Conference for Teachers of A Level English Language

Corpora and language teaching

Costas Gabrielatos
Lancaster University

1 July 2011

Outline

- Corpora: definition and types

- The role of corpora and corpus-based research: Three overviews

- Related issues:
  - Intuition & introspection vs. Attested use
  - Rules and exceptions vs. Patterns and probabilities
  - Prescription vs. Description

- Corpus research and pedagogical materials

- Corpora in the (virtual) classroom

- Corpora, teachers and learners

What is a corpus?

Loose definition

- *A body of text.*

Common definition

- *A body of machine-readable text.*
- *A text database.*

Strict definition

- *A finite collection of **machine-readable** text, **sampled** to be maximally **representative** of a language or variety.*

Definitions from McEnery & Wilson (2001: 197).

A rough-and-ready home-made corpus

- Students of class X submit essays in e-form.
- Teacher puts all term Y essays in a folder

⇩

Corpus

- What is it representative of?
- The writing of class X during term Y.

Language teaching: A core distinction

Information **given** to learners

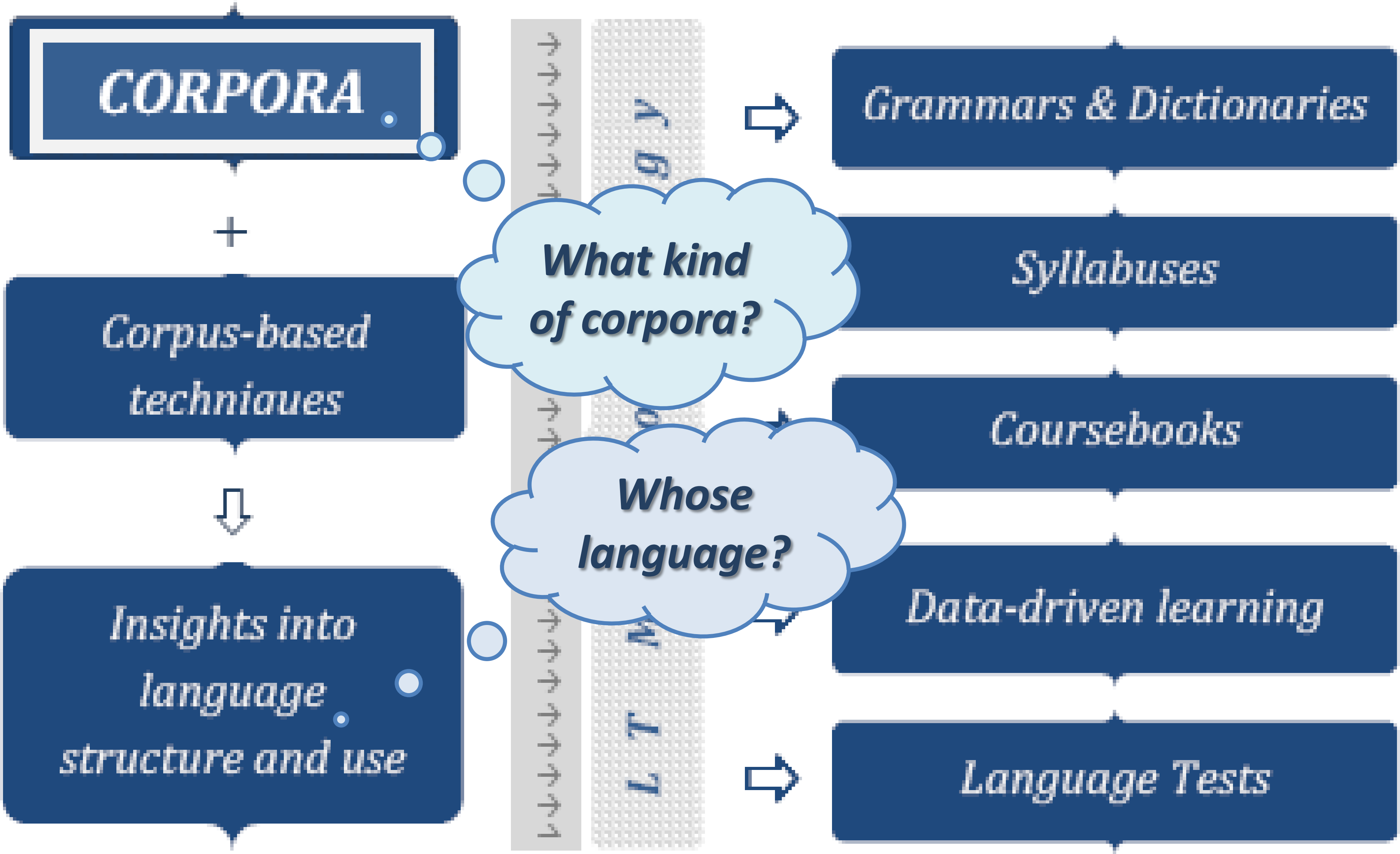
- Corpus research informs pedagogical materials:
  - dictionaries, grammars, syllabuses, exercises

Information **discovered** by learners

- Learners work with ...
  - ... corpus-derived materials (soft version)
  - ... corpora and corpus software (hard version)

(Leech, 1997: 10)

A more detailed picture



CORPORA

+

Corpus-based techniaues

Insights into language structure and use

Grammars & Dictionaries

Syllabuses

Coursebooks

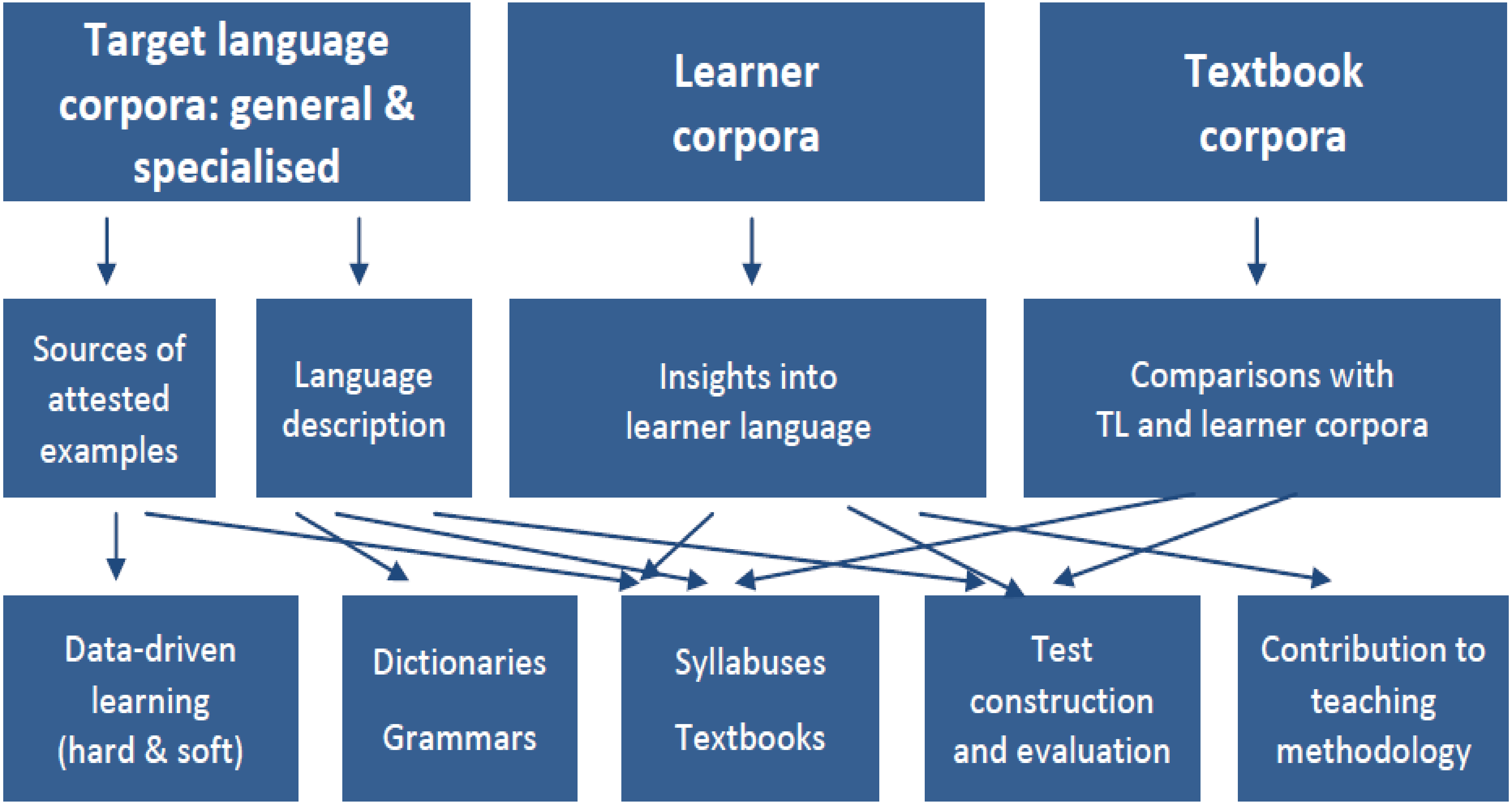
Data-driven learning

Language Tests

What kind of corpora?

Whose language?

The complex picture



Target language corpora: general & specialised

Learner corpora

Textbook corpora

Sources of attested examples

Language description

Insights into learner language

Comparisons with TL and learner corpora

Data-driven learning (hard & soft)

Dictionaries

Grammars

Syllabuses

Textbooks

Test construction and evaluation

Contribution to teaching methodology

Adapted from Gabrielatos (2005) and McEnery & Gabrielatos (2006)

Corpora and pedagogical information

What is wrong with
information, examples or exercises
based on intuition or introspection?

Intuition/introspection is not always dependable

"Question tags, along with bowler hats, mostly belong to 1960s BBC broadcasts."

(Bradford, 2002: 13)

"About every fourth question in conversation is a question tag."

(Biber et al., 1999: 211)

[Grammar based on the *Longman Spoken and Written English Corpus,* 40 million words]

Intuition/introspection is not a good frequency guide

- 1000 word types account for ??% of written texts
- 4000-5000 word types account for about ??% of written texts
- 50 high frequency function words account for about ??% of spoken language.

(Nation, 1990)

- About ??% of clauses (simple sentences) in written English have modal marking (e.g. *may, can, should*).

(Gabrielatos, 2010)

Intuition/introspection is not a good frequency guide

- 1000 word types account for 85% of written texts
- 4000-5000 word types account for about 95% of written texts
- 50 high frequency function words account for about 60% of spoken language.

(Nation, 1990)

- About 28% of clauses (simple sentences) in written English have modal marking (e.g. *may, can, should*).

(Gabrielatos, 2010)

Pattern recognition: a visual metaphor

Implications for ...

Pedagogical rules and guidelines

Discovery learning

(Gabrielatos, 2005)

Introspection - or a small data sample

A pattern

↓

Rule

X

XX

XXXX

XXXXXX

XXXXXXXX

XXXXXXXXX

XXXXXXXXXXX

XXXXXXXXXXXXX

XXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXX

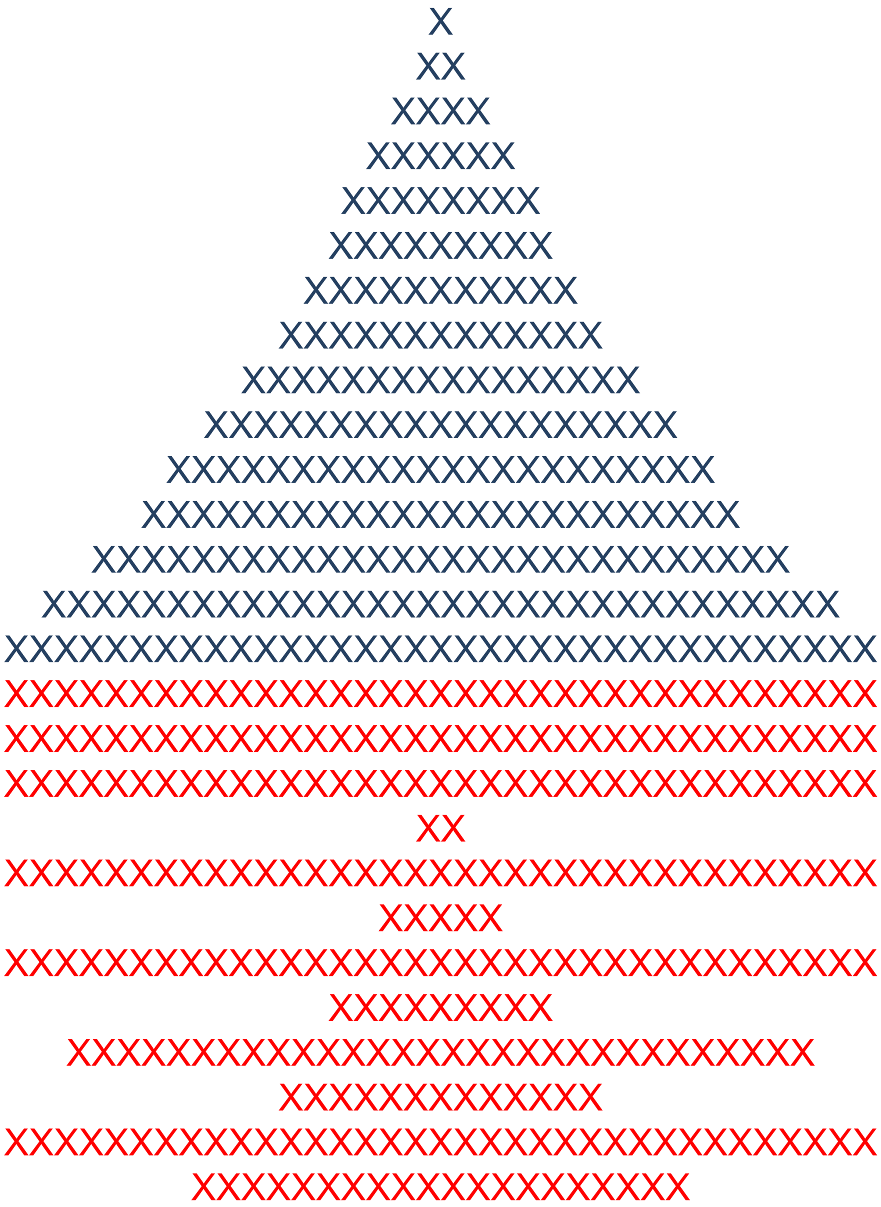
XXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

More (introspective) data

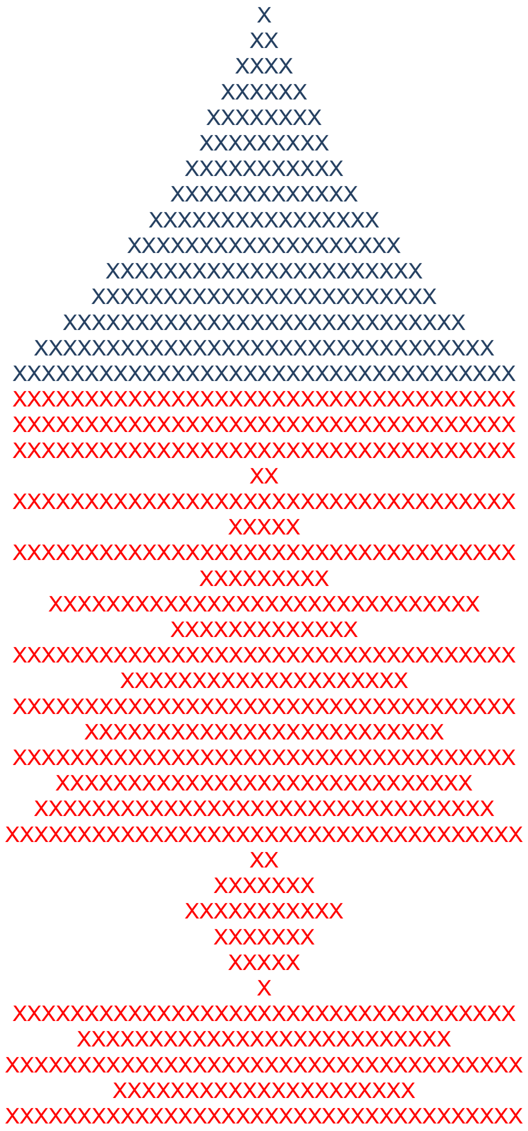


Rule

Exceptions

Some ‘irregularities’

More data



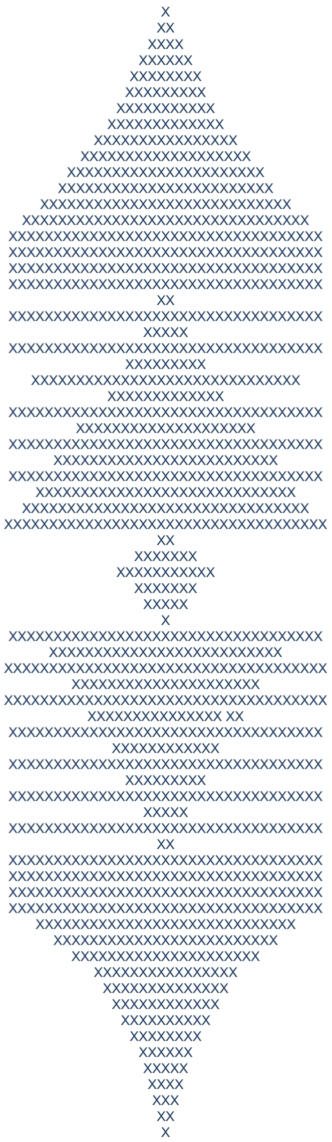
more ‘irregularities’
→ more exceptions

If a rule has a large number of exceptions,
then there is something wrong with the rule!

Are we perhaps looking at **part** of the actual pattern?

More data – a corpus sample

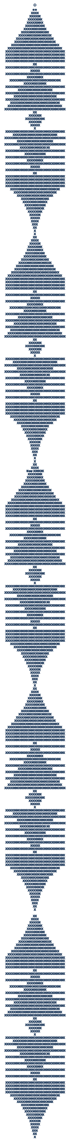
This seems to be the pattern …



… but more data will increase confidence

Even more data – a larger corpus sample

We can now be more confident that this is the pattern …



… however, adding more data may reveal more complexity.

Intuition/Introspection as sources of pedagogical information

- If the language information given to learners is based *only* on intuition/introspection …
- If the examples/texts provided are chosen or constructed to *reflect* these intuitions/introspections …

- … learners may be presented with - and accept – the personal informal observations of teachers or materials writers, or their preferred variety or idiosyncratic usage, as the only ‘truth’.
- … language myths and prescriptions can be perpetuated.
- … learners may avoid using constructions or patterns not included in introspection-based information.

Learner corpora: uses and issues

Contrastive Interlanguage Analysis

Comparisons:

- Learner use vs. expert use.
- Learners with different MT learning the same FL
- Learners with the same MT learning different FLs

(e.g. Granger 1998: 12-14, Altenberg 1997: 119)

- Error analysis
  - Error categories
  - Frequent errors
  - Patterns according to L1, age, medium, genre, task type, task context

Word frequencies (per mil.) in the Longman Learner Corpora and the BNC

(Adapted from Gillard and Gadsby, 1998: 161-162)

	BNC	LLC
	(L1)	(L2)
nice	77	**752**
happy	109	**689**
big	220	**990**
enormous	**42**	33
massive	**44**	10
huge	**82**	63

Learner discourse styles

- Ringbom (1998): learners use words of high generality – e.g. *people*, *things*. Resulting in a vague style: 'Technology has made many things easier.'
- deCock, et al (1998): learners use vague phrases LESS - e.g. 'stuff like that', 'that sort of thing'
- Lorenz (1998): learners modify adjectives more frequently – gives discourse a sense of 'overstatement' e.g. 'The sea was very clean and the people there were very happy.'
- Flowerdew (2000): learners writing shows an under-use of hedging devices – high degree of certainty.
- Petch-Tyson (1998): learners use language that shows high involvement - e.g. pronouns: *I* and *you*.
- Causes?
  - L1?
  - Pedagogical materials?
  - Teaching?

Corpora in the (virtual) classroom

Data-driven learning:
text-based vs. corpus-based

Language exposure: time and focus

- In ‘traditional’ teaching contexts, contact time is limited.
  - ➢ EFL learners lack opportunities for rich language exposure and recognition of patterns.
- Extensive reading is seen as a good way to develop intuitions in the same way that native speakers do.
- However ...
  - ➢ out-of-class time for language learning may also be limited.
  - ➢ the amount of extensive reading required may have been under-estimated.
  - ➢ Extensive reading cannot be targeted.
- Corpora offer “condensed” and targeted exposure to patterns.

Learning through reading: an example

- One page contains 500 words on average.
- The written BNC contains 90 million words
  = approx. 180,000 pages
- A six-year programme of five hourly lessons per week offers a total of about **1000 lessons**.
- To get the amount of language evidence contained in the BNC through reading over six years, a learner would need to …
  - Examine about 180 pages per lesson
    (intensive reading)
  - Read about 80 pages per day = 2-3 books per week
    (extensive reading)

Data-driven learning:
text-based vs. corpus-based

Aim: Teaching verbs combining with the noun *diet*

Text-based

Three texts on dieting = 2,250 words = 100+ concordance lines

Only 12 instances of noun*diet* → only 5 combinations with verbs

1	'I **went** on a very drastic detox	**diet**	last year, and it didn't work - I
2	heart disease are from unhealthy	**diet**	- cardiac experts are keen to stre
3		**Diet**	Guidelines Aimed at Healthy People
4	ent's suggestion they direct their	**diet**	advice to overweight Americans.
5	people, you begrudgingly **go on** a	**"diet ."**	Your initial concept of a
6	"Your initial concept of a	**diet**	is more commonly known as STARVATION
7	vicious cycle as you **went from one**	**diet**	**to the next**. Every new
8	Every new	**diet**	**started** with hope and promise, and
9	y thinking, "Oh great, another fad	**diet**	with a catchy name and empty promi
10	NO! The Eat and Burn	**diet**	identifies over 100 foods that tur
11	The Eat & Burn	**diet**	is easy to **follow**. You don't feel
12	concept behind The Eat and Burn	**diet**	is this: eat foods that safely forc

Condensed reading

- BNC: random 100 concordances of noun *diet*
- 50+ verbs

Soft version

- Teachers can prepare concordances of particular patterns:
  - verb + preposition + any word + *diet*
  - verb + article + *diet*

Hard version

- Learners identify patterns
- Focus can easily switch to *diet* as a verb
- Students may discover patterns outside lesson focus (*serendipity*)

Lexical meaning and use

Learners as lexicographers

Egregious
Dictionary definitions

- Conspicuously bad or offensive.
- Often of mistakes, extremely and noticeably bad.
- An egregious mistake, failure, problem etc is extremely bad and noticeable.
- Extraordinary in some bad way; glaring; flagrant: *an egregious mistake; an egregious liar*.
- Conspicuous ; *especially* : conspicuously bad : flagrant <*egregious* errors> <*egregious* padding of the evidence — Christopher Hitchens>

Egregious
Dictionary definitions

- Conspicuously **bad** or offensive.
- Often of mistakes, extremely and noticeably **bad**.
- An egregious mistake, failure, problem etc is extremely **bad** and noticeable.
- Extraordinary in some **bad** way; glaring; flagrant: *an egregious mistake; an egregious liar*.
- Conspicuous ; *especially* : conspicuously **bad** : flagrant <*egregious* errors> <*egregious* padding of the evidence — Christopher Hitchens>

Egregious
Dictionary definitions

- **Conspicuously bad** or offensive.
- Often of mistakes, extremely and **noticeably bad**.
- An egregious mistake, failure, problem etc is extremely **bad** and **noticeable**.
- Extraordinary in some **bad** way; **glaring**; flagrant: *an egregious mistake; an egregious liar*.
- Conspicuous ; *especially* : **conspicuously bad** : flagrant <*egregious* errors> <*egregious* padding of the evidence — Christopher Hitchens>

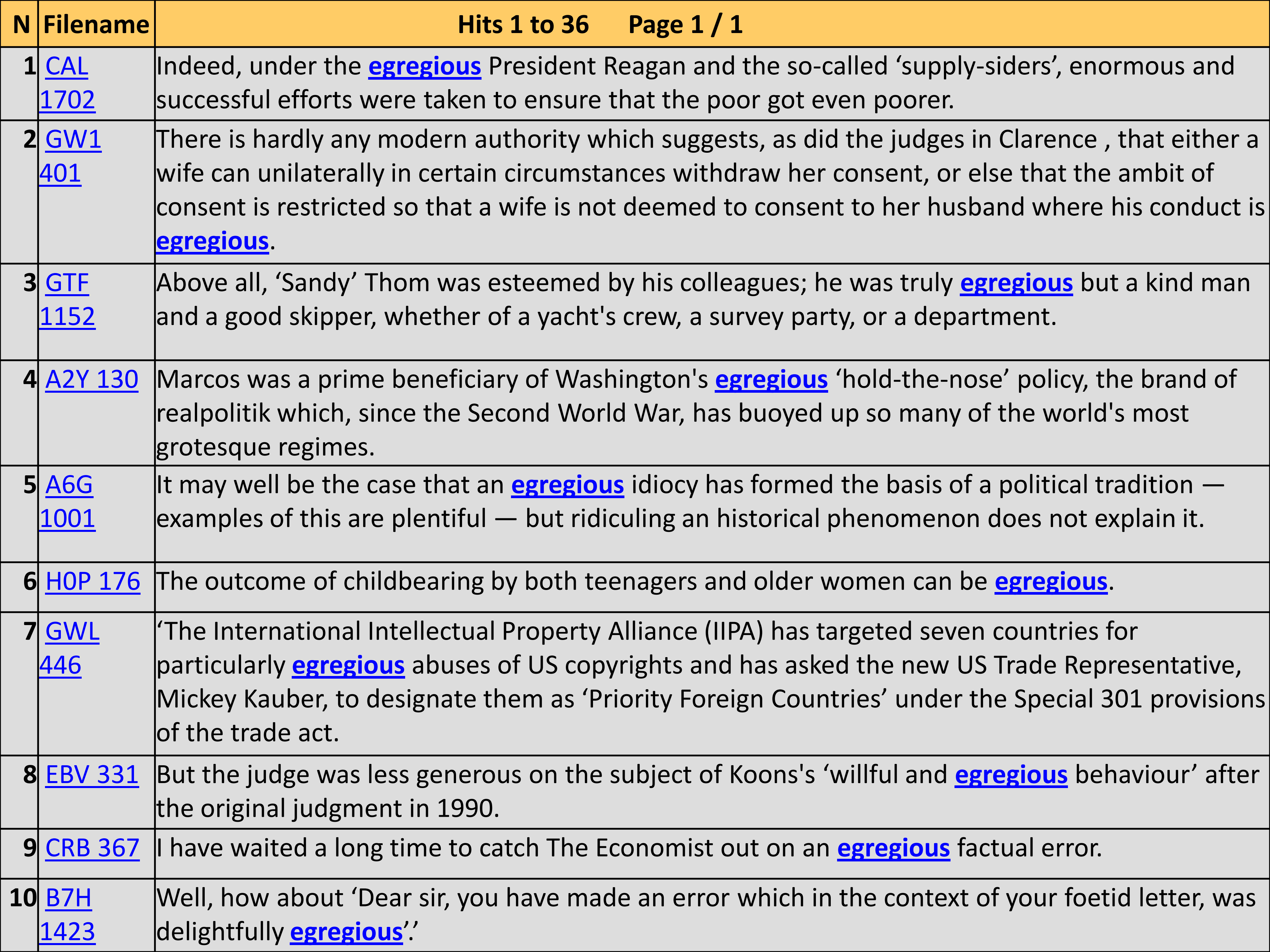
Egregious
Dictionary definitions

- Conspicuously bad or offensive.
- Often of **mistakes**, extremely and noticeably bad.
- An egregious **mistake**, **failure**, **problem** etc is extremely bad and noticeable.
- Extraordinary in some bad way; glaring; flagrant: *an egregious* ***mistake****; an egregious* ***liar***.
- Conspicuous ; *especially* **:** conspicuously bad **:** flagrant <*egregious* **errors**> <*egregious* **padding of the evidence** — Christopher Hitchens>

egregious in the BNC

- 36 instances

- Action/behaviour/event/result etc. (15)
- Person/organisation etc. (12)
- Mistake/error etc. (6)
- Object (2)
- Other (1)



N	Filename	Hits 1 to 36 Page 1 / 1
1	CAL 1702	Indeed, under the **egregious** President Reagan and the so-called ‘supply-siders’, enormous and successful efforts were taken to ensure that the poor got even poorer.
2	GW1 401	There is hardly any modern authority which suggests, as did the judges in Clarence , that either a wife can unilaterally in certain circumstances withdraw her consent, or else that the ambit of consent is restricted so that a wife is not deemed to consent to her husband where his conduct is **egregious**.
3	GTF 1152	Above all, ‘Sandy’ Thom was esteemed by his colleagues; he was truly **egregious** but a kind man and a good skipper, whether of a yacht's crew, a survey party, or a department.
4	A2Y 130	Marcos was a prime beneficiary of Washington's **egregious** ‘hold-the-nose’ policy, the brand of realpolitik which, since the Second World War, has buoyed up so many of the world's most grotesque regimes.
5	A6G 1001	It may well be the case that an **egregious** idiocy has formed the basis of a political tradition — examples of this are plentiful — but ridiculing an historical phenomenon does not explain it.
6	H0P 176	The outcome of childbearing by both teenagers and older women can be **egregious**.
7	GWL 446	‘The International Intellectual Property Alliance (IIPA) has targeted seven countries for particularly **egregious** abuses of US copyrights and has asked the new US Trade Representative, Mickey Kauber, to designate them as ‘Priority Foreign Countries’ under the Special 301 provisions of the trade act.
8	EBV 331	But the judge was less generous on the subject of Koons's ‘willful and **egregious** behaviour’ after the original judgment in 1990.
9	CRB 367	I have waited a long time to catch The Economist out on an **egregious** factual error.
10	B7H 1423	Well, how about ‘Dear sir, you have made an error which in the context of your foetid letter, was delightfully **egregious**’.’

Never begin a sentence with *because*

because in the *written* BNC

- All instances: 77,939
- Sentence-initial: 4,544 (*5.8%*)
- Rare, but used.

~~*Should I use it?*~~

When should I use it?
How should I use it?

Sentence-initial *because*

(100 random instances)

Result mentioned in the previous sentence (19)

- Time was on his side, after all. **Because**, after all, no-one had, as yet, told him that Presley City was going to be little more than blackened rubble in just two days time. [HTU 197]

Focus on cause/reason rather than result (*71*)

- In this respect it contrasted with the institution of monogamy. **Because** it was an impermanent unofficial arrangement, the pairing family held no property and at first presented no threat to the communally held property of the gens. [A6S 787]

Link to previously mentioned element (*10*)

- It is our life support system and the life support system of the whole globe. **Because** of that, it is important that we take the right steps at the right time to ensure the right results. [BN4 303]

Representation of groups

1 Investigating representations

For the Investigation task candidates should investigate how texts might produce social values and how they might contribute to maintaining or changing values.

Candidates may study texts used to represent:

- social groups (e.g. according to gender, ethnicity, disability, sexuality, age, class)
- individuals (e.g. a celebrity)
- events and issues (e.g. a war, death, work)
- institutions (e.g. the monarchy).

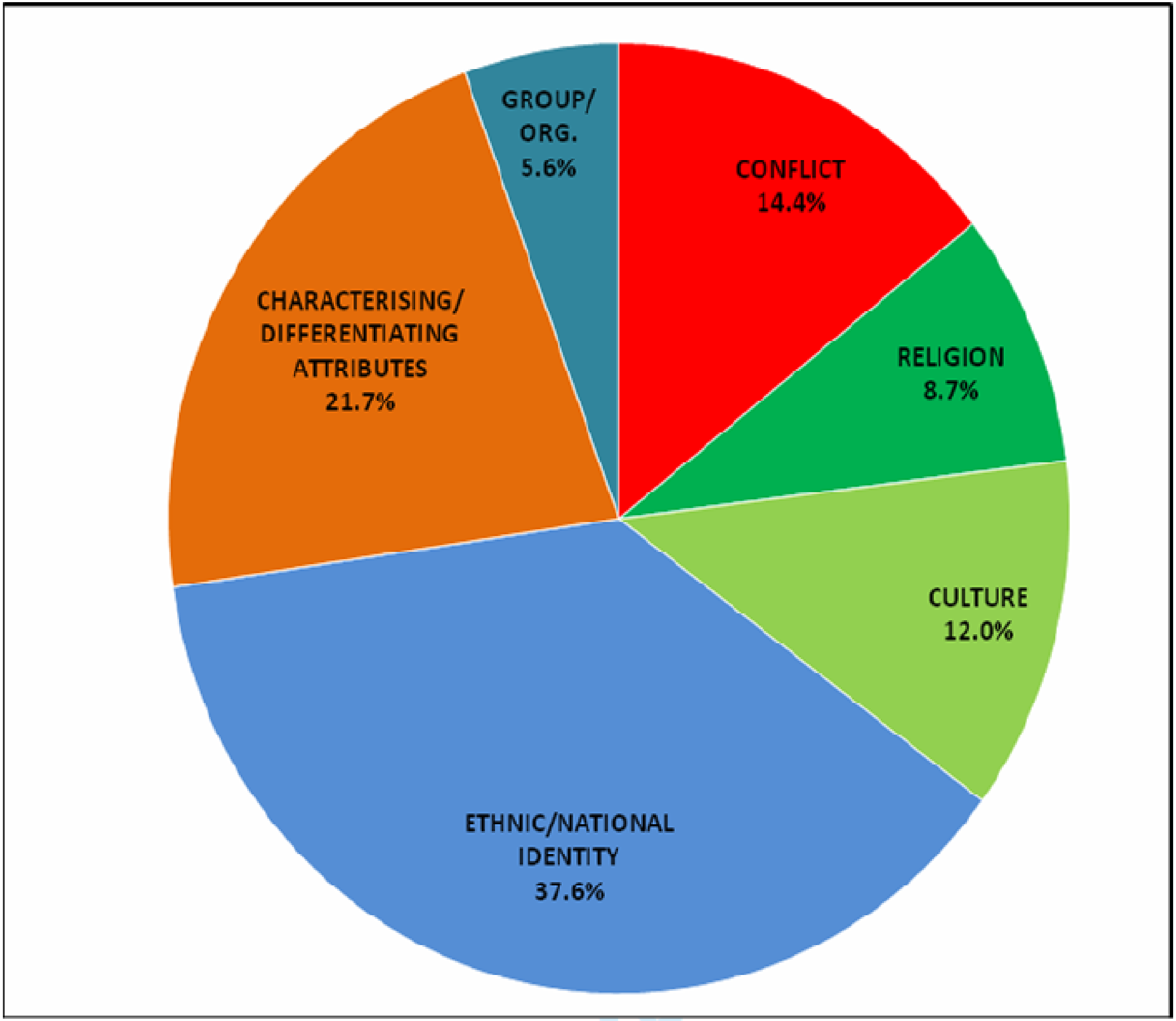
Work should focus on how lexis, grammar, semantics and discourse structure in individual texts produce representations.

[*GCE AS and A Level Specification. English Language A* (p. 8)]

Muslim + noun in 12 UK national newspapers (1998-2009): Categories

CONFLICT	*extremist, fanatic, terrorist, fundamentalist*	CONFLICT
RELIGION	*cleric, faith, month (=Ramadan), preacher*	RELIGION
CULTURE/PRACTICE	*festival, dress, culture, name, tradition*	CULTURE
EDUCATION	*school, teaching, education, college*	
VIEW/ATTITUDE/ EMOTION	*opinion, anger, voice, attitude, grievance*	
POPULATION	*community, population, nation, world*	ETHNIC/NATIONAL IDENTITY
AREA/COUNTRY	*country, state, area, region, land*	
GOVERNANCE	*leader, voter, MP, government, ruler*	
ETHICITY/NATIONALITY	*Briton, Albanian, Malay, Arab*	
GROUP/ORGANISATION	*group, organisation, association, charity*	GROUP
AGE/SEX	*woman, man, girl, youth, child, teenager*	CHARACTERISING/ DIFFERENTIATING ATTRIBUTES
FAMILY/RELATIONSHIP	*family, parent, brother, friend, wife*	
OCCUPATION/ROLE	*officer, patient, doctor, worker, assistant*	
OTHER	*house, shop*	

Baker et al. (under review); Gabrielatos et al. (2010)


GROUP/
ORG.
5.6%
CONFLICT
14.4%
CHARACTERISING/
DIFFERENTIATING
ATTRIBUTES
21.7%
RELIGION
8.7%
CULTURE
12.0%
ETHNIC/NATIONAL
IDENTITY
37.6%

Corpora and language teaching: Unresolved issues

Corpora in LT: prerequisites

Materials writers

- Informed by corpus research
- Informed by corpus-based reference books

Teachers

- Aware of nature, availability and use of corpora
- Aware of issues in data-driven language investigation.
  - ➢ Language awareness and corpus use incorporated in teacher preparation courses (e.g. O'Keefe & Farr, 2003)

Learners

- Aware of nature of corpora
- Trained in corpus use.

Language teaching institutions

- Investment in corpora, software and know-how.

Potential problems: language description

- Generalising from corpora / samples that are ...
  - ... non-representative
  - ... too small
  - ... inappropriate
- Why? → Using what is freely available rather than what is suitable.

- Corpus worship:
  - discarding intuitions
  - over-emphasis on frequency
  - lists: word forms vs. form+sense combinations

- Corpus studies depend on counting ...
  - ... which depends on labelling ...
  - ... which depends on theories ...
  - ... which depend on intuition/introspection

Potential problems: language teaching

- ‘Doing corpora’.
- New prescription (e.g. frequency worship).
- Focus on lexis and grammar – neglecting language skills (reading/listening, writing/speaking) .
- Focus on awareness – neglecting production.

Corpora and LT/LL: central characteristics

- Attested rather than constructed examples
- A wealth of examples of use in different contexts
- Amount of co-text according to needs
- From prescription to description
- From rules and exceptions to patterns and frequencies
- Idiosyncratic uses ‘diluted’ in large corpora ...
- ... but also enough data to examine idiosyncratic uses
- Facilitation of discovery learning
- Learner-centred methodology: learner as language researcher.

Resources

Free corpus interfaces

(Online access to corpora through corpus software – no download needed)

British National Corpus (BNC): http://corpus.byu.edu/bnc/

Corpus of Contemporary American English: http://corpus.byu.edu/coca/

Google Books (American English): http://googlebooks.byu.edu/

TIME Magazine (1920s-2000s): http://corpus.byu.edu/time/

WebCorp (using the web as a corpus): http://www.webcorp.org.uk/t.html

Free / affordable corpus analysis software

(You need to input the corpus)

AntConc (free): http://www.antlab.sci.waseda.ac.jp/antconc_index.html

WordSmith (single user: £50): http://www.lexically.net/wordsmith/

Corpus information and resources

ICT for ELT (Corpora): http://www.fi.muni.cz/ICT4ELT/websites/all/corpusbasedls.html

Corpus Resources: http://pioneer.chula.ac.th/~awirote/ling/corpuslst.htm

KCU 4992	**Thank you!**
KD3 5833	**Thank you!**
FPU 1944	‘**Thank you!**’
KDE 544	**Thank you!**
HGM 742	‘**Thank you!**’
FPB 1186	**Thank you!**
KB1 5413	**Thank you!**
F8M 738	**Thank you!**
KBL 4354	**Thank you!**
BMU 2243	**Thank you!**
F8B 178	**Thank you!**
AE0 2043	**Thank you!**
FX6 183	**Thank you!**
KE0 4520	**Thank you!**
KE1 2743	**Thank you!**
A0L 2978	‘**Thank you!**
KBL 2947	**Thank you!**
KB1 5403	**Thank you!**