



Learners' dictionaries: what next?

Michael Rundell, Lexicography MasterClass
UPF, 31st May 2007

Outline



- The publishing context: what's happening in the world of information
- Available resources
- Recent developments in learners' dictionaries
- Frequency data, and what to do with it
- Practical work (see handout): the frequency of word senses

Recent developments in information technology

- Fast, always-on web access
 - over 40% of households in OECD countries
- This affects:
 - How we buy holidays, theatre tickets, books, Christmas presents, music, software, TVs ...
 - How we find train times, see what's on at the cinema, check the weather, find how to get from A to B ...
 - How students learn: Internet, blogs, wikis, learning-management systems (e.g. Moodle) ...
 - **How we get information about language**

Trends: (1) unlimited information

- Answers.com: dictionary, thesaurus, translations in 20 languages, sign-language image, audio pronunciation
- Googlenews: 'Do people still say *Beam me up Scottie?*' 'Yes - most recent case was 5 hours ago.'
- Googleflight: which is commoner 'started raining' or 'started to rain'?
- Wikipedia
 - 1.8m articles in English,

Trends: (2) 'Good enough' is good enough

- Web data is adequate for most users in most situations
- *Nature* survey (2005)
 - 50 pairs of scientific articles from Wikipedia and Encyclopedia Britannica, blind review by experts
 - average errors per article: approx. 3 (EB), approx. 4 (Wikipedia)
- Convenience overrides accuracy

Trends: (3) customization, personalization

- End of the 'one-size-fits-all' model:
 - TV, radio, newspapers: e.g.
 - podcasts, video clips on mobile, 'on-demand' TV
 - personalized newspaper to print off, desktop news feeds etc
 - music: from album to personal selection
 - language learning: podcasts, iWriter etc.
- Consumer is in control, has choices

Trends: (4) Improved performance of standard software

Word 2007

1. I bought a pear of shoes yesterday.
2. I acknowledge that their may be exceptions to the rules
3. I gud to go to the store.
4. I've go to go.
5. I cannot see the mistakes you mentioned.
6. You'd be happy to no that our sales have increased.
7. You told me too go to the first street on the right, correct?
8. What did you tell me to do?
9. He told me that he wanted to leave the convention sew badly.
10. Let me know if you can chat some time today.
11. He was loosing too much time.
12. You cannot associate this account on more then one mailbox.
13. Do you eat breakfast everyday?
14. Where can I get a stand alone version of Microsoft Office Word 2007?
15. The require us to submit an application before Thursday.

Word 2003

1. I bought a pear of shoes yesterday.
2. I acknowledge that their may be exceptions to the rules
3. I uan to go to the store.
4. I've go to go.
5. I cannot see the mistakes you mentioned.
6. You'd be happy to no that our sales have increased.
7. You told me too go to the first street on the right, correct?
8. What did you tell me to do?
9. He told me that he wanted to leave the convention sew badly.
10. Let me know if you can chat some time today.
11. He was loosing too much time.
12. You cannot associate this account on more than one mailbox.
13. Do you eat breakfast everyday?
14. Where can I get a stand alone version of Microsoft Office Word 2007?
15. The require us to submit an application before Thursday.

And ...

- *Everything is free* (the default assumption)
 - Why would consumers pay to get something they can get for nothing?
- We (dictionary publishers) are no longer the 'gatekeepers'
- So: how should we respond?

Possible responses by publishers

- **Wrong**
 - pile in more information, regardless of type
 - leads to overload and clutter, loss of focus
- **Better**
 - figure out: what we can provide that learners *need*, and that isn't available elsewhere?
 - better, fuller language description
 - in learner-friendly form
 - clear focus: materials that meet known learners' needs

Available resources: (1) language data

- Massive corpora: data sparseness a thing of the past
 - general purpose corpora (like new 1 billion word Oxford English Corpus – OEC)
 - special-subject corpora: build it yourself with WebBootCat (www.sketchengine.co.uk)
- Sophisticated software
 - concordances
 - Word Sketches
 - Sketch difference, etc etc

Available resources: (2) new delivery platforms, unlimited space

- CD-ROM, DVD-ROM
- Online versions
- Portable devices
 - handheld dictionaries
 - handheld e-readers (e.g. Iliad: <http://www.irextechnologies.com/products/iliad>)
 - screen and keyboard, no onboard data but wireless link to Web: download whatever you want (novels, reference resources etc)

Know your user: user profiling

- Creating a user-profile
 - dictionary's language
 - monolingual, bilingual-unidirectional, bilingual-bidirectional
 - dictionary's size and scope
 - unabridged, standard, pocket
 - historical/scholarly, general-purpose, pedagogical, special-subject
 - user type, user skills
 - adult native speaker, child, learner
 - linguists and language professionals, levels of learner

Some recent developments: (1) more examples

- LDOCE4: 'Examples bank' provides additional new examples
- MED2: finds all examples of word A from dictionary entries for words B, C, D etc (see Supersearch → Example sentence search)
- Electronic editions only

Recent developments: (2) attention to metaphor (cf. Lakoff & Johnson 1980)

- MED's 60+ 'metaphor boxes'

Metaphor

When you put a lot of effort into doing something, it is like using a part of your body.

Does she have the backbone to stand up to them? Or will she just give in? ♦ You have to put your back into it. ♦ They only succeeded by using their political muscle. ♦ Put a bit more elbow grease into it. ♦ My heart's not really in it. ♦ I had to sweat my guts out to get it done in time. ♦ We must all put our shoulders to the wheel. ♦ Just try to put your best foot forward now. ♦ I've been keeping my nose to the grindstone. ♦ He was the kind of boss who liked to get his hands dirty.

Shows metaphorical links between vocabulary in same semantic field

Recent developments: (3) attention to collocation

- especially in MED & CALD: collocation boxes, drawing on information in Word Sketches

Collocation

Adjectives frequently used with **progress 1**

- considerable, great, rapid, remarkable, significant, slow, steady, substantial, tangible

Verbs frequently used with **progress 1** as the object

- assess, block, chart, check, evaluate, follow, halt, hamper, hinder, impede, monitor, obstruct, review, slow, track, watch

Recent developments: (4) using learner corpora

- Electronic collections of texts – produced by learners not by native-speakers
- Texts can be stratified according to
 - proficiency level (beginner, intermediate, advanced)
 - mother-tongue
 - task type (argumentative essay, letter writing, narrative, etc.)
 - environment (e.g. exam, untimed essay)

What learner corpora tell us

- Contrastive analysis shows
 - which words, phrases, structures regularly cause problems for learners
 - which words, phrases, structures are overused, which are underused (comparing with native-speaker texts)
- Allows us to focus on behaviour of
 - Learners from one mother-tongue background (e.g. Spanish learners), or
 - Learners from a wide range of mother-tongue backgrounds (Spanish, French, Chinese, Japanese ...)

What the data tells us

- Problems with accuracy, e.g.
 - *Although **the slavery** was abolished in the 19th century, black people still face racism in all parts of the world.*
 - *Some researchers **suggest to reformulate** the hypothesis in more general terms.*
 - *Artificial insemination has always been a controversial **problem**.*
 - *The Ministry of Education **claimed** that children should get more free time.*

Using frequency data

- Made possible by availability of large corpora
- Informs lexicographers' decisions, e.g.
 - inclusion of headwords: which ones?
 - inclusion of meanings, phrases: which ones, and in what order?
 - establishing a 'Defining Vocabulary' (e.g. 2000 common words)
 - describing syntax: which patterns
 - describing style and register (e.g. *mainly American*)

Aspects of frequency

- 'raw' frequency: occurrences of word-form
 - 'plays' occurs 3536 times in BNC
- lemmatized:
 - 'play-verb' (includes play, plays, playing, played) occurs 38,053 times in BNC
- normalized: a 'per million words' figure, allows comparison across different corpora
 - 'play-verb' occurs 380 times per million words
 - 'play-noun' occurs 82 times per million words

Aspects of frequency (continued)

- Range (or 'dispersion'): how many types of text a word appears in
 - *unfortunate*, *mucosa* have identical BNC frequency (1031 occurrences)
 - *unfortunate* appears in 648 texts, *mucosa* in 9
- Evenness of distribution
 - *goalkeeper*: 49 times per million in newspaper texts, less than 1 time per million in fictional texts

Explicit frequency information

- First attempts in LDOCE3 (1995), COBUILD2 (1995): see Kilgarriff 1997 (IJL 10:2)
- Two main aspects in current dictionaries:
 - establishing a 'core' vocabulary
 - the 'Oxford 3000', the 'red words' in MED, LDOCE
 - frequency 'banding'
 - 1-star, 2-star, 3-star words in MED, similar system in COBUILD

Does 'most frequent' equate with 'most useful to learn'? Yes, because...

- Frequency correlates with range
 - frequent words appear in all text-types
 - infrequent words are usually restricted to few text-types
- Frequency correlates with complexity
 - number of meanings
 - number of syntactic patterns
 - number of collocational and phraseological patterns