

Corpora for the coming decade

Adam Kilgarriff

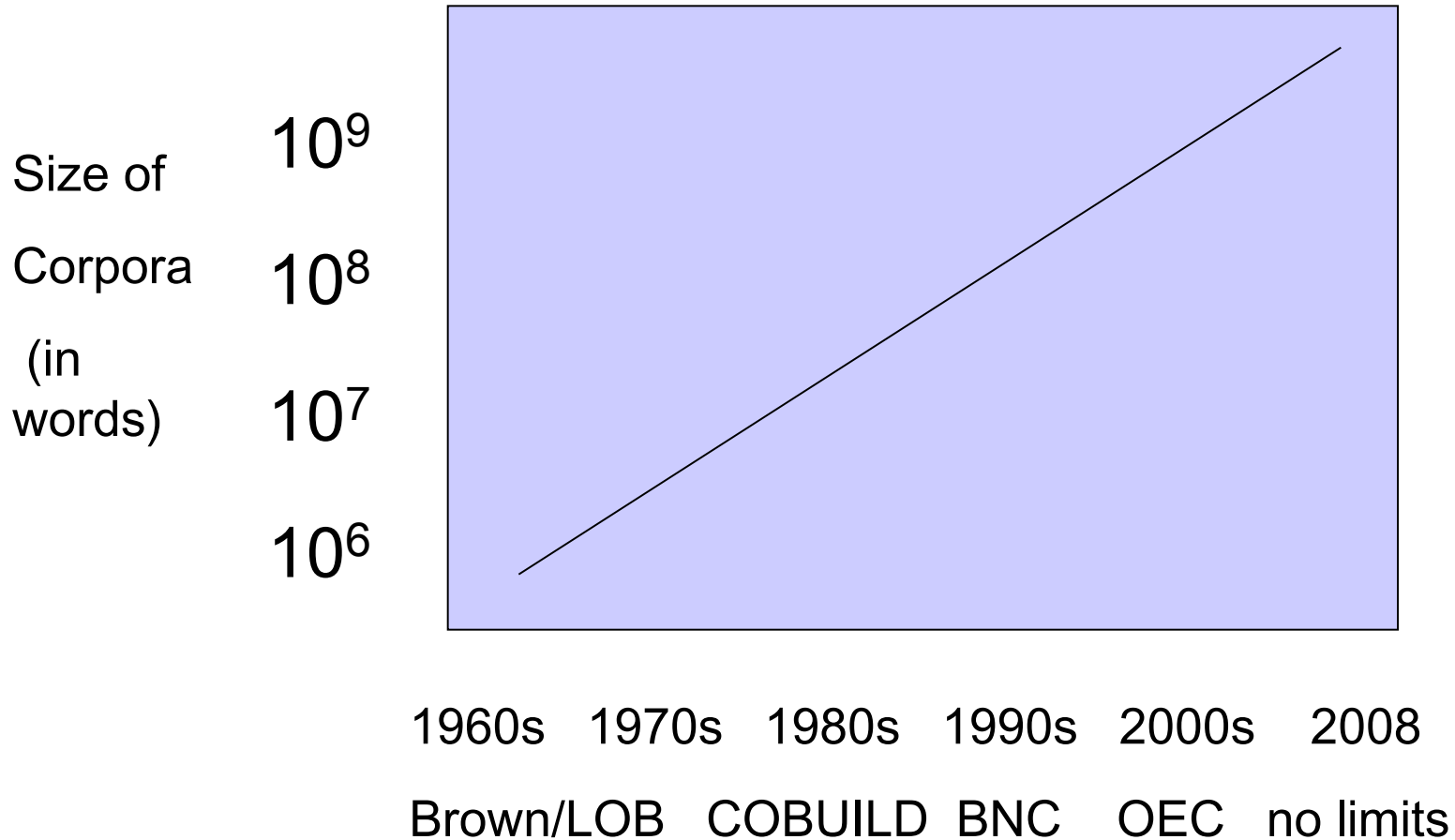
Lexical Computing Ltd

Universities of Leeds, Sussex

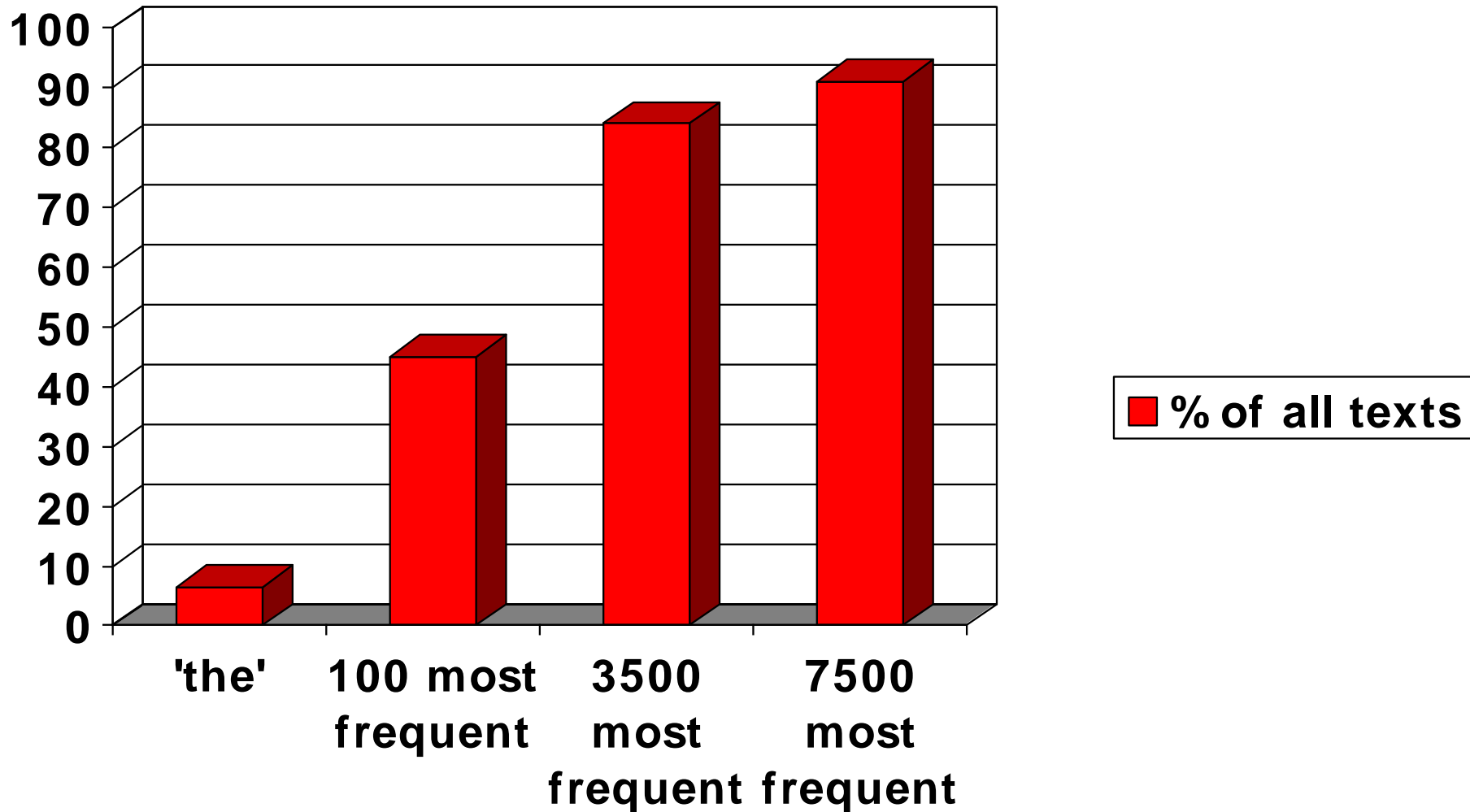
Overview

- History
- Comparing Corpora
- The Web and Corpora
 - Demo
- Corpus Factory
- Simple Maths for Keywords
- Corpora for the Coming Decade

Corpus size since the 1960s



'Zipfian' distribution of words



Comparing Corpora

- Basic science
 - Measure
 - Compare
- State of the art
 - “WSJ”, “medical abstracts”, “general”
 - atrocious

Wall St Journal vs. BNC?

Wall St Journal vs. BNC?

- Homogeneity
 - Self-similarity
- *Use same measure*
 - For homogeneity and similarity
 - (distance measure so:
 - Heterogeneity and distance
 - High number=different/ heterogeneity)

Thought experiment

	Corp1	Corp2	Distance	Interpretation
1	equal	equal	equal	same language variety/ies
2	equal	equal	high	different language varieties
3	high	high	low	impossible
4	high	low	high	corpus 2 is homogeneous and falls within the range of corpus 1
5	high	low	higher	corpus 2 is homogeneous and falls outside the range of corpus 1
6	low	low	a little higher	similar varieties
7	high	high	a little higher	overlapping; share some varieties

Measures

- Homogeneity
 - Divide randomly into halves
 - Measure distance between halves
 - Iterate, average
- Proposed measures
 - word frequency lists
 - Chi-square (normalise by DF): CBDF
 - Spearman Rank Correlation
 - From language modelling
 - *Perplexity*

How to evaluate measures

- Known-similarity corpora
 - Two text types
 - Eleven corpora
 - 100:0, 90:10, 80:20, 70:30, 60:40 ... 10:90, 0:100
 - Gold-standard judgements
 - *80:20 is-more-similar-to 70:30 than 90:10 is to 60:40*
- What percentage of gold-standard judgements does each measure get right?
 - CBDF wins
 - best with 500 DF, 500-most-freq-words

BNC 200,000-wd samples

ACC	4.6												
ART	21.4	3.4											
BMJ	20.2	23.5	3.1										
DMI	21.6	26.2	32.1	2.5									
DNB	40.6	30.1	40.1	35.2	1.9								
ENV	22.7	23.1	28.1	34.7	41.5	2.6							
FAC	20.5	25.1	31.1	7.8	36.9	36.9	3.4						
GRA	27.8	30	33.5	31.4	45.3	29	34.4	2.2					
GUA	14.1	18.4	22.7	11.4	31.1	23.2	12	32.3	3.9				
HAN	24.1	33.8	33	32.1	52.3	32	31.2	36.2	22.6	3.7			
IND	12.8	17.8	23	14	30.1	21.7	14.5	28.1	4.1	23.3	4.4		
NME	21.2	26	30.1	9.8	39.4	34.8	5.8	31.4	15.1	33.3	16.6	3.1	
	ACC	ART	BMJ	DMI	DNB	ENV	FAC	GRA	GUA	HAN	IND	NME	

Then and now

- Work done: 1995
 - Journal article 2001
- Then:
 - Theoretical interest
 - Beggars can't be choosers
- Now
 - Any number of corpora
 - to spec, from web
 - Practical importance

The Web and Corpora

- Is the web a corpus?
- Representativeness
- What is out there?
 - Web1T
- Googleology
- Web corpus types
 - Targeted sites: Oxford English Corpus
 - General: WaC family
 - WebBootCaT

You can't help noticing

- *Replaceable or replacable?*
 - <http://googlefight.com>
 - <http://looglefight.com>

- Very very large
- Most languages
- Most language types
- Up-to-date
- Free
- Instant access

Is the web a corpus?

- Sinclair
 - in “Developing linguistic corpora, a guide to good practice. Corpus and Text – Basic Principles”
 - “...**not** a corpus because
 - dimensions unknown, constantly changing
 - not designed from a linguistic perspective
- But
 - We can find out dimensions
 - Many corpora are not designed
 - “as much chatroom dialogue as I can get”
- Def: a corpus is a collection of texts
 - when viewed as an object of language research

Is the web a corpus?

Yes

but it's not representative

Theory

A random sample of a population is representative of it.

Observations on sample support inferences about population
(within confidence bounds)

Theory

A random sample of a population is ...

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

Theory

- Population not defined
- Representative sample not possible

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

Representativeness

- 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, 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?

- What text types are there on the web?
 - some are new: chatroom
 - proportions
 - is it overwhelmed by porn? How much?
 - **Hard question**

Classifiers

Starter set of text types, with examples

Taxonomy of text types

Build text classifier

Linguist revises/extends taxonomy

Classify new samples:
Check misfits

Take new random sample

Marina Santini, Serge Sharoff

Comparing frequency lists

- Web1T vs BNC
 - Keywords of each vs other

Web-high (155 terms)

- 61 web and computing
 - *config browser spyware url www forum*
- 38 porn
- 22 US English (incl Spanish influence –/os)
- 18 business/products common on web
 - *poker viagra lingerie ringtone dvd casino rental collectible tiffany*
 - NB: BNC is old
- 4 legal
 - *trademarks pursuant accordance herein*

Web-low

- Exclude British English, transcription/tokenisation anomalies
 - *herself stood seemed she looked yesterday sat considerable had council felt perhaps walked round her towards claimed knew obviously remained himself he him*

Observations

- Pronouns and past tense verbs
 - *Fiction*
- Masc vs fem
- *Yesterday*
 - Probably daily newspapers
- Constancy of ratios:
 - He/him/himself
 - She/her/herself

- 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
- Let's
 - study the web
 - source of language data
 - apply our tools for web use (dictionaries, MT)
 - use the web as infrastructure

Web corpus types

- Large, general corpora
- Small, specialised corpora
 - Specially for translators
 - BootCaT, WebBootCaT

Basic steps

- Gather pages
 - Google hits
 - Select and gather whole sites
 - General crawl
- Filter
- De-duplicate
- Linguistic processing
- Load into corpus tool

Filtering

- Non-text (sound, image etc) files
- Boilerplate (within file)
 - Copyright notices, navigation bars
 - “high markup” heuristic
- Not “text in sentences”
 - Look for function words
 - Lists?? Sports results?? Crossword puzzles??
- Spam, pornography
 - Tough
- De-duplication (also tough)

Corpus Factory

- Many languages
- General corpus, 100m+ words
 - Fast
 - High quality
 - Comparable across languages

Gather Seed Words

- Sharoff: used word lists from preexisting corpora
 - BNC for English
 - RNC for Russian
- Bottleneck: No pre-existing large general corpora for many languages.
 - That is why we are building them!
 - Seed words from many domains required.

Gather Seed words

- Wikipedia (Wiki) Corpora
 - many domains
 - free
 - 265 languages covered, more to come
- Extract text from Wiki.
 - Wikipedia2Text
- Tokenise the text.
 - Morphology of the language is important
 - Can use the existing word tokeniser tools.

Gather Seed words

- Thai Word Segmentation
 - Before tokenization
 - ปัญหาของประเทศพม่าในภูมิภาคคืออะไร
 - (Gloss: Burma's problems in the region)
 - After tokenization
 - ปัญหา/ ของ/ ประเทศ/ พม่า/ ใน/ ภูมิภาค/ คือ/ อะไร
 - problem/ of/ Country/ Burma/ in/ Region/ is / ?
- Used Swath word Segmentor.

Gather Seed words

- Most frequent are function words
 - Top 500 (roughly)
 - Use to identify connected text.
- Mid frequency as seeds
 - 1000th to 6000th words (roughly)

Query Generation: cont..

Table 2: Query length, hit counts at 90th percentile and Best Query Length

	length= 1	2	3	4	5	Best
Dutch	1,300,000	3,580	74	5	-	3
Hindi	30,600	86	1	-	-	2
Telugu	668	2	-	-	-	2
Thai	724,000	1,800	193	5	-	3
Vietnamese	1,100,000	15,400	422	39	5	4

Collection

- 30,000 queries
- Retrieve top 10 search hits of each query.
 - Yahoo Search API
- Download

Cleaning

- Body Text Extraction (Finn et al. 2001)
 - Boilerplate: rich in markup
 - Body text: middle of page, light in markup
 - 3 zones: High-low-high
 - Retain low

Filtering

- Wanted: “stuff in sentences”
 - Connected text
- Not wanted: anything else
 - Menus, directories, catalogues...
- Connected text
 - half of all tokens are very common words
- Discard pages failing test

Near Duplicate Detection

- Broder et al (1997) 'shingling'
- To be replaced by Pomikalek's methods (Pomikalek 2009)

Web Corpus Statistics

	Unique URLs collected	After filtering	After de- duplication	Web corpus size	
				MB	Words
Dutch	97,584	22,424	19,708	739 MB	108.6 m
Hindi	71,613	20,051	13,321	424 MB	30.6 m
Telugu	37,864	6,178	5,131	107 MB	3.4 m
Thai	120,314	23,320	20,998	1.2 GB	81.8 m
Vietnamese	106,076	27,728	19,646	1.2 GB	149 m

Evaluation

- For each of the languages, two corpora available:
 - Web and Wiki
 - Dutch: also a carefully designed lexicographic corpus.
- Hypothesis: Wiki corpora are ‘informational’
 - Informational --> typical written
 - Interactional --> typical spoken

Evaluation

- 1st, 2nd person pronouns
 - strong indicators of interactional language.
 - English: *I me my mine you your yours we us our*
- For each languages
 - Ratio: web:wiki

Results

Thai			
Word	Web	Wiki	Ratio
ผม	2935	366	8.00
ดิฉัน	133	19	7.00
ฉัน	770	97	7.87
คุณ	1722	320	5.36
ท่าน	2390	855	2.79
กระผม	21	6	3.20
ข้าพเจ้า	434	66	6.54
ตัว	2108	2070	1.01
กู	179	148	1.20
ฉัน	431	677	0.63
Total	11123	4624	2.40

Table : 1st and 2nd person pronouns in Web and Wiki corpora per million words

Corpora for the coming decade

How should they be different?

- Bigger

- Better

Bigger

- Motivation
 - Ample data for rare phenomena
 - Big subcorpora
 - For language modelling
- More like Google-scale
 - but without Google disadvantages
 - See *Googleology is Bad Science*, CL 2007

Better

- Less noise
- Fewer duplicates
- Richer markup
 - At word, sentence level
 - At document level (text type, subcorpora)

Divide and rule

- Bigger (+ cleaning + deduplication)
 - Big Web Corpus (BiWeC)
 - Currently 5.5b fully processed
 - Target 20b words
 - Jan Pomikalek, Pavel Rychly
- Better
 - New Model Corpus

New Model Corpus

- model
 1. small version: *model train*
 2. design: *data model*
- New Model Corpus
 - 1:100 scale model
 - To replace BNC as design model

BNC design model

- Most often used
 - Eg for other languages
- pre-web
 - $f(\text{blog})=0$
- Corpora now bigger, far quicker, far cheaper, different issues
- *BNC design model past its sell-by*
 - Kilgarriff Atkins Rundell, Corpus Lg 2007

New model

- Data
- Markup

Data

- From the web
- 100m words
- Small sample size
 - Copyright
 - ??Creative Commons Licence

Composition

- General crawl 50
- Targeted
 - Fiction 7
 - Blog 7
 - Newspaper (RSS feed) 7
 - Speech 10
 - Film transcripts, chatshow
 - Domain-specific 19
 - Business, medical, law

Markup

- Collaborative
 - We distribute data
 - Anyone applies their tools
 - Pos-tagger, parser, co-ref resolution, domain classifier, WSD, semantic classifier, time phrases, named entities...
 - We integrate, display in Sketch Engine
 - Research potential from multiple markup

Two strands

- Apply methods with good accuracy (and ***fast***) to BiWeC

Two strands

- Apply methods with good accuracy (and *fast*) to BiWeC
- *Bigger*
- *Better*

Some plans

- Corpus similarity/homogeneity
 - Web service for measuring
- New General Service List
 - Replacing West (1953)
 - Words (English) you *always* need
 - Many corpora of different text types
 - 2000-wd samples
 - Which words occur in 95% of docs in every text type

Hierarchy of Domains

- Domains are in hierarchies
 - Science, physics, subatomic physics
- Domains: represented by corpora
- Can we find correlates in wordlists
- What we *could* find

	Core	science	physics	subatomic	
Science		70	30	0	0
Physics		70	5	25	0
Subatomic		70	5	5	20