO?
• Providing language data and tools as services
• Linking repositories
• Linking tools and data
• Training, advice, support
• Promoting the use of digital text and speech among scholars across the humanities and social sciences

HOW MUCH?
• Preparatory phase c. €4m from FP7 to 2011
• Varying national infrastructure investments including c. €1m p.a. for central coordination
• Project funding (including closed calls for research infrastructures)

LINKS WITH THE UK?
• Key partners in the preparatory phase (Lancaster, Oxford, Sheffield)
• Oxford Text Archive registered as a CLARIN Centre
• Director of User Involvement (Martin Wynne, p/t secondment)
• But important resources, tools, services, centres of expertise not integrated (yet?)
The CLARIN Vision

For digital research with language resources and tools, to facilitate the transition

...from...

• Language data in digital silos
• A (highly advanced) cottage industry of scholarly tool development
• Short-term project funding
• A panoply of formats and standards
• A multitude of licences
• Ad hoc arrangements for access and collaboration

...to...

• A network of technical centres supporting infrastructure services
• Professional tool development
Highly available, persistent and sustainable services
Interoperable data and tools
Secure cross-border federated identity management
Training, support and advice
A researcher in Vienna, from his desktop computer, can:

- Log in via a single sign-on, with local authentication, and then:
- search for, find and obtain authorization to use resources in Oxford, Prague and Berlin
- select the precise dataset to work on, and save that selection
- run grammatical analysis tools from Sheffield, semantic analysis tools from Lancaster and statistical tools from Tübingen over the dataset
- use computational power from the local, national or other computing centre where necessary
- obtain advice and support for carrying out all technical and methodological procedures
- save the workflow and results of the analysis, and share those results with collaborators in Paris, Edinburgh and Zagreb
- Collaborators can adapt and re-run the analyses and share and discuss results
Zukunft

in Zukunft soll die Produktion von synthetischen Fasern gesteigert werden
evtl. hat Zukunft eventuell lässt eine günstige Entwicklung erwarten ist
folgerichtig entsprechend
diese Technologie hat (keine) Zukunft

Sprachwissenschaft Futur
in dem Satz sich versteckt steht (sich) in der Zukunft

Dazu
- Zukunft ist Sicht, Zukunftssicht, Zukunftsbild, Zukunftsforschung, Zukunftsmöglichkeiten, Zukunftswissenschaft
- Zukunftsforschung, Zukunftskonzepte, Zukunftslösungen, Zukunftswirtschaft, Zukunftsperspektiven, Zukunftspolitik, Zukunftstransformation, Zukunftstheorie
- Zukunftstheorien, Zukunftstrends, Zukunftsperspektiven, Zukunftsforschung, Zukunftsfaktoren, Zukunftspotenzial

ähnliche Zukunft

Zukunft

Version: 0.4.23 | Quelle: WGD | Artikeltyp: Vollartikel | Kompakt | Details

Deutsches Textarchiv (wöchentlich aktualisiert)

Treffers: 4823

<table>
<thead>
<tr>
<th>KWIC</th>
<th>Datum</th>
<th>Datum</th>
<th>Zufällig</th>
<th>Links</th>
<th>Rechts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1905</td>
<td>1905</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1902</td>
<td>1902</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Andere Zukunft

Die Zukunft ist keine Welt, sondern eine Perspektive. (S. 42)

Deutsche Zukunft

Die Zukunft ist kein Traum, sondern ein realer Ort, der bereit ist, sich zu entfalten. (S. 43)

Kernkorpus 20

Treffers: 13444, davon anzeigbar: 11790

<table>
<thead>
<tr>
<th>KWIC</th>
<th>Datum</th>
<th>Datum</th>
<th>Zufällig</th>
<th>Links</th>
<th>Rechts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>1999</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Allerdings beginnt die Zukunft mit jedem Augenblick neu

Verantwortung Zukunft

Die Zukunft ist immer schon dabei, als wir jetzt hier stehen und überlegen, wie wir die Welt gestalten möchten. (S. 44)

Verantwortung Zukunft

Die Zukunft ist immer schon dabei, als wir jetzt hier stehen und überlegen, wie wir die Welt gestalten möchten. (S. 44)

Verantwortung Zukunft

Die Zukunft ist immer schon dabei, als wir jetzt hier stehen und überlegen, wie wir die Welt gestalten möchten. (S. 44)
<table>
<thead>
<tr>
<th>token ID</th>
<th>tokens</th>
<th>POSTags</th>
<th>lemmas</th>
<th>namedEntities</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Damals</td>
<td>ADV</td>
<td>damals</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>in</td>
<td>APPRK</td>
<td>in</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Odessa</td>
<td>NE</td>
<td>Odessa</td>
<td>B-GPE</td>
</tr>
<tr>
<td>14</td>
<td>war</td>
<td>VAFIN</td>
<td>sein</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>os</td>
<td>PPER</td>
<td>os</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>sehr</td>
<td>ADV</td>
<td>sehr</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>kalt</td>
<td>ADJD</td>
<td>kalt</td>
<td></td>
</tr>
</tbody>
</table>
Real life gotchas

- I need a reference corpus of modern popular fiction in order to trace norms and deviations from norms in the literary texts which I am researching.
- I've constructed a small corpus and I want to POS-tag it, add semantic annotation and my own stylistic tags, then search and explore the corpus and all of the annotations.
- I want to view concordances, collocations, wordlists, keywords making use of the standardised (e.g. TEI) annotations in my corpus.
- I want to use BNCweb, BYU-BNC, Intellitext, SketchEngine, the Oxford English Corpus, the Bank of English and a number of specialized corpora for my research and teaching, and all of my students and collaborators need access too.
- I want to use Antconc in a class in the computing lab and need to show the technicians the licence under which it is made available.
- I want to do semantic tagging and analyse the results ... in French.
- I want to get some useful visualizations of my linguistic data.
In fact, for which real-life research question, who is likely to be able to access all of the following?

- The relevant corpus of texts in a reliable, well documented, easily accessible and usable form
- Wordclass tagging or the tools to do it (and other forms of tagging)?
- Tools to exploit the corpus and annotations
- Metadata to interpret results, make internal comparisons etc
- A set of relevant comparable reference corpora allowing comparison with other genres, times, places?
- Training on how to use the data and tools (which I will be able to use when I get home)
- An existing literature, body of knowledge, academic community with whom to share, publish, dispute, build on results?
'Research centres' and 'Service centres'
### Research Centres
- Research projects
- Creation of resources and tools
- Creation of interfaces
- Disseminating results
- Community engagement
- Obtaining and managing research funding

### Service Centres
- Hosting (systems administration, maintenance, updates)
- Data curation and long-term preservation
- Software repository services
- Access and identity management
- Career development for data scientists, digital specialists, technical staff
- Long-term funding commitments

Research and development, transition to service
Research infrastructure

• Research and research infrastructure – the railway metaphor

• Passengers, trains, tracks, rail companies, stations, ticketing, payment systems, security, standards, customer support, international aspects, etc.

• Lots of analogies to be made but the key point is that in order to make a journey you don't want to have to design build and run a railway system (or negotiate your own way through a maze of different services, or hike through the wilderness)
What added value could CLARIN bring to a centre like CASS?

- Model and expertise for the separation, clarification and professionalization of infrastructure functions
- Providing services for wider pool of users
- Potential additional funding for infrastructure services
- Opportunity for embedding in long-term e-infrastructure
- Opportunity to play a leading role in UK and European e-infrastructure
- Increased opportunities for international collaborative research projects
CLARIN and the UK

- No official CLARIN national initiative,
- No UK official response to invitations to join the consortium
- No single agency takes responsibility for possible participation – no 'Ministry' or 'National Funding Agency'

But

- Lots of existing funding for digital research, resources, centres of expertise, generic e-infrastructure, etc.
- A loose network of researchers, developers and resource creators, and clarin-uk@jiscmail.ac.uk
- Strong desire from CLARIN and European funders for UK to join!
Summary

• Many areas of linguistics (and related disciplines) have been thoroughly transformed by digital data, tools and methods
• Many areas of the humanities and social sciences are currently experiencing the digital turn (with great differences in the pace and impact of changes)
• One aspect of the digital turn across is the introduction of digital language technologies and resources – this is a great opportunity for us!
• There are numerous barriers to the successful deployment of these resources in real-life research scenarios
• CLARIN is an attempt to address these problems and more fully realize the potential for the use of language resources (across the humanities and social sciences)
• CLARIN is building the infrastructure to underpin, support and sustain research
• Is there a way forward for CLARIN in the UK?
'CLARIN for Beginners'
http://blogs.it.ox.ac.uk/martinw/2014/04/11/clarin-for-beginners/

'Silos or Fishtanks?'
http://blogs.it.ox.ac.uk/martinw/2012/04/06/silos-or-fishtanks/

'The Role of CLARIN in Digital Transformations in the Humanities'
Martin Wynne
DOI: 10.3366/ijhac.2013.0083
Edinburgh University Press 2013
Thank you