Best Practice for Language Corpora

Martin Wynne
Head of the Oxford Text Archive
Oxford e-Research Centre &
Oxford University Computing Services &
Faculty of Linguistics, Philology & Phonetics
University of Oxford
martin.wynne@oucs.ox.ac.uk

What is a corpus

“A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.”

(John Sinclair, 2005)

Sampling a language

- Defining the population (e.g. modern British English, literary French 1842-1848, blogs about the Gulf War, etc.)
- Design criteria (time, space, socio-economic...)
- Modes and media (speech, writing, audio, video, images)
- Sampling (principled, pragmatic and opportunistic, factors)
- Size (as big as possible? Other practical considerations)
- Representativeness (and consequences for your interpretations of data)
- Balance (of sections representing different elements)

Types of corpus

- General reference / specialised
- Synchronic / diachronic / monitor
- Writing / speech / other modes
- Text / audio / other media
- Multilingual: translation, parallel, comparable
- Native / non-native / learner corpora
- Sampled / complete

Standards and good practice

Standards for:

- text encoding
- annotation
- Metadata

De facto standards for many different communities...

But there is a convergence around XML.

Corpus availability

- Licensing: permission from rights holders, licences for users
- Arrangements for allowing access to the corpus
- Download / online access services
- Sustainability of arrangements? Will it work in ten years time?
- Using a corpus data centre

An example: British National Corpus

- The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written
- TEI-conformant XML
- Lots of metadata
- Linguistic annotation for wordclass, lemma and some multi-word units
- Available on DVD, with a small charge for the licence to use it
- Various online services (at British Library, Brigham Young University, Lancaster University, etc.)
- Digital audio aligned with text transcriptions, coming soon!
- http://www.natcorp.ox.ac.uk/corpus/index.xml

BNC: if we were doing it now...

- Bigger
- More speech?
- New genres
- Digital audio aligned with text
- Digital video aligned with text
- TEI XML
- Off-the-shelf licence, e.g. Creative Commons
- Online service
- Copyright material and secure access?

An example

http://corpus.byu.edu/bnc/

Reuse of corpora

- Can the corpus be downloaded?
- Is there an online access service?
- Is the access free, or does it need to be paid for?
- Is individual registration necessary?
- Does user registration use an external single sign-on service?
- Does the online service remember my sessions?
- Do I need a specific operating system or software applications?
- Do licence forms need to be signed?
- Does the access service allow adding my own data?
- Can I connect the corpus to other online tools?
- Can I make available a new service built on the corpus?
- Can I make extended or interpreted data available?

The problems facing users

- Many archives known only to certain communities
- Archives are mostly unconnected, and data difficult to find
- Every archive has its own standards for storage and access
- Resources are in different formats, follow different standards, are described in differing ways
- Tools are hard to use for non-specialist
- Tools and data are not available to use online (or with limited functionality and processing capacity)
- Many researchers are not aware of the potential benefits of using language and speech technology tools

The Vision

A researcher in Oxford from his desktop computer can:

- do a single sign-on with local authentication, and then:
- search for, find and obtain authorization to use corpora 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 or national 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, Vienna and Zagreb
- discuss and iteratively adopt and re-run the analyses with collaborators

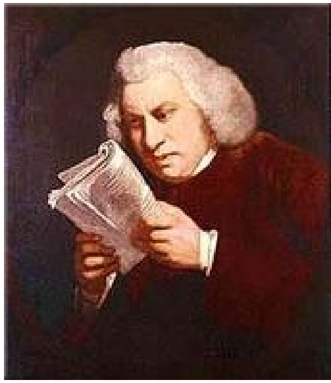
Future considerations

Things to think about for the...

- Creator of tools
- Creator of data
- Research Centre or Faculty
- Funders and Policy-makers
- Library
- Computing service
- Individual researcher

Creator of corpora

- **Standards:** how do I make my corpus fit into the infrastructure for finding and using it?
- **Sustainability:** will the arrangements for reusing the corpus still operate in 1, 2, 5 and 10 years' time?
- **Connecting and linking:** how can I facilitate research using my data along with other ones?
- **Creating a finished product:** when should I stop developing and make it available?
- **Impact:** how can I create the maximum impact with my data, in my academic field and beyond?



Best Practice

Best practice is creating your corpus conformant to relevant standards to allow:

- deposit in an appropriate archive;
- re-use by other researchers;
- interoperability with tools and other corpora;
- discovery;
- long-term preservation.

But is the field mature enough to say exactly how to do this?

Part 2: the corpus in the era of the data deluge

There were good reasons for creating carefully crafted language corpora in the past.

Nowadays, with masses of language data in electronic form at our fingertips, do we still need to do this?

Doesn't it make more sense now to use the web as a corpus?

The traditional case for the corpus

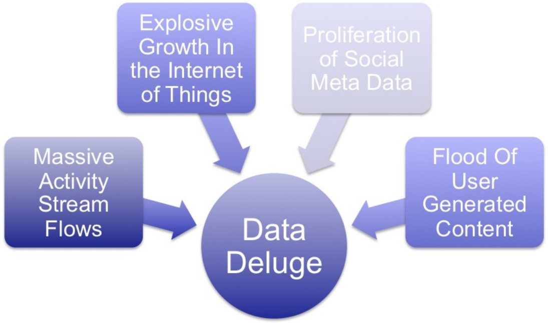
A desire to find out new things about language by looking at a lot of it at the same time.

- Need for electronic language data
- Need for control of contextual and linguistic variables
- Need for metadata
- Need for linguistic annotation

Uses of the corpus

- Show me lots of examples of *x*
- Can I find evidence in the corpus to support existing hypotheses?
- Are there features of a language variety (or text) which have not been noticed by scholars?
- How frequent is *x* (in comparison to *y*)?
- What is the distribution of *x* in different types of text?
- What is the earliest occurrence of *x*?
- How did the frequency and usage of *x* change over time?
- Starting with the data, how is the language best described?

The data deluge


Explosive Growth In the Internet of Things

Proliferation of Social Meta Data

Massive Activity Stream Flows

Data Deluge

Flood Of User Generated Content

The Economist

Obama the warrior

Misgoverning Argentina

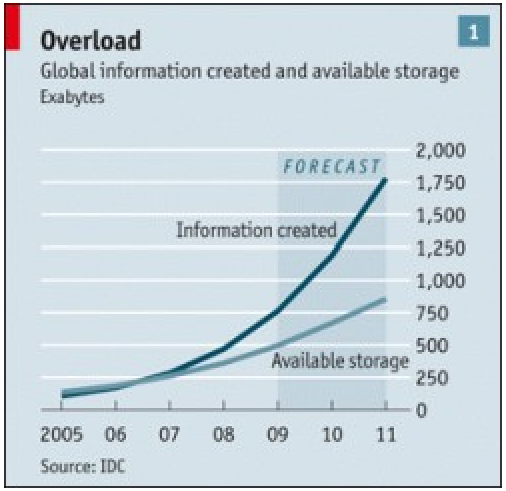
The economic shift from West to East

Genetically modified crops blossom

The right to eat cats and dogs

The data deluge

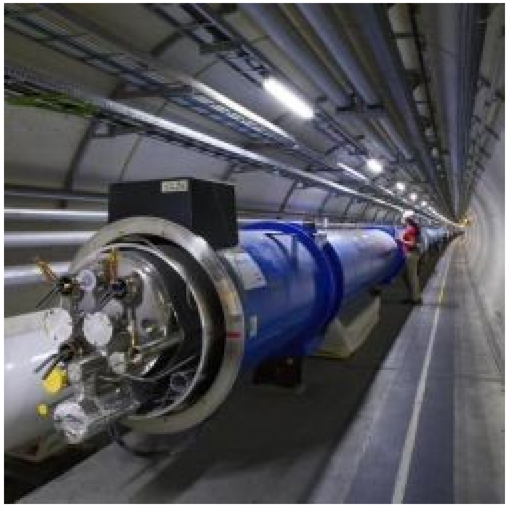
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT


Overload

Global information created and available storage

Exabytes

Source: IDC



Web for corpus

Filtering and organising the data deluge to make corpora from web texts

The case for corpus building today

- Some forms of language are not represented on the web:
  - Spoken data
  - Historical data
  - Published texts in copyright
  - Other?
- Controlling and understanding provenance, size, balance and representativeness are important in order to be able to interpret data
- To answer some questions we need annotated corpora (tokenized, lemmatized, POS-tagged, parsed, etc.)

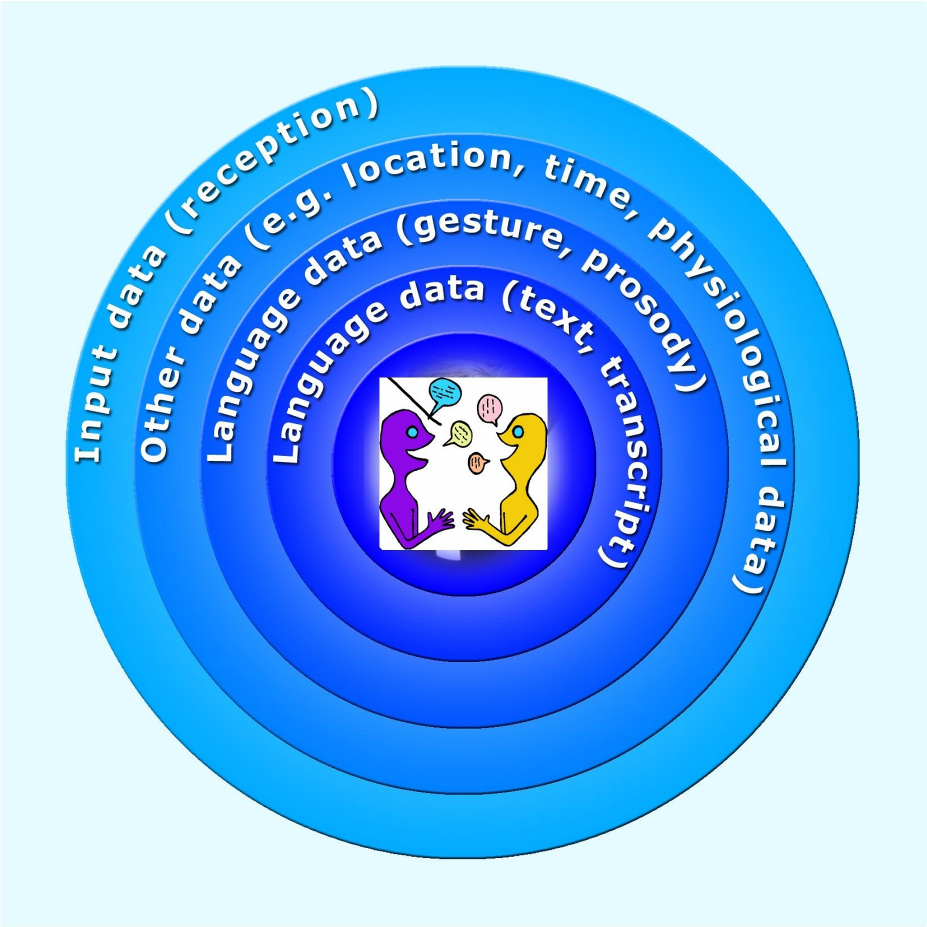
Beyond the finite text corpus

How can we start to capture, filter and make usable the wealth of language data out there, from:

- spoken interactions
- sign language
- computer-mediated communications
- electronic publishing
- large-scale digitization projects
- public, open data

Furthermore, can we collect data on the web, via crowd-sourcing?

Capturing the context


Input data (reception)
Other data (e.g. location, time, physiological data)
Language data (gesture, prosody)
Language data (text, transcript)

- Svenja Adolphs and Dawn Knight, University of Nottingham

The Web **as** Corpus

http://www.webcorp.org.uk

Problems with language on the web

- Biased distribution of text-types and genres (journalism, computing, pornography v. academic writing, novels, speech)
- Unknown provenance (who, when, why?)
- Native and non-native producers of language
- Mixture of varieties
- Unclear separation of linguistic elements of the webpage
- Accessing the hidden web
- Repeated text
- Lack of persistence of source data
- Unknown (or undesirable) sampling and ranking strategies of search engines

Or is this too normative and prescriptive?

Problems with language in the corpus

- Limited size
- Expensive, time-consuming and slow to make
- Usually limited to out of copyright texts
- Not up to date
- Design decisions were made by someone else
- Not comparable to other corpora
- Access restrictions
- Need specialist applications
- Often, only restricted online access with limited functionality, processing, and not connected to other resources

Tools and methods

Do we have the tools, services, infrastructures, virtual research environments, methods and procedures for working with unrestricted data?

References

- Developing Linguistic Corpora http://www.ota.ox.ac.uk/documents/creating/dlc/
- Resource Discovery Task Force http://rdtf.jiscinvolve.org/wp/
- Anderson, W. and Corbett, J. (2009), *Exploring English with Online Corpora*, Palgrave.
- CLARIN http://www.clarin.eu/
- Hey, A. J. G. and Trefethen, A. E. (2003) The Data Deluge: An e-Science Perspective. In: Grid Computing - Making the Global Infrastructure a Reality, pp. 809-824, Wiley and Sons. ISBN 0470853190
- The Data Deluge, http://www.economist.com/node/15579717?Story_ID=1557971
- WebCorp http://www.webcorp.org.uk