# Corpora & Corpus-Informed Learning in EAP

Alex Gilmore
Department of Civil Engineering

---

*October 1, Building 10, 4F, Komaba Campus, University Of Tokyo, 15:30 ~ 17:00*

# What are corpora & what can they tell us?

- A *principled* collection of texts (spoken or written), stored on computer, which can be analyzed both *quantitatively* and *qualitatively* with the help of analytical software.

- *Representative* of whatever the corpus is designed to show:

  e.g. Corpus of classroom language would need to capture design variables such as student age & gender, location, level, class size, teacher characteristics (gender, qualifications, experience, nationality), etc. Size?

# What are corpora & what can they tell us?

- Synchronic corpora ~ show what a target language is like at a given time (usually present)

- Diachronic corpora ~ show how a target language changes through time (historical linguistics)

# Wordlists & word frequency

Top 40 most frequent words:
5m written

| | | | |
|---|---|---|---|
| 1 | THE | 21 | AS |
| 2 | TO | 22 | AT |
| 3 | AND | 23 | BUT |
| 4 | OF | 24 | BE |
| 5 | A | 25 | HAVE |
| 6 | IN | 26 | FROM |
| 7 | WAS | 27 | NOT |
| 8 | IT | 28 | THEY |
| 9 | I | 29 | BY |
| 10 | HE | 30 | THIS |
| 11 | THAT | 31 | ARE |
| 12 | SHE | 32 | WERE |
| 13 | FOR | 33 | ALL |
| 14 | ON | 34 | HIM |
| 15 | HER | 35 | UP |
| 16 | YOU | 36 | AN |
| 17 | IS | 37 | SAID |
| 18 | WITH | 38 | THERE |
| 19 | HIS | 39 | ONE |
| 20 | HAD | 40 | BEEN |

Top 40 most frequent words:
5m spoken

| | | | |
|---|---|---|---|
| 1 | THE | 21 | ON |
| 2 | I | 22 | OH |
| 3 | AND | 23 | WE |
| 4 | YOU | 24 | HAVE |
| 5 | IT | 25 | NO |
| 6 | TO | 26 | LAUGHS |
| 7 | A | 27 | WELL |
| 8 | YEAH | 28 | LIKE |
| 9 | THAT | 29 | WHAT |
| 10 | OF | 30 | DO |
| 11 | IN | 31 | RIGHT |
| 12 | WAS | 32 | JUST |
| 13 | IT'S | 33 | HE |
| 14 | KNOW | 34 | FOR |
| 15 | MM | 35 | ERM |
| 16 | IS | 36 | BE |
| 17 | ER | 37 | THIS |
| 18 | BUT | 38 | ALL |
| 19 | SO | 39 | THERE |
| 20 | THEY | 40 | GOT |

# Wordlists & word frequency

- Written corpus:

(a) *a* & *the* = high incidence of noun phrases

(b) *of* = post-modified noun phrases (The house *of* his aunt)

(c) *that* = subordinator (X reports *that*...; It is claimed *that*...)
*that* = relative pronoun (The house *that* Jack built)

(d) *to*, *for*, *in* = prepositional phrases (They drove *to* Scotland *for* a few days)

- Spoken corpus:

(a) Markers of interactivity: *I*, *you*, *yeah*, *know*, *mm*, *right*, *er*, *laughs*

# Wordlists & word frequency


- Compleat Lexical Tutor 'Vocabprofile' tool allows teachers to get a frequency breakdown for words in a target text, based on BNC's 1,000 - 20,000 word frequency levels

- Engineering journal article vs. Graded reader (Intermediate)

# Vocabprofile tool: word frequency breakdown

## Journal article

| Freq. Level | Families (%) | Types (%) | Tokens (%) | Cumul. token % |
|---|---|---|---|---|
| K-1 Words : | 378 (44.42) | 549 (42.04) | 7097 (62.86) | 62.86 |
| K-2 Words : | 175 (20.56) | 239 (18.30) | 994 (8.80) | 71.66 |
| K-3 Words : | 66 (7.76) | 79 (6.05) | 271 (2.40) | 74.06 |
| K-4 Words : | 72 (8.46) | 89 (6.81) | 408 (3.61) | 77.67 |
| K-5 Words : | 40 (4.70) | 47 (3.60) | 181 (1.60) | 79.27 |
| K-6 Words : | 24 (2.82) | 28 (2.14) | 276 (2.44) | 81.71 |
| K-7 Words : | 17 (2.00) | 19 (1.45) | 192 (1.70) | 83.41 |
| K-8 Words : | 12 (1.41) | 12 (0.92) | 187 (1.66) | 85.07 |
| K-9 Words : | 13 (1.53) | 16 (1.23) | 81 (0.72) | 85.79 |
| K-10 Words : | 14 (1.65) | 17 (1.30) | 220 (1.95) | 87.74 |
| K-11 Words : | 5 (0.59) | 5 (0.38) | 9 (0.08) | 87.82 |
| K-12 Words : | 13 (1.53) | 14 (1.07) | 153 (1.36) | 89.18 |
| K-13 Words : | 5 (0.59) | 5 (0.38) | 15 (0.13) | 89.31 |
| K-14 Words : | 10 (1.18) | 10 (0.77) | 24 (0.21) | 89.52 |
| K-15 Words : | | | | |
| K-16 Words : | 1 (0.12) | 1 (0.08) | 1 (0.01) | 89.53 |
| K-17 Words : | | | | |
| K-18 Words : | 4 (0.47) | 5 (0.38) | 30 (0.27) | 89.80 |
| K-19 Words : | 2 (0.24) | 2 (0.15) | 4 (0.04) | 89.84 |
| K-20 Words : | | | | |
| Off-List: | ?? | 242 (18.53) | 1147 (10.16) | 100.00 |
| Total (unrounded) | 851+? | 1306 (100) | 11290 (100) | 100.00 |

| *Pertaining to whole text* | |
|---|---|
| Words in text (tokens): | 11290 |
| Different words (types): | 1306 |
| Type-token ratio: | 0.12 |
| Tokens per type: | 8.64 |
| *Pertaining to onlist only* | |
| Tokens: | 10143 |
| Types: | 1064 |
| Families: | 851 |
| Tokens per family: | 11.92 |
| Types per family: | 1.25 |

**BNC-13,000 types: [ fams 5 : types 5 : tokens 15 ]**

nordic_[2] outcrop_[2] phosphorus_[7] synoptic_[3] uv_[1]

**BNC-14,000 types: [ fams 10 : types 10 : tokens 24 ]**

climatology_[4] con_[1] contour_[1] geophysics_[1] glacial_[1] hereinafter_[4] hydrological_[3] runoff_[2] seawater_[2] southernmost_[1] subsurface_[4]

**BNC-15,000 types: [ fams : types : tokens ]**

**BNC-16,000 types: [ fams 1 : types 1 : tokens 1 ]**

labile_[1]

**BNC-17,000 types: [ fams : types : tokens ]**

**BNC-18,000 types: [ fams 4 : types 5 : tokens 30 ]**

denitrification_[11] gyre_[12] hydrographic_[6] isotherm_[1]

**BNC-19,000 types: [ fams 2 : types 2 : tokens 4 ]**

biogenic_[2] fluvial_[2]

## Graded Reader

| Freq. Level | Families (%) | Types (%) | Tokens (%) | Cumul. token % |
|---|---|---|---|---|
| K-1 Words : | 310 (72.77) | 391 (73.08) | 1803 (85.73) | 85.73 |
| K-2 Words : | 65 (15.26) | 73 (13.64) | 115 (5.47) | 91.20 |
| K-3 Words : | 27 (6.34) | 29 (5.42) | 41 (1.95) | 93.15 |
| K-4 Words : | 9 (2.11) | 9 (1.68) | 11 (0.52) | 93.67 |
| K-5 Words : | 2 (0.47) | 2 (0.37) | 2 (0.10) | 93.77 |
| K-6 Words : | 4 (0.94) | 4 (0.75) | 12 (0.57) | 94.34 |
| K-7 Words : | 2 (0.47) | 2 (0.37) | 2 (0.10) | 94.44 |
| K-8 Words : | 2 (0.47) | 2 (0.37) | 2 (0.10) | 94.54 |
| K-9 Words : | 2 (0.47) | 3 (0.56) | 29 (1.38) | 95.92 |
| K-10 Words : | 2 (0.47) | 2 (0.37) | 7 (0.33) | 96.25 |
| K-11 Words : | 1 (0.23) | 1 (0.19) | 2 (0.10) | 96.35 |
| K-12 Words : | | | | |
| K-13 Words : | | | | |
| K-14 Words : | | | | |
| K-15 Words : | | | | |
| K-16 Words : | | | | |
| K-17 Words : | | | | |
| K-18 Words : | | | | |
| K-19 Words : | | | | |
| K-20 Words : | | | | |
| Off-List: | ?? | 39 (7.29) | 77 (3.66) | 100.00 |
| Total (unrounded) | 426+? | 535 (100) | 2103 (100) | 100.00 |

| *Pertaining to whole text* | |
|---|---|
| Words in text (tokens): | 2103 |
| Different words (types): | 535 |
| Type-token ratio: | 0.25 |
| Tokens per type: | 3.93 |
| *Pertaining to onlist only* | |
| Tokens: | 2026 |
| Types: | 496 |
| Families: | 426 |
| Tokens per family: | 4.76 |
| Types per family: | 1.16 |

**BNC-6,000 types: [ fams 4 : types 4 : tokens 12 ]**

VP-negative: k-6

blossom_[1] greenhouse_[2] housekeeper_[6] malaria_[1]

**BNC-7,000 types: [ fams 2 : types 2 : tokens 2 ]**

VP-negative: k-7

moreover_[1] shrug_[1]

**BNC-8,000 types: [ fams 2 : types 2 : tokens 2 ]**

VP-negative: k-8

indie_[1] malaysia_[1]

**BNC-9,000 types: [ fams 2 : types 3 : tokens 29 ]**

VP-negative: k-9

orchid_[27] thermometer_[2]

**BNC-10,000 types: [ fams 2 : types 2 : tokens 7 ]**

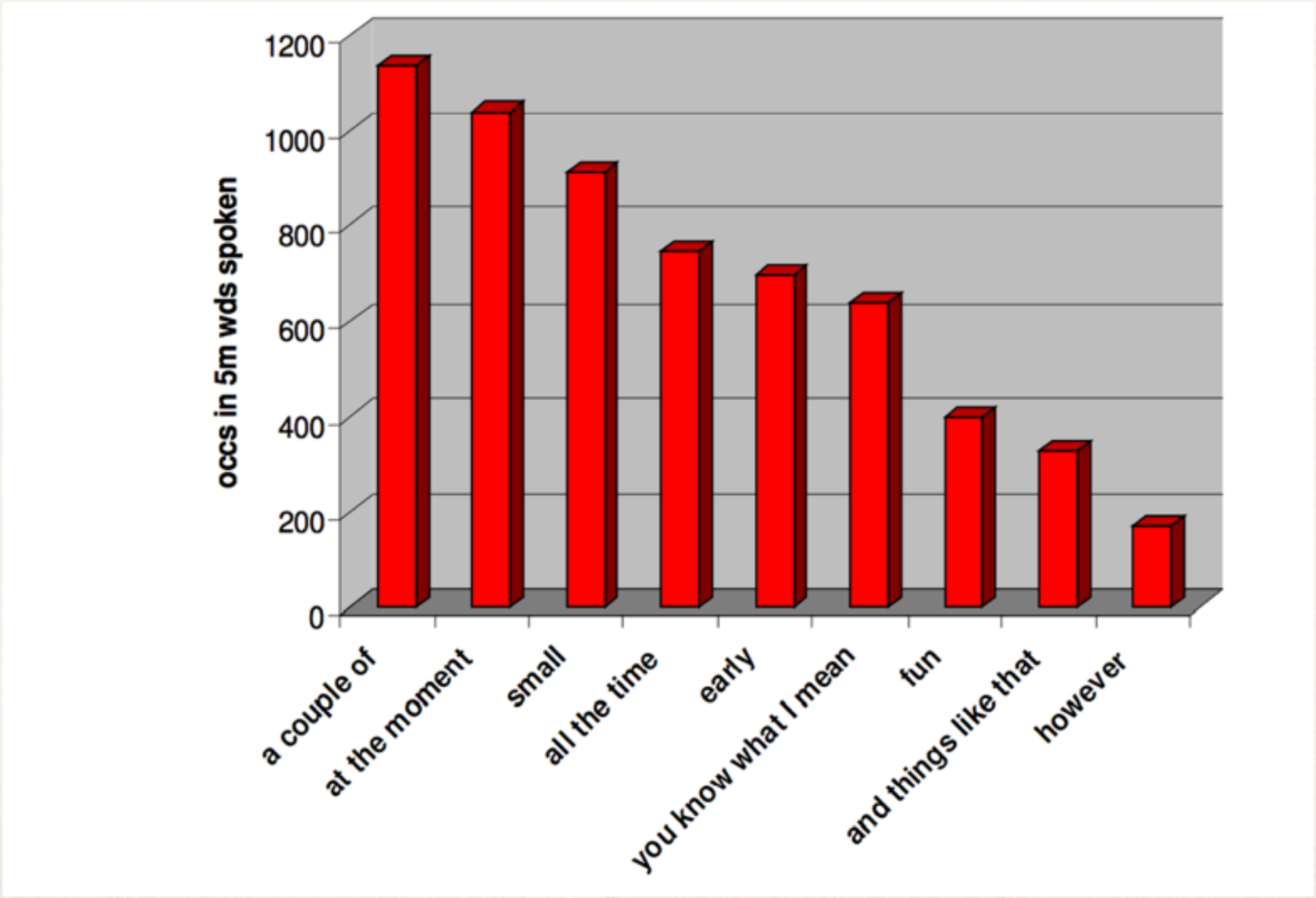
VP-negative: k-10

hothouse_[4] leech_[3]

**BNC-11,000 types: [ fams 1 : types 1 : tokens 2 ]**

batten_[2]

# Word-cluster frequency



# Sinclair's 'idiom principle' vs. 'open choice principle'

- The production of texts involves alternation between word-for-word combinations (open choice principle) & pre-constructed, multi-word combinations  (idiom principle).

- A large amount of language (~50%) is prefabricated

- (Chunks larger than 6 or 7 rare)

  (Erman & Warren, 2009)

# Concordance lines (Key Word in Context - KWIC)

Node

| |
|---|
| delivery of key speeches in the Supreme Soviet, for example, there was no **evidence** that prominent groups were gaining proportionately more influence vis-a-vis the party, which may demonstrate |
| latter's consent, unless there is a specific arrangement to the contrary, or **evidence** that the parties intended otherwise. This draft also has a separate provision relating to |
| in Butcher v Dowlen 1981RTR 24. 1.7 If the judge is undecided on the other **evidence** after he has heard both sides then the conviction is conclusive, but the judge |
| fait accompli of' socialist' reforms to enlist their total support. However the **evidence** suggests (see pages 75-6) that it is more likely that this decision was |
| the appearance of the national curriculum legislation in England and Wales can be interpreted as **evidence** of the government's determination to bring the educational system of Scotland into close alignment |
| life is his appointment to the Muftilik, it is necessary here to mention the **evidence** behind the tradition, while leaving till later the discussion of the validity of the |
| Any officer, servant or agent of the Bank may, on producing if required **evidence** of his authority, enter any premises occupied by any person on whom a notice |
| consistently recorded higher levels of total approval. At the same time there was some **evidence** that those who had a more general interest in politics, as distinct from those |
| of maintaining the standards reached in the new services. Some are beginning to show **evidence** of decay in the activity levels achieved and the associated staff performance. Other work |
| (The problems these women faced as workers are considered in Part II.) **Evidence** suggests that working class wives were prepared to put up with occasional drinking bouts by |
| the matter should be dealt with -- that is to say, by hearing oral **evidence**. Accordingly, Lord Meston submits that the mother's removal of the child on |
| a correlation between the presence of active inflammation and PEG absorption. There was little **evidence** to support the presence of a primary defect in the colonic barrier in patients with |
| model requires that general practitioners and managers develop new skills particularly in contracting. The **evidence** that contracting is an efficient mechanism in the NHS is still limited. General practitioners |
| minstrel songs of the tenth and eleventh centuries we know exceedingly little. There is **evidence**, however, of a steady development in oral tradition between the death of Count |
| and others' related to a time when he was in secure accommodation. The **evidence** in support of that is to be found in the evidence of Helen Taylor. |
| own bodies. # (Griffin 1984: 74-5) Now this is certainly intriguing **evidence** of the ability of some species to adapt to mirror images (problems gleefully exploited |
| acquired distinctiveness and equivalence solely in terms of associative mechanisms. Is there any positive **evidence** that might prompt us to adopt the more complex position that differentiation occurs as well |
| of the employment status of either mother or father and without proof of need or **evidence** of contributions. The acceptance of such a scheme meant that the government had accepted |
| the ruins of the buildings. A more likely suggestion would be that it is **evidence** of the estate workers continuing to live on the site and work the land long |

(BNC academic sub-corpus)

# Concordance lines

- Read from the key word (node) outwards, rather than from left to right
- Identify the central group of words which form a phrase or can stand alone
- Sorting alphabetically left or right of the node facilitates identification of common lexico-grammatical patterns in the concordance lines

# Concordance lines: Adjectives commonly used with *evidence*

| Type/Quality | Quantity | Time sequence |
|---|---|---|
| conclusive | enough | present |
| experimental | some | recent |
| scientific | little | |
| clear | no | |
| available | | |

# Collocation & colligation patterns

- Lexical items may be primed to co-occur with other words (collocation), e.g. lean + meat:

| |
|---|
| , red and green peppers, onions, chickpeas, kidney beans, tomatoes and **lean** diced ham, wholemeal roll with low-fat spread. Dinner Chicken casserole, boiled potatoes |
| sliced tomatoes, fruit with virtually fat-free fromage frais. Dinner Bolognese sauce made with **lean** mince, and wholemeal spaghetti. Virtually fat-free yoghurt for dessert. # FRIDAY |
| cottage cheese, wholemeal toast with low-fat spread and jam or yeast extract. Lunch **Lean** roast meat (visible fat removed), jacket-baked potato, lightly steamed green vegetables |
| # very low-fat cottage cheese # low-fat hard and soft cheeses. # MEAT # Very **lean** minced beef # Meat for roasting (very lean, with visible fat removed) |
| soft cheeses. # MEAT # Very lean minced beef # Meat for roasting (very **lean**, with visible fat removed). # FISH # Kipper -- or other oily |
| stick finely chopped celery and cook 5 minutes. Add 225g (8oz) very **lean** mince to pan, stirring until browned. Sprinkle 15ml (1 tbsp) flour |
| or yeast extract, very-low-fat curd or cottage cheese, mashed banana. Two rashers **lean** back bacon (visible fat removed) plus two tomatoes and 50g (2oz) |

- Lexical items may be primed to co-occur with a grammar words (colligation), e.g. possesive adj. + true feelings:

| |
|---|
| . Hope smiled to himself: the smile broadened, and to disguise his **true feelings** he turned the smile on Mr Crump; who was greatly encouraged as he had |
| before the Armistice was signed, she realised that she had been denying her **true feelings** for years. She had loved Connor from the day she set eyes on him |
| ; you do whatever you wish.' Stephen's controlled voice disguised his **true feelings**, but Christina sensed his jealousy and changed the subject.' Why don't |
| you're probably a very good husband, but you like to hide your **true feelings**." Oh, don't be so serious, Basil,' smiled |
| a statement of fact. Dexter was surprised that the TV presenter revealed her **true feelings** towards Nicola so quickly: most people in his experience, when first interviewed by |
| 'd ever seen, but which she always felt quite at odds with her **true feelings**. In fact, there were lots of things she'd like to change about |
| turned, looking at the girl, smiling at her reassuringly, keeping his **true feelings** from showing. Games. It was all one big game to DeVore. He |
| . It had been then, perhaps, that he had first realised his **true feelings** for her. Then that he had first articulated it inside his head. I |
| What else will happen?' she mumbled.' You'll discover my **true feelings** for you,' he said in a low tone.' Sounds ominous. |

# Collocation & colligation patterns

Blonde

| | | | |
|---|---|---|---|
| 1 | ☐ | HAIR | 141 |
| 2 | ☐ | GIRL | 19 |
| 3 | ☐ | WOMAN | 14 |
| 4 | ☐ | HEAD | 14 |
| 5 | ☐ | CURLS | 9 |
| 6 | ☐ | STRANDS | 5 |
| 7 | ☐ | HAIR-DYE | 3 |
| 8 | ☐ | BOMBSHELL | 3 |

(BNC fiction sub-corpus: common collocates of ???)

# Part of speech (POS)-tagging

```
^ JOURNAL_NN1 OF_IO GEOPHYSICAL_JJ RESEARCH_NN1 :_: OCEANS_NN2 ,_, VOL._NN1
118_MC ,_, 16251644Export_FO of_IO nutrients_NN2 from_II the_AT Arctic_JJ
Ocean_NN1 &lsqb;_( 1_MC1 &rsqb;_) This_DD1 study_NN1 provides_VVZ the_AT
first_MD physically_RR based_VVN mass-balanced_JJ transport_NN1 estimates_NN2
of_IO dissolved_JJ@ inorganic_JJ nutrients_NN2 (_( nitrate_NN1 ,_,
phosphate_NN1 ,_, and_CC silicate_NN1 )_) for_IF the_AT Arctic_JJ Ocean_NN1
```

CLAWS POS-tagging for a journal article:

NN1 = singular common noun

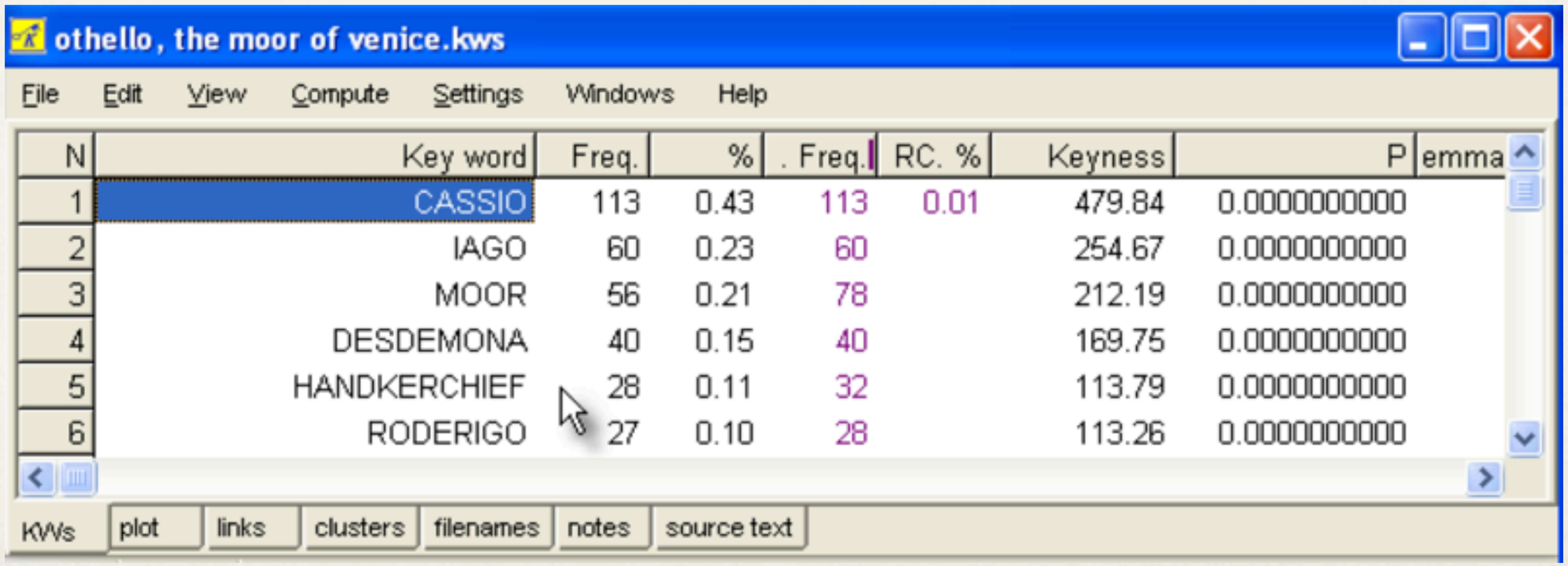
IO = of (as preposition)

JJ = general adjective

# Key words (e.g. WordSmith)

- Compares a target corpus with a (larger) reference corpus to identify *unusually frequent* words

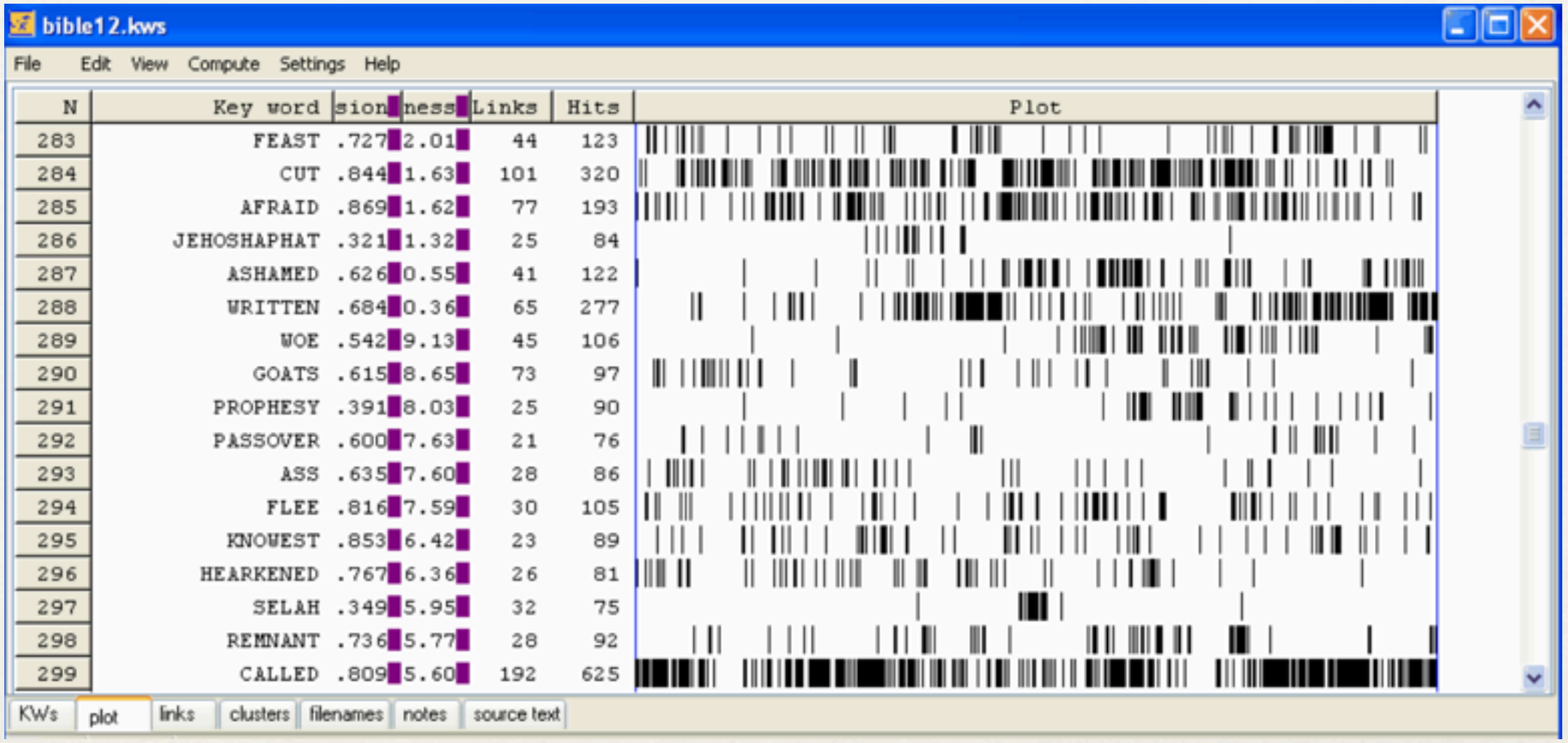
- Usually identifies 'what a text is about':




| N | Key word | Freq. | % | . Freq. | RC. % | Keyness | P | emma |
|---|---|---|---|---|---|---|---|---|
| 1 | CASSIO | 113 | 0.43 | 113 | 0.01 | 479.84 | 0.0000000000 | |
| 2 | IAGO | 60 | 0.23 | 60 | | 254.67 | 0.0000000000 | |
| 3 | MOOR | 56 | 0.21 | 78 | | 212.19 | 0.0000000000 | |
| 4 | DESDEMONA | 40 | 0.15 | 40 | | 169.75 | 0.0000000000 | |
| 5 | HANDKERCHIEF | 28 | 0.11 | 32 | | 113.79 | 0.0000000000 | |
| 6 | RODERIGO | 27 | 0.10 | 28 | | 113.26 | 0.0000000000 | |

# Dispersion

- Visual insights into where key words crop up in a text (WordSmith):



# Type/token ratios

- A measure of the lexical variety (or vocabulary load) in a text
- TTR = number types/number tokens x 100
- TTRs for speech (~ 50%) tend to be lower than those for writing (~ 70%) because of limited processing time

# Semantic prosody

- The way in which seemingly neutral words can take on positive or negative associations through frequent co-occurrence with particular collocations (BYU-BNC):

## Cause

| | | CONTEXT | ALL |
|---|---|---|---|
| 1 | | PROBLEMS | 453 |
| 2 | | DEATH | 370 |
| 3 | | CONCERN | 298 |
| 4 | | DAMAGE | 291 |
| 5 | | EFFECT | 210 |
| 6 | | ACTION | 208 |
| 7 | | TROUBLE | 160 |
| 8 | | PROBLEM | 141 |
| 9 | | HARM | 117 |
| 10 | | SERIOUS | 102 |
| 11 | | INJURY | 90 |
| 12 | | DISEASE | 90 |

## Provide

| | | CONTEXT | ALL |
|---|---|---|---|
| 1 | | INFORMATION | 847 |
| 2 | | SERVICES | 468 |
| 3 | | SERVICE | 430 |
| 4 | | SUPPORT | 423 |
| 5 | | EVIDENCE | 337 |
| 6 | | BASIS | 284 |
| 7 | | OPPORTUNITY | 228 |
| 8 | | CARE | 209 |
| 9 | | DETAILS | 205 |
| 10 | | USEFUL | 192 |
| 11 | | ADVICE | 188 |
| 12 | | TRAINING | 180 |

# Semantic prosody: wave

## BNC: Fiction sub-corpus

| | | CONTEXT | ALL |
|---|---|---|---|
| 1 | ☐ | HAND | 31 |
| 2 | ☐ | SWEPT | 13 |
| 3 | ☐ | WAVE | 12 |
| 4 | ☐ | NAUSEA | 10 |
| 5 | ☐ | GOODBYE | 9 |
| 6 | ☐ | BROKE | 9 |
| 7 | ☐ | ARMS | 9 |
| 8 | ☐ | WASHED | 7 |
| 9 | ☐ | DIZZINESS | 6 |
| 10 | ☐ | PANIC | 6 |
| 11 | ☐ | ANGER | 6 |
| 12 | ☐ | HIT | 6 |

## BNC: Newspaper sub-corpus

| | | CONTEXT | ALL |
|---|---|---|---|
| 1 | ☐ | VIOLENCE | 9 |
| 2 | ☐ | STRIKES | 8 |
| 3 | ☐ | ATTACKS | 6 |
| 4 | ☐ | WAND | 5 |
| 5 | ☐ | SWEPT | 5 |
| 6 | ☐ | BUYING | 5 |
| 7 | ☐ | BOMBINGS | 4 |
| 8 | ☐ | MAGIC | 4 |
| 9 | ☐ | POLITICAL | 4 |
| 10 | ☐ | WATER | 4 |
| 11 | ☐ | JOB | 4 |
| 12 | ☐ | SUPPORT | 4 |

# Corpora accessible (free) online

- British National Corpus (BNC): 100 million words of spoken (10%) & written (90%) language, representing a wide cross-section of 20th century British English

  http://www.natcorp.ox.ac.uk/ or *http://corpus.byu.edu/bnc/

- Corpus of Contemporary American English (COCA): 450 million word million words of spoken (21%) & written (79%) language, representing a wide cross-section of contemporary American English

  http://corpus.byu.edu/coca/

- Michigan Corpus of Academic Spoken English (MICASE): 1.8 million words of academic speech, recorded at University of Michigan

  http://quod.lib.umich.edu/m/micase/

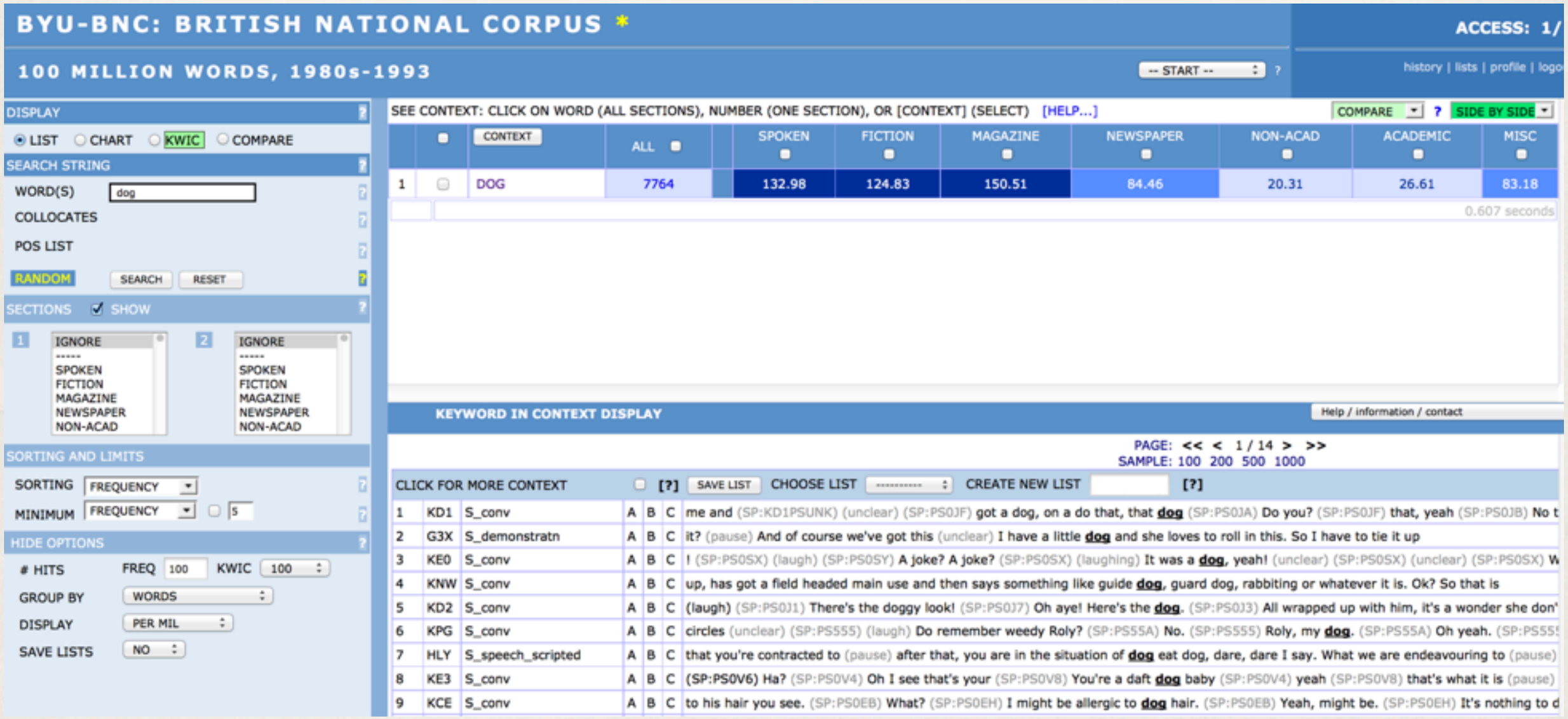
- Hong Kong Engineering Corpus (HKEC): 9.2 million words of (mostly written) English from the engineering sector in Hong Kong

  http://rcpce.engl.polyu.edu.hk/HKEC/

- Compleat Lexical Tutor: Provides a variety of corpus analysis tools for learners, researchers & teachers, using BNC, COCA & Brown corpora, or user's own text files

  http://www.lextutor.ca/

# Exploring the BYU corpora



# Approaches to corpus-informed language learning

- Indirect approaches: Use corpora to inform materials design & ensure that language models presented are  both naturalistic & pedagogically useful

- Direct approaches: Use corpora in the classroom & allow teachers & learners to discover language patterns for themselves

# Indirect approaches

- “[The ELT profession] has been rather slow to incorporate corpus methods into its working practices. It is still the case that the majority of ELT materials-writers rely on a combination of their own intuitions and teaching experience, and a well-established canon of apparently self-evident 'facts' about the language which have, more or less, the status of tradition.”  (Rundell 1996)

# Burton (2012), *Corpora* 7.1

- Questionnaire: only 8 out of 13 professional textbook writers had ever used corpora to inform their materials design (extent not reported)

- Decision to query corpora came from authors themselves, rather than publishers

- Reasons cited for *not* referring to corpora: lack of: (a) expertise; (b) access; or (c) time

(Burton, G. (2012). 'Corpora and coursebooks: destined to be strangers forever?' *Corpora* 7.1, 91-108)

# Direct approaches

- Widening gap between corpus-linguistics research and classroom teaching (Römer 2006; McCarthy 2008; Zhang 2008; Aijmer 2009).

- This is despite the recognized pedagogic value of 'data driven learning' (DDL) (e.g. Johns 1991):

(i) Encourages learners to discover language patterns for themselves (inductively) => greater cognitive processing/deeper learning

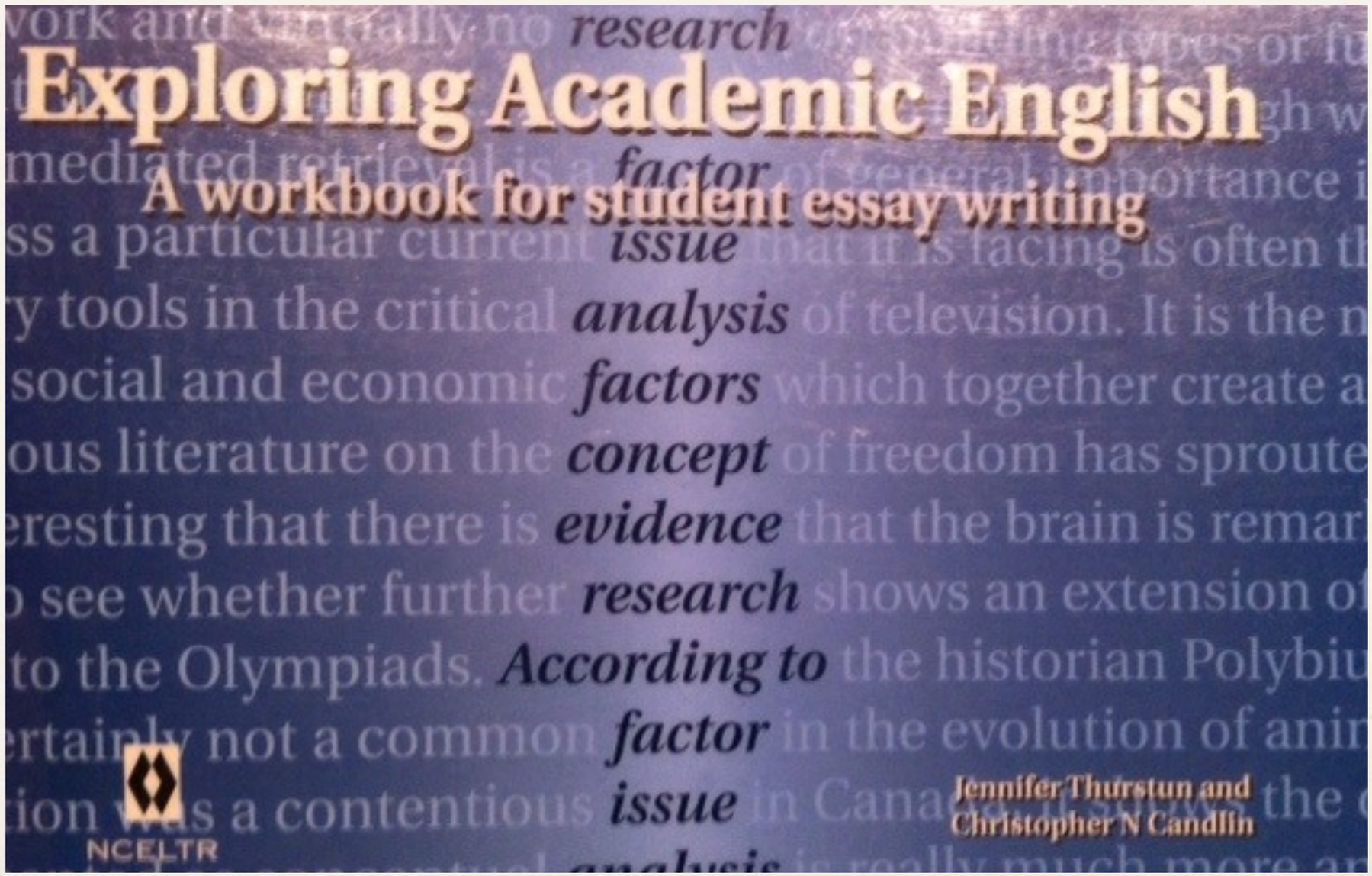
(ii) More learner-centred, allowing learners to focus on items that best fit their particular stage of interlanguage development

(iii) Allows learners to notice language patterns omitted from textbooks or reference books (too complicated or overlooked?)

# Reasons for poor take-up of DDL?

- Lack of technical resources or corpus analysis tools (use paper-based worksheets instead?)
- Time-consuming nature of inductive learning
- Destroys learners' world of clear cut grammatical rules
- Unfamiliar terminology (parsing, tagging, type-token ratios, etc.)
- Unpredictability engendered by DDL approach

# Example of DDL: *Exploring Academic English*



# Example of DDL: *Exploring Academic English*

- Innovative concordance-based workbook for use in EAP classes or for independent learning (see Thurstun & Candlin 1998)

- Deals with 6 important rhetorical functions in EAP: (a) stating the topic; (b) referring to the literature; (c) reporting others' research; (d) discussing processes; (e) hedging; (f) drawing conclusions

- 3 or 4 high frequency words identified from each rhetorical category (indirect approach) & concordance lines analysed by students (direct approach)

# Example of DDL: *Exploring Academic English*

- 4 stages in each unit:

(i) 'Look' stage: concordance lines presented (right-sorted)

(ii) 'Familiarise' stage: identification of lexico-grammatical patterns around key words

(iii) 'Practice' stage: gap-fill & matching exercises

(iv) 'Create' stage: students practice using key words in own texts

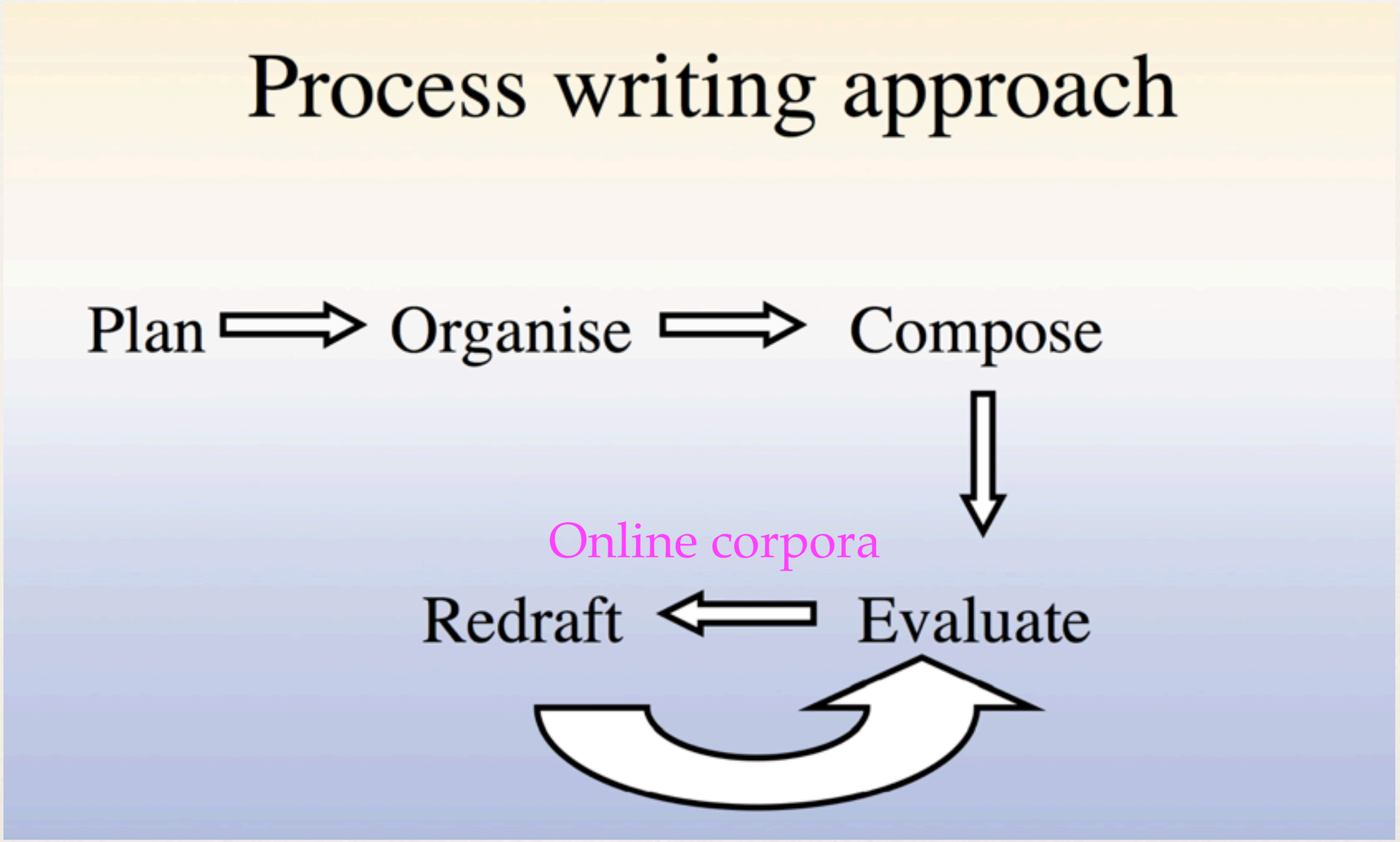
# *Exploring Academic English: according to…*

- Try the DDL tasks for the target phrase, *according to*

- Compare the workbook tasks with the information available for *according to* online (with the BYU-corpora)

- Can you see any advantages or disadvantages for choosing paper-based vs. web-based modes of learning in the classroom?

# Using online corpora to develop students' writing skills (Gilmore 2009)



# Can online corpora help learners in the re-drafting process?

- 45 2nd year students in compulsory academic writing class
- 1st draft of factual report on the topic of 'obsession'
- Problem areas in essays highlighted by teacher
- 30-minute training session on using online corpora
- Students (in pairs) evaluated mistakes in 1st drafts using BNC & COBUILD corpora
- Redrafting completed out of class
- 1st & 2nd drafts assessed for naturalness (blind rated by 2 NS teachers)
- Students commented on usefulness of online corpora

# Training session

- Typical student errors analysed using online corpora:

(i) Since then, he started to go.

(ii)... but we cannot make it worth.

(iii) My confidence changed.

(iv) X died for a car accident.

# Blind rating of students' writing

**Student Writing Samples**

**The following extracts are taken from university students' academic essays. Please indicate which version you consider to be more natural by placing a cross next to it, for example:**

**I started to associate with my girlfriend a year ago. __**
**I started going out with my girlfriend a year ago. X__**
**No difference __**

**If you do not consider one to be more natural than the other, please put a cross next to 'No difference X '.**

**Thank you very much for your help with this research.**
**KO (051324)**

1. Today, his fantastic activity allow himself to be the children's hero. __
   Today, his fantastic plays make himself be the children's hero. __
   No difference __

# Results

- Total number modifications 1st => 2nd draft = 350 (Range 1-17 per essay)
- Improved = 67.3%
- No difference = 27.4%
- Worse = 5.3%

# Example modifications

**Improved:**

1st draft: He became popular in the USA not only Japan.

2nd draft: He became popular not only Japan but also in the USA.

**No difference:**

1st draft: Her activity will attract people as before.

2nd draft: Her activity will attract people same as before.

**Worse:**

1st draft: Underage smoking was prohibited in Japan, so she couldn't avoid fired.

2nd draft: Underage smoking was prohibited in Japan, so she couldn't evade displacement.

# Student feedback

- Were online corpora useful for redrafting your essays?
- Yes = 95%

  " I think it is very useful for me because I can know the native speaker's sentences. In my dictionary there are many sentences but they are not natural sentences".

- No = 5%

  "If I don't know what I should type, I can't find out".

  "Corpus is not useful for me because it is complicated".

# Conclusion

- Corpora and corpus analysis tools can make a valuable contribution to EAP classes
- Online access and functionality has improved considerably in recent years
- Challenges exist but can be overcome with adequate training & motivation

Thank you!

ありがとうございました

# References


- Aijmer, K. (ed.) (2009). Corpora and language teaching. Amsterdam: John Benjamins.
- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Burton, G. (2012). 'Corpora and coursebooks: destined to be strangers forever?' *Corpora* 7.1, 91– 108
- Erman, B. & B. Warren (2009). 'The idiom principle and the open choice principle'. *Text* 20.1, 29 - 62.
- Gilmore, A. (2009). Using on-line corpora to develop students' writing skills. *English Language Teaching Journal* 63/4: 363-372.
- Johns, T. (1991). 'Should you be persuaded – Two samples of data-driven learning materials'. In T. Johns & P. King (eds.), 'Classroom Concordancing'. *ELR Journal*, 4, 1–16.
- McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context. *Language Teaching* 41.4, 563–574.
- Römer, U. (2006). 'Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments'. *Zeitsschrift fürAnglistik und Amerikanistik* 54.2, 121–134.
- Rundell, M. (1996). 'The corpus of the future, and the future of the corpus'. Talk at Exeter, special conference on 'New Trends in Reference Science', 29/3/96. http://web.archive.org/web/20040211202044/http://www.ruf.rice.edu/~barlow/futcrp.html
- Thurstun, J., & C. Candlin (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes* 17(3), pp. 267-280.
- Zhang, S. (2008). 'The necessities, feasibilities and principles for EFL teachers to build a learner-oriented mini-corpus for practical classroom uses'. *Asian EFL Journal* 29, 1–15.