

Corpora by Web Services

Adam Kilgarriff

Lexical Computing Ltd

Lexicography MasterClass Ltd

Universities of Leeds and Sussex

Starting a PhD in NLP

- Then
 - Prolog
 - Type in a few
 - grammar rules
 - Lexical entries
 - Example sentences
 - We're off!

Now

- Corpus
 - Which?
 - Budget/schedule
 - How much can we afford?
 - Hard disk space
- Access software
 - Build
 - Big job, making it fast is hard – or
 - Research, acquire, install, maintain ...

-
- Research question
 - Morphology, syntax, discourse structure, semantics, anaphor
 - First six months at least
 - Acquiring data, software
 - Complications



Malta, May 2010

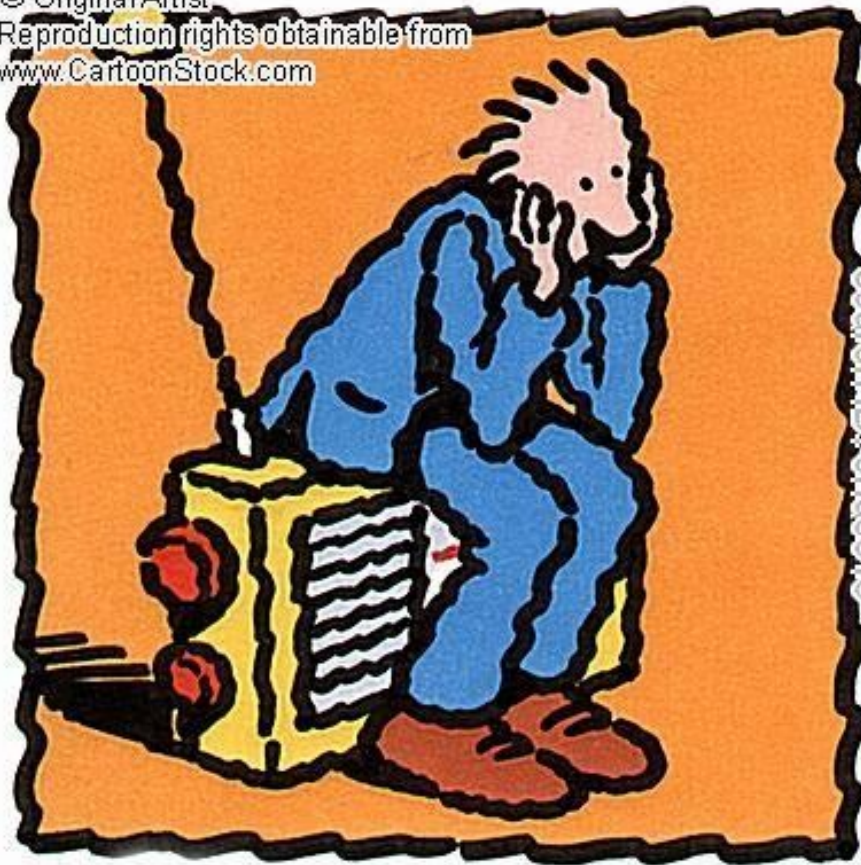
Kilgarriff: Corpora by Web Services

If you're not super-geeky

- Did I do it properly?
- Dumbing down
 - Let's choose an easier question
- Looking over shoulder

Disappointment

© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com



search ID: ena0013

Making it easy

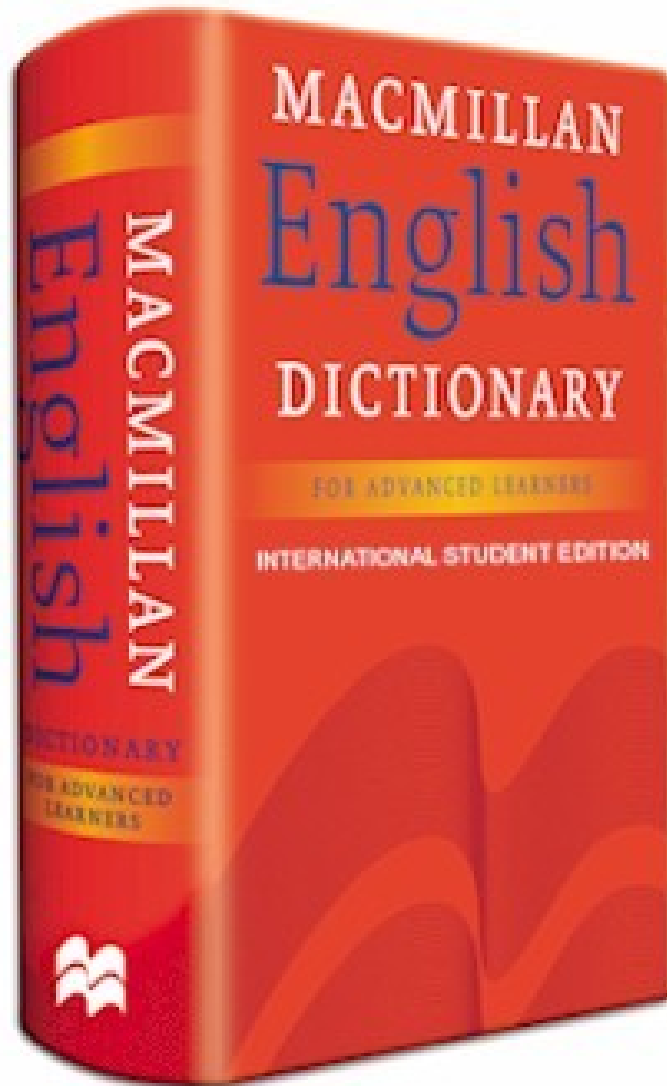
- Like picking up a hire car

Corpora by web services

- Possible?
- Already available

Sketch Engine

- Corpus querying
- Fast
- Handles large corpora
- In use for lexicography at
 - OUP, CUP, Macmillan, Collins, Le Robert
- Word sketches
 - Data-driven summary of a word's grammatical and collocational behaviour



Corpora

Arabic	174	Hindi	31	Russian	188
Chinese	456	Indonesian	102	Slovak	536
Czech	800	Irish	34	Slovene	738
Dutch	128	Italian	1910	Spanish	117
English	5508	Japanese	409	Swedish	114
French	126	Norwegian	95	Telugu	5
German	1627	Persian	6	Thai	108
Greek	149	Portuguese	66	Vietnamese	174
		Romanian	53	Welsh	63

Big, High Quality corpora

- Big
 - Performance
 - Banko and Brill 2004
 - There's no data like more data
 - Ample data for rare phenomena
 - Big subcorpora
 - 5b
 - Medical: 30m

Quality

- Bad data
 - Spam
 - Navigation-bars
 - Duplicates
 - Lists
 - Bungled formatting
 - Wrong language
 - ...
- Less discussed
 - Maybe a footnote
- Quick fixes and run



The Google/Yahoo/Bing option

- Appeal
 - Not setup costs
 - Start googling today

but

- Limited hits-per-query
- Limited hits-per-day
- Sort order
 - 'unsorted' not possible
- Snippets too short for research
- No (documented) morphology
- Limited query syntax

and

-
- At mercy of commercial company
 - Might change at any time
 - Not replicable

So

- Appeal
 - No setup costs
- Serious research
 - Many difficult practical issues
 - **Not** a tool designed for linguists
- Conclusion
 - If only SE indexes are big enough
 - Yes
 - Else no

Strategy

- More languages
 - Corpus Factory, as Sharoff
- Bigger and better (English)
 - Big Web Corpus (BiWeC)
 - 5.5b fully processed
 - Rich markup
 - New Model Corpus
 - Collaboration model

TEDDCLOG

- Taiwan English Data-Driven CLOze Generation
- with Simon Smith and colleagues, Taipei
- API case study

Cloze

□ 'fill-the gap'

■ *Several metal _____ violently with cold water*

□ A: behave

□ B: react

□ C: realise

□ D: respond

□ Popular with students, teachers, testers

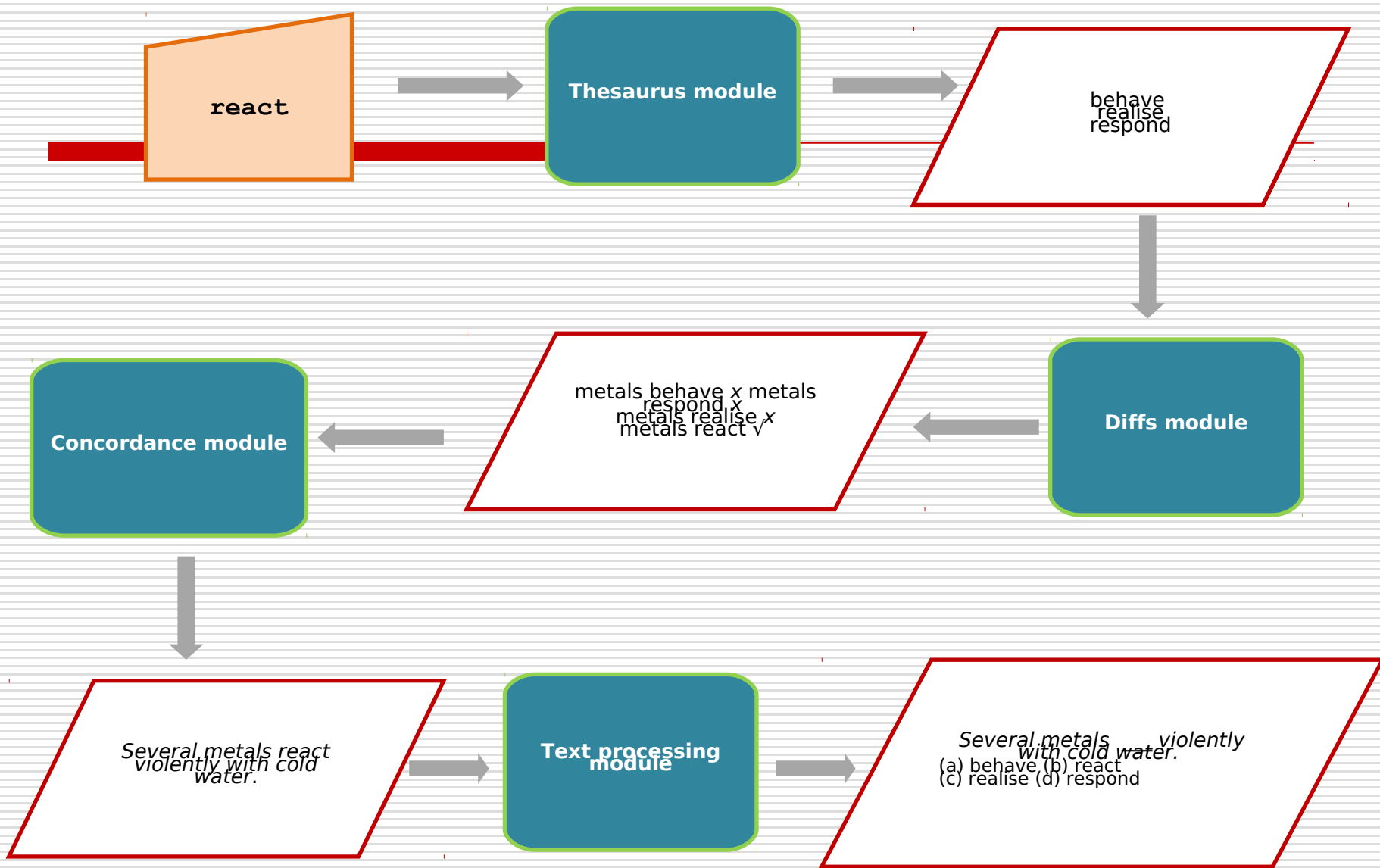
■ Unpopular with theorists :-)

One objection

- Test item writers make them up
- Not naturally-occurring language
 - *The Sinclair-Johns critique*

 - Also: expensive

- TEDDCLOG
 - Uses corpus sentences and distractors



API calls

- Find distractorts
 - thesaurus
- Find key-only collocate
 - Sketch diffs
 - Needs optimising
- Find carrier sentence
 - Concordance with GDEX module
 - Good Dictionary Example Finder

Current status

- TEDDCLOG
 - Next phase: producing decent results
- Corpora by Web Services
 - Increasing server capacity
 - Looking for users

Not just

like picking up a hire car

Not just

like picking up a hire car

more like *picking up a Ferrari*

Another announcement: DANTE

- **Lexical database for English**
 - Detailed Accurate Extensive of English
 - Highly corpus-driven
 - 3 yr project
 - 18 expert lexicographers
 - Led by Sue Atkins
 - BNC, FrameNet, Euralex, COBUILD...
- English side, New English-Irish dictionary
- Available for NLP research imminently