Bigger and better and bigger and **better**

A computational, corpus-based research programme for linguistics

Adam Kilgarriff
Lexical Computing Ltd
Lexicography MasterClass Ltd
University of Sussex

Overview

- Research programme
- Example: corpus lexicography
  - Word sketches
- Web as corpus
- Le voyage long du texte au sens
- *If time:*
  - Linking dictionary and corpus

- What is language?

- What is language?
- In our heads

- What is language?
- In our heads
- In texts and sound signals

- What is language?
- In our heads
- In texts and sound signals
- **Both**

Methodology

- Study language in our heads
  - Competence
  - Chomsky
  - “rationalist” (Descartes, Leibniz)

Methodology

- Study language in our heads
  - Competence
  - Chomsky
  - “rationalist” (Descartes, Leibniz)
  - Odd method for objective science
  - Practical problems: coverage, arbitrariness

Methodology

- Study text
  - “empiricist” (Locke, Hume)

  - Physics: forces, matter
  - Chemistry: chemicals, bonds
  - Language: text, speech signals

It goes against the grain

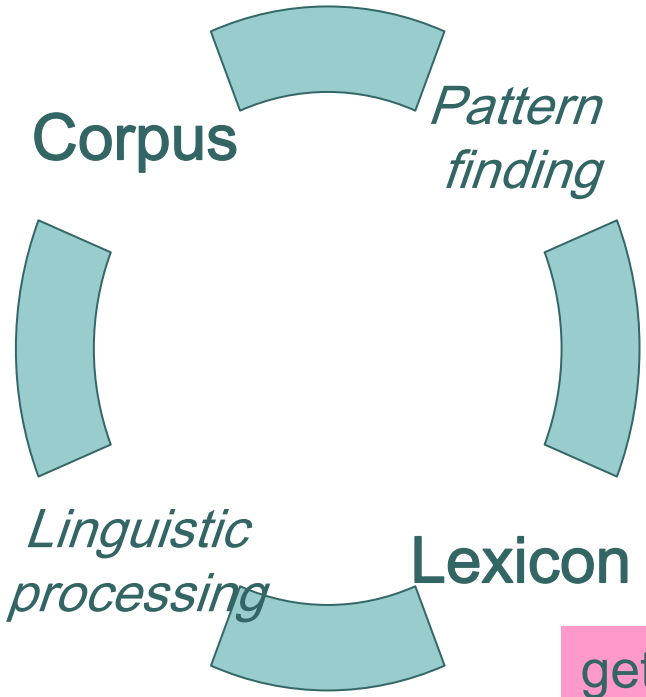
- What is important about a sentence?
  - *its meaning*
- Corpus methodology:
  - **Throw away** individual sentence meaning
  - Find patterns

Computers and corpora

- Machine learning
  - finds patterns in data sets
- Corpora
  - bigger and bigger data sets
- Language technology tools
  - lemmatizers, POS-taggers, parsers
- A new way to find out about language
  - *15 years of rapid ascent*

A virtuous circle

More data →

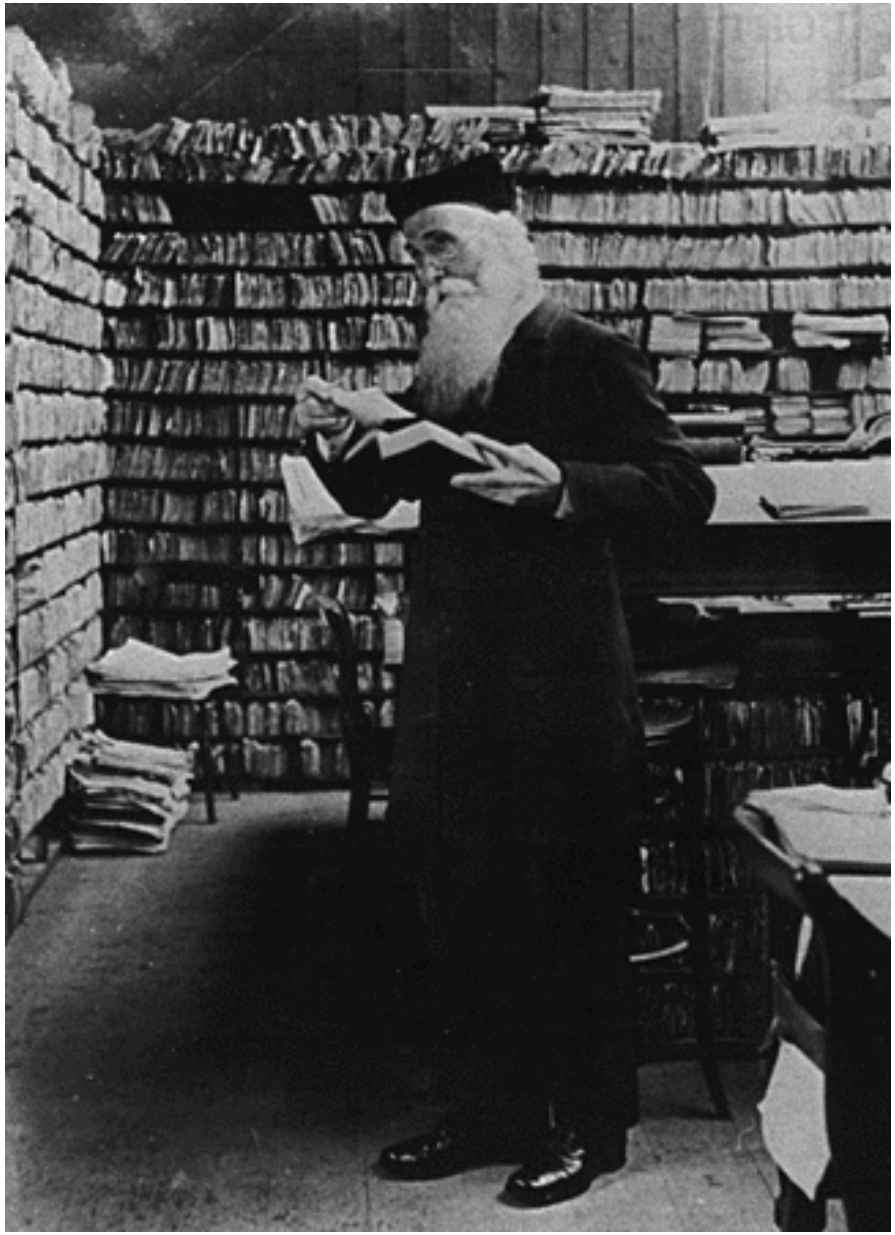


- Part-of-speech tagging
- Parsing
- Lemmatizing

gets richer each time round

Example: corpus lexicography

- four ages



Age 1: Pre-computer

Oxford English Dictionary:

- 20 million index cards

Age 2: KWIC Concordances

- From 1980
- Computerised

Age 2: KWIC Concordance

```
arity, which will be used to take a party  of under-privileged child
 from outside. You are invited to a party and after a couple of drin
tion, we believe politicians of all parties will listen to our views
ould be reaching agreement with all parties  concerned,  as to which
lack people.  I have certainly been party to one or two discussions
  These should be discussed by both parties before entering into the
presents They had hosted a cocktail party at Kensington palace,  for
akes. By midnight the end-of-course party  is in full swing,  but mo
e should be a right for the injured party to terminate the contract.
 by the Safran Peoples ' Liberation Party.  This presents the powerf
s. Ahead I could see the rest of my party plodding towards the final
cial ethic.  The two main political parties - the Tories and the Lib
ritish successes in Perth The small party of British players competi
to help control.  One member of the party went to summon the rescue
rket society fashion magazine.  The party was held at his flat which
 security and secrecy than any Tory Party Conference : it seems that
```

Age 2: KWIC Concordances

- From 1980
- Computerised
- COBUILD project was innovator
- the coloured-pens method

The coloured pens method

1 arity, which will be used to take a party of under-privileged children to
2 from outside. You are invited to a party and after a couple of drinks you
3 tion, we believe politicians of all parties will listen to our views. &eq
4 ould be reaching agreement with all parties concerned, as to which event
5 lack people. I have certainly been party to one or two discussions amongs
6 . These should be discussed by both parties before entering into the relat
7 presents They had hosted a cocktail party at Kensington palace, for examp
8 akes. By midnight the end-of-course party is in full swing, but most cad
9 e should be a right for the injured party to terminate the contract. A ma
10 by the Safran Peoples ' Liberation Party. This presents the powerful nei
11 s. Ahead I could see the rest of my party plodding towards the final slope
12 cial ethic. The two main political parties - the Tories and the Liberals
13 ritish successes in Perth The small party of British players competing in
14 to help control. One member of the party went to summon the rescue team a
15 rket society fashion magazine. The party was held at his flat which was a
16 security and secrecy than any Tory Party Conference : it seems that bootl

1 political association

2 social event

3 group of people

4 person in an agreement/dispute

5 to **be party to** something...

Age 2: limitations

as corpora get bigger:

too much data

- 50 lines for a word: read all
- 500 lines: *could* read all, takes a long time
- 5000 lines: **no**

Age 3: Collocation statistics

- **Problem:**
  too much data - how to summarise?
- **Solution:**
  list of words occurring in neighbourhood of headword, with frequencies
- Sorted by salience

Collocation listing

For collocates of *save* (>5 hits), window 1-5 words to right of nodeword

word		***word***
forests		life
$1.2		dollars
lives		costs
enormous		thousands
annually		face
jobs		estimated
money		your

Age 4: The word sketch

A corpus-derived one-page summary of a word's grammatical and collocational behaviour

Age 4: The word sketch

- Large well-balanced corpus
- *Parse* to find
  - subjects, objects, heads, modifiers etc
- *One list for each grammatical relation*
- Statistics to sort each list, as before

English word sketches

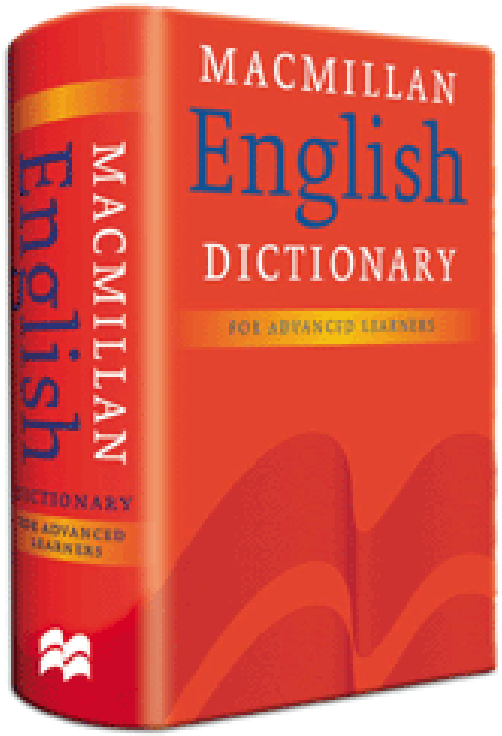
- British National Corpus (BNC)
  - 100 M words, already POS-tagged
- lemmatized
  - *assisting => assist (v)*
- parsed
- database of 70 million triples

<object, sip, coffee> <subject, arrive, coffee>

<and-or, tea, coffee> <modifier, coffee, instant>

Macmillan English Dictionary For Advanced Learners

Ed: Rundell, 2002



Euralex 2002

Euralex 2002

- Je les voudrais pour ma langue, s'il vous plaît

The Sketch Engine

- Input:
  - any corpus, *any language*
    - Lemmatised, part-of-speech tagged
  - specification of grammatical relations
- Word sketches *integrated with*
- Corpus query system
  - Supports complex searching, sorting etc
  - IMS-Stuttgart formalism (also for corpus input)
    - Corpus searches *and* grammar writing
    - Christ and Schulze 1994

Grammar writing

- Uses CQL (Corpus query language)
  - Christ and Schulze, U. Stuttgart, 1994
- defining an object:

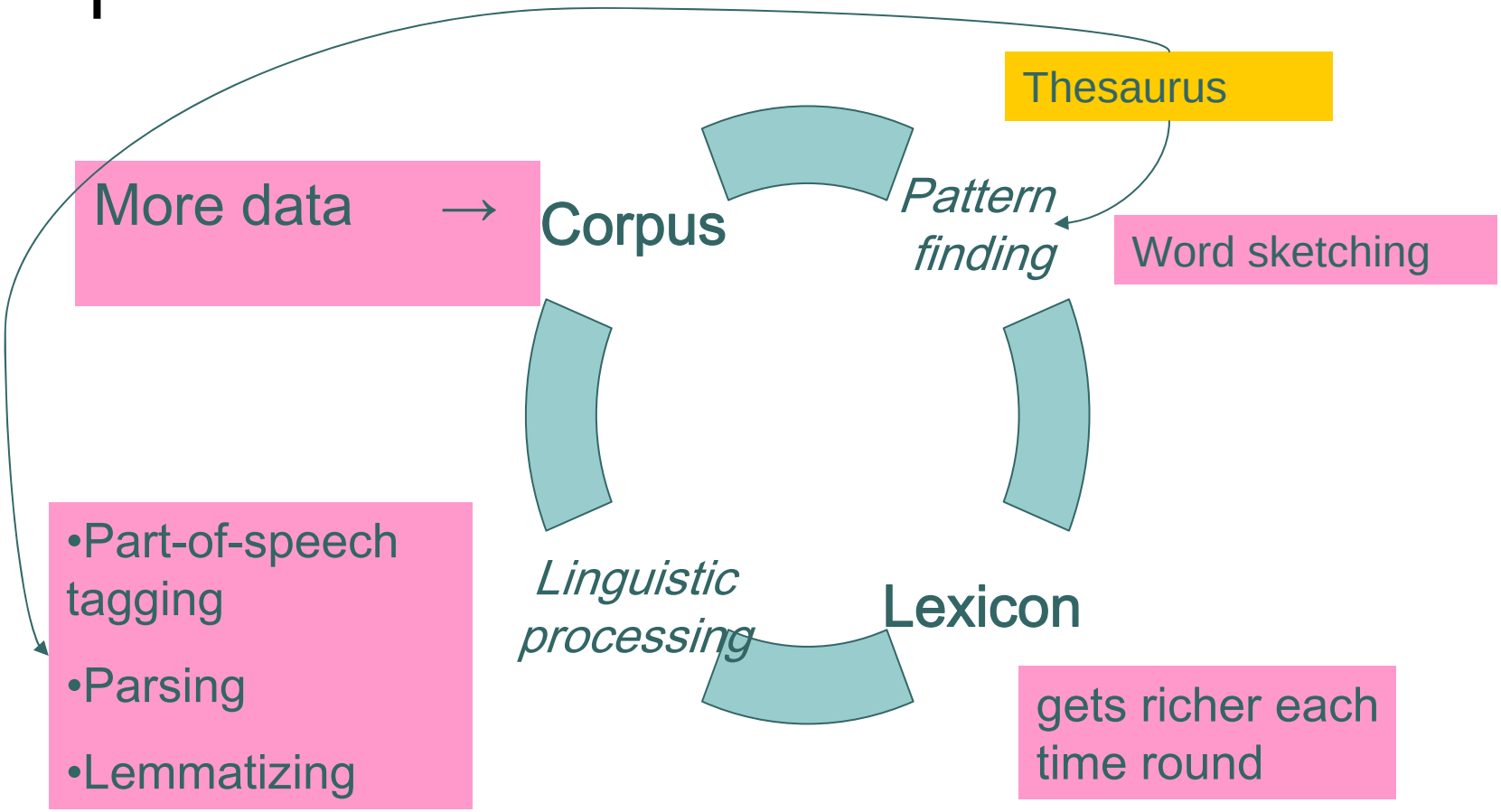
```
v (adj|n|det|num|adv)* n
```

- Developer: Pavel Rychly, Brno
- Users:
  - OUP, CUP, Macmillan for lexicography
  - Universities for teaching and research
  - ELT textbook authors
- Demo:
  - http://www.sketchengine.co.uk/
    - Self-registration for free account

Demo

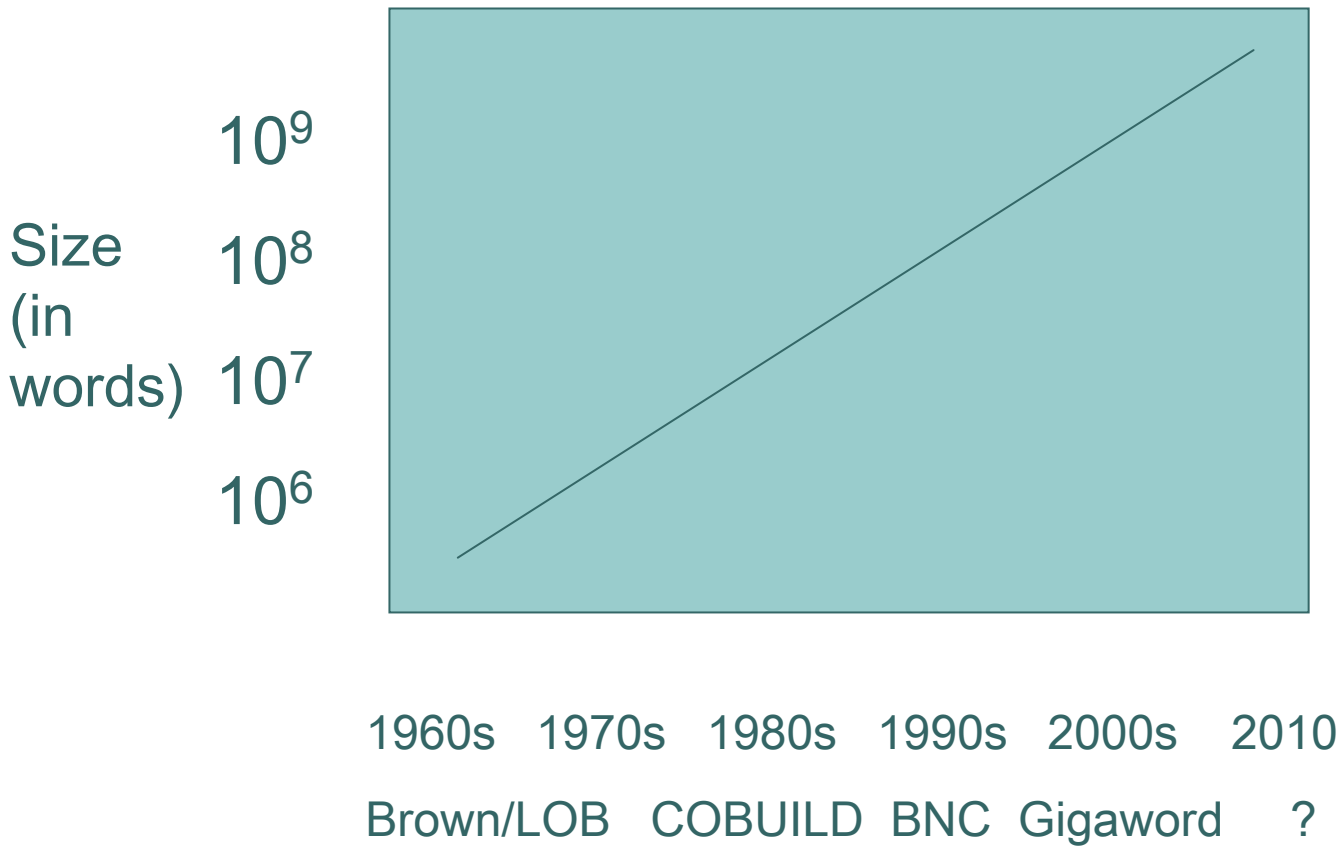
- A French web corpus (130M words) in the Sketch Engine
  - (see other powerpoint)

A virtuous circle



Web as Corpus

Corpora


Size (in words)
10^9
10^8
10^7
10^6
1960s 1970s 1980s 1990s 2000s 2010
Brown/LOB COBUILD BNC Gigaword ?

Corpora within CL/TALN

- 1989: corpora arrive on scene
- 1989-1993: “too dirty”: **battles**
- 1993: CL Special Issue: consummation
- …
- 1999: web arrives on scene
- 1999-2003: “too dirty”
- 2003: CL Special Issue

Web as linguistic resource

You can't help noticing

- *speculater* or *speculator*?

Higher tech

- Webcorp: http://www.webcorp.org.uk/
- Resnik's *Linguistic Search Engine*
- Fletcher's *Kwicfinder*

  http://www.kwicfinder.com/KWiCFinder.html
- Baroni and Bernadini's *BootCat*
  - *WebBootCaT*
- Supporting technology: Google API

BootCat:

- Put in seed terms
- Google search
- Retrieve Google hits
- (get pages that those pages point to)
  - Big corpora (1.7B words of German)
  - Small instant corpora
    - Web service
- (load into corpus query tool)

Seed words lessicografia lessicologici lessicologia fraseologia collocazioni terminologia

Enclose multiwords expressions into quotes (").

Your Google key pm8eaihQFHJWh/L79fYk+15t+9iGMOnB

This application uses Google API. You need a Google key to be able to use it. If you do not have one please go to https://www.google.com/accounts/NewAccount.

Language Italian

Select the language of the corpus to be built.

Select URLs ☑

Check this box if you would like to manualy filter found URLs before the webpages will be downloaded.

Tag corpus ☑

Your corpus will be POS-tagged and lemmatised using the TreeTagger. Following languages are currently supported: English, French, German, Italian, Spanish. This options has no effect if used with any other languages.

Corpus name Italian-lexicography

Chose a name for your corpus.

Your email address adam@lexmasterclass.com

The time needed for building a corpus is highly variable and may take minutes, or hours. If you enter your email address you will be notified when the corpus is ready to use.

Build a corpus!

BootCaT alpha

www version

Corpus built

Your corpus was built successfully.

Corpus name	Italian-lexicography
Size	1051 kB
Words count	155700
Web pages retrieved	27
Build time	01:29
Access URL	http://corpora.fi.muni.cz/bootcat/corp/VPyJwaQQ/

Download the corpus in raw format

Download the corpus in vertical format

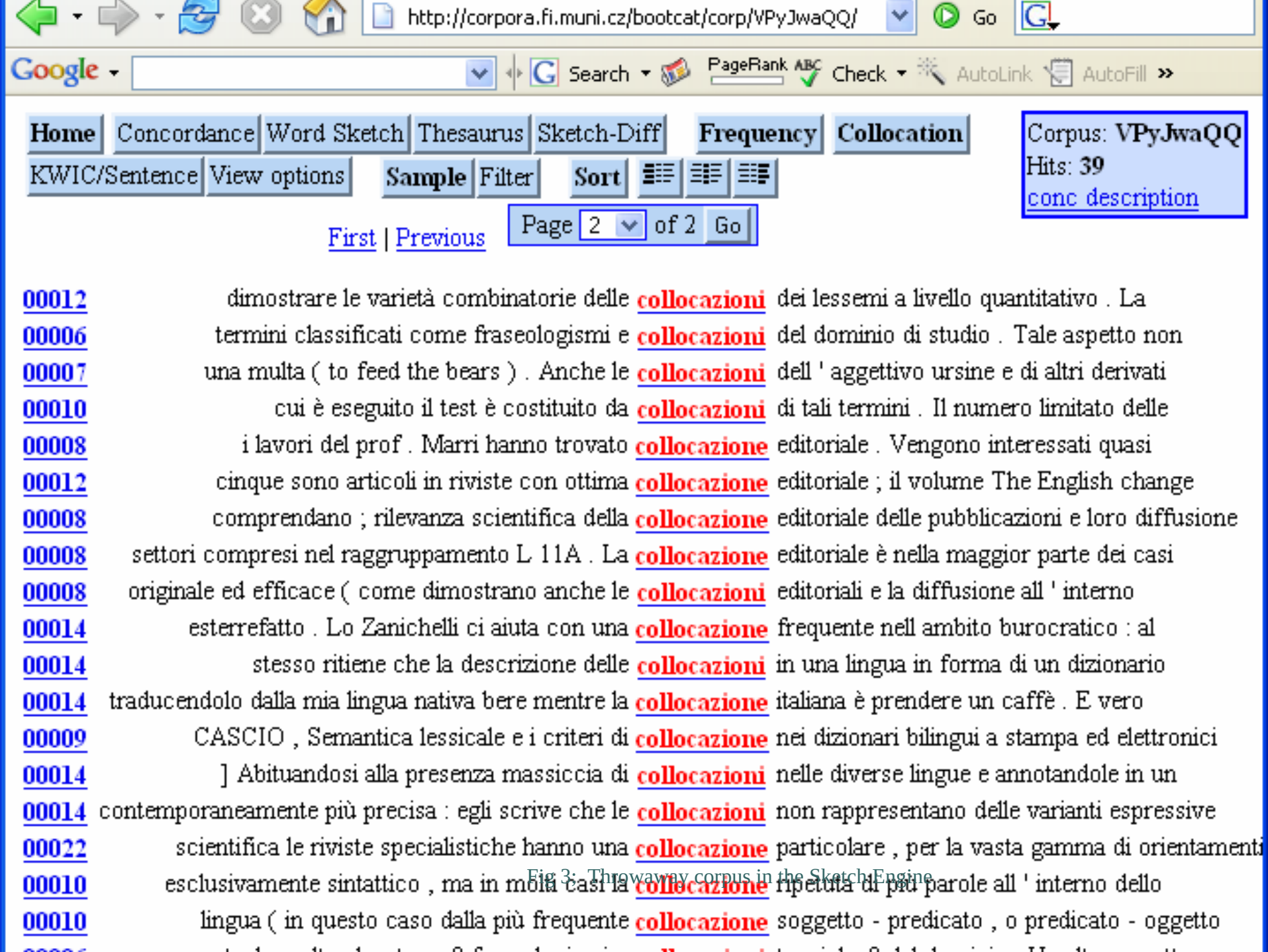


Fig 3: Throwaway corpus in the Sketch Engine

but it’s not representative

Theory

A random sample of a population is representative of it. Observations on the sample support inferences about the population (within confidence bounds).

- ***What is the population?***
  - production and reception
  - speech and text
  - background language
  - copying

sublanguage

- Language = core + sublanguages
- Options for corpus construction
  - none
  - some
  - all
- None
  - impoverished view of language
- Some: BNC
  - cake recipes and gastro-uterine disease
  - *not* car repair manuals or astronomy or …
- All: until recently, not viable

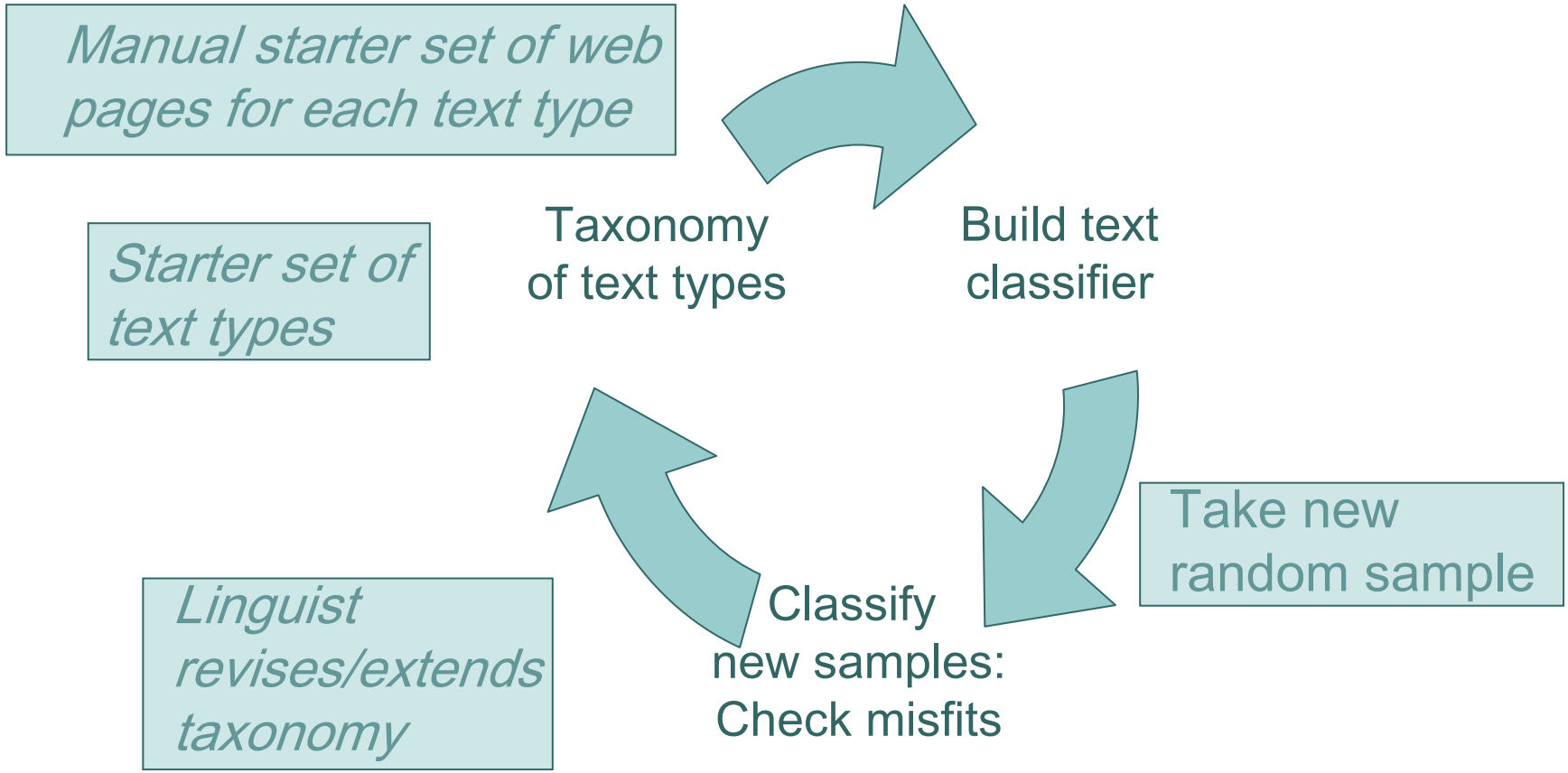
Summary

- The web is not representative
- ***but nor is anything else***
- Text type variation
  - under-researched, lacking in theory
    - Atkins Clear Ostler 1993 on design brief for BNC; Biber 1988, Baayen 2001, Kilgarriff 2001
- Text type is an issue across NLP
  - Web: issue is acute because, as against BNC or WSJ, we simply don't know what is there

What is out there?

- Sampling the web is non-trivial
  - Henziger et al, WWW9, 2000
- What text types are there on the web?
  - some are new: chatroom
  - proportions
    - is it overwhelmed by porn? How much?
    - **Hard question**

Text type classifiers



Link analysis

- WWW research community
  - Link analysis
    - Hubs and authorities, Kleinberg
    - CLEVER, Chakrabarti et al
    - Google's PageRank
    - Kumar et al: *Trawling the Web for emerging cybercommunities* (over 100,000 found)
  - Hypothesis
    - Communities identified by links correspond to communities identified by sublanguage/lexicon

The web as an object of study

- The web
  - a social, cultural, political phenomenon
  - new, little understood
  - *a legitimate object of science*
  - mostly language
    - we are well placed
  - a lot of people will be interested

Numerous roles

- study the web
- web as corpus
  - source of data for studying language
- apply language technology for web use
  - IR, QA, MT
- Infrastructure
  - finding, getting, sharing

The Trouble with Google

- not enough instances (max 1000)
- not enough context
  - ca 10-word snippet around search term
- ridiculous sort order
  - search term in titles and headings
- linguistically dumb
  - not lemmatised
    - *aime/aimer/aimes/aimons/aimez/aiment* …
  - not POS-tagged
  - and why not parsed

DIY

- Language researchers not a mighty commercial interest
- Google won't prioritise doing things our way
- let's do it ourselves

Components

1. web crawler
2. filter/classifier
3. linguistic processor
4. database
5. statistical summariser
6. user interface

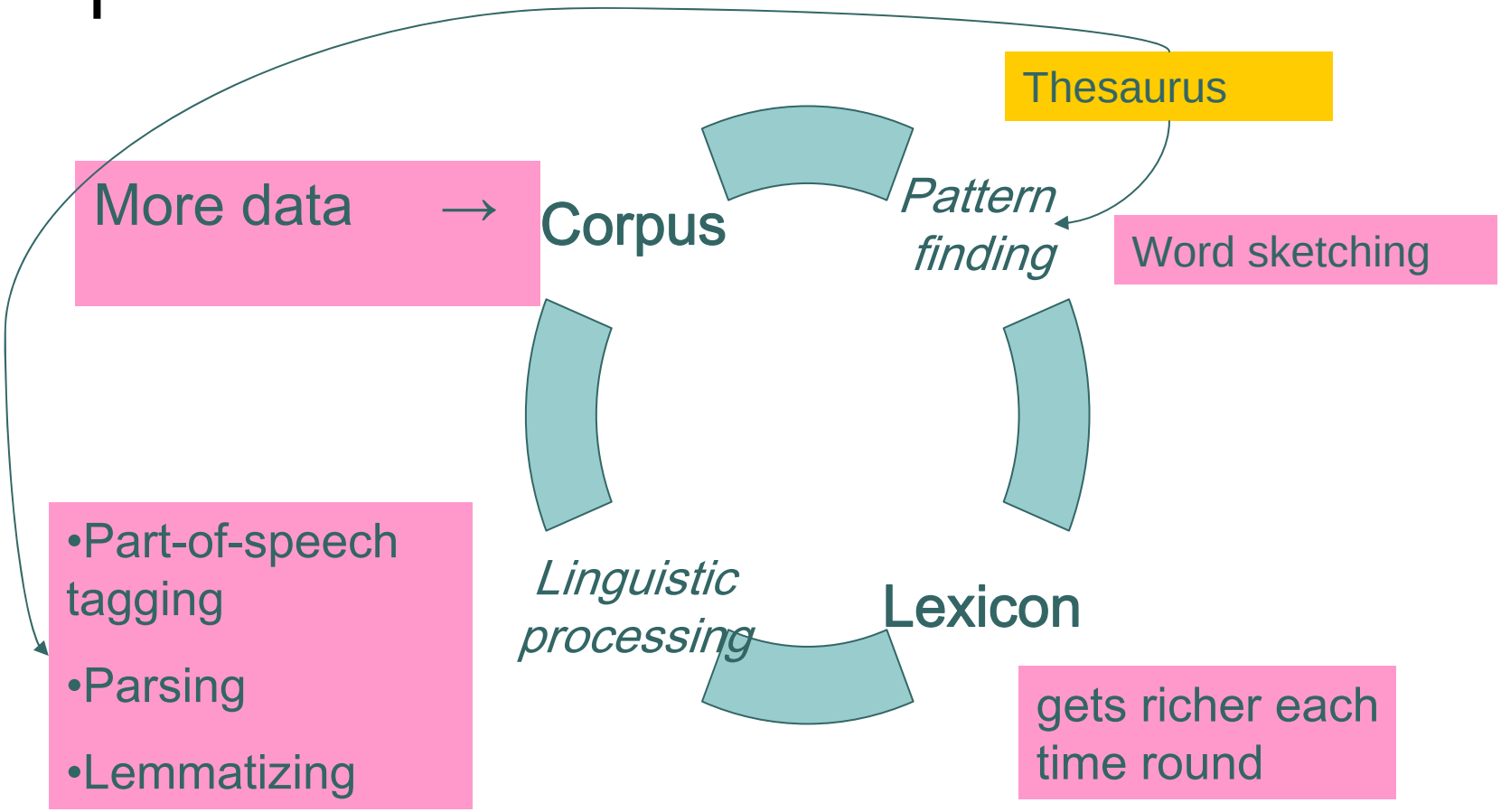
Examples

- DeWaC: 1.7 b words German
  - Baroni and Kilgarriff, EACL 2006
- ItWaC: 1.3 b words Italian
- Serge Sharoff, Leeds Univ UK
  - English Chinese Russian English French Spanish, all searchable online
- Newest: WWW-Fr 130m words

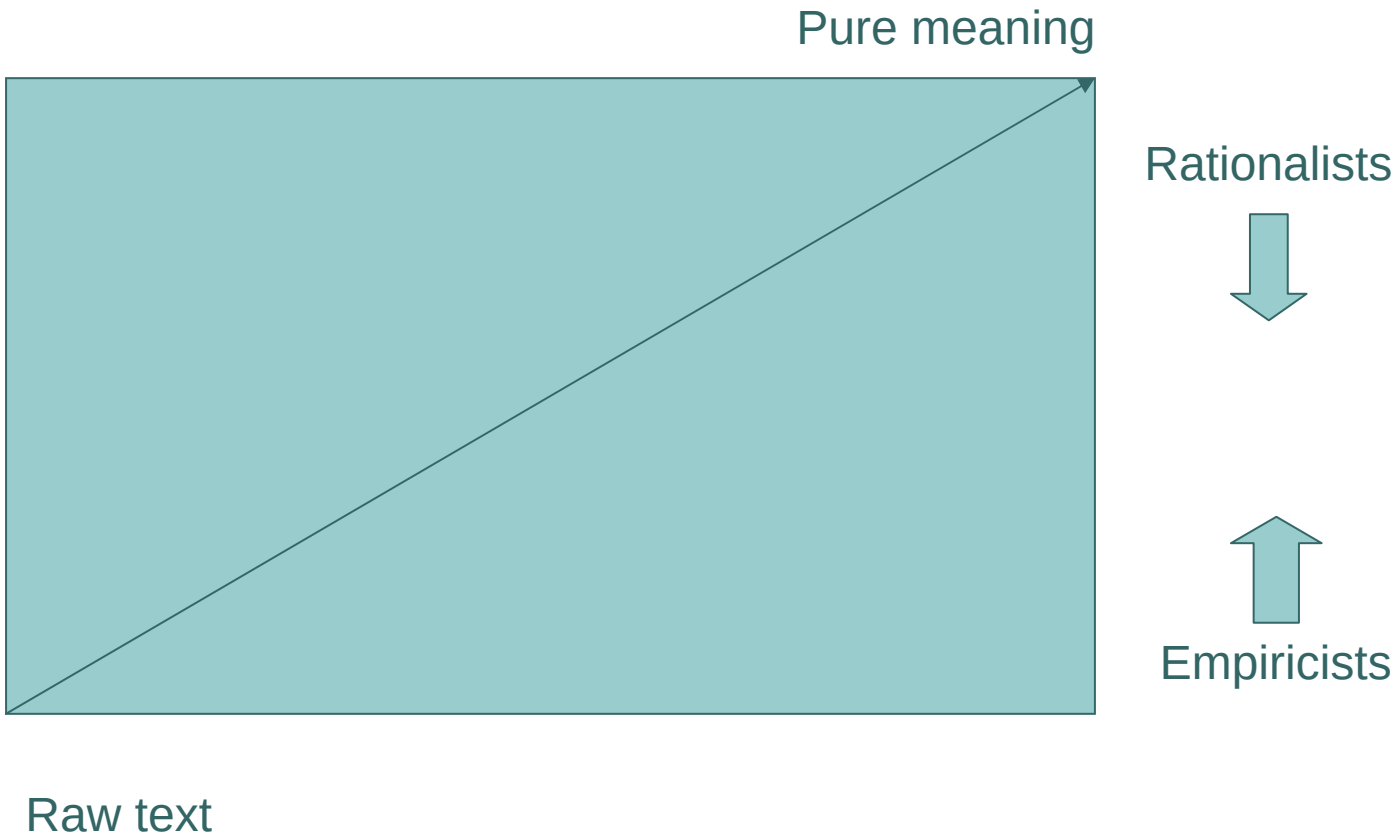
Community-building

- WAC Kool Ynitiative (WaCKY)
  - mailing list
  - convenor: Marco Baroni
- WAC workshops
  - WAC1, Birmingham 2005
  - WAC2, Trento (EACL), April 2006
- ACL SIG proposal
- Relations to WWW community

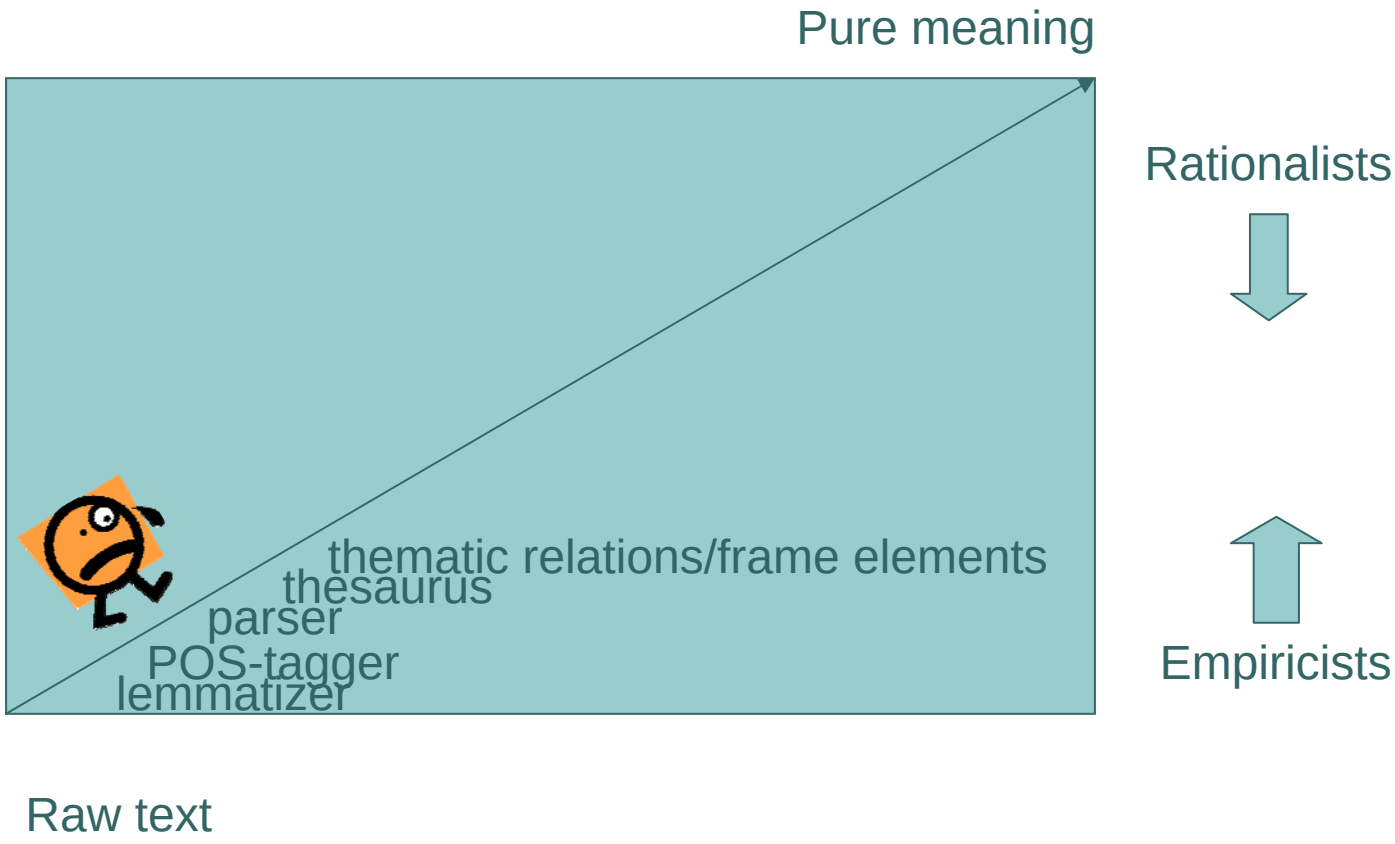
A virtuous circle



The long journey from text towards meaning



The long journey from text towards meaning



Linking dictionary and corpus

Observation:

- corpus: arbitrary sample
- dictionary: systematic account

Children

- encounter arbitrary samples
- develop systematic account

Conclusion: a corpus should be

- provisional, dispensable
- used to develop lexicon

Review of WSD

- SEMCOR
  - Based on WordNet
  - All corpus words tagged with WordNet senses
  - Widely used in WSD
  - *standard model*
- “Putting the dictionary into the corpus”

Disadvantages

- Dictionary enriches corpus
- Corpus is not dispensable

  wrong way round

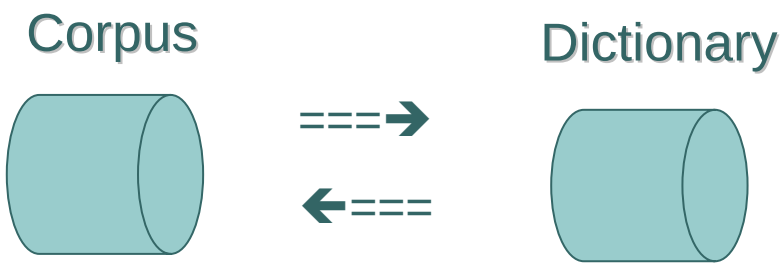
  We want to

  put the corpus into the dictionary

- Make richer dictionaries (like children), not richer corpora

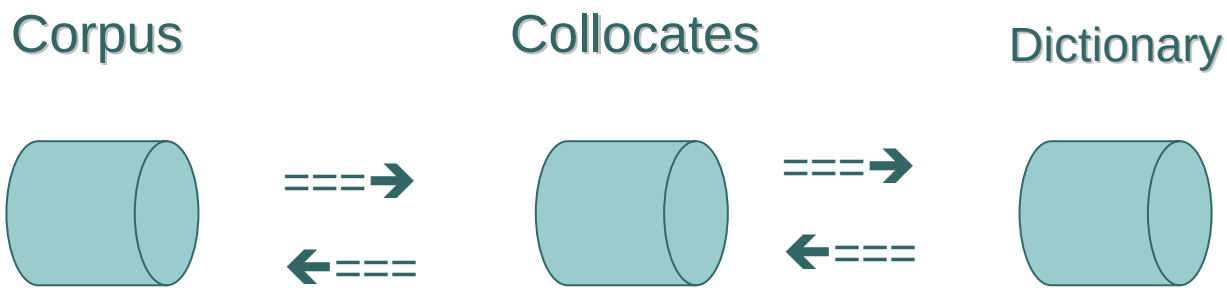
Levels of abstraction

- Direct linkage:



- Fragile
  - Updates (to C or D) break links
- Dictionary: abstract
- Corpus: raw
    - *Intermediate level needed*

Intermediate database

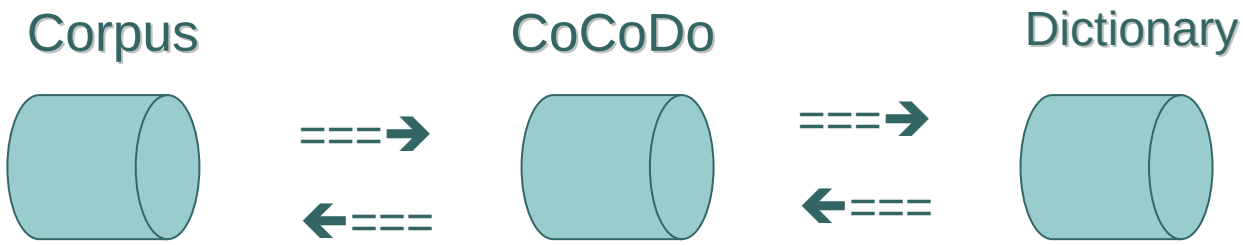


- How most WSD works
  - Analyse dictionary to give set of collocates
  - Match to collocates in a corpus
- Dispensable corpus
- Example: *word sketches*

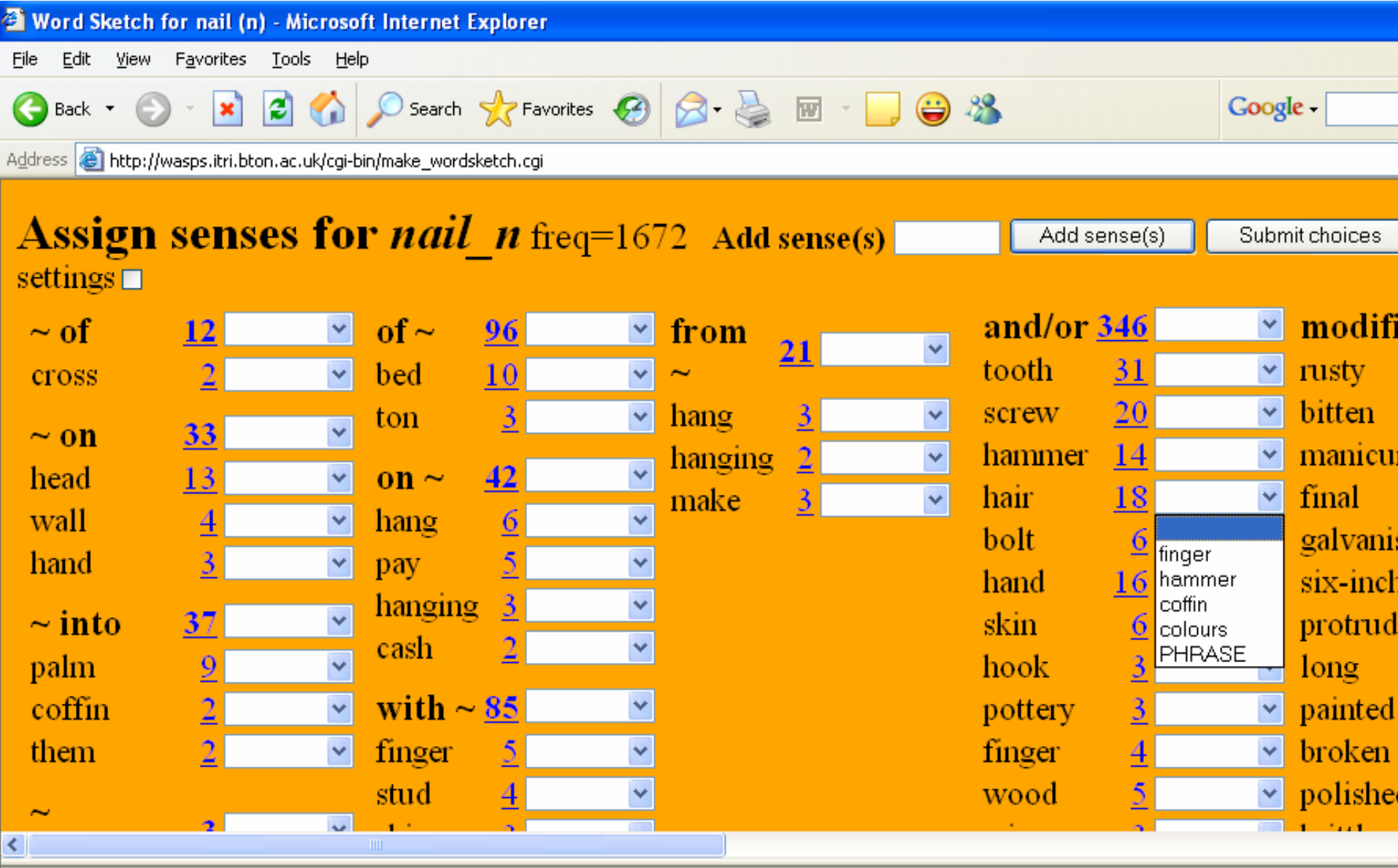
Not just collocates: *CoCoDo*

- collocates
  - <object, drink (v), tea (n)>
- constructions
  - <v+obj+ing, hear (v)> *I hear him singing*
- domain-based clues
  - <domain=computing, mouse (n)>
- **Co**llocates, **Co**nstructions, **Do**mains

Linking CoCoDo's to senses



- Automatically extract CoCoDos from corpus
- How linked to senses?
  - Automatic (WSD techniques)
- Manual
  - WASPS project
  - "dictionary-free": ideal for new dictionaries
  - Labour costs
- Mixed
  - WSD with manual confirmation/correction



concordance window

Output and benefits

- Dictionary directly carries corpus evidence
- Integrated dictionary and corpus
- Rich resource for
  - Lexicographer
  - Dictionary user (learner, translator, …)
  - Machine translation
  - Natural language processing
- Sound model of D-C relations

Automatic dictionary drafting

- CoCoDo functions being added to Sketch Engine
  - WASPS-type semi-automatic
  - Database of mappings
  - Automatic (pure corpus-driven) mode
- Discussions with UK publishers

ADD: Sketch Engine work in progress

- Clustered word sketches
  - http://corpora.fi.muni.cz/bnc/run.cgi/wsclust_form
- Text type sensitivity
  - http://corpora.fi.muni.cz/bnc/run.cgi/hwsketch?ttattrs=bncdoc.genre&corpname=bnc&lemma=run&lpos=v
- Multiword sketches
  - http://corpora.fi.muni.cz/bnc/run.cgi/wsketch?showmwlink=1&corpname=bnc&lemma=run&lpos=v
- Collocationality (= entropy)

Text type

run bnc freq = 38882

object	12051	3.1
risk	354	39.03
errand	55	38.84
gauntlet	46	37.52
counter	102	35.08
finger	236	33.62
usually W_fict_prose,		
marathon	52	31.71
gamut	26	31.36
course	353	31.03
business	389	30.35
riot	70	30.21
program	126	29.21
race	115	25.58
length	107	24.58
mile	100	24.52
Solaris	21	23.23

subject	11762	6.0
tear	78	26.58
usually W_fict_prose,		
sweat	41	26.37
usually W_fict_prose,		
train	111	26.29
thread	37	23.71
shiver	18	22.7
stream	55	22.64
blood	86	22.48
usually W_fict_prose,		
road	143	22.3
engine	68	21.0
brook	14	20.97
bus	55	20.02
tremor	14	19.68
workstation	28	19.45
usually W_non_ac_tech_engin,		

modifier	6798	3.4
away	833	57.56
parallel	117	56.66
smoothly	119	50.6
amok	39	48.36
aground	42	47.73
fast	122	42.73
concurrently	37	42.42
upstairs	67	39.12
usually W_fict_prose,		
straight	105	38.92
faster	55	37.06
efficiently	45	36.73
right	83	32.85
now	266	32.58
currently	73	31.85
long	75	31.14

and/or
run
own
turn
usually W_
scream
jump
swim
walk
organis
shout
cycle
install
cry
laugh
load
dance

Multi word sketches

run Word Sketch - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Google

Address http://corpora.fi.muni.cz/bnc/run.cgi/wsketch?showmwlink=1&corpname=bnc&lemma=run&lpos=v Go

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff

run bnc freq = 38882

change opt

object	12051	3.1	subject	11762	6.0	modifier	6798	3.4	and/or	806	0.2	part_trans 1
risk >	354	39.03	tear >	78	26.58	away >	833	57.56	run >	104	34.4	down >
errand >	55	38.84	sweat >	41	26.37	parallel >	117	56.66	own >	39	30.15	along >
gauntlet >	46	37.52	train >	111	26.29	smoothly >	119	50.6	turn >	74	29.93	up >
counter >	102	35.08	thread >	37	23.71	amok >	39	48.36	scream >	14	23.8	over >
finger >	236	33.62	shiver >	18	22.7	aground >	42	47.73	jump >	17	23.68	out >
marathon >	52	31.71	stream >	55	22.64	fast >	122	42.73	swim >	10	20.97	off >
gamut >	26	31.36	blood >	86	22.48	concurrently >	37	42.42	walk >	19	19.48	around >
course >	353	31.03	road >	143	22.3	upstairs >	67	39.12	organise >	13	17.71	back >
business >	389	30.35	engine >	68	21.0	straight >	105	38.92	shout >	10	17.15	round >
riot >	70	30.21	brook >	14	20.97	faster >	55	37.06	cycle >	5	17.04	through >
program >	126	29.21	bus >	55	20.02	efficiently >	45	36.73	install >	9	16.54	in >
race >	115	25.58	tremor >	14	19.68	right >	83	32.85	cry >	8	15.52	on >
length >	107	24.58	workstation >	28	19.45	now >	266	32.58	laugh >	7	13.93	
mile >	100	24.52	application >	96	18.93	currently >	73	31.85	load >	5	12.7	
Solaris >	21	23.23				long >	75	31.14				

run bnc freq = 38882

cl

subject	247	-16.5
tear	59	23.26
sweat	21	18.82
blood	13	6.09
rain	7	4.45
water	15	1.93

modifier	33	-20.8
almost	5	3.28

part_trans	532	9.1
down	532	53.62

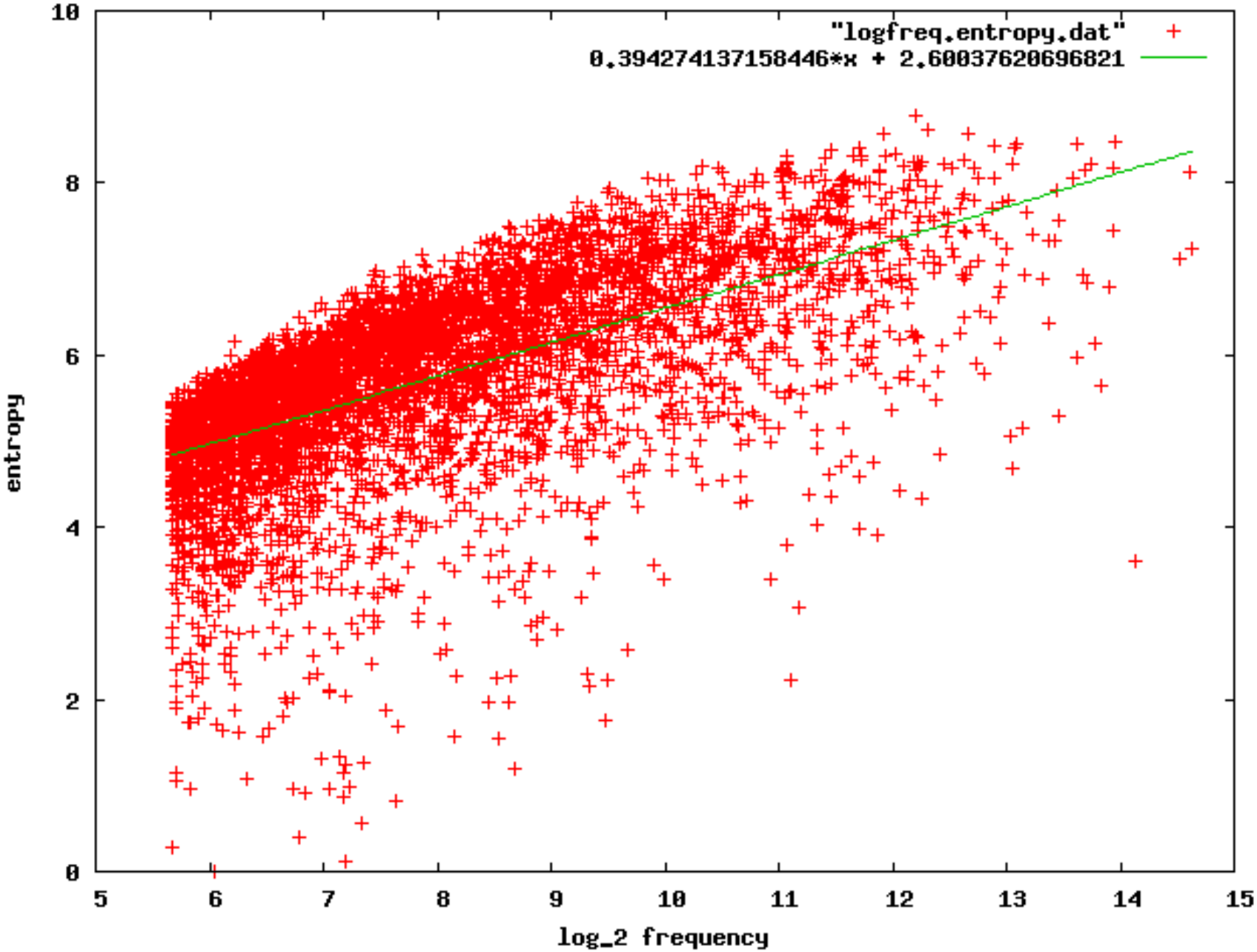
part_down-a_obj	426	31.1
cheek	31	33.05
face	63	30.41
spine	13	25.17
back	24	21.85
chin	10	21.44
slope	5	13.35
passage	6	13.06
neck	5	11.01
flight	5	10.99
side	10	10.74
street	6	10.61
bank	5	7.63
centre	6	7.61
industry	5	7.17

Collocationality

- Which words are most ‘collocational’
- Dictionary publishers
  - Where to put ‘collocation boxes’
- Language learners

Calculation of entropy for *advantage (object relation).*

Verb	Freq	MLE-prob (freq/3730)	Log	Prob x log
Take	2084	.5587	-0.84	-.469
Gain	131	.0351	-4.83	-.169
Offer	117	.0314	-4.99	-.157
See	110	.0295	-5.08	-.150
Enjoy	67	.0180	-5.79	-.104
Obtain	58	.0155	-6.01	-.093
…	…	…	…	…
Clarify	1	.000268	-11.86	-0.0031
…	…	…	…	…
Total	**3730**	**1.000**		**-3.909**


"logfreq.entropy.dat"
0.394274137158446*x + 2.60037620696821
entropy
log_2 frequency
0
2
4
6
8
10
5
6
7
8
9
10
11
12
13
14
15

place (17881), attention (8476), door (8426), care (4884), step (4277), advantage (3730), rise (3334), attempt (2825), impression (2596), notice (2462), chapter (2318), mistake (2205), breath (2140), hold (1949), birth (1016), living (953), indication (812), tribute (720), debut (714), button (661), eyebrow (649), anniversary (637), mention (615), glimpse (531), suicide (486), toll (472), refuge (470), spokesman (453), sigh (436), birthday (429), wicket (412), appendix (410), pardon (399), precaution (396), temptation (374), goodbye (372), fuss (366), resemblance (350), goodness (288), precedence (285), havoc (270), tennis (266), comeback (260), farewell (228), prominence (228), go-ahead (202), sip (198),