

Automating the creation of dictionaries

how far have we come, and what are the prospects for the future

Michael Rundell
Macmillan Dictionaries and Lexicography MasterClass

Outline

- Eliminating ‘drudgery’: 1960-1998
- Towards automation: the last 10 years
- Discussion and outlook: lexicographer’s changing role

Automating the creation of dictionaries

- Key tasks in dictionary-making are:
 - Data collection: corpora and other forms of evidence
 - Data analysis: examining the evidence to discover relevant facts
 - Synthesis: creating dictionary entries

Making a dictionary ... the old way



James Murray and his colleagues on the Oxford English Dictionary

Data collection: developments

□ Pre-1980

- Hand-gathered citations (Johnson, OED)
- Low volumes of data, bias towards the atypical

□ Early corpora (for English)

- Brown (1962), BCET/COBUILD (1982), BNC (1992)
- Steady growth: 1m → 10m → 100m
- But: labour-intensive, expensive, not big enough

Data collection (contd)

□ Post-2000

- Web as data-source: arrival of mega corpora
 - 2bn normal (for English), 20bn on horizon
- Data collection and annotation largely automated: a 'one-stop' operation (WebBootCat and similar: Baroni et al. 2006)

Data collection: beyond the corpus

□ Google resources

■ n-gram viewer:

<http://www.ngrams.googlelabs.com/>

□ Twitter feeds, e.g.

■ <http://apps.buradayiz.webfactional.com/twitter/gender/query/>

■ shows gender preferences in vocabulary choices

□ How to use this new data?

Data collection: outcomes

- ❑ Time/effort: 99% less
- ❑ Costs: *much* lower, reduces 'entry fee' to corpus lexicography
- ❑ Diversity of content: some gaps on Web, but *mainly* positive (e.g. Keller & Lapata 2003, Fletcher 2004)
- ❑ Data volume: effectively unlimited, end of data-sparseness

Data analysis: developments

- Read/sort citations (Johnson to 1980):
manual, labour-intensive
- Scan concordances (COBUILD1: hard copy): revolution for lexicography but
 - sparse data
 - absence of linguistic processing
- From late 1980s: annotated corpora
 - lemmatization, POS-tagging

Extract from original COBUILD (BCET) corpus: data for *seal* (1983)

m br 132 ook and cranny in the vessel where even a stray seal could be hiding and you can take my word
 m br 25 er. "every wave on the atlantic was like a dead seal drag- ing its driftwood artillery from h
 8 seal

r br 127 e king's taster?' i looked at the unbroken lead seal. 'not uness you think some- one has brought
 r br 129 church until 1835. years later, galileo put his seal on copernicus's discovery, wvas hauled up b
 r br 140 her lover to assuage her inner doubts, set the seal on her femininity, provide her with psychic
 m br 172 brooding darkncss is lifted? could the seventh seal or winter light have been conceived in anot
 m br 14 that never cleaned anything away, heavy thermal seal over diesel fuel, mildew, garbage, excremen
 a br 34 s foot in it. lynn tried to be gracious but the seal was set on her dislike of him, and somethin
 a br 127 plant aboard.' 'i've checked.' smithy broke the seal. 'we talked last night. at least, i did. yo
 a br 82 ce she discovered that, lynn thought, the final seal would be set on jane's hatred and rebellion
 10 sealed

r br 133 ingenue. both their fates were, to some extent, sealed. after "bunty" closed he went sady back t
 r br 129 place strips of the paper in a thin rubber tube sealed at one end and connected to suction at th
 a br 151 d as superior and knowledgeable. a partnership. sealed by why? so many exquisite little symmetri
 a br 138 g was led., the europea party swept to dover, in sealed cars through back streets. "you were a lo
 a br 135 ss asked "thank you. i am not fond of salad." a sealed envelope passed to the prime minister wit
 a br 86 c forms and filled them out. i put those in one sealed envelope, the signed affidavit - i just
 a br 132 ottage he would flee to when all was signed and sealed. he hadn't had a proper night's sleep for
 r br 80 ed. "on a night like this? no fear. the gash is sealed in polythene bags, then they're punctured
 r br 84 lions of years but his doom, paradoxically, was sealed in the very fact that he became too perfe
 r br 199 ote out the telegram, put it into its envelope, sealed it and handed it over to dolly. the four
 2 sealing

m br 48 gon stream thinning and trickling out: frontier sealing, cencus grievance, black operations (pre
 r br 21 m each other. our once one-flesh divided again, sealing me into me, him into him. he is now a te
 3 seals

m br 3 saw a row of old houses, huddled together like seals on a rock. then there was a long field tha
 m br 35 ang we'd get stone together and keep the lurps, seals, recondos, green-beret bushmasters redunda
 r br 101 omen serves only their own artificial needs and seals them off in their folie a deux from the re

Data analysis: outcomes

- Volume of data
 - BCET corpus **23** hits for *seal* (n & vb)
 - UKWaC corpus has **38,237**
- Ease of use: corpus query systems
 - BCET: printed, fixed order
 - Now: multiple views, sophisticated functionality
- Better data – but much more of it (so still poses analysis problems)

Entry writing: developments

Two aspects

- Applying labels: register, region, domain, time, etc

- Inserting and aligning cross-references

- shut/close the stable door after the horse has bolted*

- where does it go? what cross-refs are needed?

Labels: extract from (old) OED style guide

Order and punctuation of labels

Single labels are followed by a full stop.

Where there is more than one label in an entry, the order is:

regional, subject, register, usage, status ...

Usage labels such as *derog.* or *humorous* are placed in parentheses when they follow other labels, although not when they are used on their own....

Where you wish to use two labels from the same category, e.g. two status labels, they are joined using a roman ‘and’, not ‘or’, unless there is actual doubt.

Examples:

Physics. rare.

S.Afr. Mining.

N.Amer. Mil. slang

U.S. Pol. (offensive)

<Entry><DEnt><FwkSenCnt><LabelGp><STYLE>

FwkSenCnt [5]

POS [N]

LABELGP

REGIST [INF]

DOMAIN [BASKB]

STYLE [JOURN]

MEANING a jump shot in basketball

EXCNT

EX Perry was fouled by Steve Kerr while attempting a **jump**er and sank both foul shots .

EXCNT

EX He sank 6 of 7 shots in the half , setting off the loudest cheers of the game with hook shots , baseline **jump**ers and turnarounds over outmatched Atlanta center Jon Koncak .

FwkSenCnt [6]

POS [N]

LABELGP

REGION [HIBE]

DOMAIN [REL]

MEANING a person who converts from Catholicism to Protestantism

EXCNT

EX "A local man said that the Jeningses were 'jumpers', and he explained what it meant: 'The word comes from the Irish - d'iompaigh siad ina bProtastúnaigh (they turned Protestant). And they say that's why they dropped an 'n' from the name after that, y'know? Because all the Jenningses in Mayo are Catholics.'"

FwkSenCnt [7]

VARCNT

VAR jumper wire

ATTRIBUTES

Element: STYLE	
Name	Value
label	journ
	- undefined -
	child
	drugs
	euph
	fig
	For
	Fr
	Ger
	hum
	iro
	It
	journ
	Lat
	leg
	lit
	pc
	prov
	Span
	spok
	tech
	TM
	youth

Cross-references

- ❑ Old model (e.g. OED1): completely manual
 - labour-intensive, unreliable
- ❑ Next: computer provides error report (shows unmatched x-refs), lexicographer fixes
- ❑ Now: completely handled by DWS:
efficiently, accurately, without you realizing it

Entry writing: outcomes

- From 1960s (Random House): dictionary as database, each entry component has own field
 - Some data-types not intended for end-user (e.g. semantic codes in LDOCE1)
- Publishers develop home-grown systems
- From 1990s: dedicated DWS - single package handles
 - text origination, editing, database functions, workflow, output

Summary: 1960-1998

□ Positives

- Corpora: well-annotated, relatively inexpensive, much larger
- Corpus-querying software: fast, sophisticated, less noise
- Entry writing: dedicated DWS facilitates main tasks, from entry writing to publishing
- All: less drudgery, *and* computers do it better

Summary: 1960-1998

□ Limitations

- Ever-growing volume of data: analysis process increasingly demanding for lexicographers
- Core tasks still depend on human effort, human judgment

Newer developments: 1998-

- Word Sketches
- GDEX
- Preferences: colligational and text-type information
- Quality control
- Tickbox lexicography

Word Sketches

- First version: MEDAL1 project, 1998-2001
 - needed systematic account of collocation
 - sketches presented as standalone html files

- Unintended consequences
 - first place to look: Word Sketch *then* concordance
 - solves 'data overload' problem
 - for humans: how to read and process all that data?
 - for computers: more data is better data, higher chance of separating signal from noise

Changing role of technology

□ Before: supportive

- facilitates lexicographer's work

□ Now: proactive

- identifies salient facts, presents them to the lexicographer
- lexicographer makes final selection

Word Sketches: further developments

- Sketch grammars customized to specific projects
 - for DANTE project: gramrel names in Sketches conform to DANTE styles
 - ‘constructions’ shown first in Word Sketch
 - PP types (many) on a separate page
 - new layout maximizes efficiency
 - ‘More/Less data’ buttons
 - ‘One-click copying’: from corpus to DWS

'Constructions' for DANTE project



[Home](#) [Settings](#) [Log out](#)

Search in [Help](#)

user: Michael Rundell corpus: LEXMCI

Search in LEXMCI

- [Concordance](#)
- [Word List](#)
- [Word Sketch](#)
- [Thesaurus](#)
- [Sketch-Diff](#)
- [? Help on main menu](#)

- [Save](#)
- [Change options](#)
- [Turn on clustering](#)
- [More data](#)
- [Less data](#)
- [Switch menu position](#)

remember *(verb)* LEXMCI freq = 250722

Constructions		
that_0	47469	14.4
Ving	13900	10.5
Vinf_to	13330	3.4
wh	13223	7.7
NP_NP	5425	8.3
NP_Ving	4040	15.4

PP_cl_wh	291	0.9
though	12	1.02

PP_Ving	724	0.8
once	12	2.41
whilst	9	1.25
while	23	0.79

NP_PP_Ving	1060	3.4
------------	----------------------	-----

NP	65745	3.9
anything	677	5.89
name	1295	5.71
word	805	5.38
everything	405	5.33
occasion	234	5.29
day	2007	5.23
password	141	5.17
feeling	223	5.1
victim	180	5.06
standing	128	5.06
conversation	139	5.0
incident	203	4.94
excitement	84	4.82
thing	1168	4.81
moment	249	4.75

Part	67	0.1
along	22	3.37

NP_Part	93	0.2
along	16	2.91
off	32	0.95

Part_PP	58	0.2
along_with	22	5.2
up_to	18	1.95
out_of	10	0.54

NP_PP	20381	3.3
as	2561	3.53
about	641	3.4
because	122	2.99
like	209	2.92
so	58	2.88
before	161	2.68
unto	6	2.48
once	15	2.4
down	13	2.35
outside	28	2.34
from	1023	2.25
beside	6	2.24
at	1054	2.2
of	6138	2.1
except	8	2.05

AJP	1322	0.3
being	186	8.05
....	24	5.02
.....	9	4.89
mum	7	4.48
most	236	4.39
loving	15	4.19
..	15	3.87
	12	3.5
	24	2.96
everyday	7	2.47
	5	2.32
more	183	1.98
little	57	1.97
	9	1.85
least	7	1.56

'GDEX': good example software

- First use, Macmillan project: attach examples to collocations
 - required 8000 new examples (suitable for learner's dictionary)
 - 'traditional' method expensive – can it be automated?
 - GDEX: selects and promotes 'best' examples
 - lexicographers *usually* select from top ten – streamlines example-collection process

GDEX heuristics

- Sentence length (10-26 words): 10 words enough for Czech?
- Mostly common words: good
- Rare words: not so good
- Full sentences only
- Not too many pronouns (anaphoric reference)
- Not too many capitals (proper names)
- Typicality: *third collocate* is a plus

GDEX weightings

- For each sentence
 - Score on each heuristic
 - Weight scores
 - Add together weighted score
- How to set weights?
- Further development (e.g. Slovenian project)
 - users select best examples manually
 - system 'learns' from human choices
 - weightings refined iteratively

Labels: colligation, text-type

- Dictionaries use labels to identify preferences
 - Colligational preferences (cf. Hoey 2005), e.g.
 - *usually passive, never before noun, usually plural, always imperative*
 - preferred position in sentence: *at the same time*
 - Text-type preferences, e.g.
 - style, register (*formal, informal, journalistic, etc*)
 - region (*American English, Indian English etc*)
 - domain (*IT, chemistry, business, medicine etc*)

Applying Labels

- Currently: a manual process
 - lexicographer applies label if s/he notices a strong preference for a given word, phrase etc
 - unreliable, unsystematic
- A blunt instrument: limited range of categories, scope for more granularity, e.g.
 - *always sentence-final, mainly in narrative/descriptive writing, etc*
- How to (a) automate (b) make more systematic?

E.g. find nouns that are *usually plural*

- For each noun in the corpus
 - Count, under condition 1
 - all plural instances
 - Count, under condition 2
 - all instances
 - Compute ratio
 - Sort all words according to ratio
 - Words at top of list are best candidates
- Mutatis mutandis, similar process for other types of label (*usually passive* etc)

In development: automating domain labelling

- Classifying web genres: big research area
- Compare general and specialised corpora to identify ‘keywords’ in each domain
 - e.g. keywords for Botany: *rhizome, anther, pistil, stamen, integument, stigma, etc*
- Software applies labels to keywords automatically

Quality control

- Critical feature
 - big, long-term projects
 - large teams (often geographically dispersed)

- Traditional method
 - Senior Editor scans text, identifies problems
 - if trivial, fixes them
 - if systematic (has lexicographer not understood an aspect of editorial policy?), gives feedback

Quality control

Newer approach

As above, but also

create inventory of recurrent problems

use search scripts in the DWS to identify all cases of these

fix them in a single operation:

some manually, some by program

Using search scripts for quality control

□ Common problem: distinguishing

- *We wanted to go*: + Vinf_to
- *We wanted her to go*: + NP Vinf_to
- *We were advised to go*: ?

□ Use these scripts in DWS search system

- `<FwkStrCnt:(%<strv@code=(Vinf_to)),<hwd:(^#[a-m].*)`
- `<FwkStrCnt:(%<strv@code=(NP-Vinf_to)),<hwd:(^#[a-m].*)`

SkXmlBox Evidence finder

Evidence finder Tutorial Close

in as: michael.rundell (Michael Rundell)

Personal information

rCnt:(%<strv@code=(NP Vinf_to),<hwd:^(^#[a-e].^)

Go! Query builder Databases

r query: <FwkStrCnt:(%<strv@code=(NP Vinf_to),<hwd:^(^#[a-e].^)

Results layout

corpus: New English Irish Dictionary for LexMC (128 result(s) Summary Print Go to list module Export

context: 30, fixed

28 matches in 96 documents . Go to page: 1 2 3 4 5 6 7

DocID	label	<input type="checkbox"/> xml_0
Select All		
dit View	<input type="checkbox"/> combine	</POS><MEANING>join, merge</MEANING><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>Several fa
dit View	<input type="checkbox"/> command	oxen that drew him out of the city.</EX></ExCnt></FwkStrCnt><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>And so Gr
dit View	<input type="checkbox"/> commandeer	</POS><MEANING>enlist sb to aid in a task</MEANING><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>For some
dit View	<input type="checkbox"/> commission	γ : Subject to certain reservations .</EX></ExCnt></FwkStrCnt><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>Perhaps f
dit View	<input type="checkbox"/> commit	mitting yourselves to buying .</EX></ExCnt></FwkStrCnt><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>The docu
dit View	<input type="checkbox"/> compel	</POS><MEANING>make someone do sthg</MEANING><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>The meas
dit View	<input type="checkbox"/> concede	could have been gentler I guess .</EX></ExCnt></FwkStrCnt><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX> The initia
dit View	<input type="checkbox"/> condemn	eath in Panama ?</EX></ExCnt></FwkCollocCnt></FwkStrCnt><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>It will nev
dit View	<input type="checkbox"/> condemn	, disease , starvation and death .</EX></ExCnt></FwkStrCnt><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>In his illn
dit View	<input type="checkbox"/> configure	s , mail , scripts and macros scan .</EX></ExCnt></FwkStrCnt><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>Various o
dit View	<input type="checkbox"/> conjure	</MEANING><LabelGp><TIME label="obs"></TIME></LabelGp><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>In the na
dit View	<input type="checkbox"/> conscript	> for life long service in the army .</EX></ExCnt></FwkStrCnt><FwkStrCnt><STRV code="NP Vinf_to"></STRV><ExCnt><EX>Trained s

Quality control: outcomes

- More complete, more systematic quality control
 - run routine checks at regular intervals
- Further development: program checks to
 - run automatically
 - fix problems automatically

Tickbox lexicography (TBL)

- Combines Word Sketch and GDEX, e.g.
 - version of Word Sketch with tickboxes beside each collocate
 - tick required collocations, click 'next' button
 - system offers six possible examples (filtered by GDEX)
 - tick required examples, then click 'copy to clipboard' button
 - system builds XML structure (according to DTD of target dictionary)

Concordance

Word List

Word Sketch

Thesaurus

Sketch-Diff

Help on main menu

Save

Change options

Turn on

clustering

More data

Less data

Switch menu position

advice (noun) ukWaC (MCD gramrels) freq = 296155

1 : an opinion that someone gives you about the best thing to do in a particular situation

v+N	112412	2.5
<input checked="" type="checkbox"/> provide	20188	8.59
<input checked="" type="checkbox"/> give	17570	8.56
<input checked="" type="checkbox"/> offer	12651	8.98
<input checked="" type="checkbox"/> seek	10519	10.13
<input checked="" type="checkbox"/> take	5375	6.26
<input checked="" type="checkbox"/> get	4900	6.65
<input checked="" type="checkbox"/> need	3352	7.23
<input type="checkbox"/> be	3310	1.54
<input checked="" type="checkbox"/> follow	3283	6.73
<input type="checkbox"/> include	2387	5.5
<input checked="" type="checkbox"/> receive	2284	6.86
<input type="checkbox"/> have	1741	2.9
<input checked="" type="checkbox"/> obtain	1629	7.47
<input type="checkbox"/> require	1173	5.8
<input type="checkbox"/> contain	990	5.86
<input checked="" type="checkbox"/> want	912	6.22
<input type="checkbox"/> like	829	6.74

N+v	21293	0.8
<input type="checkbox"/> be	11862	3.39
<input type="checkbox"/> regard	1101	7.31
<input type="checkbox"/> have	1062	2.2
<input type="checkbox"/> please	691	6.51
<input type="checkbox"/> do	424	2.77
<input type="checkbox"/> relate	366	5.06
<input type="checkbox"/> include	349	2.8
<input type="checkbox"/> concern	322	5.57
<input type="checkbox"/> come	194	2.55
<input type="checkbox"/> follow	143	2.34
<input type="checkbox"/> provide	141	1.51
<input type="checkbox"/> need	129	2.72
<input type="checkbox"/> receive	127	2.92
<input type="checkbox"/> cover	123	3.13
<input type="checkbox"/> go	121	1.65
<input type="checkbox"/> apply	109	3.3
<input type="checkbox"/> centre	100	5.54

n+N	45711	1.6
<input checked="" type="checkbox"/> expert	4021	9.14
<input type="checkbox"/> specialist	2655	8.46
<input type="checkbox"/> career	2155	7.81
<input type="checkbox"/> business	1391	5.42
<input type="checkbox"/> health	1135	5.46
<input type="checkbox"/> offer	1065	7.14
<input type="checkbox"/> safety	924	6.48
<input type="checkbox"/> housing	700	6.53
<input type="checkbox"/> telephone	690	6.97
<input type="checkbox"/> offering	674	7.66
<input type="checkbox"/> consumer	624	6.72
<input type="checkbox"/> travel	590	6.44
<input type="checkbox"/> policy	553	4.47
<input type="checkbox"/> tax	543	5.9
<input type="checkbox"/> money	501	4.92
<input type="checkbox"/> energy	464	5.33
<input type="checkbox"/> debt	462	6.6

N+n	17749	0.6
<input type="checkbox"/> service	2969	5.38
<input type="checkbox"/> line	1033	5.36
<input type="checkbox"/> centre	1002	5.64
<input type="checkbox"/> agency	646	6.12
<input type="checkbox"/> session	634	5.9
<input type="checkbox"/> page	414	4.02
<input type="checkbox"/> leaflet	358	6.94
<input type="checkbox"/> worker	324	4.85
<input type="checkbox"/> note	296	4.86
<input type="checkbox"/> bureau	252	8.27
<input type="checkbox"/> contact	235	4.04
<input type="checkbox"/> surgery	231	6.06
<input type="checkbox"/> sheet	213	5.39
<input type="checkbox"/> sector	192	3.74
<input type="checkbox"/> work	190	1.53
<input type="checkbox"/> 	171	3.1
<input type="checkbox"/> provider	167	4.55

adj+N
<input checked="" type="checkbox"/> legal
<input type="checkbox"/> good
<input checked="" type="checkbox"/> practical
<input checked="" type="checkbox"/> professional
<input type="checkbox"/> further
<input type="checkbox"/> free
<input checked="" type="checkbox"/> independent
<input checked="" type="checkbox"/> medical
<input type="checkbox"/> general
<input type="checkbox"/> available
<input checked="" type="checkbox"/> financial
<input type="checkbox"/> specific
<input checked="" type="checkbox"/> impartial
<input type="checkbox"/> technical
<input type="checkbox"/> more
<input type="checkbox"/> useful
<input type="checkbox"/> sound

aspect

- The Equal Opportunities Adviser provides guidance and **advice** on all aspects of equal opportunities.
- We offer **advice** on all aspects of travel such as insect bite protection, malaria prophylaxis, travellers diarrhoea and altitude sickness.
- Our highly experienced sales staff, engineers and link installers give you the very best possible **advice** on all aspects of the products we sell.
- 'It has saved me any number of revenue-earning hours, providing **advice** on every aspect of running a small business.' Rosemary Rowntree, International Personnel Management Ltd, Huntingdon.
- Heritage taxation Providing comprehensive **advice** on every aspect of the taxation of chattels and land, including offers in lieu of Inheritance Tax, private treaty sales, public access to exempt items and land, maintenance funds and heritage management plans.
- Advice** on legal aspects of the method in Scotland should be sought from the disposing department's Scottish solicitors or from the Office of the Solicitor to the Advocate General for Scotland (part of the Scotland Office).

matter

- These include increasing access to appropriate bank accounts, affordable credit facilities and face to face **advice** on money matters.
- But greater experience can be valuable if you're seeking **advice** on complex matters such as estate or trust planning.
- If you need help or advice The Terence Higgins Trust (THT) provide a variety of services that include support groups, counselling services, out reach workers and **advice** on health matters.
- The National Assembly receives **advice** on matters related to ancient monuments from the Ancient Monuments Board for Wales.
- The Ombudsman's decisions can be the subject of judicial review proceedings and you may wish to seek legal **advice** on that matter.
- Initial **advice** on any commercial litigation matter is free of charge and this does not mean that the advice is limited to purely sending us an email.

range

- Incentevents can also offer **advice** on a range of travel and tour related services to members to include flight reservations, sea travel, UK & Worldwide hotel booking service, vehicle hire, ferry travel, plus conference meetings and event booking service.
- For more information visit www.lgcareers.com Connexions Connexions Direct is a service for young people aged 13-19 that offers quick access to information and **advice** on a wide range of topics, including employment and training, through one easy to use website, www.connexions-direct.com, or visit the Nottinghamshire branch at
- By acting as a central point of contact for local tourism businesses, Jo will be able to offer support and **advice** on a range of schemes and activities to help develop the tourism industry at a local level.
- We offer support and **advice** on a range of youth related issues.

TBL: outcomes

- ❑ Streamlines corpus analysis process
- ❑ Gets contextual data + related examples into the DWS quickly
- ❑ First stage in creating a dictionary
- ❑ Further developments in progress

Conclusions: the story so far

- ❑ Data collection: largely automated
- ❑ Data analysis: streamlined
- ❑ Simpler lexicographic tasks (drudgery): largely automated
 - computers better than humans: faster, more accurate
- ❑ More complex tasks (e.g. labelling, constructions) significantly streamlined
- ❑ Hardest tasks (word senses, defining): more work to be done

Prospects

- Lexicographer's changing role
 - **from** scanning data to identify lexicographically-relevant facts
 - **to** validating (or rejecting) decisions made by computer

- New role
 - Identify/describe what can be automated → expand set of automatable processes
 - Identify weaknesses in support software