



# Can Linguistics Lead a Digital Revolution in the Humanities?

**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

Digital Humanities Seminar  
Nottingham  
Wednesday 13th November 2012

# Digital Humanities and (corpus) linguistics



The digital age brings the potential for fundamental transformations in the way that we do research in the humanities, but the revolution is slow in coming.

What do the transformations in research practice brought about by corpus linguistics teach us about the perils and the pleasures of the digital turn?

# Summary



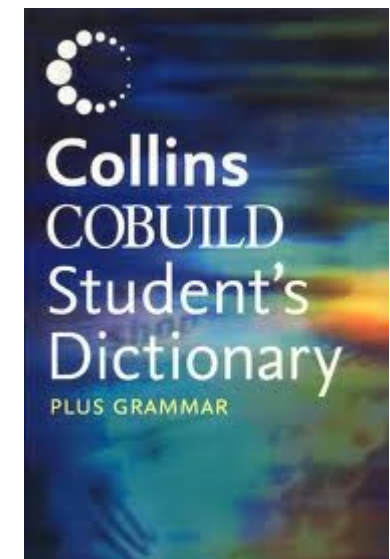
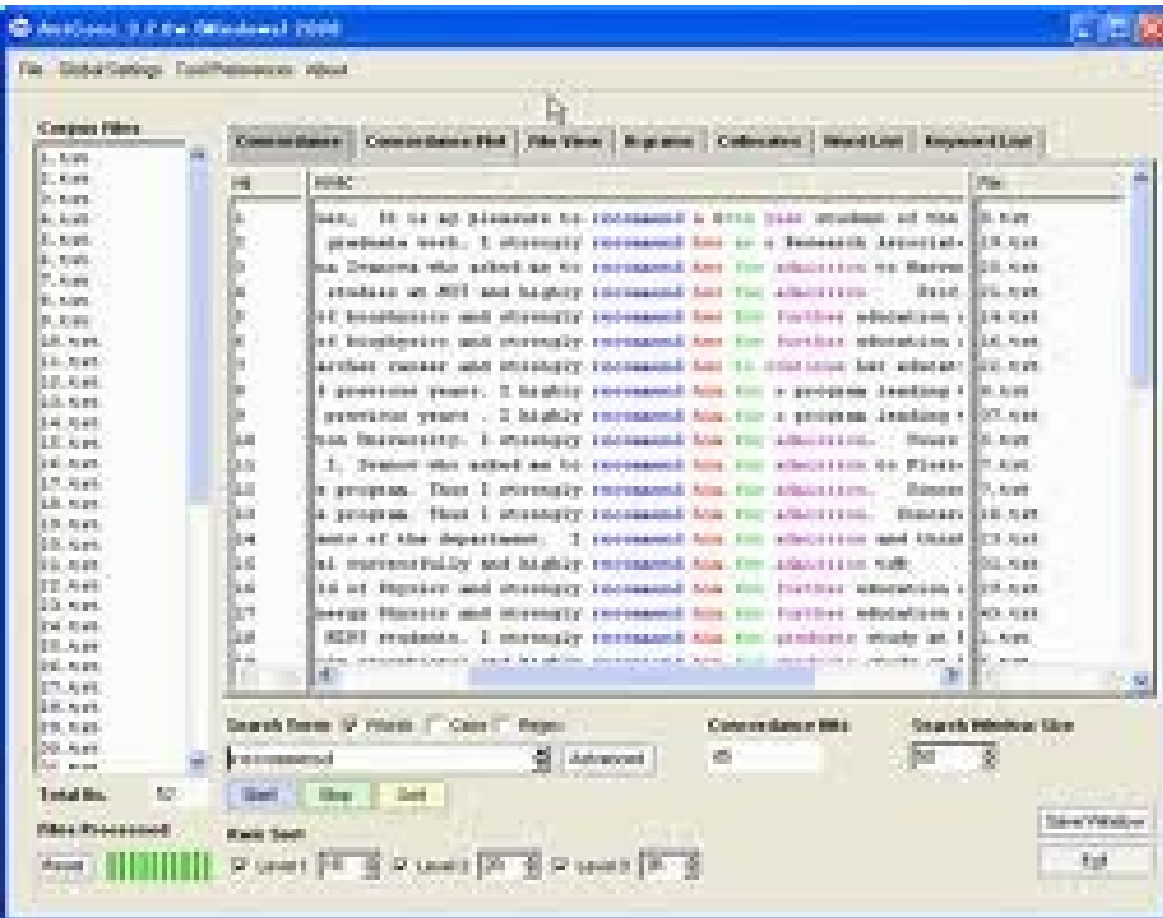
- Many areas of linguistics and some related domains have been thoroughly transformed by digital data, tools and methods
- New research questions are being asked, new communities, sub-disciplines, journals, conference series, etc. and interesting interdisciplinary activities and communities
- There are still some sticking points and problems
- 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)
- There is huge potential for other disciplines in the humanities to be transformed by digital data, tools and methods, but the revolution has not yet happened
- A dangerous opportunity: take the evidence-based, empiricist route and make the humanities more like the sciences
- Remaking the humanities after the digital turn does require some changes and compromises
- How do we go forward?

# Now...




...some observations to back up these assertions

# Corpus Linguistics



I CAME

**THE BANK OF ENGLISH** is a unique computer database which monitors and records the way in which the English language is actually used in the modern world. It is continually expanding and contains over 300 million words, from contemporary British, American, and international sources: newspapers, magazines, books, TV, radio, and real life conversations - the language as it is written and spoken today.



**BANK of ENGLISH**



**CPM** Centro de Processamento de Português Moderno

Home / Sobre / Recursos / Publicações / Contato

**Diário de Notícias**

O Centro de Processamento de Português Moderno é um projecto que visa a ser desenvolvido por métodos: verbos, nomes próprios e comuns e termos.

O projecto trabalha com o texto do verbos de um subcorpus de textos do CPM: verbos, nomes próprios e comuns e termos.

Na fase seguinte procedeu-se à análise dos verbos de actividade e à análise de 122 verbos que não se encontravam analisados no corpus textual anteriormente estudado: verbos, nomes e termos e termos.

Importante é a base de dados CPM - [Corpus de Textos de Notícias, Verbos](#) - feita a nível de análise de verbos de um subcorpus de textos do CPM: verbos, nomes e termos e termos.

Podem ainda ser consultados os dados de verbos do CPM.

• [Verbos](#)  
• [Nomes](#)  
• [Termos](#)

Para mais informações em [CPM](#).

**CORPUS.BYU.EDU**

seven online corpora | 45 - 450 million words each

corpora	resources	queries	history	researchers	publications	register	questions?	contact us
---------	-----------	---------	---------	-------------	--------------	----------	------------	------------

These corpora were created by Mark Davies, Professor of Linguistics at Brigham Young University. They have many different users, including: finding out how native speakers actually speak and write; looking at language variation and change; finding the frequency of words, phrases, and collocations; and designing authentic language teaching materials and resources.

The corpora are used by more than 100,000 people each month (more than 100,000 visits), which makes them perhaps the most widely-used corpora currently available. They also serve as the basis for an increasing number of publications by researchers from throughout the world.

In addition to the regular corpus interfaces listed below, there are also many new COCA-based resources, such as [www.worldandPhrasenet.org](#), [www.worldfrequency.info](#), [www.ngm.org](#), and [www.academicworldinfo.org](#), all of which allow you to download large amounts of corpus data for offline use. Note especially the new [130,000 integrated word list](#) from COCA, CDNC, BNC, and SOA - the largest, corrected frequency list of English.

English	# words	language / dialect	time period	compare to:
---------	---------	--------------------	-------------	-------------



# 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.

# Silos or fishtanks??



*Let's talk about fishtanks rather than silos...*

*There are lots of fishtanks out there, some very elaborate, big, pretty...*

*But they're all in different places and unconnected.*

*And if I want to keep a fish I have to build a fishtank (or put it in yours)...*

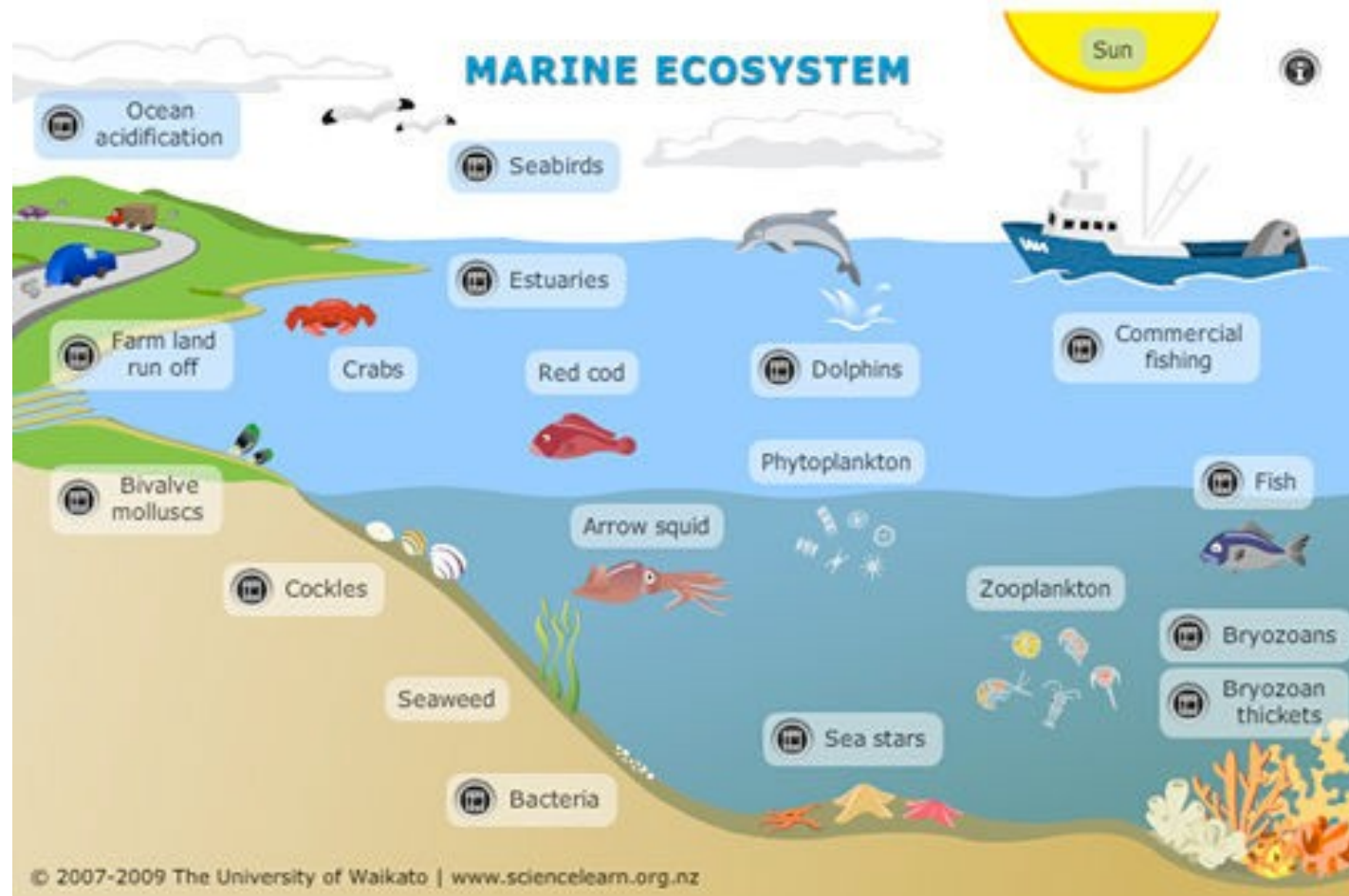
*And who's going to carry on feeding the fish?*

*Let's not all make our own fishtanks.*





Wouldn't it be better to have an **ecosystem** where we can all set our fishes free?



*You can access all of the riches of the deep and it's a lot easier to get into fish research*













# The CLARIN Vision



A researcher in Vienna, from his desktop computer, can:

- do 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 semantic analysis tools from Budapest 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
- discuss and iteratively adopt and re-run the analyses with collaborators



## Annual CLARIN Meeting 2013 in Prague

Annual CLARIN meeting The 2013 Annual CLARIN meeting will take place in Prague, Czech Republic (see information on the venue). The National

[Read more >](#)

📱 @Twitter

[CLARINERIC](#)

"'Be patient' is not a good answer for an infrastructure." -Dieter van Uytvanck at the CLARIN Annual Meeting 2013 [#clarineric](#)"

[3 days 6 hours ago.](#)



## Welcome to CLARIN!

CLARIN is the short name for the Common Language Resources and Technology Infrastructure, which aims at providing easy and sustainable access for scholars in the humanities and social sciences to digital language data (in written, spoken, video or multimodal form) and advanced tools to discover, explore, exploit, annotate, analyse or combine them, independent of where they are located. To this end CLARIN is in the process of building a networked federation of European [data repositories](#), service [centres](#) and centres of expertise, with [single sign-on](#) access for all members of the academic community in all participating countries. Tools and data from different centres will be interoperable, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work.

# Digital Transformations



- Evidence-based / data-driven / empirical research
- Real-time collaboration, wide dissemination, crowdsourcing
- Linking data: comparison, cross-searching, data mining, geo-mapping
- Linking publications with data
- Beyond the text: images, audio, video, geographical data, simulations, *in silico* experimentation
- And many more!

# How can digital data transform research?



- Finding new research questions  
(*use data at the hypothesis-forming stage of your research*)  
*“Where's a good place to look for interesting discussions of the state in the early modern period, and what do we find there?”*
- Asking new research questions  
(*use data at the analysis stage of your research*)  
*“What sort of grammatical change took place in British English across the twentieth century?”*
- Answering old questions with data more quickly, on a bigger scale, more authoritatively, comparisons to different datasets, etc.  
(*use data at the analysis stage of your research*)  
*“How much free indirect speech does Jane Austen use in her novels?”*
- Re-examine old questions which were formulated in the absence of systematic use of data  
(*use data to replicate, test and extend research findings*)  
*“Is it true that Jane Austen invented free indirect speech in the novel?”*



# But...



- How do you read a million books?
- How do you reconcile the deep knowledge of texts, and close reading of them, with broad brush overviews, statistics, digests and trends?
- Will scholars take short cuts, and do worse research?
- Have we got the infrastructure to support work like this in the humanities?



# Three barriers to the digital revolution



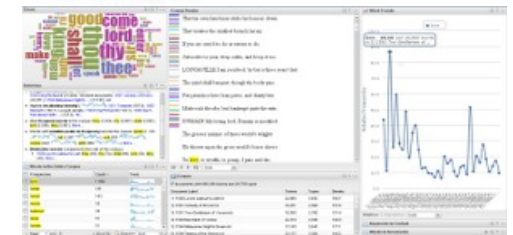
1. Technical legal and administrative barriers, and a lack of connectivity in a fragmented environment (*silos* and *fish tanks*)
2. Promoting a new discipline of 'digital humanities'
3. Methods: the methods and traditions of the Humanities make it difficult to do *e-Research*

# What are the *Digital Humanities*?



Possible answers:

- A distinct academic discipline;
- An interdisciplinary community;
- A set of resources, methods and tools;
- An infrastructure to support digital research in the humanities;
- An umbrella term for advanced digital research embedded in a humanities discipline.





Home

News & Events

People & Projects

ICT Methods

Divisions & Units

Training

Support

About

You are here: Home

## Welcome to Digital.Humanities@Oxford

This site provides a central information point about activities, resources and facilities in the digital humanities at Oxford. Use the tabs above or the following links to find out about [People and Projects](#), [Support](#) and [Training](#), and [News and Events](#). **People** and **projects** can also be viewed by [ICT Methods](#) and [Divisions and Units](#) in the University. [About](#) provides information about the site as a project, and other aspects of the digital humanities at Oxford.

### Featured Project



**Reel to Real** is an archival sound website at the Pitt Rivers Museum, funded by the Esmée Fairbairn Foundation Collections Fund. From the songs of children's games in playgrounds across Europe to Bayaka women's songs that enter people's dreams in the rainforests of the Central African Republic, [this website](#) offers an introduction to the several

thousand hours of archival sounds held by the Museum. It also includes information about the field recordists, their related collections, and a host of other resources such as films and the work of contemporary


### Latest News

#### Peter Millican spots J.K. Rowling

Peter Millican (Philosophy) has demonstrated the power of his digital Signature Stylometric System by showing signs of J.K. Rowling's style in her pseudonymously published crime novel *The Cuckoo's Calling*.

Read [More News...](#) Or see the list of [Current Events](#) in [digital humanities at Oxford](#). Oxford users can add events by [logging in](#).



Follow us on Twitter: 





## THE PUBLIC GOOD

IN EVERYONE'S INTERESTS: WHAT IT MEANS TO INVEST IN THE HUMANITIES



[HOME](#) [ABOUT](#) [RESEARCH](#) [OPPORTUNITIES](#) [WHAT'S ON](#) [PEOPLE](#) [MULTIMEDIA](#) [CONNECT](#)



### NEWS AND EVENTS

Dr Hakim Adi speaks to the Race and Resistance Network  
Jed Fazakarley reports on the group's first seminar of Michaelmas Term 2013

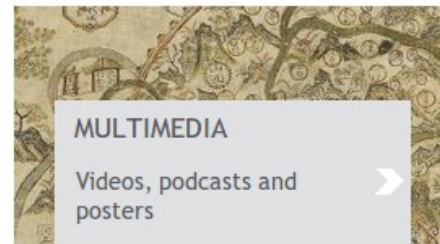


"Mind the Environmental Gap" Symposium



### RESEARCH

Networks, programmes, projects and more



### MULTIMEDIA

Videos, podcasts and posters



Welcome to the website of The Oxford Research Centre in the Humanities (TORCH).

TORCH is a major new University of Oxford initiative that seeks to stimulate and support research that transcends disciplinary and institutional boundaries.

TORCH was launched on May 12, 2013. More details about the launch and a brief video are

# What is digital scholarship in the Humanities?



These are some issues and assumptions in e-Science – do they apply in the Humanities?

- Consensus (and compromise) about funding priorities
- Adoption of technical standards
- Standards for the representation of knowledge and interpretations (agreement on concepts and categories!)
- Reproducibility and replicability of research
- Sharing of generic tools
- Curation of tools and data in professional service centres
- Support for software sustainability
- Promotion of interoperability of resources and tools
- Sharing research outputs
- Research leading to an accumulation of knowledge
- Increasingly data-driven research



# In defence of the enlightenment



"[There is] a monolithic conception of social space, according to which it would suffice to have the right information to make the right decisions. But in point of fact, information itself is far from homogenous and no purely quantitative approach is satisfying. Having ever greater amounts of information at our fingertips not only does not make us more virtuous, as Rousseau already predicted, but it does not even make us more knowledgeable."

[Tzvetan Todorov, *In Defence of the Enlightenment*, 2009]









# Steering a difficult path



- *One extreme*: Digital Humanities is like e-Science: data-driven, empirical, evidence-based, practical, based on shared facilities, tools, resources and methods
- *Another extreme*: it is in the intrinsic nature of the Humanities that we should constantly question the basic received ideas and categories, and therefore we cannot expect to have shared assumptions and methods
- Can Digital Humanities steer a route in between?

# What did linguistics do?



- More computationally advanced research often makes more sense in answering computer science research questions
- Humanistic research shedding new light on language in use is often very simple in technical terms and doesn't take full advantage of the potential for transformation (and is sometimes wrong)
- There is a tendency to justify corpus linguistics in terms of its utility outside of linguistics - in lexicography and language learning, text mining, information processing, and to support other disciplines
- (The lessons are confused because linguistics is interdisciplinary)



# The simple challenge then...



... to transform the Humanities by promoting shared digital services, facilities, resources and tools, without destroying the justification and arguments for *the Humanities for the Humanities sake*, and thus accidentally contributing to the decline and eventual destruction of civilization

# Read more...



## **'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

International Journal of Humanities and Arts Computing 7.1-2 (2013):  
89–104

DOI: 10.3366/ijhac.2013.0083

Edinburgh University Press 2013