Writing a dictionary

Fieldling - Villejuif 07/09/2020

QUINT Nicolas – LLACAN-UMR 8135 (CNRS/INALCO/USPC)

nicolas.quint@cnrs.fr
Intro. Lexicography

1. Classifying words and meanings
   - Limits and interest
   - Automatic translators
   - Theories of meaning: still temptative
   - For the time being, dictionary = best approximation

2. Study of lexicon: role in descriptive linguistics
   - One of the three pilars of standard ‘boasian’ descriptions (alongside grammar + texts)
   - As comprehensive as possible (many spoken languages = ca. 10,000 different words in common speech = f(demographics / social diversity)
Intro. Lexicography

- Bilingual dictionaries = most common for descriptions (usually described (e.g. Azérables Croissant, Koalib...) > describing (English, French). The reverse is a much harder work (question of the standard). Better avoid more than two languages.
Part I. Types of lexical works

1. Glossaries / vocabularies / wordlist
   - Low lexical information
   - Binary translations (one A term = one B term)
   - Thematic vocabularies = interesting for field inquiries (fish names, agricultural practices...) – see also surveys / comparisons

2. Dictionary
   - Richer lexical information (expected)
   - \( N \text{(meanings)} > 1 \)
   - \( N \text{(words)} \)
   - Examples of usage in A (translated into B)
   - Idioms
Part II. Structure of a lexical work - 1. The entry

- 1. What is a basic form?
  - Which form in the paradigm
    - e.g. infinitive (non finite) for verbs in Romance languages: *chanter* (French) / *cantar* (Occitan / Spanish / Portuguese) ‘sing’
    - vs. 3.M perfective (finite) for verbs in Arabic /ˈkataba/ ‘he wrote’ (the shortest form in the whole paradigm) [see also Latin]
  - Lexical families
    - e.g. standard Occitan *capbord* /ˈkaˈburt/ ‘crazy’ > entry *cap* /ˈkap/ ‘head’
  - Dialectal variants
    - e.g. Occitan *somicar* [ʂumiˈka] ‘whine’ (local variety) > entry *gemicar* /ˈchemicar/ (standard dictionary < *gemir* ‘moan’) : not easy to find/choose
Part II. Structure of a lexical work - 1. The entry

2. How to order the entries
   - Alphabetic order
     The most common (including syllabaries)
   - Thematic
     Mostly for glossaries: fish / food / domestic animals...
   - Types of signs
     Chinese (no link with the pronunciation)
   - Root
     Arabic kátabat ‘she wrote’, máktab ‘office’, kitáab ‘book’, káatib ‘writer’ > entry = KTB - كتب
     Semi-root: see above Alibert for Occitan (capbord / cap)
3. Homophones/homographs

- Same form + different parts of speech (pos)
  e.g. English tear /ˈtɛə/ (v.) vs. tear /ˈtɪə/ (n.)
  Individual or separate entries? > controversy
  Differences = pronunciation + morphology
  e.g. CVC pa [pæ] ‘for’ (PREP) + ‘so that’ (COMPL)

- Same form + same pos
  e.g. French pain /pɛ̃/ ‘bread’ (< latin PANEM)
  vs. pin /pɛ̃/ ‘pine’ (< latin PINUM)
  Two separate entries in all French dictionaries because: different written forms + etymologies
  But what about a first description (no diachronic data available): a ‘long bread’ could derive its form from an analogy with the tree...
Part II. Structure of a lexical work - 1. The entry

4. How many entries?
A high number of entries is not necessarily the ultimate aim > quality = f(detail/depth of the lexical description of each entry), see contrast wordlist vs. dictionary.

5. Spelling and phonetics
Standard spelling if existing
Always doubled with phonetic transcription (tone, stress...)

6. Can everything be translated?
e.g. souletin Basque : emaztiarentako ‘for the woman’ / French ‘pour la femme’. In a French-Basque dictionary, ‘pour’ should be present BUT explained (not translated)
e.g. Portuguese ‘a’ vs. Capeverdean ‘ZERO’
### Part II. Structure of a lexical work - 1. The entry

#### 7. Wordness and its limits

<table>
<thead>
<tr>
<th>Koalib</th>
<th>French</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>kàkró</td>
<td>poule</td>
<td>hen</td>
</tr>
<tr>
<td>ṭàkró ‘hen.DIM’</td>
<td>poussin</td>
<td>chick</td>
</tr>
</tbody>
</table>

Koalib: one or two entries? (derivation)
French: two entries (frozen derivation)
English: two entries (two roots)
Number of entries depends on both languages used in the dictionary.
Part II. Structure of a lexical work - 2. Morphological information

- **1. Part of speech (pos)**
  Compulsory

- **2. Other informations = f(language)**
  - Verb extensions
    - e.g. Koalib *nyìmì* ‘steal (tr.)’ [basic verb] > *nyímìccí* ‘steal at s.o.’s benefit’, *nyímètè* ‘steal at s.o.’s expenses’ [extensions]
  - Plurals / case / gender inflections
    - At least irregular forms
      - e.g. Arabic *kitáab* (SG) ‘book’ > *kútub* (PL)
      - e.g. Koalib *kwór* ‘man.S’ > *kwòoró* ‘man.O’
      - e.g. French *niais* ['njɛ] ‘dumb’ > *niaise* ['njɛz] ; *frais* ‘cool’ ['frɛ] > fraîche ['frɛʃ] ; *épais* [e'pɛ] > épaisse [e'pɛs]
Part II. Structure of a lexical work - 2. Morphological information

- **Doubled forms**
  - e.g. Koalib *kwâó lèerêa dêr* ‘he will shoot at the rock and hit it *(fully)*’
  - vs. *kwâòccê ñwèerêa dêr-dêr* ‘he will shoot at the rocks and hit *them* *(fully)*’

- **Intensifiers**
  - e.g. Koalib *pèeté* ‘be white’ > *pèeté péppèr* ‘be *completely* white’
  - *úñnì tìttìt* ‘be *completely* black’
  - *őorè ccèccèl* ‘be *completely* red’
  - a.s.o.

3. **Limits between dictionary and grammar**

   Fuzzy however you can’t produce a reliable dictionary without a proper analysis of its grammar (both are linked)
Part II. Structure of a lexical work – 3. Meanings

General

1. In search of the meaning unit
   Hard to grasp and therefore hard to distinguish.
   Meaning ≠ discrete (context, culture...)

2. Translation
   Lack of direct equivalents:
   Pierre *a traversé* la rivière *à la nage* (French)
   Peter *swam across* the river (Engl.)
   Pedro *cruzó* el río *nadando* (Sp.)

3. Constructions
   Good to distinguish at least one meaning / construction (e.g.
   French *manger* ‘eat (intr.)’ = meaning 1, ‘eat (tr.)’ = meaning 2
Part II. Structure of a lexical work - 4. Examples

Preferrably spontaneous
Highly desirable for all non-nominal items (not necessary in many cultures for nouns such as ‘elephant’) and for each meaning
Should be translated
Part II. Structure of a lexical work - 5. Idioms and collocations

General

1. Where should they be (in which entry)?
   Engl. ‘all that **glitter**s is not **gold**’
   French. ‘tout ce qui **brille** n’est pas de l’**or**’

2. Where should they be in a given entry?
   Ideally, to complete each meaning they are associated with
   For reduced teams: better to add them at the bottom of the entry, after the different meanings in a specific section
General

13. Semantic universals
Some body parts are clearly more easily associated with many idioms: ‘head / hand/ foot / eye...’.
Relative importance can vary f(cultures / languages). E.g. Koalib: nţà ‘head’ (120), tòkwɔ̀r ‘heart’ (81 idioms in the dictionary), lèe ‘eye’ (80), tée ‘hand/arm’ (46), láará ‘foot/leg’ (36), kááré ‘belly’ (27)

4. Use of specific abbreviations
e.g. ‘ça fait [X tps] que...’ = ‘it’s been [X time] since...’

5. Examples
Can be profitably used for idioms / meanings too...
Part III. Collecting the data

- **Different sources**
  - Immersion / participant observation
    - Keeping track of all that is (spontaneously) said > notebook at hand (if possible record > falsifiability + liability)
  - The corpus
    - Checking all terms that appear in the texts / dialogues
  - Thematic sessions
    - Fish names, cooking, music, animals (see ‘camel’)

General
Part IV. Making a lexical database

Different softwares

- Easy-to-use (Toolbox, Flex), professional...
  - Good to use a flexible tool (structure can change...)
  - You can also device your own DB (if good at computers)

- Basic principle
  - ONE TYPE OF INFORMATION = ONE FIELD IN THE DATABASE
  - The more distinctions you make the better:
    pos/number/gender/construction type/meaning....
    > retrieve info easily

- Not all fields should be published
  - E.g. ‘I don’t know, doubts, morphological/phonological (...) notes...
Part IV. Making a lexical database

- Limits to the descriptivist / fieldlinguist’s approach
  
  Funding + Time + Working force: one or a few individuals cannot build a dictionary comparable with what professional teams do for official / endowed languages
Conclusion

- A dictionary never ends
  New items / meanings / constructions / idioms always appear
- Value of a systematic work
  Compiling and scanning big corpora allow to build a statistically robust lexical database: encompass most of everyday language (tested with daily practice...)

General