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
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:

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)

Files in this item

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

Name: A00503.epub
Size: 73.94 KB
Format: Unknown
Description: Version of the work for e-book readers in the EPUB format

Name: A00503.html
Size: 166.64 KB
Format: HTML
Description: Version of the work for web browsers

Name: A00503.xml
Size: 174.1 KB

Download all files in item (428.20 KB)
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A subscription to EE will give you immediate access to 70,057 historical documents.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
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 1,202,214,531 words [44,422 texts]; frequency: 0.06 instances per million words) in 0.00 seconds.

<table>
<thead>
<tr>
<th>No</th>
<th>Filename</th>
<th>Solution 1 to 50</th>
<th>Page 1 / 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A02495</td>
<td>pineapple trees lying along the ground, which by report have lien</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A02495</td>
<td>pineapple, alder, yew, withy, and birch: Fair woods</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A02495</td>
<td>pineapple trees: few of these two kinds meet to mast a ship</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>A02794</td>
<td>Haslenut, Filbert, Almonds bitter and sweet, Chestnuts,</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>A03069</td>
<td>the Latin Strobilum, because the fruit of it something resembles the</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>A03069</td>
<td>pineapple, of the Arabian article All, and Coccoles a</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>A03362</td>
<td>pineapple, formed, after the condition of a sharp upright Pillar, in</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>A03362</td>
<td>pineapple, fashioned: dooth argue in that person, an unshamefamness, irefulness</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>A03362</td>
<td>pineapple sharp, to be unshamefast, and a boaster. The head</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>A03742</td>
<td>pineapple mean to hold the lore. Storms finest rend the sturdy stout</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>A05554</td>
<td>pineapple kernels do serve for. 4 The leaves chewed are good for</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>A05554</td>
<td>pineapple kernels in quantity of a dram, is good against the coughing</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>A05554</td>
<td>pineapple kernels, to help the cough, stinging of blood and matter</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>A05567</td>
<td>pineapple with the point upward. And such a point is called Como</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>A05567</td>
<td>pineapple but with a more finer order of scales: when it</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>A05567</td>
<td>pineapple It comes out of the Province of Santa Croce, first</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>A05567</td>
<td>pineapple trees: from thence to Mozambique are twelve miles, and to</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>A07690</td>
<td>pineapple, one amongst another; whose bodies were bound about with green</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>A07694</td>
<td>pineapple of brass, which before I said was since placed in the</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>A07694</td>
<td>pineapple tree, wherein he did engrave an Epitaph. Cap. 2</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>A07694</td>
<td>pineapple, and afterwards opening it self, spreads into many branches.</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>A07694</td>
<td>pineapple, and begin to flower below, and so upwards by degrees</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>A07694</td>
<td>pineapple sweet smelling Moly of Mompheer. 10 Moly sermitum Coniferum. The late</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>A07694</td>
<td>pineapple to be misliked. 11. Moly sermitum Coniferum. The late</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>A07694</td>
<td>pineapple Moly. This late Moly that was sent me with the last</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>A07694</td>
<td>pineapple ( from whence I gave it the name ) of the bigness</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>A07694</td>
<td>pineapple headed Knapweed. 7. Lacea Lictiniana alb Jagged white Knapweede.</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>A07694</td>
<td>pineapple headed Thistle or Knapweede. This small French plant grows sometimes in</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>A07694</td>
<td>pineapple the ends of whose scales are long, straight, sharp</td>
<td></td>
</tr>
</tbody>
</table>
Close Reading
Scalable Reading
dedicated to DATA: digitally assisted text analysis

...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 scrappy 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

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?
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
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.
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"

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

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

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


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！
<table>
<thead>
<tr>
<th>Name</th>
<th>Language</th>
<th>Purpose</th>
<th>Availability</th>
<th>License</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford Word Segmenter</td>
<td>Chinese</td>
<td>tokenization</td>
<td>free download</td>
<td>GNU Public</td>
<td>The Stanford Word Segmenter is incorporated in GATE.</td>
</tr>
<tr>
<td>GATE</td>
<td>Chinese, Korean</td>
<td>Plug-ins for numerous functions</td>
<td>free download</td>
<td>LGPLv3</td>
<td>These are plug-ins for the GATE software environment.</td>
</tr>
<tr>
<td>USAS</td>
<td>Chinese</td>
<td>semantic tagging</td>
<td>free online</td>
<td>none (limited online interface)</td>
<td>Maximum 3000 characters input</td>
</tr>
<tr>
<td>TreeTagger</td>
<td>Chinese</td>
<td>POS tagging, lemmatization</td>
<td>free download</td>
<td>ACA</td>
<td></td>
</tr>
<tr>
<td>SPPAS</td>
<td>Mandarin Chinese, Taiwanese</td>
<td>audio annotation</td>
<td>free download</td>
<td>PUB</td>
<td></td>
</tr>
<tr>
<td>Prozed</td>
<td>Chinese</td>
<td>prosody editor</td>
<td>free download</td>
<td>CC BY-NC-SA 4.0</td>
<td></td>
</tr>
<tr>
<td>Voyant</td>
<td>Japanese</td>
<td>Analysis, exploration, visualization</td>
<td>free online</td>
<td>not known</td>
<td>Suite of text analysis tools, now works with Japanese.</td>
</tr>
<tr>
<td>topic-modelling-tool</td>
<td>Japanese</td>
<td>topic modelling</td>
<td>free download</td>
<td>not known</td>
<td>A point-and-click tool for creating and analyzing topics.</td>
</tr>
<tr>
<td>i2ocr</td>
<td>Chinese, Japanese, Korean, OCR</td>
<td>OCR</td>
<td>free online</td>
<td>not known</td>
<td>Not tested</td>
</tr>
<tr>
<td>Convertio</td>
<td>Chinese, Japanese, Korean, OCR</td>
<td>OCR</td>
<td>10 free pages, then pricing applies</td>
<td>GPL v3</td>
<td>Versions exist for Linux, Mac, Windows</td>
</tr>
<tr>
<td>KoNLPy</td>
<td>Korean</td>
<td>POS tagging, corpus analysis</td>
<td>free download</td>
<td>Apache License, Version 2.0</td>
<td>Also <a href="https://github.com/pycantonese/pycantonese">https://github.com/pycantonese/pycantonese</a></td>
</tr>
<tr>
<td>awesome-korean-nlp</td>
<td>Korean</td>
<td>A curated list of resources dedicated to Natural Language Processing for Korean</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anticonc</td>
<td>Not Language specific</td>
<td>Text and corpus analysis</td>
<td>free download</td>
<td>custom licence</td>
<td>Multi-platform (Windows/Mac/Linux). A freeware tool.</td>
</tr>
<tr>
<td>SegmentAnt</td>
<td>Chinese, Japanese, Korean, tokenization</td>
<td>tokenization</td>
<td>free download</td>
<td>custom licence</td>
<td>Appears to be restricted to use with a specific data set.</td>
</tr>
<tr>
<td>UDPipe</td>
<td>Japanese</td>
<td>tokenization, tagging, lemmatization</td>
<td>Online service with restrictions</td>
<td>custom licence</td>
<td>Portal offering links to various reference, pedagogy.</td>
</tr>
<tr>
<td>On-line Chinese Tools</td>
<td>Chinese</td>
<td>tokenization, encoding detection</td>
<td>Variable: some free to download</td>
<td>not known</td>
<td>Documentation in Japanese</td>
</tr>
<tr>
<td>McCabe</td>
<td>Japanese</td>
<td>tokenization, morphological</td>
<td>not known</td>
<td>not known</td>
<td>Portal for a number of NLP web services, Japanese morphological analysis engine</td>
</tr>
</tbody>
</table>
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/content/clarin-for-researchers

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

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

https://scalablereading.northwestern.edu/

https://programminghistorian.org/