Supporting Social Scientists in the iResearch Age

Paul Thompson
Director, Centre for Corpus Research
iUser: [it’s not all] About me

- British academic
- Applied linguist
- Corpus linguist
- Journal editor
Text mining
[eg, Genia project]

Language use in specific types of discourse

Social relevance of corpus work

Tracing evolution of ideas, values and language

Identifying preferred modes of expression
In the early sixties, with development of mainframes, the first major corpora were compiled

- The Brown corpus was the model:
  1 million words, with 500 samples of 2000 words each

Noam Chomsky: the linguist must model language competence rather than performance. No corpus of evidence can represent all language.
Growing up

- The next generation of corpora:
  - COBUILD corpus 1985 = 18 million words
  - Bank of English 1991, increasing to 525 million words in 2005
  - British National Corpus 1994, 100 million words [fixed size]
  - Oxford English Corpus – now 2 billion words
Access

- **In 1960s:** access highly restricted
- **In 1990s:** corpus analysis tools limited, but some network access possible;
  first ICAME CD-ROM collection of corpora
- **In 2010:** huge corpora accessed through web interfaces;
  many tools for use on own machine;
  some owners restrict access to own researchers
<table>
<thead>
<tr>
<th>Corpus name</th>
<th>Language</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet-ZH</td>
<td>Chinese, Simplified</td>
<td>277,931,664</td>
</tr>
<tr>
<td>British National Corpus</td>
<td>English</td>
<td>112,181,850</td>
</tr>
<tr>
<td>ukWaC v1.0 old</td>
<td>English</td>
<td>1,526,599,198</td>
</tr>
<tr>
<td>French web corpus</td>
<td>French</td>
<td>126,850,281</td>
</tr>
<tr>
<td>deWaC</td>
<td>German</td>
<td>1,627,169,557</td>
</tr>
<tr>
<td>JpWaC</td>
<td>Japanese</td>
<td>409,384,405</td>
</tr>
<tr>
<td>Russian web corpus</td>
<td>Russian</td>
<td>187,965,822</td>
</tr>
<tr>
<td>Spanish web corpus</td>
<td>Spanish</td>
<td>116,900,060</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Corpus name</th>
<th>Language</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic web corpus</td>
<td>Arabic</td>
<td>174,239,600</td>
</tr>
<tr>
<td>Chinese GigaWord 2 Corpus: Mainland, simplified</td>
<td>Chinese, Simplified</td>
<td>250,124,230</td>
</tr>
<tr>
<td>Chinese GigaWord 2 Corpus: Taiwan, traditional</td>
<td>Chinese, Traditional</td>
<td>455,526,209</td>
</tr>
</tbody>
</table>
Next generation: Multimodal corpora

Mark-up of data
Linking of channels
Quantity of data
Technological developments have led to:

- New forms, and new levels of access to data and tools
- Much larger datasets
- ... and new approaches to collaboration ...
Collaboration

- Using **Web 2.0** technologies and values, e.g.:
  - Blogs, clouds, wikis, social bookmarking
  - Open source software
  - Joint construction
  - Increased levels of access for more people
- **Web 3.0** – interconnection of databases, and the Semantic Web
Archives

- Searchable archive: use keywords to locate relevant papers/information

- Rich resource but the number of options are limited
- I want to search by phrase, by word, by social variables
Key Word in Context (KWIC) searches

words occurring to left or right of key word

reaction to the change in their mother's expressions into equation (2) gives: expression This expression may be simple expression occurred and expression of (5 × 500 × expression a nonmotile strain (FDR875) regulates the expression using the expression builder again and an expression taken for granted assumptions. They are the expression flowering time in terms of morphological expression regulation (activation or repression) of gene expression. Misexpression of Wnt3a induces ectopic advantage of a second niche for replication. Expression of an autophagy-like process facilitates expression also has problems in recognising facial expression is also significantly reduced.23 Expression of BMP signalling targets are increased expression of certain genes, somatic cell mutations expression of chalcone synthase, an enzyme involved
Pattern searching

complex number systems Consider an percentage elongation (Equation 2) is an percentage elongation (Equation 2) is an promoter, we aimed to study the changing 355 was selected to allow constitutive second promoter CAB is normally drives fact that Gli3 activity is not required for graph is a QQ plot for the first gene regulation (activation or repression) of gene regulation (activation or repression) of gene sure that when this is the case, greater dermal development by directly induces expression of the form expression of the ductility expression of the ductility where If expression of the chlorophyll expression of the reporter expression of the reporter gene expression of the chlorophyll expression of the genes expression of the control expression of a variety expression of a variety of homeobox expression of the binding protein expression of the GATA

where all the are arbitrary reals where If is the distance between gauges is the distance between gauges marks when a/b binding protein over time after exposure general in a/b binding protein at time t.

depth again of homeobox containing transcription factors. A homeobox or in our case the reporter gene luciferase factors MED-1 and MED-2, which are required
Need for support

- Severe cuts in research council funds expected
- Problems of sustainability
  - End of funding for AHDS in 2008
- Quantitative databases valued
- Language databases not valued
- There is a greater role for publishers to play
What roles?

- Building repositories for data collections
- Developing interfaces and analytical tools for researchers, from a range of fields
  - Same dataset, different interfaces
- Offering increased access to these resources
- Facilitating communications between researchers within one field of research, and also researchers in different disciplines
Why and how?

- Leadership in sustaining and supporting high quality, international research work
- Ensuring sustainability
- Through partnerships with research institutions