UNIVERSITY OF BIRMINGHAM

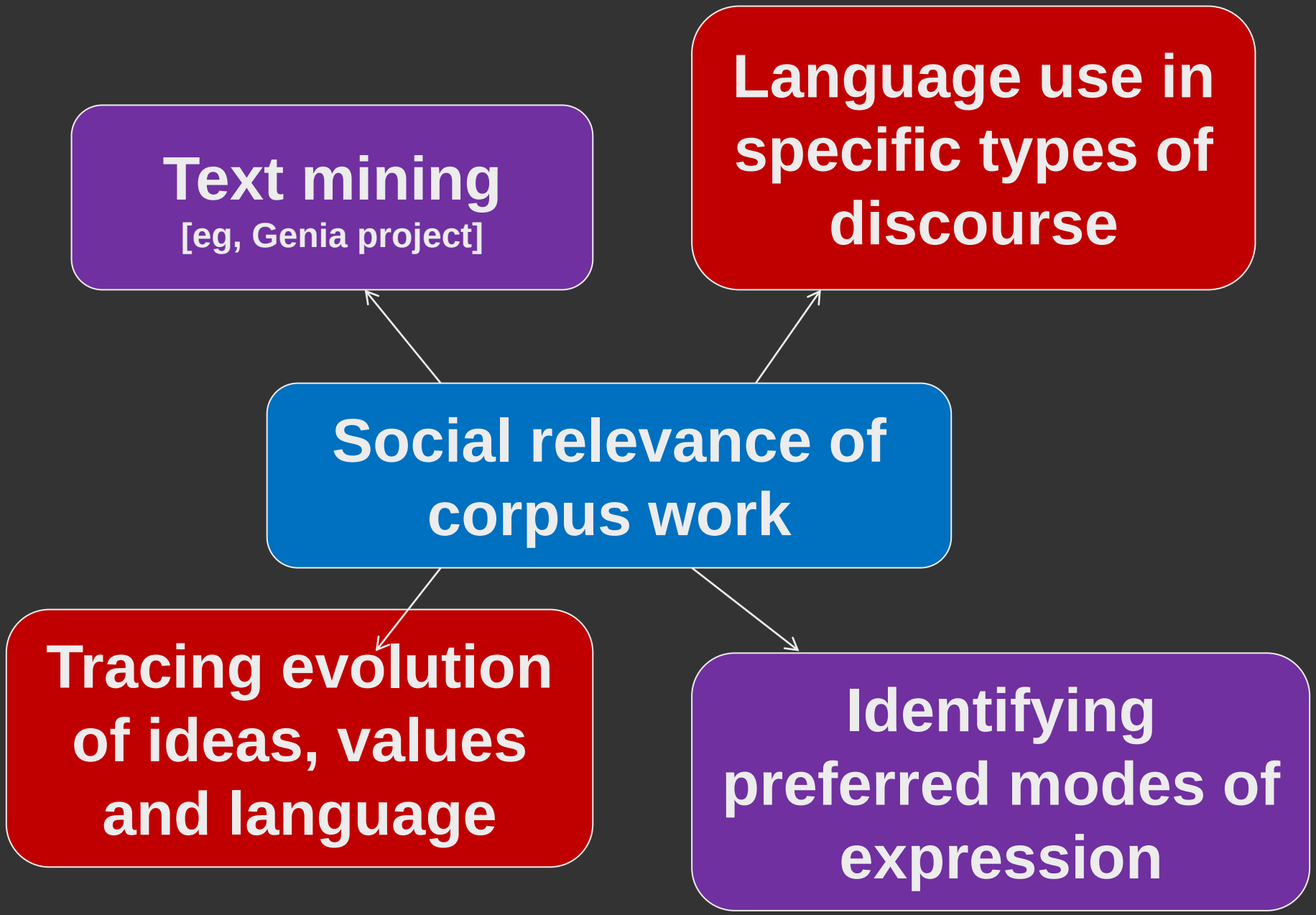
Supporting Social Scientists in the iResearch Age

Paul Thompson
Director, Centre for Corpus Research

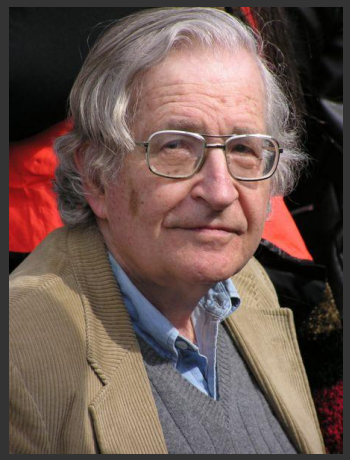
iUser: [it's not all] About me

- British academic
- Applied linguist
- Corpus linguist
- Journal editor

UNIVERSITY OF BIRMINGHAM


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

Technology and corpus linguistics



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

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**

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

UNIVERSITY OF BIRMINGHAM

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

UNIVERSITY OF BIRMINGHAM

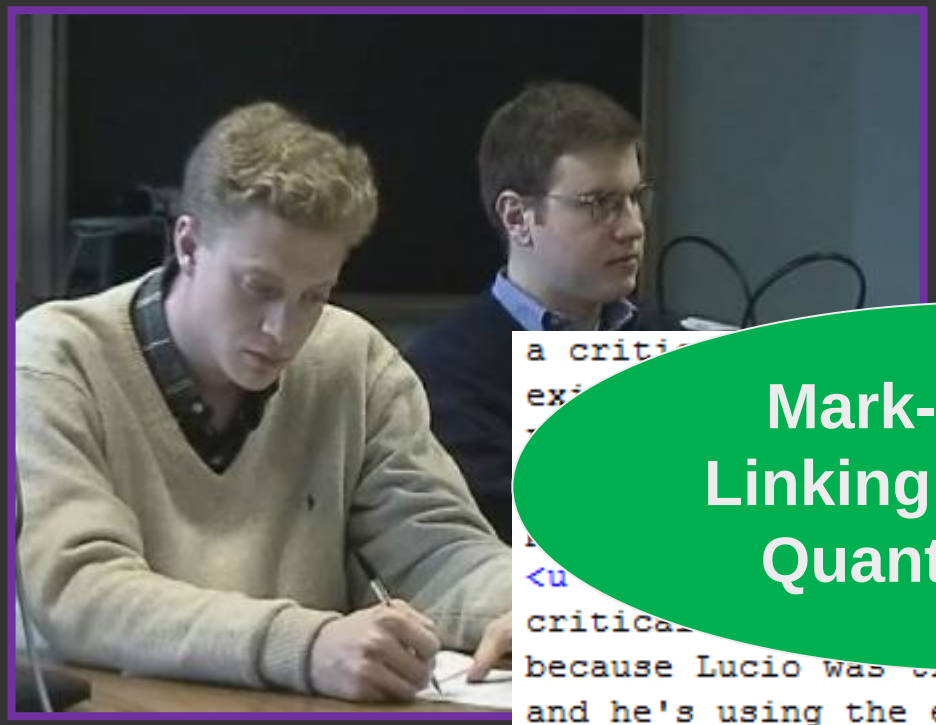
Corpus name	Language	Size
Internet-ZH	Chinese, Simplified	277,931,664
British National Corpus	English	112,181,850
ukWaC v1.0 old	English	1,526,599,198
French web corpus	French	126,850,281
deWaC	German	1,627,169,557
JpWaC	Japanese	409,384,405
Russian web corpus	Russian	187,965,822
Spanish web corpus	Spanish	116,900,060

Sh

Corpus name	Language	Size
Arabic web corpus	Arabic	174,239,600
Chinese GigaWord 2 Corpus: Mainland, simplified	Chinese, Simplified	250,124,230
Chinese GigaWord 2 Corpus: Taiwan, traditional	Chinese, Traditional	455,526,209

Sketch Engine (used by several publishers)

Next generation: Multimodal corpora



Mark-up of data
Linking of channels
Quantity of data

```
a criti                                    e the system as it
exi                                        ing system and
                                           way the system
                                           olve the
<u                                         myself as a
critica                                    i'll explain why
because Lucio was trying to solve the problem of development
and he's using the existing theories the existing theories
to develop and critical theory is exactly getting out of
those theories right and maybe like developing a new problem
solving theory and that's critical thinking it's not problem
solving</u>
<u who="sm5442"><gap reason="inaudible" extent="1 sec"/>
your kind of # imperialistic notion by</u>
```

Technological developments have led to:

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

UNIVERSITY OF BIRMINGHAM

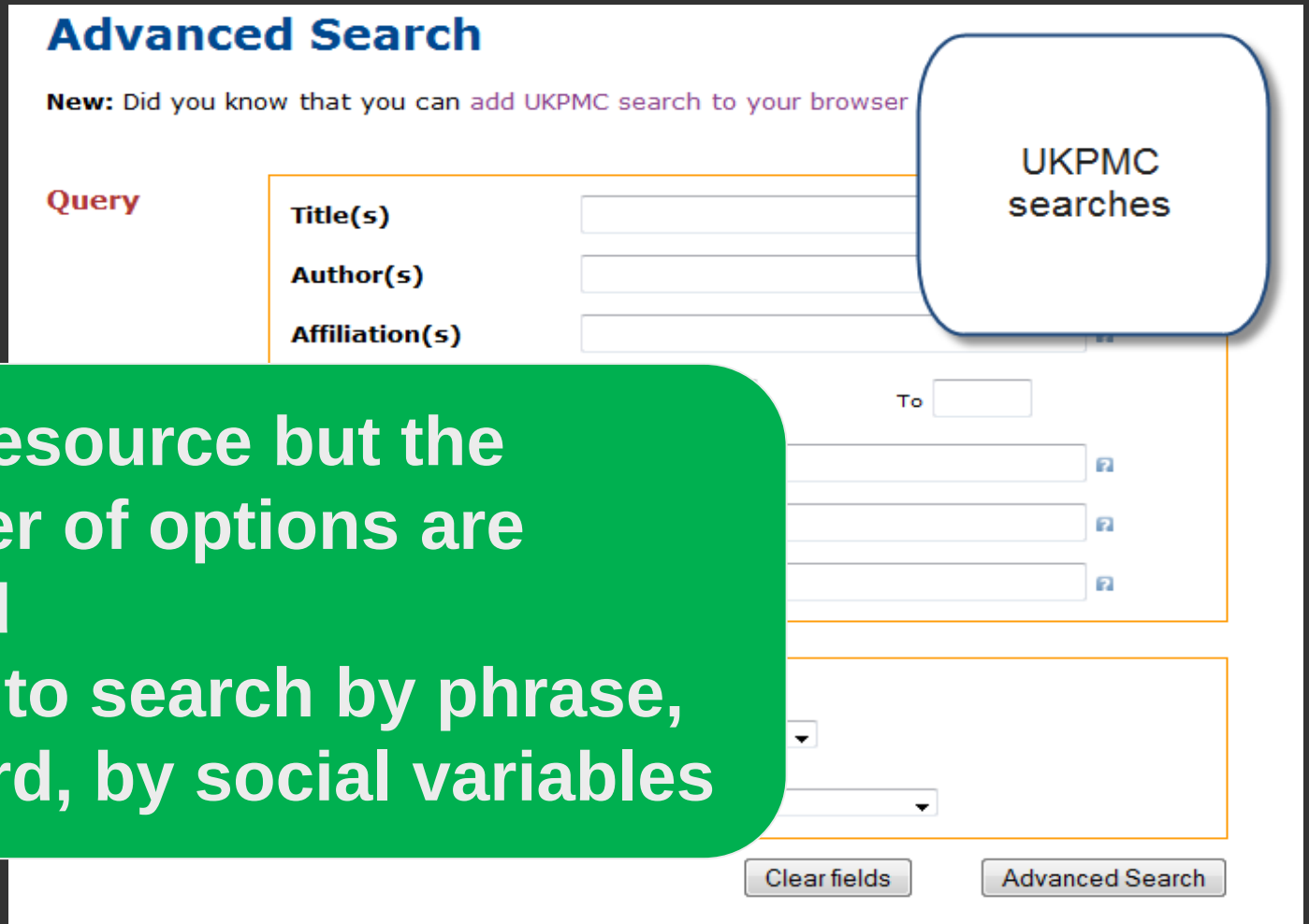
Collaboration

- 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

UNIVERSITY OF BIRMINGHAM

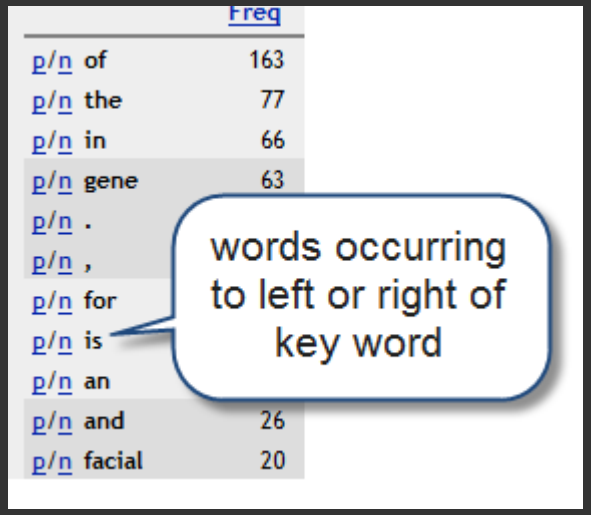
Archives

- **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



reaction to the change in their mother's	expression	makes it very hard for their mothers to
expressions into equation (2) gives: This	expression	may be simpl
transcribed from the phyA promoter this	expression	occurred at
created from expression builder window and an	expression	of (5 × 500 ×
nonmotile strain (FDR875) regulates the	expression	of 2363 gene
using the expression builder again and an	expression	of [Torsiona
taken for granted assumptions. They are the	expression	of a change
flowering time in terms of morphological	expression	of a plant body controlled by particular
regulation (activation or repression) of gene	expression	of a variety of homeobox containing transcript n
. Misexpression of Wnt3a induces ectopic	expression	of AER-specific genes, and application
advantage of a second niche for replication.	Expression	of an autophagy-like process facilitates
. The spatial and temporal regulation of	expression	of any given maternal effect gene often
also has problems in recognising facial	expressions	of basic emotions and my find it difficult
expression is also significantly reduced.23	Expression	of BMP signalling targets are increased
chromosomal abnormalities will prevent the	expression	of certain genes, somatic cell mutations
cotyledons in our CHS:Gus plants, as CHS drives	expression	of chalcone synthase, an enzyme involved

Key word/phrase placed in centre and lines sorted by the first word to the right - this helps the analyst to identify patterns

Pattern searching

omplex number systems Consider an	expression of the form	where all the are arbitrary reals
ercentage elongation (Equation 2) is an	expression of the ductility	where lf is the distance between gauges
ercentage elongation (Equation 2) is an	expression of the ductility where lf	is the distance between gauges marks when
moter, we aimed to study the changing	expression of the chlorophyll	a/b binding protein over time after exposure
r 35S was selected to allow constitutive	expression of the reporter	gene
r 35S was selected to allow constitutive	expression of the reporter gene	in a
second promoter CAB is normally drives	expression of the chlorophyll	a/b
fact that Gli3 activity is not required for	expression of the genes	do
ing graph is a QQ plot for the first gene	expression of the control	agai
lation (activation or repression) of gene	expression of a variety	of hom
lation (activation or repression) of gene	expression of a variety of homeobox	containing transcription factors. A homeobox
sure that when this is the case, greater	expression of the binding protein	or in our case the reporter gene luciferase
dermal development by directly induces	expression of the GATA	factors MED-1 and MED-2, which are required

the pattern:
"expression of" + 0-2 words + a noun

UNIVERSITY OF BIRMINGHAM

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

UNIVERSITY OF BIRMINGHAM

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

UNIVERSITY OF BIRMINGHAM

Why and how?

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

UNIVERSITY OF BIRMINGHAM