CLARIN

Common Language Resources and Technology Infrastructure

A European Research Infrastructure

# Tools for analyzing language corpora

**Martin Wynne**
Head of the Oxford Text Archive
Oxford e-Research Centre and
Oxford University Computing Services
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)

# Corpus analysis functions

- Searching
  - words
  - lemmas
  - wildcards
  - regular expressions
  - annotations
- Concordancing
- Collocations
- Expanding co-text
- Displaying annotations
- Using metadata
- Wordlists
- Keywords
- Identifying multi-word units
- Beyond the written, monlingual text corpus...

http://www.pala.ac.uk/resources/sigs/corpus-style/searching/handbook.html

# Typology of tools

- Is it installed locally or accessed on the web?
- Are there limits on the number of searches or results?
- Is it platform dependent?
- For use with one corpus (or a specific set) or generic?
- For use with particular types of encoding?
- Use directly on a corpus, or does it require pre-processing (e.g. indexing)?
- What range of functions does it have?
- What can it do with annotation?
- What can it do with metadata?
- What languages can it handle?
- Is it free? What are the licensing arrangements?

# Demos

**Antconc**

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

**http://phrasesinenglish.org/**

**http://taporware.ualberta.ca/**

**http://bncweb.info/**

**http://www.webcorp.org.uk/**

**http://ngrams.googlelabs.com/**

**http://www.dwds.de/**

**http://weblicht.sfs.uni-tuebingen.de/englisch/weblicht.shtml**

# 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

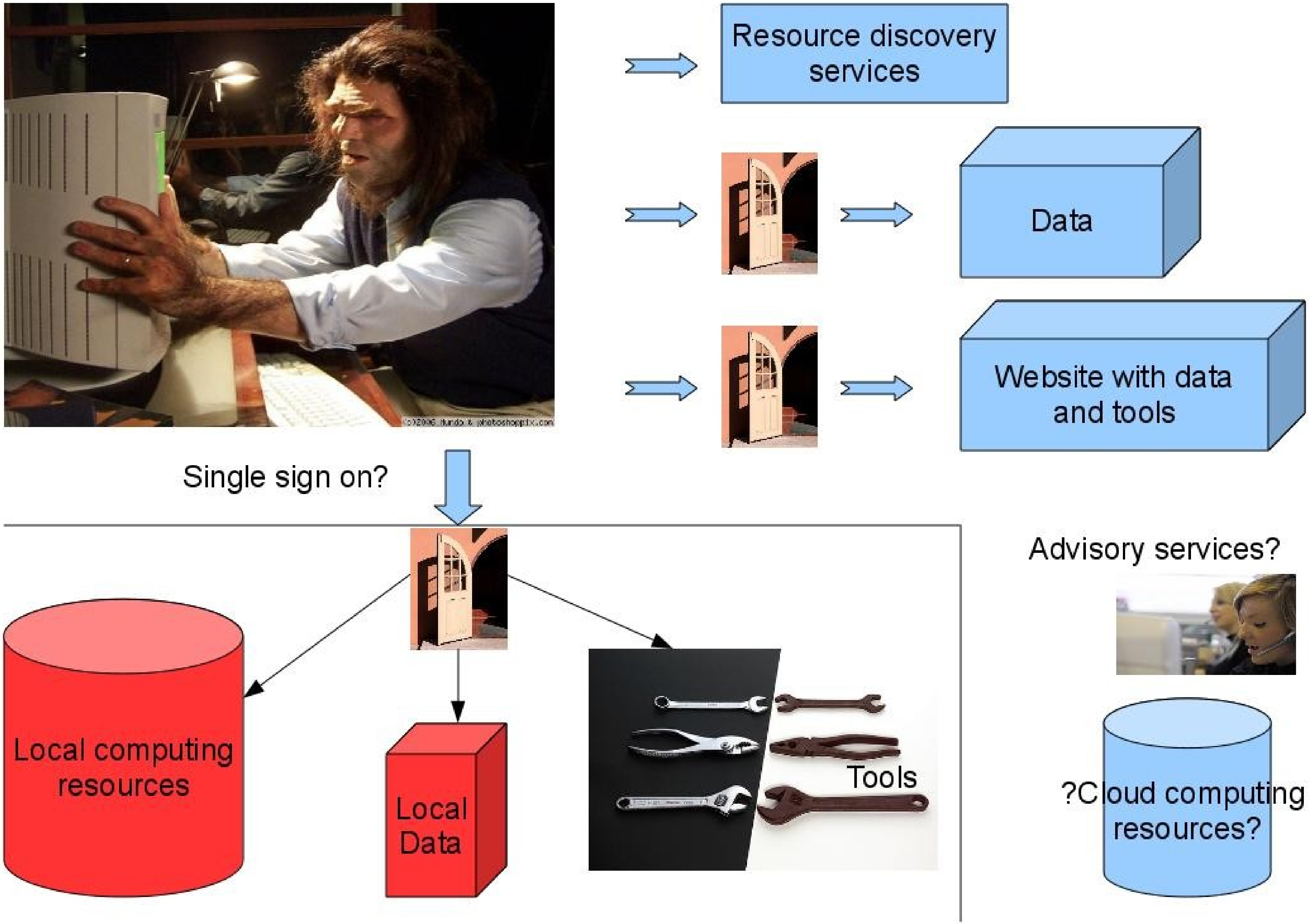
Do we have the tools, services, infrastructures, virtual research environments, methods and procedures for working with unrestricted data?

# Availability 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



(c)2006 Hundo & photoshoppix.com

Resource discovery services

Data

Website with data and tools

Single sign on?

Local computing resources

Local Data

Tools

Advisory services?

?Cloud computing resources?

# The fragmented environment

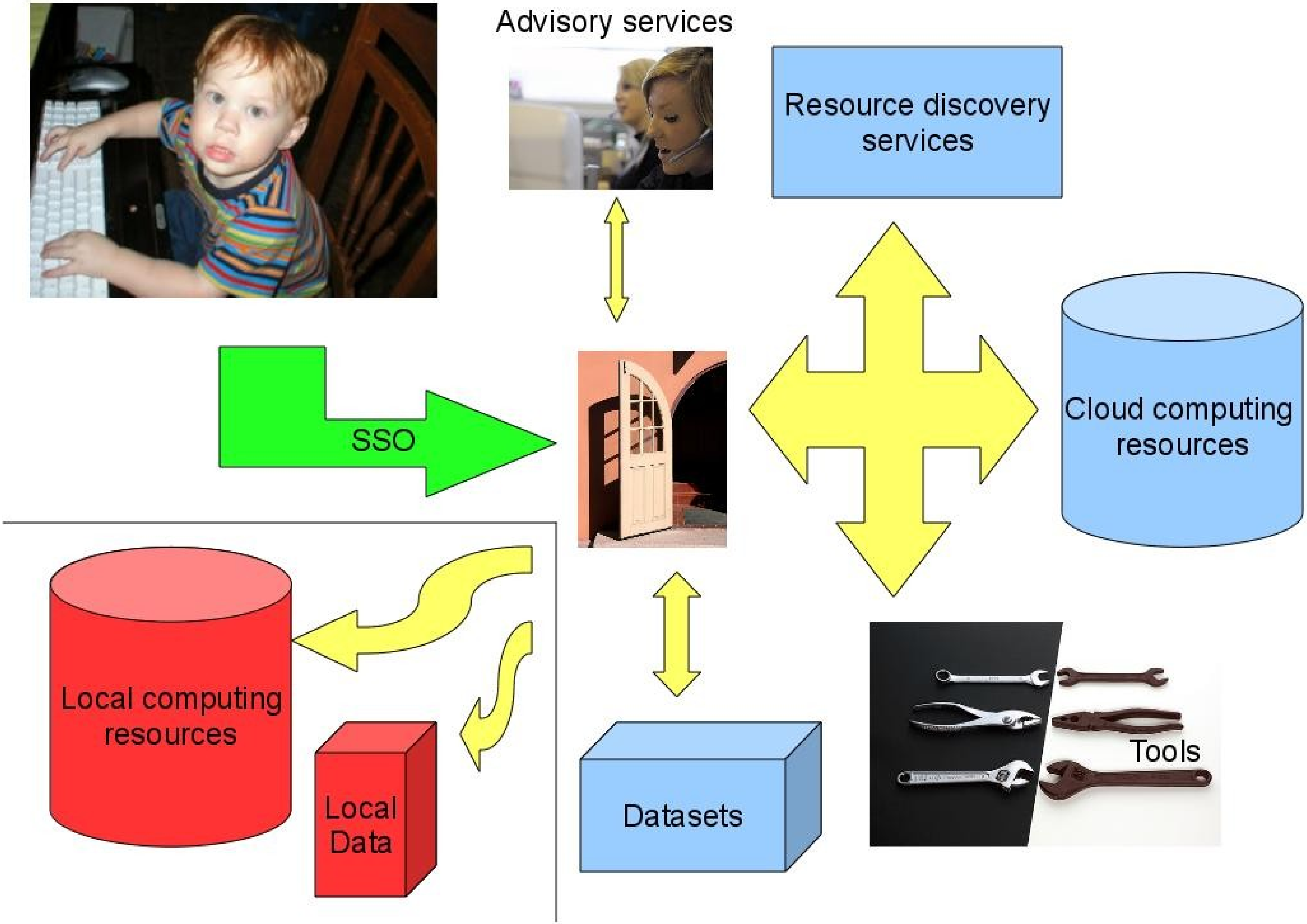
Today's Virtual Research Environments (VREs) exist in a confusing landscape, involving:

- research projects,
- academic associations,
- repositories,
- digital libraries,
- generic computing infrastructure (including commercial services)
- Grid, supercomputing and cloud computing facilities,
- institutional repositories
- standards bodies
- networks, projects, infrastructure initiatives...

# The CLARIN 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



Advisory services

Resource discovery services

SSO

Cloud computing resources

Local computing resources

Local Data

Datasets

Tools

# Future considerations

Things to think about for the...

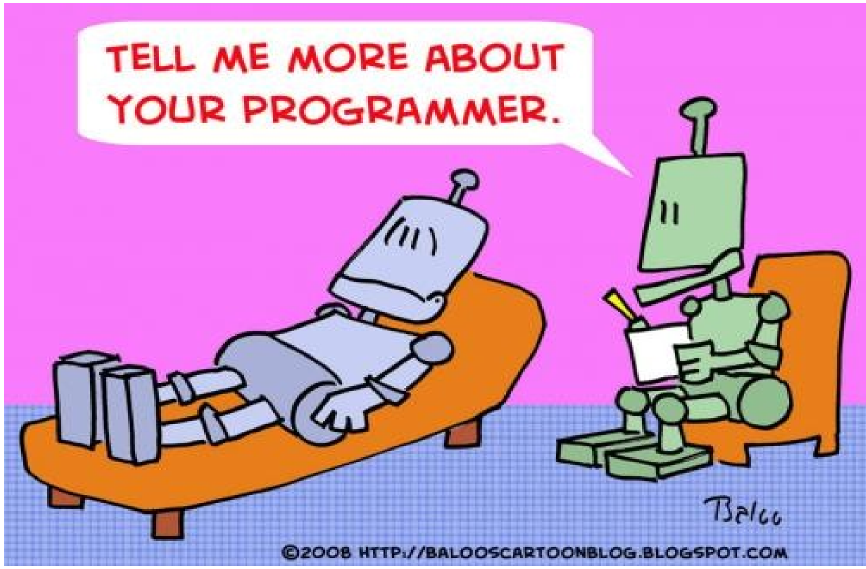
- Creator of tools
- Creator of data
- Research Centre or Faculty
- Funders and Policy-makers
- Library
- Computing service
- Individual researcher

# Creator of tools

Think beyond the needs of your current tasks and project.

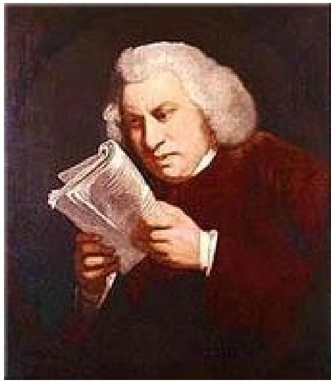
- **Standards:** what standards and guidelines do I need to follow to allow my tools to be used in the infrastructure;
- **Interoperability:** what datasets, tools and services will my tools need to be able to work with
- **Sustainability:** what do I need to do to make sure that this will be working in 10 years' time?

**Production services** not *ad hoc* **tools**


TELL ME MORE ABOUT YOUR PROGRAMMER.
Balcc
©2008 HTTP://BALOOSCARTOONBLOG.BLOGSPOT.COM

# 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?

# References

- Developing Linguistic Corpora http://www.ahds.ac.uk/creating/guides/linguistic-corpora/
- Wynne, M, *Searching and Concordancing* http://www.pala.ac.uk/resources/sigs/corpus-style/searching/ha
- Anderson, W. and Corbett, J. (2009), *Exploring English with Online Corpora*, Palgrave.
- CLARIN http://www.clarin.eu/

# Discuss

- How can we make better use of corpora? Do we need:
    - More corpora
    - Bigger corpora
    - A better range
    - More standardization
    - Better ways to use the web as a corpus
    - Better tools
    - To put them all in one place
    - To be able to find them all from place
    - To be able to search them all from one place
- Given finite resources, which should be the priority?