

Introduction to CLARIN Tools and Resources for Digital Data and East Asian Languages

Martin Wynne

Martin.wynne@bodleian.ox.ac.uk

Bodleian Libraries &
Faculty of Linguistics, Philology and Phonetics,
University of Oxford

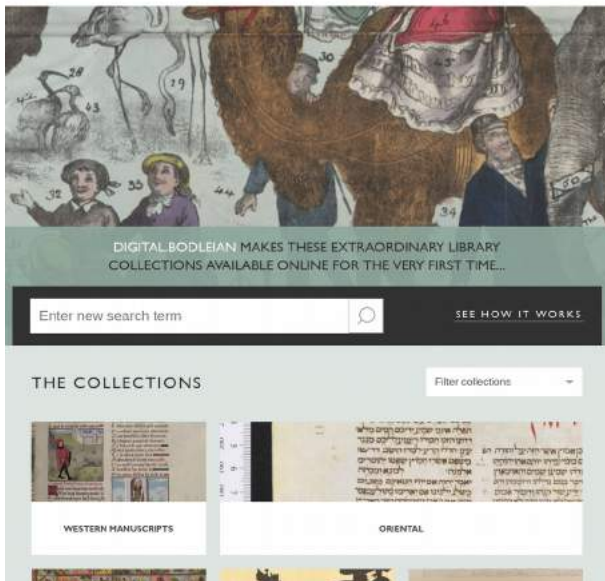
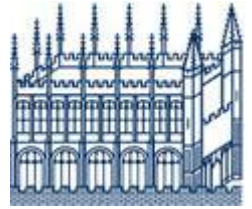
National Coordinator, CLARIN-UK

How to Analyse Large Volumes of Online Text

Université Libre de Bruxelles

Friday 13th October 2017

Oxford Text Archive, 40 years on and in a new home

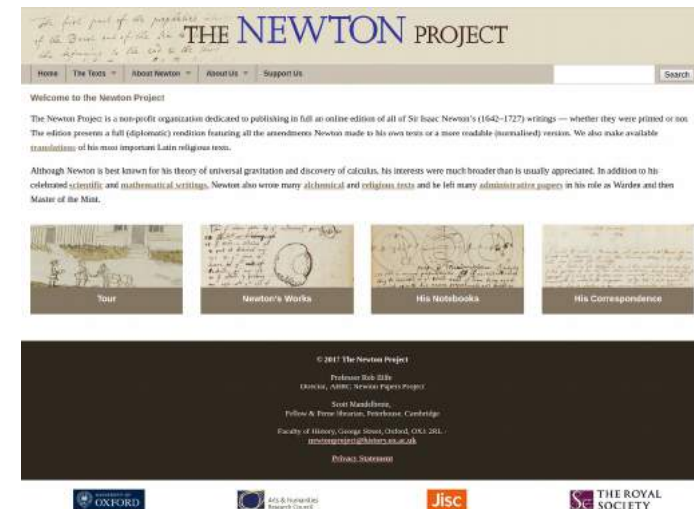
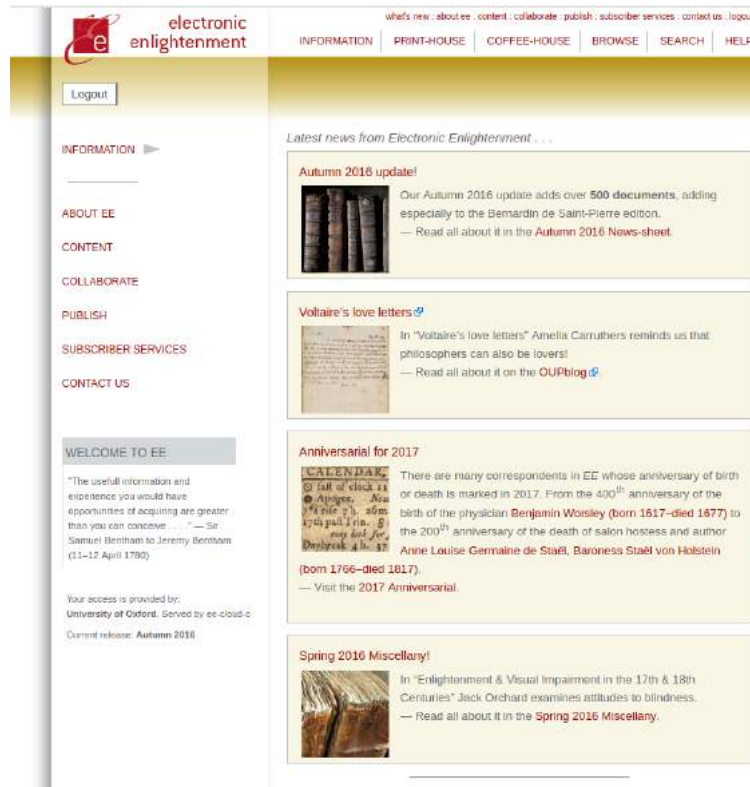


Networking the Republic of Letters, 1550-1750

Established in 2005, and now in our **fourth phase** (April 2017 – September 2018), we are a collaborative, interdisciplinary research project based at the University of Oxford with funding from **The Andrew W. Mellon Foundation**. We are using digital methods to reassemble and interpret the correspondence networks of the early modern period.



The Letter Carriers by *Amelie Story* for Cultures of Knowledge (with inspiration from *Pier Francesco Mida*)



Duell-ease A worde with. valiant spiritts shewing the abuse of duells, that valour, refuseth challenges and priuate combates. sett foorth by G.F. a defendour of Christian valoure.

“

Please use the following text to cite this item or export to a predefined format:

”

BIBTEX

CMDI

G. F., defendour of Christian valoure., 2013, *Duell-ease A worde with. valiant spiritts shewing the abuse of duells, that valour, refuseth challenges and priuate combates. sett foorth by G.F. a defendour of Christian valoure.*, Oxford Text Archive, <http://hdl.handle.net/20.500.12024/13791>.

Share:

f

t

g+

✎ Authors	G. F., defendour of Christian valoure.
📅 Date issued	2013-12
📄 Type	Text
🚩 Language(s)	English
🏢 Publisher	University of Oxford
👤 Collection(s)	Early English Books Online (Phase 2)

Show full item record

📁 Files in this item

⬇️ Download all files in item (428.28 KB)

This item is **Academic Use** and licensed under:
Text Creation Partnership

Name

Size

Format

Description

A00503.epub

73.94 KB

Unknown

Version of the work for e-book readers in the EPUB format

📄

01001
10011

Download file

Name

Size

Format

Description

A00503.html

168.64 KB

HTML

Version of the work for web browsers

🌐

Download file

Preview


Name

Size

A00503.xml


174.1 KB

📄



Bodleian Libraries

UNIVERSITY OF OXFORD



UNIVERSITY OF OXFORD

What can you do?

DEPOSIT

CITE

Browse

> All of the Repository

My Account

Login

Statistics

Statistics **BETA**

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

Help Desk

3



Logout



immensely
resource for
2017 Anniversary
A listing of some of those
correspondents whose
birth or death is marked
in 2017.

... read it now!



Bodleian Libraries
UNIVERSITY OF OXFORD



Your access is provided by:
University of Oxford. Served by
ee-cloud-c

Current release: Autumn 2016

Electronic Enlightenment — letters & lives online

... reconnecting the first global social network!

Electronic Enlightenment is the most wide-ranging online collection of edited correspondence of the early modern period, linking people across Europe, the Americas and Asia from the early 17th to the mid-19th century — reconstructing one of the world's great historical "conversations".

A **subscription** to *EE* will give you immediate access to **70,057** historical documents. Listen in on the first global social network as **8,560** historical figures discuss everything from religious tolerance to animal rights, vulcanology to classical archeology, economic modelling to celebrity culture.

EE update Autumn 2016

Electronic Enlightenment Scholarly Edition of Correspondence

Now over 70,000 items of edited correspondence! Our latest update adds over **500 documents**, adding especially to the Bernardin de Saint-Pierre edition.

Electronic Enlightenment Biographical Dictionary

With over 100 additions and revisions our Biographical Dictionary now has over 8,500 entries!

For an overview of the update please read the latest **News-sheet**.



Not sure where to start? Then why not:

- take a **Guided tour**;
- **Browse** the list of correspondents;
- read our latest **News-sheet**;
- brush up on our **Fact sheet**.

Outline

1. What are *natural language processing* and *text analysis*?
2. What is CLARIN?
3. Focus on tools for analysing social media and CMC
4. Focus on tools for East Asian languages

Outline

1. What are *natural language processing* and *text analysis*?
2. What is CLARIN?
3. Focus on tools for analysing social media and CMC
4. Focus on tools for East Asian languages

Exploring digital text

- 1) Linguistic approaches e.g. discourse analysis,
- 2) 'Qualitative text analysis' in the social sciences, and
- 3) Computational approaches, e.g. sentiment analysis, topic modelling)

Corpus Linguistics

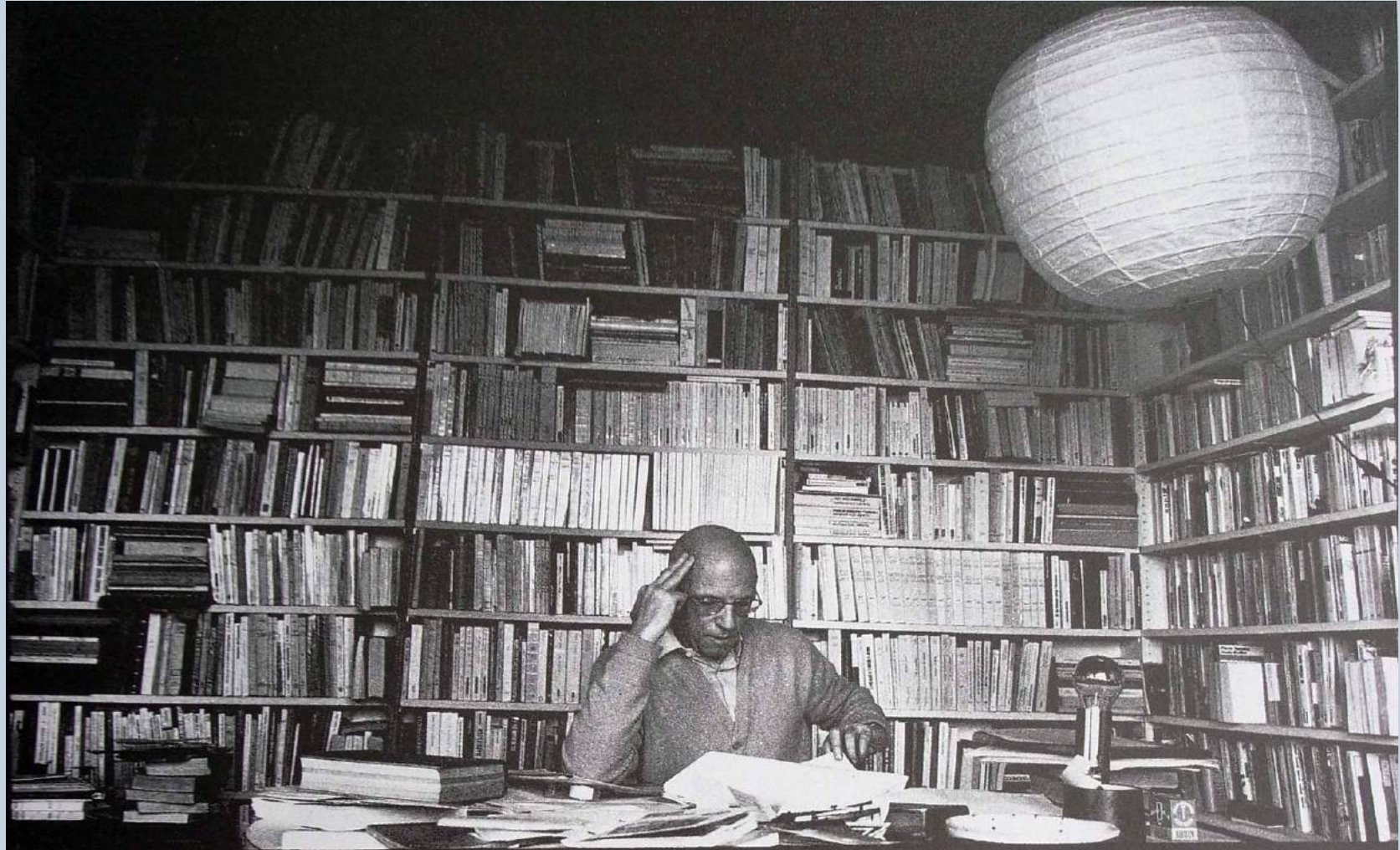


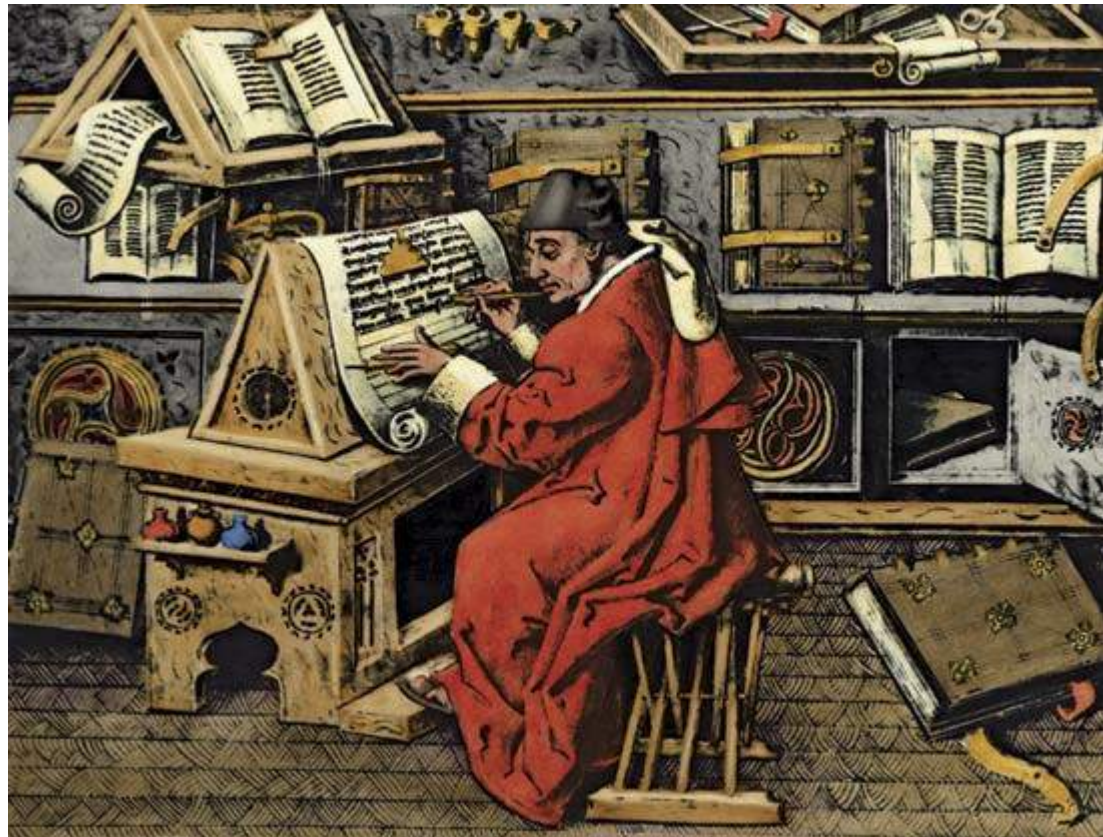
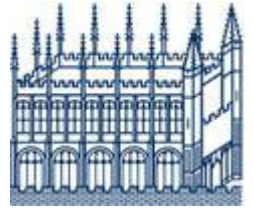
Your query "pineapple" returned 75 matches in 41 different texts (in 1,202,214,511 words [44,422 texts]; frequency: 0.06 instances per million words) [0.00 seconds]

< << >> > Show Page: 1 Line View Show in random order New query Go!

No	Filename	Solution 1 to 50	Page 1 / 2
1	A02495	high land , great rocks of Alabaster , great woods , and	Pineapple trees lying along within the ground , which by report have lien
2	A02495	into the country , and found that the woods were fur ,	pineapple , alder , yew , withy , and birch : Fair woods
3	A02495	as grass . The timber is most Firr , yet plenty of	Pineapple trees : few of these two kinds meet to mast a ship
4	A02758	Haslenut , Filberts , Almonds bitter and sweet , Chestnuts ,	Pineapple , and Fisticknut . IN the beginning of this chapter , we
5	A03069	the Latins Strobilum , because the fruit of it something resembles the	Pineapple . The Frenchmen call it Alticocalum , of the Arabic article All
6	A03069	it Alticocalum , of the Arabic article All , and Cocalos a	Pineapple , whereof it is corruptly called Artichault , in Italian and Spanish
7	A03362	and desert , is the same laudable and commended . The head	Pineapple formed , after the condition of a sharp upright Pillar , in
8	A03362	be big and round , but the upper part sharp to a	Pineapple fashioned : dooth argue in that person , an unshamefastness , irefulness
9	A03362	of bigness , doth argue a good wit naturally . The head	Pineapple sharp , to be unshamefast , and a boaster . The head
10	A03742	mean to hold the lore . Storms rifest rend the sturdy stout	pineapple te . Of lofty ruing towers the false the feller be .
11	A05054	good to eat , and do serve for almost all purposes ye	Pineapple kernels do serve for . 4 The leaves chewed are good for
12	A05054	heat , and ach being applied . 14 The root taken with	Pineapple kernels in quantity of a dram , is good against the coughing
13	A05054	good to be laid upon old sores . Drink one dram with	Pineapple kernels , to help the cough , spetting of blood and matter
14	A05237	thereof arises round and sharp upward as a top , or a	Pineapple with the point upward . And such a point is called Conon
15	A05237	lower . The colour is green , the fruit not unlike the	Pineapple : but with a more finer order of scales : when it
16	A05569	because of a certain resemblance which the fruit hath with the	Pineapple . It comes out of the Province of Sancta Croce , first
17	A05569	upon it along by the shore there are some trees , like	Pineapple trees : from thence to Mozambique are twelve miles , and to
18	A07650	Elm , the Olive , the prickie Chestnut , & the high	Pineapple , one amongst another ; whose bodies were bound about with green
19	A07834	Sepulcher of the Emperor Adrian , upon the top whereof was the	Pineapple of brass , which before I said was since placed in the
20	A08548	after he was dead , did entomb his corpse near unto a	Pineapple tree , wherein he did engrave an Epitaph . Cap . 2
21	A09010	which at the first appearing , is like unto a Cone or	Pineapple , and afterwards opening it self , spreads into many branches ,
22	A09010	they be blown or separated , very like to a Cone or	Pineapple , and begin to flower below , and so upwards by degrees
23	A09010	sweet smelling Moly of Mompelier . 10 Moly sertinum Coniferum . The late	Pineapple Moly . depiction of flowers bulbosum , or bulbed Violet before described
24	A09010	to be misliked . 11 . Moly serotinum Coniferum . The late	Pineapple Moly . This late Moly that was sent me with the last
25	A09010	thick thrust together , fashioned very like unto the form of a	Pineapple (from whence I gave it the name) of the bigness
26	A09011	but lean down to the 6 . Iacea pumila Narbonensis .	Pineapple headed Knapweede . 7 . Iacea Liciniata alb Jagged white Knapweede .
27	A09011	thick and long . 6 . Iacea alia pumilae Narbonensis . The	Pineapple headed Thistle or Knapweede . This small French plant grows sometimes in
28	A09011	many scales sharp up to the top like unto a	Pineapple the ends of whose scales are long straight sharp

Close Reading



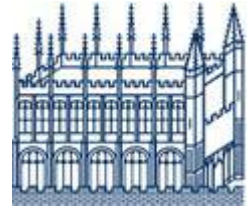




Scalable Reading

dedicated to DATA: digitally assisted text analysis

SEARCH



...the broad circumference
Hung on his shoulders like the Moon, whose Orb
Through Optic Glass the Tuscan Artist views
At Ev'ning from the top of Fesole,
Or in Valdarno, to descry new Lands,
Rivers or Mountains in her spotty Globe.
(*Paradise Lost*, 1. 286-91)

SCALABLE READING

LATEST ENTRIES



UNCATEGORIZED

What is a digital combo?

by MMUELLER on Apr 25, 2017 • No Comments

How should an old book live in the digital environment of the 21st century? My answer is "as a digital combo that brings together three data streams, each a surrogate that represents and contextualizes aspect of the original object. Call them the bibliographical, material, and textual streams. This scrawny diagram illustrates their interaction in the...

HUMANITIES IN A DIGITAL AGE

Whither TEI? The Next Thirty Years

by MMUELLER on Sep 20, 2016 • 2 Comments

In the next fifty years the entirety of our inherited archive of cultural works will have to be re-edited within a network of digital storage, access, and dissemination (Jerome McGann, 2001) You have to put the corn where the hogs can get at it (Bill Clinton) Only the paranoid survive (Andrew Grove) Introduction At...

EEBO-TCP

RECENT POSTS

What is a digital combo?

Apr 25, 2017

Whither TEI? The Next Thirty Years

Sep 20, 2016

Freebo, Free Lunch, and Crowdfunding New EEBO Images

Dec 19, 2015

New release of Shakespeare His Contemporaries

Nov 24, 2015

Hannah, Kate, and Lydia at work

Oct 27, 2015

CATEGORIES

crowdsourcing (12)

data curation (15)

EEBO-TCP (1)

Humanities in a Digital Age (4)

Close, distant
and scalable
reading, or
digitally
assisted text
analysis

Martin
Mueller,
Northwestern
University

Using natural language processing

- **What** is my data: can I work with an existing datasets, or do I need to monitor and capture new streams of data? What contextual information can I preserve, and how?
- **Where** do explore my data: can I do my research on the online platforms, or do I upload my data to an analysis platform or engine?
- **How** to I analyse my data: do I need to install and use software locally? (And **who** will do it?)
- **Share**: how will I record my methods, and share my data and code?

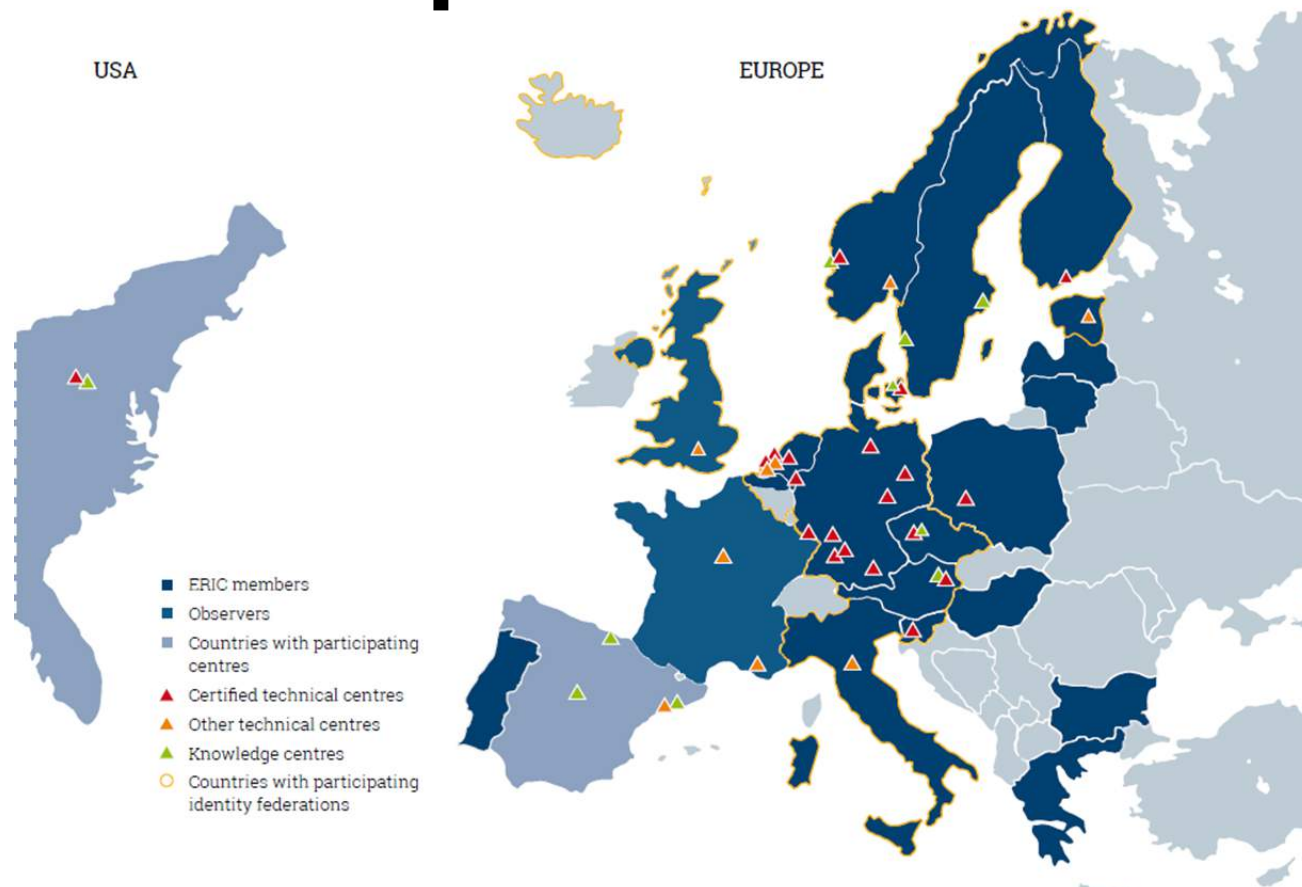
Outline

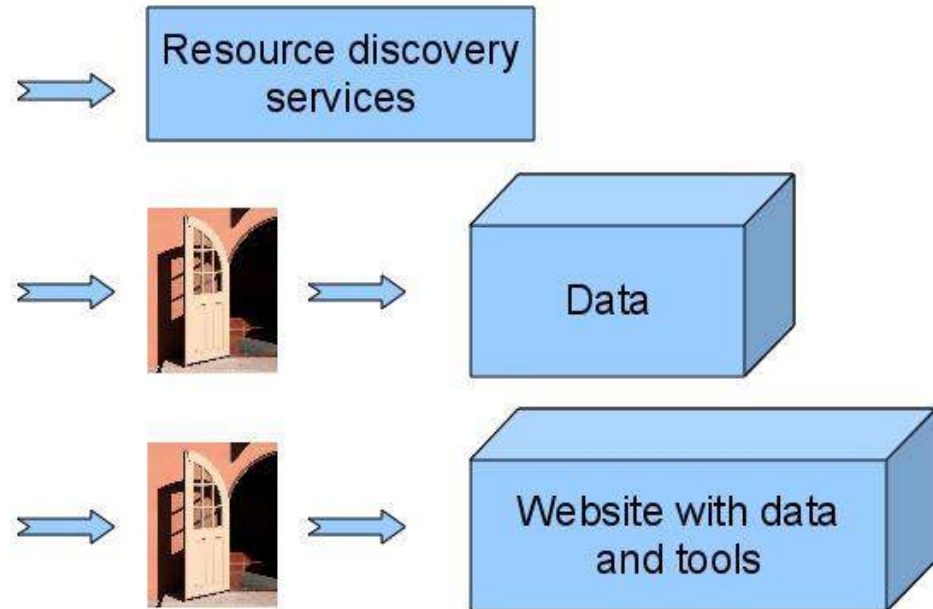
1. What are *natural language processing* and *text analysis*?
- 2. What is CLARIN?**
3. Focus on tools for analysing social media and CMC
4. Focus on tools for East Asian languages

CLARIN in five bullets

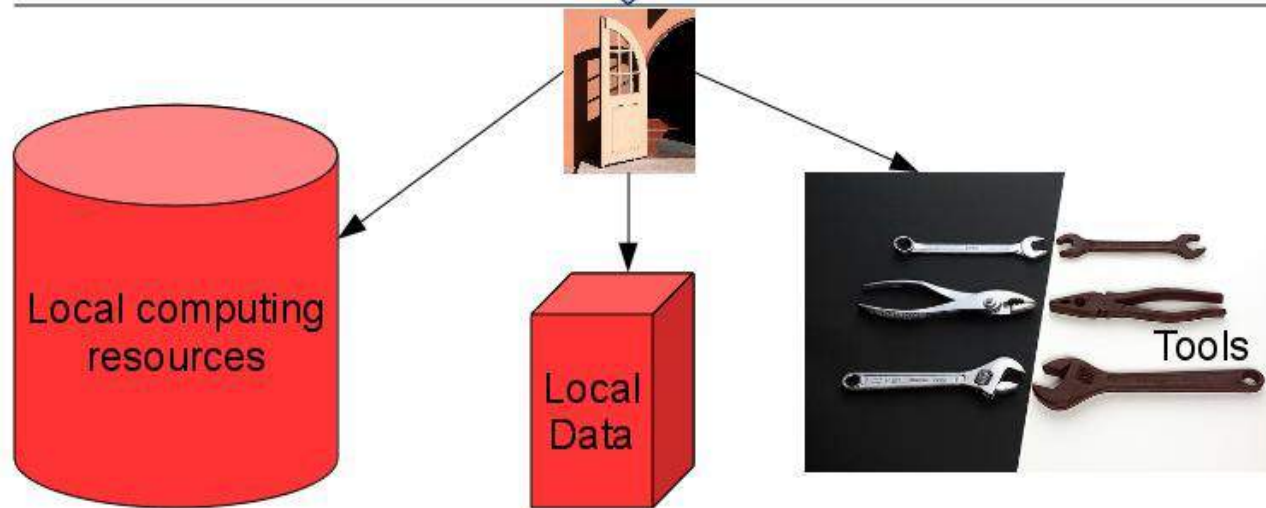
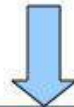
- **CLARIN** is the Common Language Resources and Technology Infrastructure;
- it provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form),
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** online environment.

CLARIN ERIC: 19 members, 2 observers, 1 associated partner

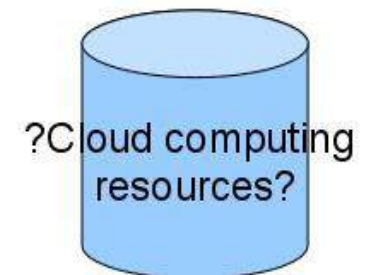




Single sign on?



Advisory services?



The CLARIN Vision

A researcher in Brussels, from his desktop computer, will be able to:

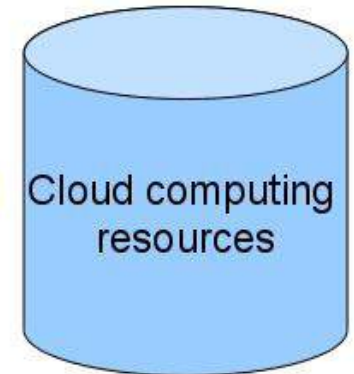
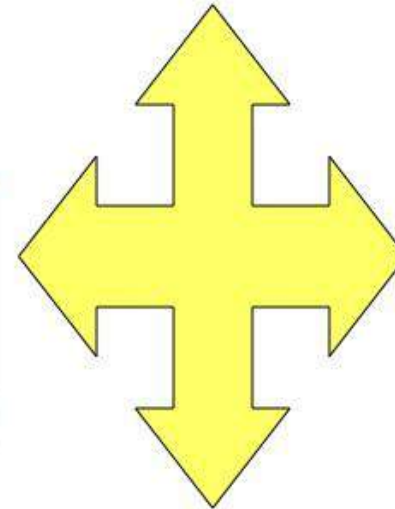
- log in locally at their local institution,
- search for, find and obtain authorization to use resources in Oxford, Prague and Berlin,
- select the precise dataset to work on, and save that selection,
- run semantic analysis tools from Budapest and statistical tools from Tübingen over the dataset,
- use computational power from local, national or other computing centres (if and when necessary),
- obtain advice and support for carrying out all technical and methodological procedures,
- save the workflow and results of the analysis in a citable form,
- share the results with collaborators in Paris, Edinburgh and Zagreb,
- discuss online with collaborators,
- iteratively adapt and re-run the analyses.



Advisory services



Resource discovery services



Outline

1. What are *natural language processing* and *text analysis*?
2. What is CLARIN?
- 3. Focus on tools for analysing social media and CMC**
4. Focus on tools for East Asian languages

CLARIN-PLUS workshop "Creation and Use of Social Media Resources"



Thursday, 18 May, 2017 - 09:00 to Friday, 19 May, 2017 - 14:30

Goals of the workshop

Background

With the increasing volume and impact of communication on social media, social media analysis has become one of the most trending topics in natural language research, which can be observed in a growing number of workshops and conferences dedicated to this topic, projects funded, and research centers established. As a result, a number of social media resources containing chats, online commentaries, reviews, blogs, emails, forums, etc., as well as audio and video recordings, have been accumulated in the repositories of CLARIN centers. What is more, due to their distinct communicative characteristics, they pose new technical challenges for the standard natural language processing tools as well as new legal and ethical challenges for the dissemination of such resources, which has also been addressed by CLARIN, making the available infrastructure an important means for attracting new users to the CLARIN community.

Aims

The aims of the workshop are: to demonstrate the possibilities of social media resources and natural language processing tools for researchers with a diverse research background who are interested in empirical research of language and social practices in computer-mediated communication; to promote interdisciplinary cooperation possibilities; to initiate a discussion on the various approaches to social media data collection and processing.

CLARIN-PLUS receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 676529.



Long-term vision

- easy access to social media material
- services suited for CMC data can easily be found and employed
- encourage researchers to develop and address discipline-specific hypotheses and scholarly questions

Challenges and multidisciplinary potential

Social media data sets are considered a rich data type

- suited for both *close reading* and *distant reading*
- messy or noisy data
- links with data in other modalities than text
- context-dependent meaning

Social media data sets have a big potential for reuse and repurposing within many fields of study in the humanities and social sciences (and beyond):

- *Humanities*: language variation and change, discourse analysis, ...
- *Social sciences*: social and cultural dynamics, political sciences, economics, ...

Overview of corpora (1/2)

Lang	Name of corpus	Data types	Size	Period
German	Dortmund Chat Corpus	Chats	1,06m	/
German	DEREKO subcorpus	News & German Wiki	670m	/
German, English	Monitor corpus of tweets from Austrian users	Tweets	30-40m	2007-2017
German	DWDS subcorpus – Blogs	Blogs	102m	/
Estonian	Mixed Corpus: New Media	Forums, chats, comments	25m	2000-2008
Finnish	Suomi 24	Forums	2,600m	2001-2016
Lithuanian	LITIS v.1	News	190k cmnts	2010-2014
Dutch	SoNaR New Media Corpus	Tweets, chats, SMS	35m	2005-2012
Dutch	Flemish online teenage talk	Facebook, Whatsapp	2.9m	2015-2016
Welsh	Corpus of Welsh Language Tweets	Tweets	7m tweets	/

Overview of corpora (2/2)

Lang	Name	Data types	Size	Period
French	CoMeRe Repository	Emails, forums, chats, tweets, Wiki, etc.	75-80m	Various
Italian	Web2Corpus_it	Forums, Blogs, Newsgroups, social networks, chats	/	/
Slovenian	JANES	Slovene CMC	200m	2013-2016
Norwegian, English, French	NTAP climate change blog corpora	Blogs related to climate change	5,000m	2000-2014
Polish	Corpus Highly Emotive	Tweets	160m	/

Problems

- Missing metadata
 - Unknown temporal span for e.g. DWDS, DEREKO, Dortmund Chat, Corpus of Welsh Language Tweets
 - Unknown annotation process for DWDS, LITIS, Corpus of Welsh Language Tweets
- Licensing
 - Unclear for most of the surveyed corpora

Datasets

- 17 datasets identified
 - by language:
 - 9 different languages (cz, dk, el, de, it , es ,se, si, uk)
 - 1 multilingual
 - most for Slovene (6), English (3) and Italian (3)
 - by data type:
 - Tweets (10)
 - Facebook comments (2)
 - mixed (3)
 - blogs (1)
 - Reddit (1)
 - by task: sentiment analysis (5), NER (1), entity linking (1), rest miscellaneous
- 8 of these integrated in the CLARIN infrastructure

Tools

- **Within the CLARIN infrastructure:**

- GATE tools (CLARIN-UK)
- JANES tools (Clarin.si)

- **Elsewhere:**

- Hunaccent (Hungarian)
 - Accentizer of Hungarian text
- Twython (language-independent)
 - Python wrapper for the Twitter API
- dmi-tcat
 - A set of tools to retrieve and collect tweets from Twitter for statistical analysis
- Tweet NLP
 - A tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets, along with annotated corpora and web-based annotation tools.

Outline

1. What are *natural language processing* and *text analysis*?
2. What is CLARIN?
3. Focus on tools for analysing social media and CMC
4. **Focus on tools for East Asian languages**

Software for the Analysis of East Asian Languages

Martin Wynne, University of Oxford
martin.wynne@bodleian.ox.ac.uk



言語分析のための
あなたのソフトウ
ェアツールについ
て教えてください
い！

Please help CLARIN build an
overview of LRTs for
Japanese, Chinese, and other
East Asian languages. Get in
touch, or add a comment to:

<http://bit.ly/CLARINEastAsian>

Name	Language	Purpose	Availability	License	Comments
Stanford Word Segmenter	Chinese	tokenization	free download	GNU Public	The Stanford Word Segmenter is incorporated i
GATE	Chinese, Korean	Plug-ins for numerous functio	free download	LGPLv3	These are plug-ins for the GATE software envir
USAS	Chinese	semantic tagging	free online	none (limited online interface	Maximum 3000 characters input
TreeTagger	Chinese	POS tagging, lemmatization		ACA	
SPPAS	Mandarin Chinese, Taiwanese	audio annotation	free download	PUB	
Prozed	Chinese	prosody editor	free download	CC BY-NC-SA 4.0	
Voyant	Japanese	Analysis, exploration, visuali	free online	none (online service)	Suite of text analysis tools, now works with Jap
topic-modelling-tool	Japanese	topic modelling	free download	not known	A point-and-click tool for creating and analyzing
i2ocr	Chinese, Japanese, Korean,	OCR	free online	none (online)	
Converio	Chinese, Japanese, Korean,	OCR	10 free pages, then pricing p	none (online)	Not tested
KoNLPy	Korean	POS tagging, corpus analysis	free download	GPL v3	Versions exist for Linux, Mac, Windows
awesome-korean-nlp	Korean	A curated list of resources dedicated to Natural Language Processing for Korean			
pycantonese	Cantonese	Python corpus search functio	free download	Apache License, Version 2.0	Also https://github.com/pycantonese/pycantone
Antconc	Not language specific	Text and corpus analysis	free download	custom licence	Widely used with Japanese and Chinese. See
SegmentAnt	Chinese, Japanese, Korean,	tokenization	free download	custom licence	Multi-platform (Windows/Mac/Linux). A freeware
UDPipe	Japanese	tokenization, tagging, lemma	Online service with restricte	UDPipe is free software dist	Appears to be restricted to use with a specific d
On-line Chinese Tools	Chinese	tokenization, encoding detect	Variable: some free to download		Portal offering links to various reference, pedag
MeCab	Japanese	tokenization, morphological	not known	not known	
ChaSen	Japanese	tokenization	not known	not known	Documentation in Japanese
Language Grid	Japanese	web services orchestration		not known	Portal for a number of NLP web services,
janome	Japanese	tokenization, POS tagging	free download	Apache License, Version	Japanese morphological analysis engine v

Your help to improve and enlarge the catalogue of tools is needed!

Please email suggestions to:
martin.wynne@bodleian.ox.ac.uk
clarin@clarin.eu

(Subject: *East Asian language resources*)

Links

This presentation: <http://www.slideshare.com/martin-wynne/CLARIN-CMC-EastAsian> [to be confirmed]

<http://ota.ox.ac.uk/>

<http://www.e-enlightenment.com/>

<http://www.clarin.eu/>

<https://www.clarin.eu/event/2017/clarin-plus-workshop-creation-and-use-social-media-resources>

<https://www.clarin.eu/content/clarin-for-researchers>

<http://cass.lancs.ac.uk/>

<https://cqpweb.lancs.ac.uk/>

<https://scalablereading.northwestern.edu/>

<https://programminghistorian.org/>

<http://bit.ly/CLARINEastAsian>