CLARIN

Common Language Resources and Technology Infrastructure

A European Research Infrastructure

TEI or not TEI? Some anecdotal evidence from CLARIN and the UK

Martin Wynne

Oxford University Computing Services,

Oxford e-Research Centre &

Faculty of Linguistics, Philology and Phonetics,

University of Oxford

martin.wynne@oucs.ox.ac.uk

Problems

"We have communities of people all annotating the same basic data for a range of different phenomena, and annotating the same data in competing ways (comparing methods for speech act markup, or comparing the results of automatic processing using a range of methods to hand annotation). Even if you could easily pack all this stuff into one XML file, it would be a bad idea - you need version control, dependency management, and the ability for several people to work at the same time."

[computational linguist, working with speech data in XML, but not TEI]

TEI or not TEI?

“[in CLARIN] most web services are not using TEI because inside of WebLicht, we need *processing information* which can not - or not easily – be stored inside a TEI file. Therefore, we developed [our own format]”

Although note that Weblicht does have a converter for TEI text, at the beginning and end of the chain

Not TEI :-(

“I gave up on it a long time ago. It provides far too much detail for what I normally wish to do. Plain text files get you a long way in analysing texts. Beyond that, XML and other more public, general, standards seem to work just fine. I really cannot see the point of the TEI anymore.”

[a leading corpus linguist in the UK]

TEI or not TEI?

“Mixing the mark-up which represents the physical structure of the codex and the linguistic information in one xml file makes no sense, it results in a messy xml file, overlapping, etc. Generally, I think, mixing two different types of mark-up information is not really possible.”

[historical linguist, who uses TEI texts at certain points in his workflow]

TEI!

“You'd be crazy not to use the TEI today. Everyone should be forced to use it. I'm so glad that when I phoned up the Arts and Humanities Data Service all those years ago, they told me to use the TEI. I didn't know what they were talking about, but they were right. People who don't want to impose standards are too liberal.”

[technologist who works with corpus linguists in the UK]

Summary of Technical Problems

- Annotation tools don't accept TEI as an input format, e.g. TreeTagger, CLAWS, all syntactic parsers (although TEI XML might still be useful as interchange or pivot format in a processing workflow which requires use of multiple formats at different stages)
- There are no mature, stable, usable standards and tools available for using stand-off annotation in a production environment
- Performance issues: plain text files are smaller than XML for processes which go through texts serially,and processing is simpler; parsing XML is slow; database queries are much faster (and there is not much expertise around in using eXist)

Organizational Problems

- There is a long learning curve for the average linguist to get to terms with TEI guidelines and technologies
- There isn't necessarily a lot to gain by using the TEI (in the short term), since there aren't hardly any tools and services for analysing TEI corpora
- The advantages of a TEI text are for others – for interchange, re-use, long-term preservation, use in other disciplines, etc. The creator does a lot more work and doesn't get rewarded for that.

A solution?

The Stick

Funders, instititutions, repositories, communities, etc. insist on standards conformance, possibly with penalties for non-conformance

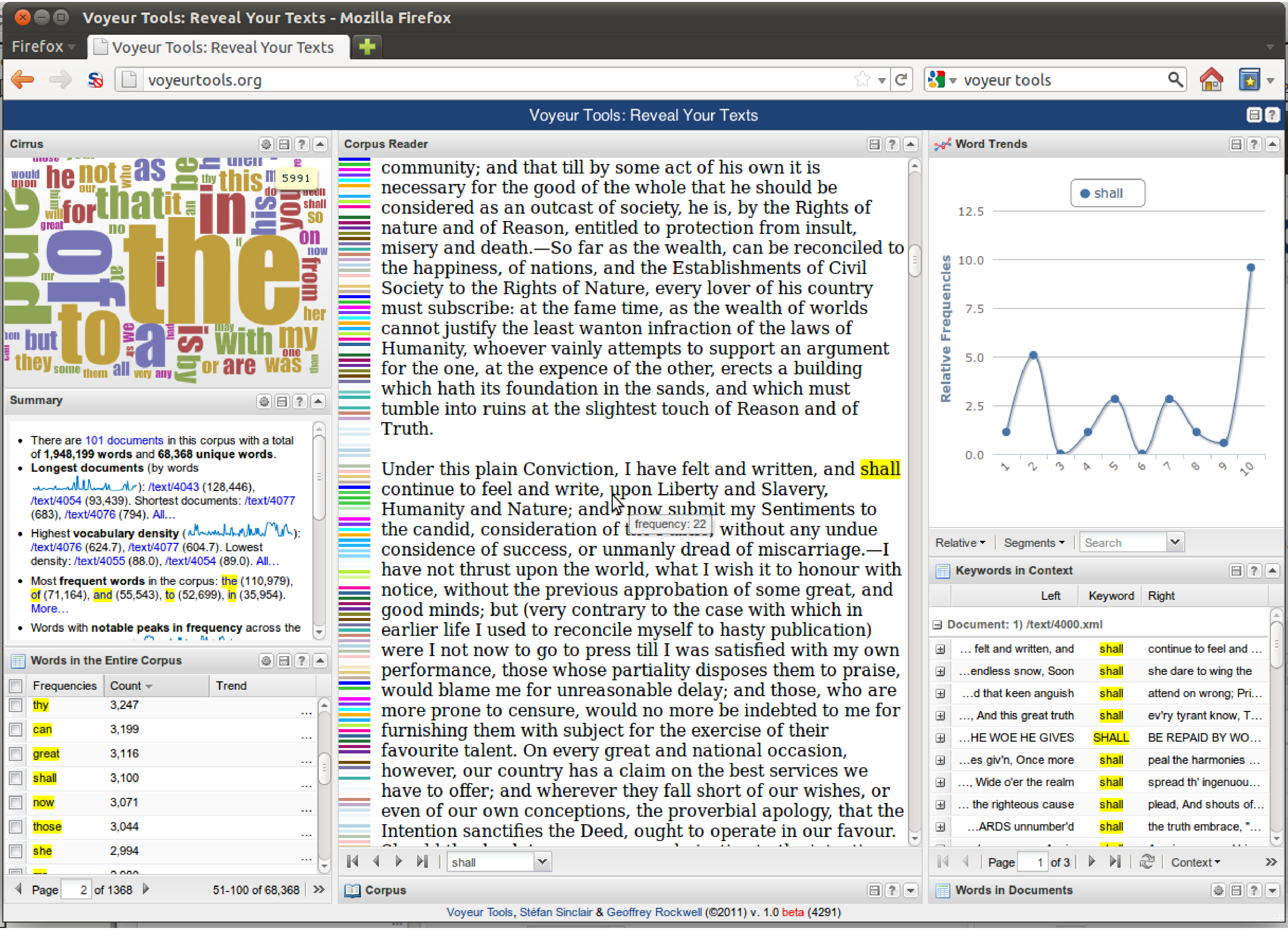
The Carrot

Offer tools with the functions that linguists want that work with TEI corpora, and infrastructure services that the corpus can easily be plugged into

In this case, it looks like offering the carrot of services is more effective of the stick of compulsion. The TEI should therefore focus on offering repositories and tools for using TEI texts and corpora.

The Oxford Text Archive...

- ...is happy to discuss the deposit of texts and corpora in the archive!
- We are currently developing an enhanced service where TEI texts are made available at persistent locations with consistent metadata (available via OAI-PMH), so that web services can be more easily built on top of and around them
- Usually there is a small charge (c. £500 GBP) to cover some of the costs of processing of new deposits and making them available, but *pro bono* work is possible in the case of hardship!
- An example on the following (and final slide) – Voyant tools hosted at the University of Alberta performing analysis of TEI-encoded ECCO texts hosted at the Oxford Text Archive
(with thanks to Geoffrey Rockwell, Stefan Sinclair and Sebastian Rahtz)



Summary

- There are 101 documents in this corpus with a total of **1,948,199 words** and **68,368 unique words**.
- **Longest documents** (by words): /text/4043 (128,446), /text/4054 (93,439). Shortest documents: /text/4077 (683), /text/4076 (794). All...
- Highest **vocabulary density** ( ): /text/4076 (624.7), /text/4077 (604.7). Lowest density: /text/4055 (88.0), /text/4054 (89.0). All...
- Most **frequent words** in the corpus: the (110,979), of (71,164), and (55,543), to (52,699), in (35,954). More...
- Words with **notable peaks in frequency** across the

Words in the Entire Corpus

□	Frequencies	Count	Trend
□	thy	3,247	...
□	can	3,199	...
□	great	3,116	...
□	shall	3,100	...
□	now	3,071	...
□	those	3,044	...
□	she	2,994	...

Page 2 of 1368 51-100 of 68,368

community; and that till by some act of his own it is necessary for the good of the whole that he should be considered as an outcast of society, he is, by the Rights of nature and of Reason, entitled to protection from insult, misery and death.—So far as the wealth, can be reconciled to the happiness, of nations, and the Establishments of Civil Society to the Rights of Nature, every lover of his country must subscribe: at the fame time, as the wealth of worlds cannot justify the least wanton infraction of the laws of Humanity, whoever vainly attempts to support an argument for the one, at the expence of the other, erects a building which hath its foundation in the sands, and which must tumble into ruins at the slightest touch of Reason and of Truth.

Under this plain Conviction, I have felt and written, and shall continue to feel and write, upon Liberty and Slavery, Humanity and Nature; and now submit my Sentiments to the candid, consideration of the [illegible], without any undue considence of success, or unmanly dread of miscarriage.—I have not thrust upon the world, what I wish it to honour with notice, without the previous approbation of some great, and good minds; but (very contrary to the case with which in earlier life I used to reconcile myself to hasty publication) were I not now to go to press till I was satisfied with my own performance, those whose partiality disposes them to praise, would blame me for unreasonable delay; and those, who are more prone to censure, would no more be indebted to me for furnishing them with subject for the exercise of their favourite talent. On every great and national occasion, however, our country has a claim on the best services we have to offer; and wherever they fall short of our wishes, or even of our own conceptions, the proverbial apology, that the Intention sanctifies the Deed, ought to operate in our favour.

Keywords in Context

	Left	Keyword	Right
⊟	Document: 1) /text/4000.xml		
⊞	... felt and written, and	shall	continue to feel and ...
⊞	...endless snow, Soon	shall	she dare to wing the
⊞	...d that keen anguish	shall	attend on wrong; Pri...
⊞	..., And this great truth	shall	ev'ry tyrant know, T...
⊞	...HE WOE HE GIVES	SHALL	BE REPAID BY WO...
⊞	...es giv'n, Once more	shall	peal the harmonies ...
⊞	..., Wide o'er the realm	shall	spread th' ingenuou...
⊞	... the righteous cause	shall	plead, And shouts of...
⊞	...ARDS unnumber'd	shall	the truth embrace, "...

Page 1 of 3 Context

Voyeur Tools, Stéfan Sinclair & Geoffrey Rockwell (©2011) v. 1.0 beta (4291)