From data to dictionary: corpus-based lexicography, past, present - and future?

Michael Rundell
Lexicography MasterClass
lexmasterclass.com
and Macmillan Dictionary
macmillandictionary.com

Outline

1. Background: two news stories and what they tell us
2. Revolution 1: the arrival of corpora
3. Corpus data and how to use it
4. Revolution 2: from print to digital
5. Crowdsourcing, wikis, user-generated content, and their potential
6. What's coming next?

Warm up

What's the difference between

jargon

and

terminology?

Verbs and adjectives that go with…

terminology *(noun)*

LEXMCI freq = 12,402 (7.2 per million)

Constructions		
wh	307	2.2
that_0	232	0.9
Vinf_to	193	0.6

object_of	3,720	2.7
standardise	52	7.05
demystify	7	5.48
clarify	33	5.11
standardize	3	4.52
encode	14	4.48
harmonise	5	4.37
specialise	27	4.22
assimilate	4	3.95
use	1,311	3.74
understand	148	3.72
borrow	13	3.67
associate	54	3.59
adopt	53	3.57

AJ_premod	3,980	3.0
grammatical	41	7.04
multilingual	21	6.55
non-logical	11	6.47
unfamiliar	22	5.75
Aristotelian	8	5.55
confusing	27	5.53
obscure	23	5.43
monolingual	6	5.34
archaic	9	5.33
standardized	8	5.32
object-oriented	8	5.32
linguistic	27	5.2
correct	103	5.1

Verbs and adjectives that go with…

jargon *(noun)*

LEXMCI freq = 6,491 (3.8 per million)

Constructions		
wh	167	2.3
that_0	124	0.9
Vinf_to	95	0.6

object_of	1,719	2.4
demystify	19	7.63
spout	14	7.0
de-mystify	4	6.16
junk	3	5.51
avoid	244	5.4
decipher	5	5.22
banish	6	5.03
debunk	3	4.93
unravel	5	4.59
eschew	3	4.44

AJ_premod	2,223	3.3
incomprehensible	22	6.97
marketing	9	6.95
impenetrable	13	6.46
Avoid	11	6.37
meaningless	24	6.26
arcane	9	6.05
technical	285	5.85
unnecessary	51	5.78
pseudo-scientific	4	5.73
pretentious	8	5.73
cp	4	5.72
confusing	28	5.72
off-putting	4	5.34
legal	252	5.09

Two news stories
(1) the power of dictionaries

Oxford Junior Dictionary's replacement of 'natural' words with 21st-century terms sparks outcry

"A" should be for acorn, "B" for buttercup and "C" for conker, not attachment, blog and chatroom, according to a group of authors including Margaret Atwood and Andrew Motion who are "profoundly alarmed"….. "There is a proven connection between the decline in natural play and the decline in children's wellbeing," they write, …"Obesity, anti-social behaviour, friendlessness and fear are the known consequences," they say…. The Oxford Dictionaries have a rightful authority and a leading place in cultural life … (13 Jan 2015)

What this tells us

- people care about what dictionaries include
- people see dictionaries as having "authority" and influence

Two news stories
(2) Public engagement with language

He might be a pedantic oddity, but Wikipedia's grammar crusader is my modern-day hero

I wish I had the dedication to remove from public gaze all incidences of "hopefully" being used incorrectly, and I'd love to have the nerve and moral courage to correct people when they use split infinitives … (Simon Kelner, *Independent*)

What this tells us

- huge public interest in language - and partisanship: people take sides
- everyone has a view - but most are misguided

http://www.macmillandictionaryblog.com/hopefully
http://www.macmillandictionaryblog.com/real-grammar-quiz-question-4-is-it-ok-to-split-an-infinitive

Two revolutions in dictionaries (1) The “Corpus Revolution” (1980s)

- John Sinclair and the COBUILD project
  - started 1981, University of Birmingham
  - first lexicographic corpus of English
  - > 7 million words, order of magnitude bigger than Brown, LOB
  - underpins first corpus-based dictionary (COBUILD 1987)
  - See Sinclair (Ed.) 1987

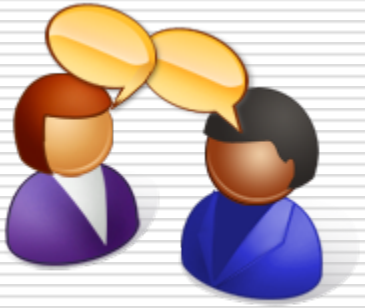
Extract from the COBUILD corpus: data for *seal*

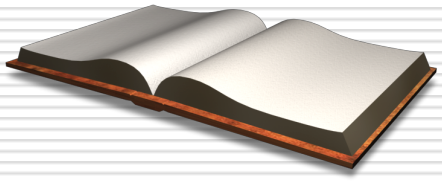
```
   br l32    ook and cranny in the vessel where even a stray seagull could be hiding and you can take my word
m br  25     er. "every wave on the atlantic was like a dead seagull drag- ing its driftwood artillery from h

                                                      8   seal

   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
   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
   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
   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 181     omen serves only their own artificial needs and seals them off in their folie a deux from the re
```

The “Corpus Revolution”: what happened next

- Steady development
  - More data
    - COBUILD corpus **7 million** words, 23 hits for *seal*
    - current corpora **2 billion words+**, 50,000 hits for *seal*
  - Smarter corpus querying software (CQS)
    - from static printouts to multiple querying options
  - UK dictionaries become corpus-based
- Effects profound — but mostly “internal”
  - revolutionises dictionary-making
  - but do users notice?

Corpus-based dictionaries: from corpus to dictionary

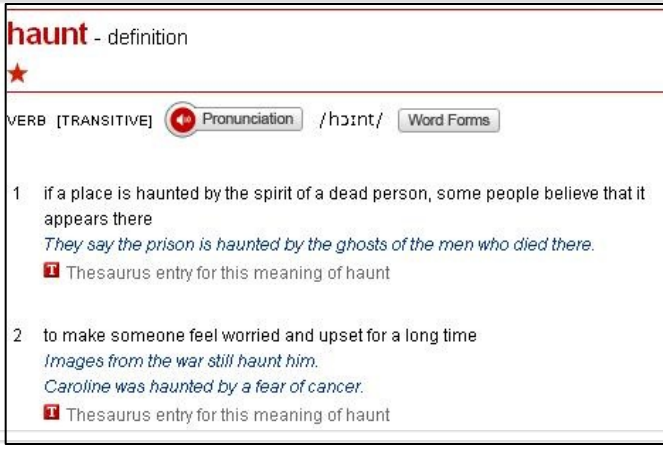




Language in use
written and spoken

Corpus: a
sample

etween - to put it another way - those **haunted** by the spectre of a bad harvest and con
vest and consequent famine, and those **haunted** by its opposite, overproduction or sudd
een even poorer, but who were always **haunted** by the spectre of poverty (as they unde
l intercourse. The bourgeois world was **haunted** by sex, but not necessarily sexual promi
r the grass in confidence. Its ancestors **haunt** the mixing bowl, stirred to a gallop by
surprised that the Tory party's . It'll **haunt** the Tory party for years. It'll be a bigger
e them again. Like the lost chord, they **haunted** me. I wandered through Biomorph Land,
e from his collaborators? A composer is **haunted** by certain themes, subjects or moods, a
gs of fear and apprehension which had **haunted** her during the last week. It wou
ight, and now seemed to be practically **haunting** the place, and obviously very much in lo
e, the memories came flooding back to **haunt** her. Rationally she knew it was



Dictionary: a
distillation

What do we need to know about words?

- That they're **real**: what's the evidence?
- What they **mean**
  - how many meanings do they have?
  - are there *implicatures* (indirectly expressed meanings)?
- How they **combine** with other words
  - syntactic behaviour, collocations, phraseology
- What **contexts** they are used in
  - formal/informal/technical etc – or *any context*?
  - British/American/Australian etc – or *anywhere*?

Combinations: which syntax patterns do you find with *decide*?

- A concordance for *decide*

brother and her aunt. What does she	**decide**	on in both cases? What is a "Sonderangebot
in a much better state. So we had to	**decide**	what to do. Abandon it - or try and make
previous September. What happens if I	**decide**	to reduce my hours? There will be an effect
t would have been like 100 years ago.	**Decide**	if the changes are better or worse. Do
ated. One summer my people	**decided**	to send me to college. This is how it happen
the way ahead. Mr Speaker, we have	**decided**	to accept these principal recommendations
land study. More elaborate tables for	**deciding**	who might be eligible for treatment are
teers. The direction of WDM's work is	**decided**	by our elected Council. Two thirds of WDM
Arakawa and Gins, in many ways, have	**decided**	(it's obvious!) not to be unhappy. They
tarters and six main courses. We both	**decided**	on the ogen melon for the starter. The
home. In January 1998, the Appellant	**decided**	to go to India for four weeks. She asked
final interview the Sub-Committee will	**decide**	whether or not to re-instate. The Secretary
e with the robotic approach. We also	**decided**	to investigate ways to reduce the programmi

Which verbs collocate with *deadline*?

ability to keep work organised and meet	**deadlines**	is essential. AS level General Studies
for some reason they are waiting until the	**deadline**	passes. And if they did have genuine processing
can learn new skills quickly, can work to	**deadlines**	and follow instructions. • A good communicator
billion. Since then there have been missed	**deadlines**	and deferment of expected dates for announcer
next thing I know I'm getting emails about	**deadlines**	! Anyway, throughout the afternoon the emphas
is completed accurately and to required	**deadlines**	and deal with queries or errors quickly
bodies are in overdrive trying to meet a	**deadline**	. These reactions are emotional (anxiety
Breathnach (proinnsias.breathnach@may.ie).	**Deadline**	for applying for 2004/5 is April 30, 2004
exchange, please contact Miss L. Cole. The	**deadline**	for investing in your 2005/06 ISA is 5
deliver communications solutions to tight	**deadlines**	, as a basis for specialisation in public
attitude to meeting national targets and	**deadlines**	and the creative and sensitive promotion
your order and you have complied with our	**deadlines**	a 50% refund can be paid to you within
ions Foreign Affairs Committee to propose a	**deadline**	. The suggestion comes as British officials
him proceed with his research but set a	**deadline**	for the delivery of the first outcome.

The trouble with concordances

- Too much data!
  - *deadline*: corpus has > 35,000 examples
  - *decide*: corpus has > 300,000 examples
- **You can't read them all…solution?**
- Starting point: Church & Hanks 1990
  - "Mutual Information" metric (MI) identifies pairs of words that co-occur with high probability
  - Leads to….

The solution: Word Sketches

- Automated one-page summary of most frequent word combinations
- Corpus is “shallow parsed”, software finds data for specific “grammatical relations”
  - e.g. find all verbs for which *deadline* is the object
- Initial goal: device for finding collocations
- Unintended consequence: primary tool for writing dictionaries (Kilgarriff & Rundell 2002)

Part of Word Sketch for *deadline*

- Second column ('object_of') shows the verbs that most often have *deadline* as their object
- Much quicker!

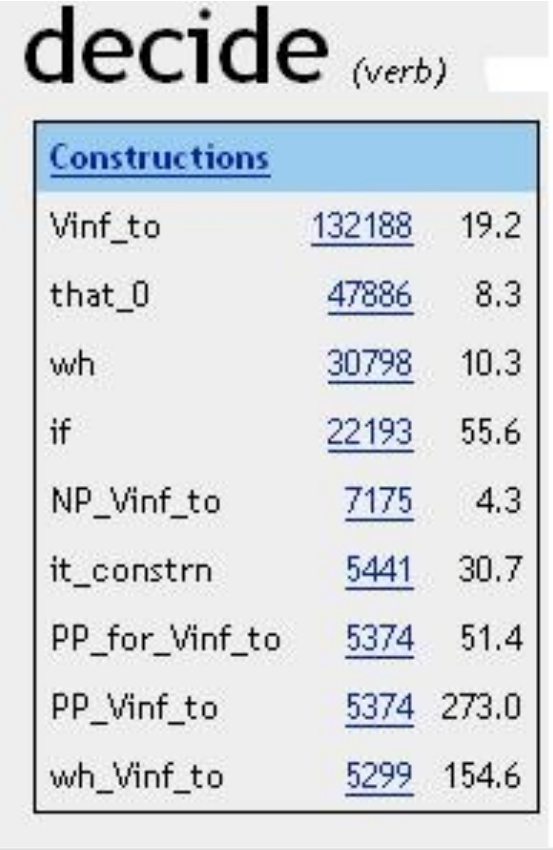
deadline *(noun)* LEXMCI freq = 3883

Constructions		
PP_Ving	1629	7.0
Vinf_to	1088	1.1
PP_for_Vinf_to	543	37.3

object_of	10912	2.7
meet	2258	7.03
set	1353	6.0
miss	805	7.17
extend	560	6.73
have	540	0.45
give	363	2.22
impose	177	5.99
agree	155	3.67
include	108	0.78
fix	86	4.46
face	85	3.37
publish	84	2.69
beat	80	4.17
follow	78	1.16
achieve	69	3.01

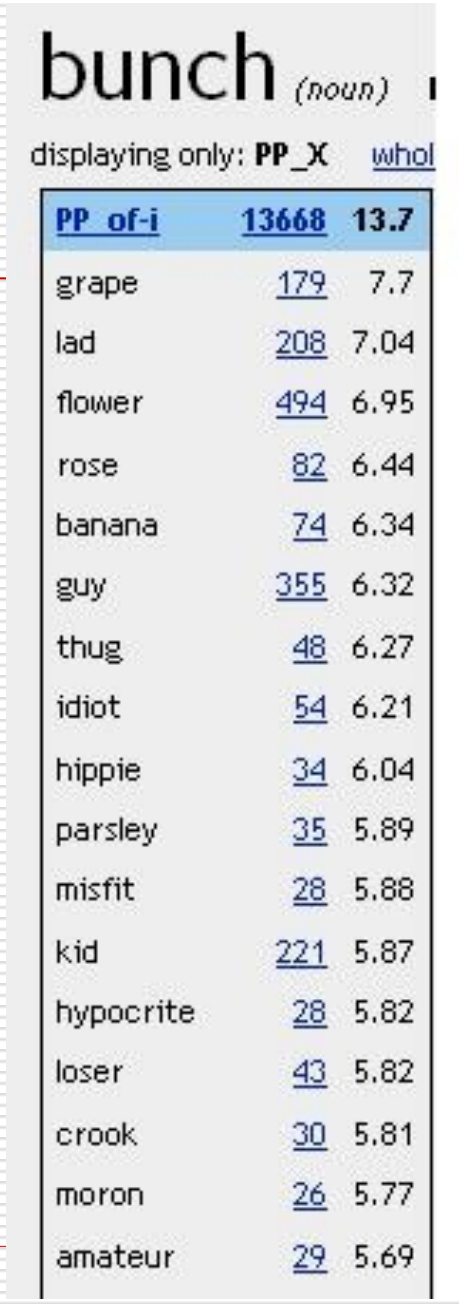
Part of Word Sketch for *decide*

- Which constructions go with *decide*?
- And which are the most frequent?
  1. with an infinitive (Vinf_to)
  2. with a that-clause (that_0)
  3. with a wh- word
  4. etc



Speaker attitude (implicatures): how we use *bunch*

- Word Sketch for ***+of***
- Shows most frequent collocates, ordered by saliency not frequency
  - *grapes, flowers, keys …*
  - *people, guys, kids …*
  - *losers, idiots, morons …*



bunch *(noun)*

displaying only: **PP_X** whol

PP_of-i	13668	13.7
grape	179	7.7
lad	208	7.04
flower	494	6.95
rose	82	6.44
banana	74	6.34
guy	355	6.32
thug	48	6.27
idiot	54	6.21
hippie	34	6.04
parsley	35	5.89
misfit	28	5.88
kid	221	5.87
hypocrite	28	5.82
loser	43	5.82
crook	30	5.81
moron	26	5.77
amateur	29	5.69

What corpora tell us about language (1)

- The word (on its own) is not a unit of meaning: meanings are constructed through **context**, through words in combination
- "So strong are the co-occurrence tendencies of words …that we must widen our horizons and expect the units of meaning to be much more extensive and varied than is seen in a single word". Sinclair 2004.

What corpora tell us about language (2)

- In language, anything is *possible*, but what matters is what is typical, normal, and recurrent
- "Although the number of *possible* combinations may in principle be limitless …the number of *probable* combinations …is rather limited" Hanks 2013.399.
- —> Knowing what to focus on in mass of data

Two revolutions in dictionaries (2) migration from print to digital

- A slow start, from early 1990s
  - dictionaries on CD-ROMs or handheld devices
  - changes are mostly cosmetic (“books in digital form”)
- Rapid acceleration, from about 2008
  - central role of the Web, rise of mobile devices
  - this time, effects are “external”: how information is produced, published, and used
  - completely new paradigm…still emerging

Consequences: redefining *dictionary*

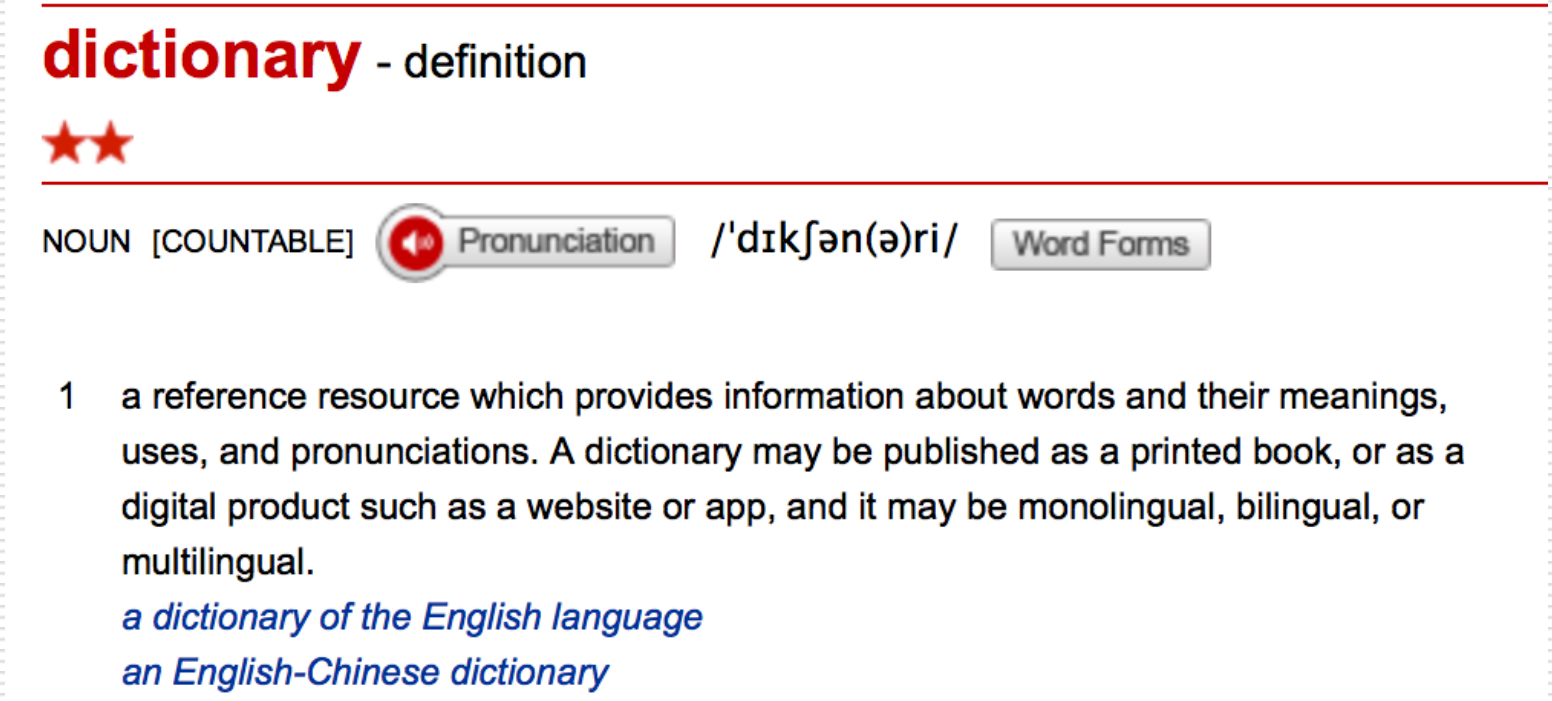
diction /ˈdɪkʃ(ə)n/ noun [U] **1** the way that you pronounce words, especially whether or not you speak or sing clearly **2** the choice of words used in a speech or piece of writing

dictionary /ˈdɪkʃən(ə)ri/ noun [C] ★★
1 a book that gives a list of words in alphabetical order and explains what they mean: *a dictionary of the English language* **1a.** a book that lists words in one language and gives translations in another: *a German-English dictionary*
2 a book about a particular subject that gives an alphabetical list of words, phrases, or names with information about them: *a dictionary of art/music*

dictum /ˈdɪktəm/ noun [C] an expression or statement that people often repeat because it says something interesting or wise about a subject

Definition of “dictionary” in 2007 (MED2, paper)

Redefining *dictionary*



dictionary - definition

★★

NOUN [COUNTABLE] Pronunciation /ˈdɪkʃən(ə)ri/ Word Forms

1 a reference resource which provides information about words and their meanings, uses, and pronunciations. A dictionary may be published as a printed book, or as a digital product such as a website or app, and it may be monolingual, bilingual, or multilingual.

a dictionary of the English language

an English-Chinese dictionary

Definition of “dictionary” in 2015 (MED online)

Redefining *dictionary* - again

- A one-stop reference and language-awareness resource
  - dictionary **and** thesaurus, monolingual **and** multilingual
  - other language resources: e.g. blog, grammar tips, resources for teachers, games, videos …
  - everything is linked: boundaries between dictionary and Web (corpus data, other Web data) more porous
  - plus “user-generated content” (UGC)

Staying up to date - not as obvious as it sounds

- Definitions of *meeting, marriage*, *dictionary*
- Definitions: what to do about *cassette, fax, floppy disk, video recorder…?*
- *phone*: what is the default reading (mobile? smart?)
  - add new entries for *dumbphone, feature phone, landline*

Staying up to date: what does *camera* mean?

- In 2002

- In 2015, almost always digital
- Most photos taken with phones or tablets, not dedicated cameras
- What to do?
  - add to definition: “either as part of a mobile device or as a separate item”

The new lexicography: who does the work?

- Lexicographers (but not as much as before)
- Machines: automating key tasks, e.g.
  - corpus creation (see WebBootCat)
  - term extraction (=finding headwords for a dictionary)
  - identifying significant combinations (syntax, collocation)
  - finding “good” example sentences (GDEX: Kilgarriff et al 2008)
  - identifying contextual preferences (e.g. “mainly found in journalism”)
- …and the general public: UGC, crowdsourcing

Does crowdsourcing have any value for dictionaries?

- Three subtypes
- The wiki model
  - collaborative, self-regulating, reflects the idea of "the wisdom of crowds"
- UGC (=user-generated content)
  - users share knowledge and expertise
- Crowdsourcing (strictly speaking)
  - very large-scale tasks achieved through mass participation: "many hands make light work"

Some examples

- Wiki model
  - Wikipedia: a great success. Can this model work for dictionaries?
- UGC
  - sharing expertise: e.g. “how-to” videos
  - having your say: e.g. Comments and conversations on blogs, news sites
  - Macmillan’s *Open Dictionary, the Urban Dictionary*

UGC: Macmillan's *Open Dictionary*
macmillandictionary.com/open-dictionary/

- >3000 items contributed by users
- Many are neologisms
  - contributes to updating main dictionary
- Good for "long tail" vocabulary, e.g.
  - Terminology, language of specialized domains
  - World Englishes
- Now fully integrated into main dictionary

Crowdsourcing

- *OED*: “reading programmes” to collect citations
  - early example of crowdsourcing (1857 —>), thousands of contributors
  - http://public.oed.com/history-of-the-oed/reading-programme/
- Lancaster/IBM spoken corpus, 1984
  - (partly) transcribed and annotated by Lancaster students as assignments

Involving users: pros and cons

- Wiki model: Wiktionary
  - less successful for lexical information (is anyone an "expert" on *decide* or *bright*?)
  - non-corpus-based, old-school lexicography
- UGC model
  - provides valuable data - but random (many common words missing from *Urban Dictionary*)
  - Collins' experiment: many submissions just made up!
  - http://www.macmillandictionaryblog.com/what-goes-in-the-dictionary-when-the-dictionary-is-online

Involving users: pros and cons

- Crowdsourced model
  - goal-oriented: often a clear objective
  - managed: the “crowd” aren’t experts - but experts do post-processing
  - great potential: e.g. Doug Higby’s work in Ghana
  - http://www.macmillandictionaryblog.com/what-goes-in-the-dictionary-when-the-dictionary-is-online
  - and his video at http://rapidwords.net: recommended

What next? A few speculations…

- Disappearance of dictionaries (as standalone objects)
  - "dictionary" becomes part of "search"
  - dictionary embedded in other applications: e-readers, news sites
  - lexical data underpins grammar checkers, text-remediation software, adaptive learning tools
- More automation (Rundell & Kilgarriff 2011)
  - specialised corpora
  - definitions? word sense disambiguation?
- Intelligent use of crowdsourcing.