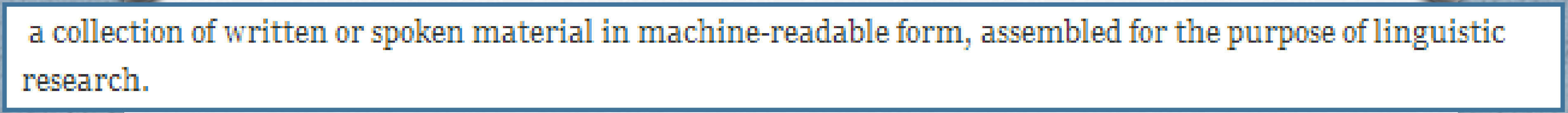
Corpus Research and BEAR

Dr Paul Thompson

Centre for Corpus Research



a collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research.

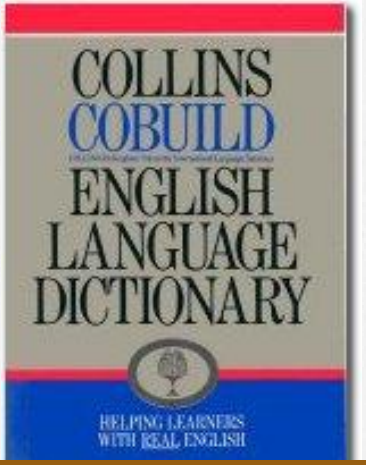
Corpus linguistics is an evidence-based approach to the study of language

Corpus linguistics is typically concerned with patterning and frequency in large collections of text

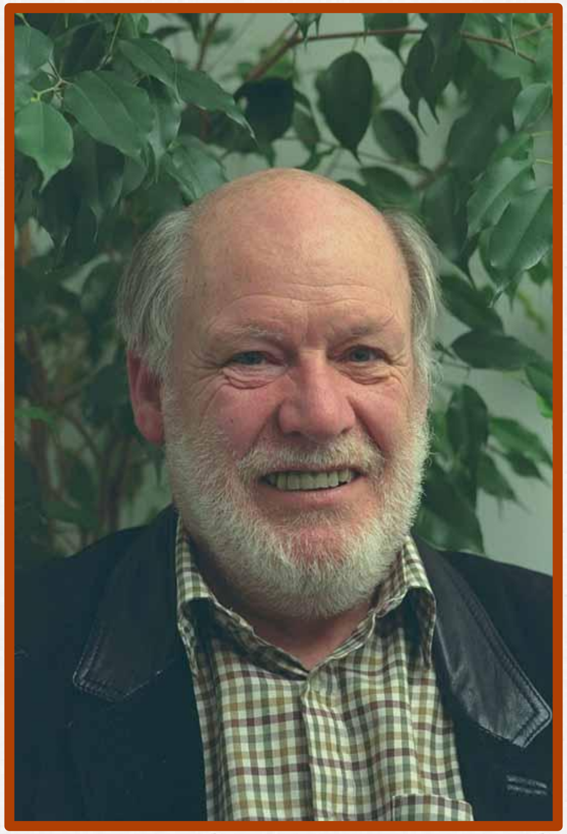
In early days, corpus linguists mainly worked with orthographic forms of data but now some are working with multimodal data

Corpus linguistics at Bham

- COBUILD project
  - John Sinclair
  - Began in 1980



Collins Birmingham University International Language Database

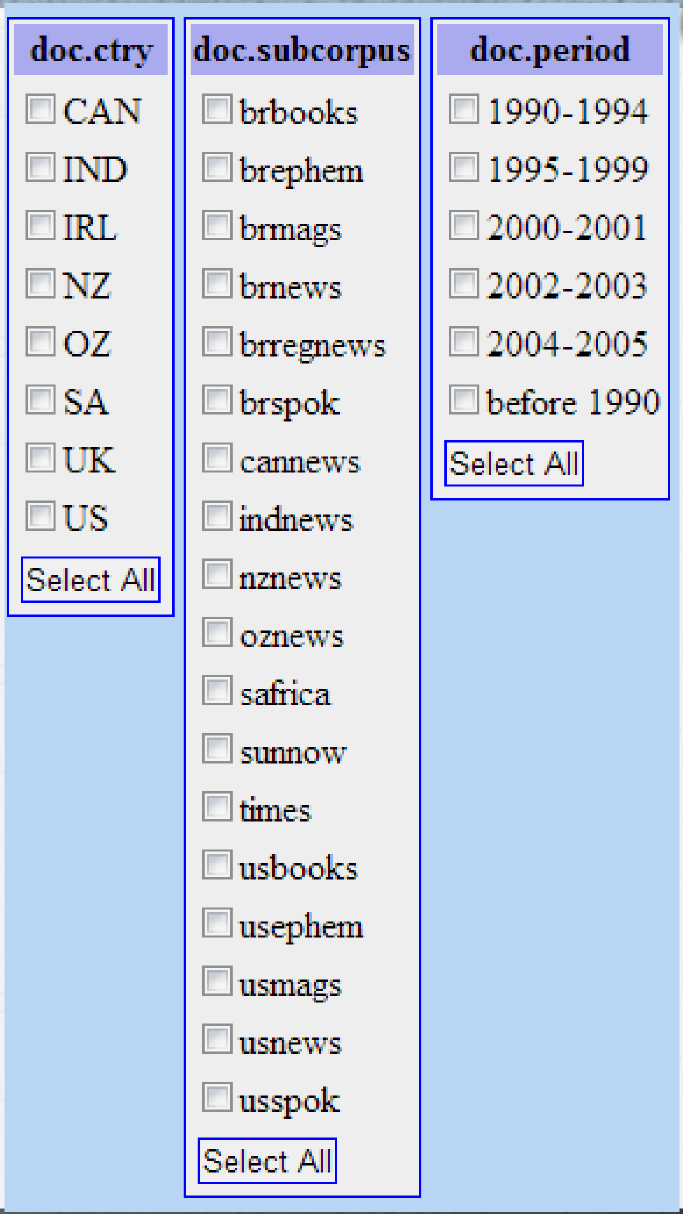


Bank of English

Texts from newspapers, fiction, magazines, etc.

Regional variation:
UK, US, Canada, Australia, South Africa

Always growing – the 'monitor' corpus



doc.ctry

□ CAN
□ IND
□ IRL
□ NZ
□ OZ
□ SA
□ UK
□ US

Select All

doc.subcorpus

□ brbooks
□ brephem
□ brmags
□ brnews
□ brregnews
□ brspok
□ cannews
□ indnews
□ nznews
□ oznews
□ safrica
□ sunnow
□ times
□ usbooks
□ usephem
□ usmags
□ usnews
□ usspok

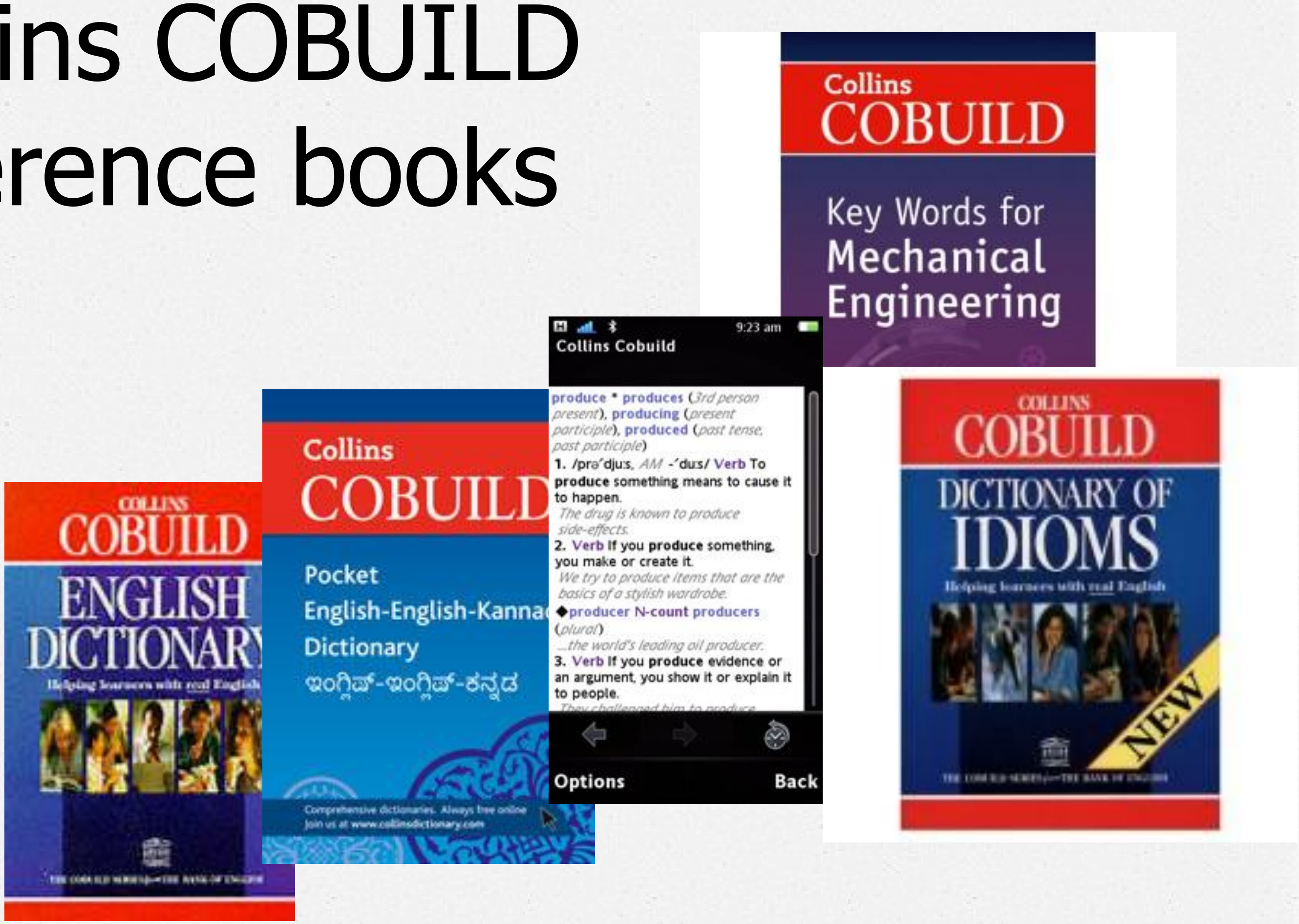
Select All

doc.period

□ 1990-1994
□ 1995-1999
□ 2000-2001
□ 2002-2003
□ 2004-2005
□ before 1990

Select All

Collins COBUILD reference books


COLLINS
COBUILD
ENGLISH DICTIONARY
Helping learners with real English
THE COBUILD SERIES from THE BANK OF ENGLISH
Collins
COBUILD
Pocket
English-English-Kannad
Dictionary
ಇಂಗ್ಲಿಷ್-ಇಂಗ್ಲಿಷ್-ಕನ್ನಡ
Comprehensive dictionaries. Always free online
Join us at www.collinsdictionary.com
Collins
COBUILD
Key Words for
Mechanical
Engineering
COLLINS
COBUILD
DICTIONARY OF
IDIOMS
Helping learners with real English
NEW
THE COBUILD SERIES from THE BANK OF ENGLISH

9:23 am

Collins Cobuild

produce * **produces** (*3rd person present*), **producing** (*present participle*), **produced** (*past tense, past participle*)

1. /prəˈdjuːs, *AM* -ˈduːs/ Verb To **produce** something means to cause it to happen.
The drug is known to produce side-effects.

2. Verb If you **produce** something, you make or create it.
We try to produce items that are the basics of a stylish wardrobe.

◆**producer** N-count **producers** (*plural*)
...the world's leading oil producer.

3. Verb If you **produce** evidence or an argument, you show it or explain it to people.
They challenged him to produce

New developments @ccr.bham

1. Elsevier journal article corpus
2. Social interaction corpus
3. ‘Open resource’ monitor corpus

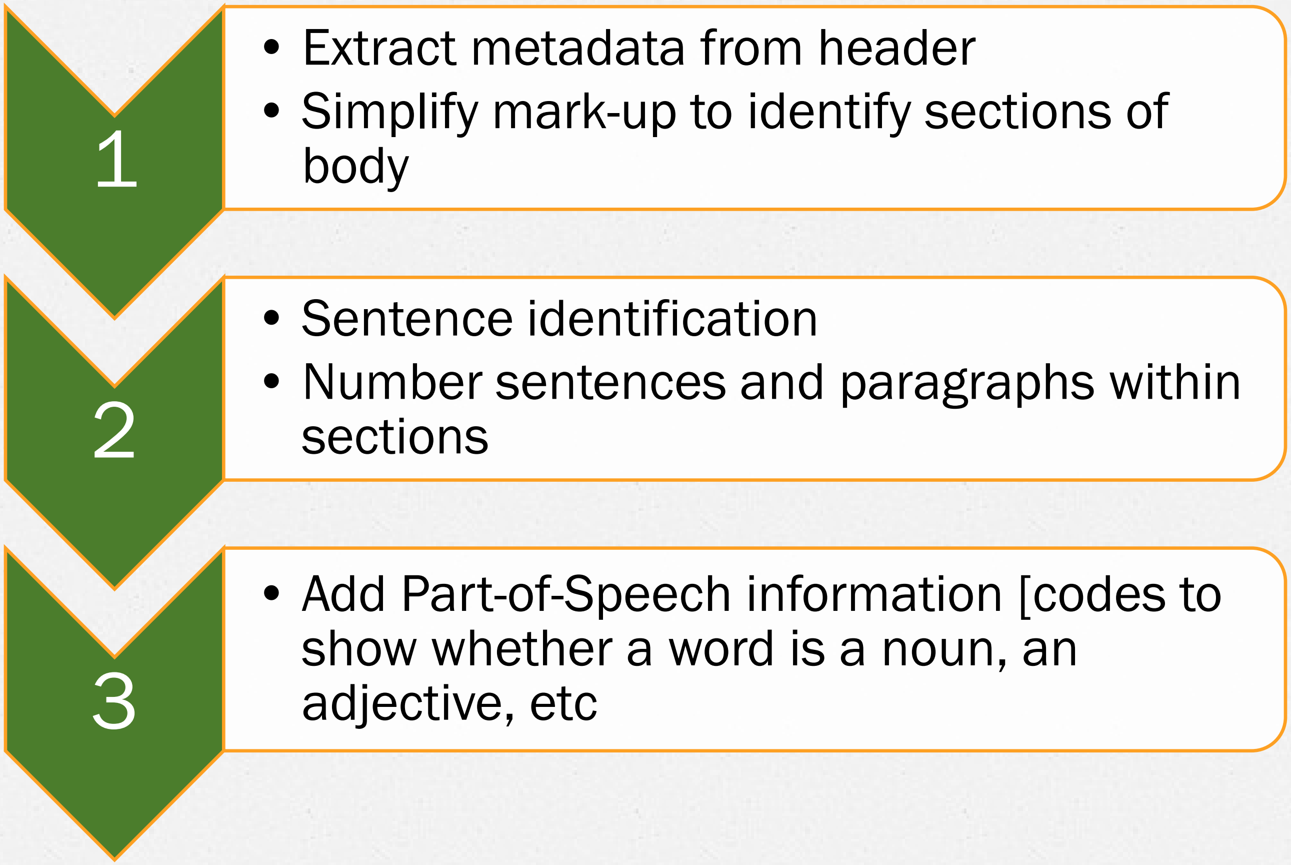
1

Elsevier journal article corpus

- Access to thousands of files
- All articles are XML files
- From Elsevier’s catalogue of c. 2000 scientific journals
- One aim in preparing texts for corpus treatment:
  - Preserving ‘textness’ of texts using ‘textual coordinates’

1

Processing XML files



1

- Extract metadata from header
- Simplify mark-up to identify sections of body

2

- Sentence identification
- Number sentences and paragraphs within sections

3

- Add Part-of-Speech information [codes to show whether a word is a noun, an adjective, etc

1

Sentence 1 out of 3, in the 5th paragraph out of 26 paragraphs

```
<p n="p5.26">
<s n="s1.3;p5.26">Demographic changes illustrate that
marri                                                          s
being
<s n=
as ha
<s n=
conse
not a
does
</p>
```

With POS tags

```
<s n="s1.3;p5.26">
<w pos="AJ0">Demographic</w> <w
pos="NN2">changes</w>
<w pos="VVB">illustrate</w> <w pos="DT0">that</w>
<w pos="NN1">marriage</w> <pos="VBZ">is</w>
<w pos="AV0">increasingly</w> <w
pos="VVG">becoming</w>
<w pos="VVN">seen</w> <w pos="CJS">as</w>
<w pos="AJ0">outdated</w> <w pos="CJC">and</w>
<w pos="VBZ">is</w> <w pos="VBG">being</w>
<w pos="VVN">supplanted</w> <w pos="PRP">by</w>
 <w pos="NN1">cohabitation</w> <w pos="SENT">.</w></s>
```

Types of investigation

- What linguistic features are characteristic of:
  - Introductions
  - Methods
  - Results
  - Discussion sections?
- What are the most common phrases and semantic structures and whereabouts in paragraphs and section do they tend to occur?

- Early days of corpus work: description of language use in general
- A more functional account of language – how language is used to express meanings and how texts are structured

Applications

- Helping apprentice writers
- … and apprentice readers
- Adapted approach can be used in literary studies, to follow themes

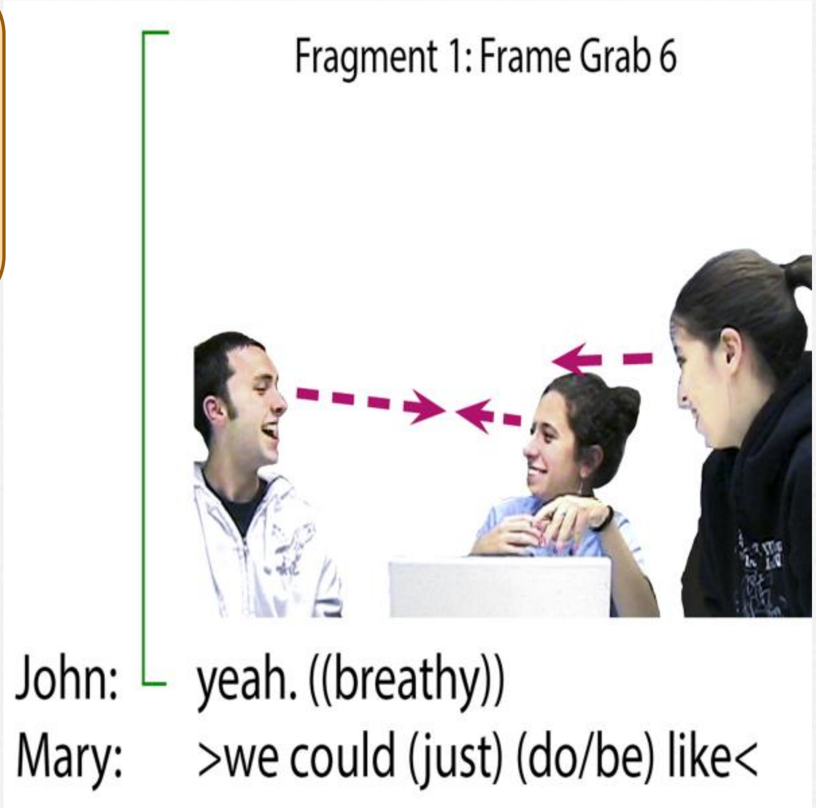
Theory

- More fine-grained discourse models
- New models of language knowledge

Multimodal corpora

Collaborative project with Numa Markee, University of Illinois at Urbana-Champaign

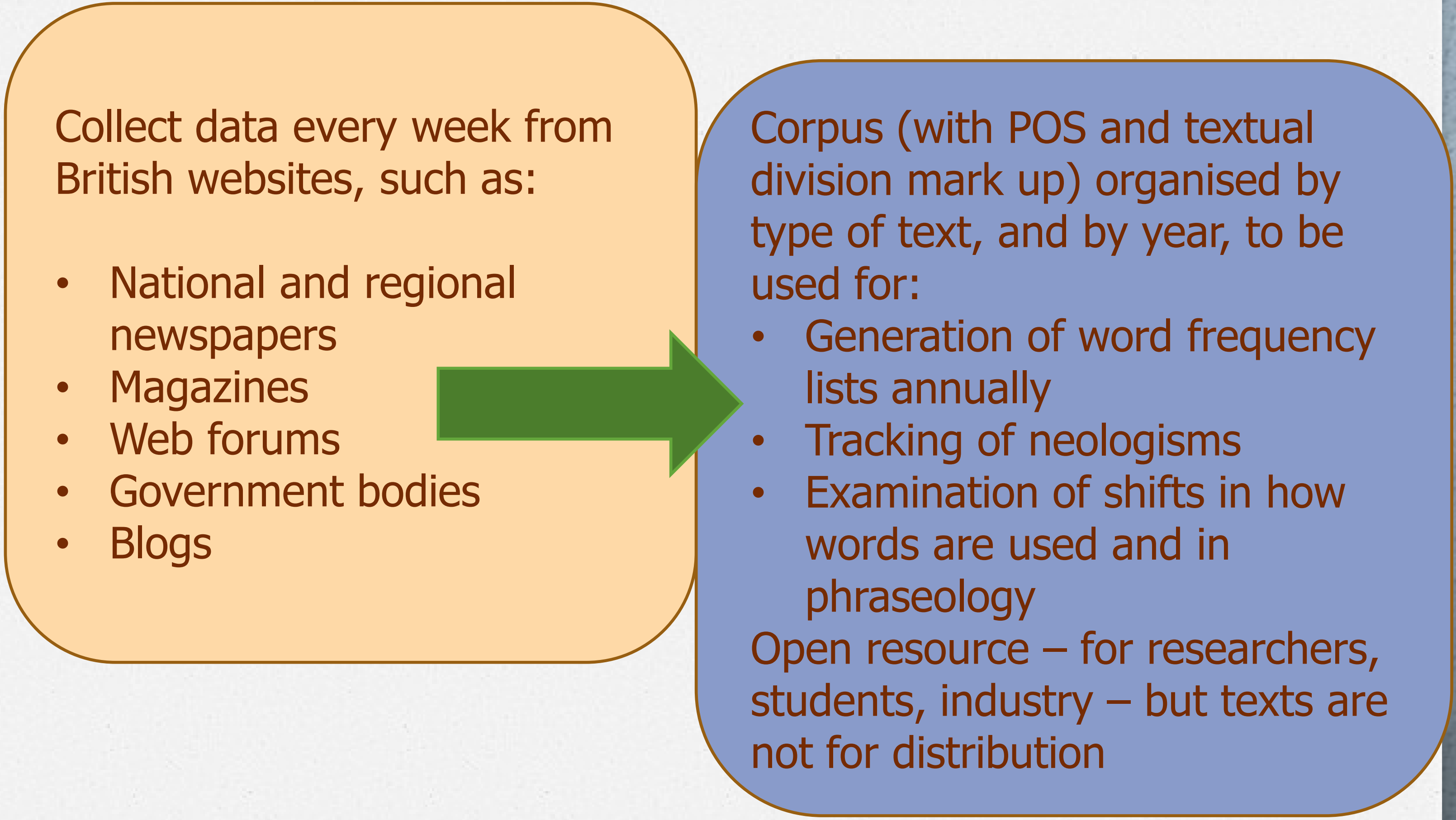
Linking annotated orthographic transcription of spoken interactions (task-related) to audio and video tracks of event



Fragment 1: Frame Grab 6

John: yeah. ((breathy))
Mary: >we could (just) (do/be) like<

New monitor corpus



Collect data every week from British websites, such as:

- National and regional newspapers
- Magazines
- Web forums
- Government bodies
- Blogs

Corpus (with POS and textual division mark up) organised by type of text, and by year, to be used for:

- Generation of word frequency lists annually
- Tracking of neologisms
- Examination of shifts in how words are used and in phraseology

Open resource – for researchers, students, industry – but texts are not for distribution

Challenges

- How to make sense of large quantities of data?
  - New ways of visualizing the dispersion of language features over texts, over time, between text types
  - Identifying correlations between linguistic features and other contextual features