LEXICOGRAPHY 2

Claire Bowern and Arienne Dwyer

CoLang: Institute for Collaborative Language Research
June 19-22, 2012

Ways to order information

- Frequency [most common orders first]
- Alphabetical order
- Cultural considerations (language-internal hierarchies)
- Prestige
- Meaning type [concrete meaning before metaphorical meaning]

Today:

- Choices to make in designing a dictionary.
- What goes into a dictionary?

SOME CHOICES

Choice: Which Language(s)? Which direction?

- English – Language or Language – English?
- Language – English:
  - Often easier for researchers to compile.
  - Can be used for interlinearizing, translation from Language to English.
  - *Assumed* (only dictionary model in SIL products)

- English – Language:
  - Often more useful for heritage languages, endangered languages with lots of language learners, non-fluent speakers
  - monolingual, bilingual, or multilingual?

Example

Could you learn French from this? Why or why not?

what [wɔt]. I. *a.* **1.** (*Relative*) (Ce, la chose) que, **He took away from me w. little I had left,** il m'a le peu qui me restait. **He traded with w. capit had,** il faisait le commerce avec ce qu'il posséda capital. **2.** (*Interrogative*) Quel, *f.* quelle. **W. ti it?** quelle heure est-il? **Tell me w. time it is,** moi l'heure qu'il est. **W. right has he to give or** de quel droit donne-t-il des ordres? **W. good is** à quoi cela est-il bon? **W. news?** quoi de nouv **W. day of the month is it?** le combien sommes-n **3.** (*Exclamatory*) **W. an idea!** quelle idée! **fool he is!** qu'il est bête! **W. silly fools we ha been!** comme nous avons tous été bêtes! **W. a people!** que de gens! II. **what,** *pron.* **1.** (*Rel* = *that which*) Ce qui, ce que. **W. is done cann undone,** ce qui est fait est fait. **W. I like is n** ce que j'aime c'est la musique. **W. is most remar is that . . .,** ce qu'il y a de plus remarquable que. . . . **He had a key, and w. is more, he has it** il avait une clef, et qui plus est, il l'a encore. **T w. it is all about,** voici ce dont il s'agit. **Come w.**

Multilingual dictionary

Choice: Media/format

- Print?
  - publisher/cost
  - …
- Web?
  - host?
  - …
- Both?
  - Can you easily produce both from your underlying data?

Choice: Audience?

- Who is going to use the dictionary?
- What information will they need (or not need)?
- What is their familiarity with the language?
- What is their familiarity with reading, alphabetization, etc?
- Do you need to cater to multiple audiences? If so, who's primary?

WHAT GOES INTO A DICTIONARY?

Information in a dictionary entry

- **Lexical/Semantic** – about the word and its meaning, how the word relates to other words in the language
- **Phonological/Phonetic** – pronunciation information
- **Grammatical** – paradigm forms, suppletion, gender/class information, etc
- **Social** – usage contexts, register, dialect, etc
- **Encyclopedic** – information about the item in the real world (e.g. how it's made, where it lives, etc)
- **Historical** – etymology of the word, is it a loan, etc
- **Sources** – where were the words recorded from?

Things that can be included in a dictionary

- the **headword** sorted by semantic field or alphabetical order – many database programs allow for variable sorting
- **parts of speech** (be very careful about creating parts of speech labels on the basis of the gloss of the word. This is very misleading.)
- phonological **irregularities**; **pronunciation** if unpredictable from the standard orthography
- a single-word **gloss** for interlinearization
- a more detailed **definition**
- morphological **paradigmatic** information, such as gender, class or conjugation.
- any notes on comments your consultant made about the semantics of an item
- **encyclopaedic** information, e.g., information on an item's usage or ethnographic information about the cultural importance of the item. This could be accompanied by a picture
- synonyms and antonyms, hyponyms or other information about how the word relates to other items in the lexicon
- **example sentences** illustrating usage, with translations

Dictionary entries (2)

- the source of the word (e.g., if a borrowing; etymology, if known)[2]
- **semantic field(s)** of the item
- any **usage** information – e.g., if it is slang or taboo
- a **reversal** field (so that you can compile an English–Language finderlist from your data)
- **sound clip(s)**, and example sentences
- derived words
- the source of the information (e.g., who told you the word)
- questions for further research.

MORE CHOICES

Headwords

- What should the citation form be?
  - Easy choice in a language without much morphology (where the words don't change much)
  - Harder choice if some word parts don't exist on their own.
  - eg: Bardi verbs: e.g. -jarrala- 'run':
    - iyarralan 'he/she is running'
    - iny**jarrala**gal 'he/she ran'
    - nganyjarralagal 'I ran'
    - arra oolarrala 'he/she isn't running'
    - irrjarrala 'they are running'

Headwords

- Solution for Bardi: use the third person singular past form, which always shows the full root.
- (see also http://sydney.edu.au/arts/linguistics/research/wagiman/dict/dict.html for similar problem in Wagiman)

Headwords

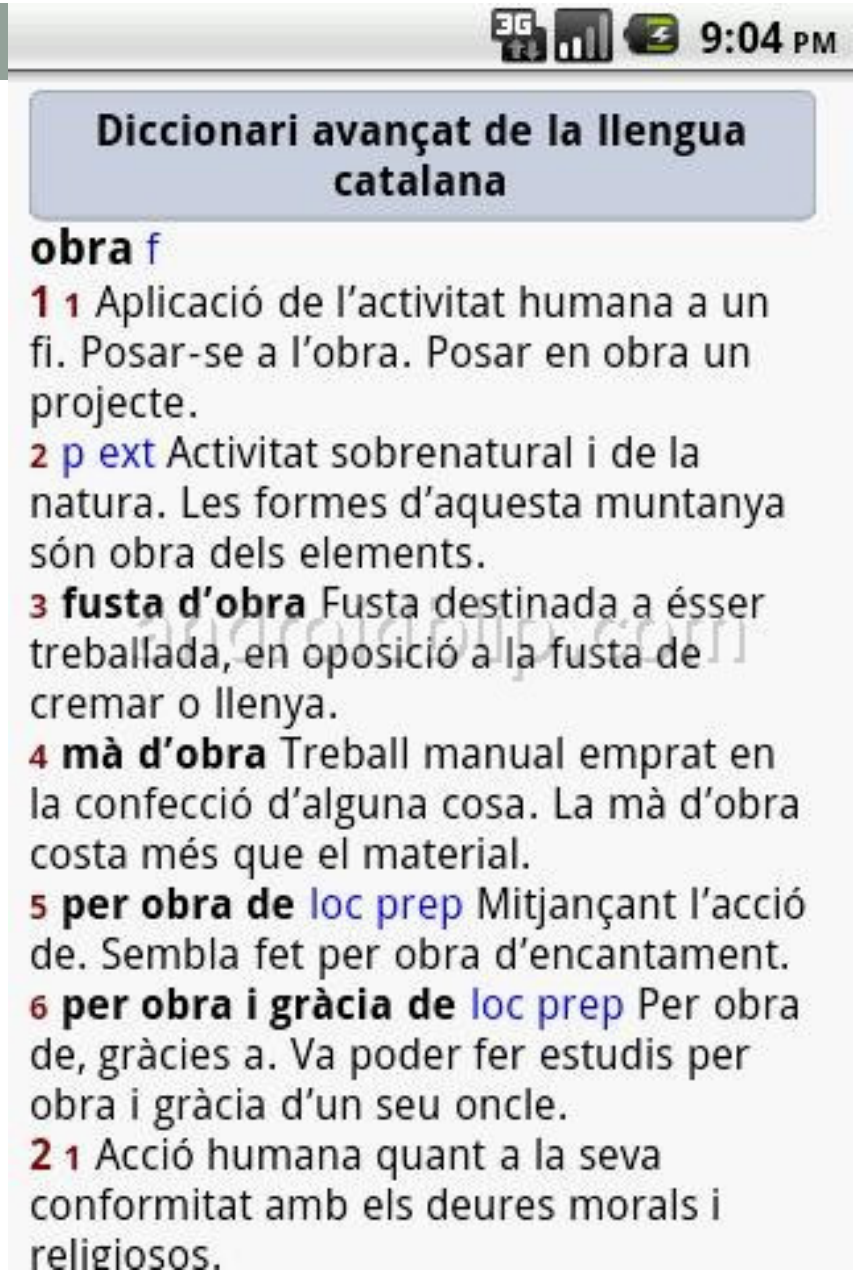
- How different do words have to be before they count as separate headwords?
- Discuss: how many headwords?
  - Bank
  - Count
  - On
  - Field

Headwords

- Morphologically related forms?
- Often (e.g. in corporate dictionaries):
  - Inflectional morphology (e.g. singular vs plural) not separate headwords [and not listed] unless
    - Very different semantics [brother ~ brethren]
    - Irregular forms [child ~ children; bring ~ brought]
  - Derivational morphology (e.g. augmentatives, diminutives, etc) not listed unless
    - Not productive
    - Accompanied by meaning change
  - Phrasal compounds
    - Usually listed (come on, come off, come over, etc)

Examples

- Headwords vs subentries



- Headwords vs 'return all items':
- http://chamacoco.swarthmore.edu/?fields=all&q=dog

Headwords

- What spelling system to use?
  - Practical orthography?
  - IPA?
  - Established orthography?
- Guiding principle: what will be most useful to dictionary users.

Ordering of entries

- Alphabetical
  - Easiest for large dictionaries but can be hard for those new to literacy; adult readers often find alphabetical order very unintuitive. (Can be ameliorated by printing the order across the top or bottom of the page.)
- Semantic field
  - Great for browsing,
  - Good for learners
  - But can be hard to look up a word (semantic fields are arbitrary; e.g. ‘eat’ under **body funtion, food, verb, home,** etc?

- (web/e-dictionaries often allow fuzzy searching, making ordering less important)
  - Example: Mi'kmaq online dictionary: http://www.mikmaqonline.org/
  - Example 2: Yiddish: http://www.cs.uky.edu/~raphael/yiddish/dictionary.cgi
- Root-based sorting (word roots + derivatives)

Illustrating entries

- Adding pictures
- Where from? (copyright issues, cf. wikipedia creative commons license)
- How to pick which entries to illustrate?
  - Cultural items?
  - Flora/fauna?
  - Anything that you have pictures for? (sourcing illustrations can be a good way to get others involved in the dictionary)

Dictionary Scope?

- How big a dictionary do you plan?
  - Everything available
  - What we can do before the money runs out
  - First draft in 6 months with what we have by then, second draft in 12 months
  - Launch web site at 500 entries, then continue adding.
  - Start with plants and animals book, then a series of leaflets on different semantic domains, then combine and expand into dictionary
  - Compile all words and glosses, then add definitions, examples, etc as possible
  - …

Dictionary scope

- How much grammatical information to include in an entry?
  - Everything available? (nice to be comprehensive, but might overwhelm learners)
  - Include paradigms? (takes up space, not needed for fluent speakers, but helpful for learners)

What other sorts of information to include?

- Dialectal representation? One dialect or several?
- Separating dialectal information:
  - http://www.pledari.ch/mypledari/index2.php (Surmiran)
- Including words from all dialects as headwords:
  - http://203.122.249.186/Lexicons/Walmajarri/Walmajarri%20Lexicon/lexicon/mainintro.htm (Walmajarri)

Words to include/leave out?

- Words to include/leave out?
  - Include everything?
  - Swear words
  - Taboo words
  - words used only by some sections of the community [cf intellectual property rights discussed by Marsha and Alice on Monday]
  - Loanwords (when does a loan become native?)
  - bound forms (word pieces)
  - productively formed words [e.g. compounds]
  - Idioms
  - Personal names? Place names? (in Appendix instead?)

EXAMPLE/PRACTICUM

Task 1

- Work in small groups, working with someone who has a dictionary project in mind.
- Go through some of the considerations we've talked about today and make notes on things you need to think about for each topic.
- We will discuss questions/comments at the end of the session.

Task 2

- With a partner, have a look at the excerpt from the Bardi dictionary in tlTerm.
- Identify the types of information that are in the entries.
- Identify some of the problems with the current dictionary.
  - NOTE! This is a draft dictionary where I haven't yet done much editing to make it useful for people who aren't me. I have a long list of things that need to be done! My feelings won't be hurt if you're very critical ☺
- Identify some of the things you like about the dictionary.

Homework!

- Revisit the list of ten words from last night's homework, equipped with your new set of questions about dictionary formats, content, etc.
- Plan a full entry out of one item.