Lexical Tools for the Documentation of Endangered Languages: Requirements Analysis Checklist

RFC 1.0

Dafydd Gibbon (U Bielefeld, Germany)
Bruce Connell (U Yorktown, Canada)
Firmin Ahoua (U Cocody, Abidjan, Côte d'Ivoire)

DOBES Technical Report *4* (Ega)
(Status: RFC draft, January 2001 — printed January 14, 2001)

Contents

Objectives

The goal of this Technical Report is to provide an initial framework for the DOBES Lexicon Working Group[1]which was set up at the DOBES Hannover Workshop, 12-13 January 2001, in specifying lexical construction and access tools for use in the efficient documentation of endangered languages. In order to fulfil this goal, the properties of such a lexical specification are formulated in the form of a systematic check–list type of questionnaire. Please send completed questionnaires (any common format, paper or electronic) to:

Prof. Dr. Dafydd Gibbon
Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Postfach 100131
D-33501 Bielefeld

At the DOBES Hannover Workshop, part of the specifications for annotation and encoding in the documentation of endangered languages were defined. These are prerequisites for the definition of a corpus-based lexicon, as the lexicon must be compatible with empirical corpus criteria. It may also be considered desirable to have a non–corpus–based lexicon be compatible with corpus annotations.

However, the questionnaire is not intended to pertain to the needs of your own project alone, but also to the needs of yourselves and others as potential linguistic, ethnographic and other users of lexical information on endangered languages, whether for linguistic analysis or other purposes.

In this document, technical terms are only explained by straightforward examples in order to avoid overloading the questions with explanations. If definitions are felt to be needed, they can be found in [Gibbon 2001].

Please give detailed descriptive answers as far as possible, not just yes/no answers.

[1]Various contributions to the present checklist were made explicitly or implicitly by all participants at the DOBES Hannover meeting.

1 Types of 'lexicon'

The word *lexicon* is used in a very general sense throughout this questionnaire. Please clarify your own usage:

1. How many types of 'lexicon' do you use?
    - paper book
    - electronic book form
    - database
    - other
2. Which other types would you like to use?
3. What are your criteria for defining the following (perhaps including different subtypes)?
    - wordlist
    - lexicon
    - dictionary
    - hyperlexicon
    - concordance
    - lexical database
    - other
4. Comments

2 Techniques for collating lexicon material

There are many ways of putting together the basic materials from which lexica are made. Please outline your own techniques:

1. What kind of and which sources of materials for lexica do you use?
   - generally known wordlists
   - your own wordlists
   - other lexica of various kinds
   - terminology domain studies
   - extraction from corpora
   - other
2. What techniques do you currently use for organising lexicon material?
   - paper, card index
   - word processing software (WordPerfect, Word, StarOffice, ...)
   - database or spreadsheet software (Shoebox, Access, Lotus, Excel, StarOffice, FileMaker, ...)
   - other
3. For each of the types of lexical material organising software that you use, esp. the software,
   - What are the advantages of the system for your current purposes?
   - What are the disadvantages of the system for your current purposes?
   - What would you like to do with it that you can't do now?
4. Is only one person concerned with collating material for a given lexicon, or do several work on one lexicon simultaneously?
5. Comments

3 Techniques for making the finished product

The end user of your dictionary may require a paper document, or some form of lexicon on another medium. Please describe the techniques you use in order to produce your output:

1. Do you use book production techniques, and if so, which?
   - word processor
   - PageMaker type software
   - automatic generation from a lexical database
   - other
2. Do provide a Database Management System (DBMS) with a user interface, and if so, which DBMS?
   - PC or Mac based DBMS (Shoebox, Access, Lotus, StarOffice, Filemaker, Oracle, ...)
   - web-based DBMS (Java, CGI, ...)
   - custom DBMS
   - other
3. For each of the types of lexical materials oranisation that you use, esp. the software,
   - What are the advantages of the system for your current purposes?
   - What are the disadvantages of the system for your current purposes?
   - What would you like to do with it that you can't do now?
4. Comments

4 Information in the lexicon (1)

A key issue in specifying the lexicon is analogous to the definition of annotation types and encodings. Please note the kinds of macrostructure and microstructure you work with:

1. Macrostructure 1: Which kinds of lexical entry or headword type do you currently handle? E.g. morph/morpheme, word (simplex/derived/compound word), phrase (idiom, fixed expression, ...
2. Macrostructure 2: How do you handle the multilingual aspect of your lexicon E.g. definitions in the indigenous language, translation dictionaries, ...
3. Microstructure 1: Which kinds of linguistic lexical information do you currently handle for each lexical entry? E.g. orthography, phonemic transcription, fine phonetic transcription, prosody, morphemic decomposition, Lieb/Drude Advanced Glossing ([Lieb & Drude 2001]), Dwyer tier grouping & optionality specifications ([Dwyer 2001]), polysemy of different kinds, full definitions, native definitions, sufficient morphological information to be able to define full paradigms, ...
4. Microstructure 2: Which kinds of other linguistic, encyclopaedic, ethnographic, non-linguistic information do you currently handle for each lexical entry? E.g. etymology, dialect variants, stylistic variants, terminological or other definitions, illustrative contexts, cross-references to other entries, ...
5. Microstructure 3: Which kinds of media information would you want to handle for each lexical entry (audio, photo, graphics, video clip, ...)
6. Microstructure 4: Which kinds of housekeeping information do you currently handle for each lexical entry? E.g. date of creation, date(s) of modification, responsible lexicographer, actual lexicographer, source(s) of information, ...
7. Comments

5 Information in the lexicon (2)

Specifications for lexicon contents are changing as requirements on corpus archiving, language documentation and linguistic analysis change. Please specify what additional kinds of information, over and above your present practice, you would like to see in a lexicon for endangered languages:

1. Which macrostructure units don't you handle yet but would like to?
2. Which kinds of microstructure units don't you handle yet but would like to?
3. Which of the tier types specified for Annotation and Encoding at the DOBES January workshop would you need for your lexicon?
4. Which of the Lieb/Drude Advanced Glossing tiers would you use?
5. Would you want to apply the Dwyer tier grouping & optionality specifications to the lexicon? If so, give details.
6. Are you familiar with Dafydd Gibbon's HyprLex? If so, describe which aspects of its functionality you consider useful for linguistic documentation.
7. Are you familiar with Steven Bird's Hyperlex? If so, describe which aspects of its functionality you consider useful for linguistic documentation.
8. Comments

6 Lexical metadata

A lexicon is in one sense metadata about corpora; in another it is itself a document which requires characterisation. Please describe the kinds of information you use in order to identify and describe your lexicon, the information it contains, and its uses.

1. How do you currently document your lexicon?
2. Which lexical metadata levels do you currently use, and which kinds of lexical metadata do you use at each level?
    - Metadata pertaining to the whole lexicon? E.g. dates of production, sources, lexicographers, sources, media, ...
    - Macrostructural metadata pertaining to each type of entry contained in the lexicon? E.g. characterisations of words, fixed expressions, ...
    - Microstructural metadata pertaining to each type of lexical information associated with entries? E.g. explanations of fields in a lexical database, ...
    - Metadata pertaining to each lexical entry? E.g. when and where found, ...
    - Metadata pertaining to each item of information for each entry? E.g. when and where entered, ...
3. Which kinds of linguistic generalisation for reference in lexical entries do you use? E.g. thesaurus domains, sketch grammar, sketch morphology, sketch phonology, ...
4. How much of the corpus metadata discussed at the DOBES January workshop would you want to see applied to the documentation of a lexicon?
5. How do you see the relation between a lexicon and a corpus?
6. Who uses your lexicon?
7. Comments

7 Search functionality

The main point of making a lexicon is to support search for information about lexical entries. Please outline the kinds of search that you currently use:

1. Paper lexica:
    - Search for meanings by looking up lexical forms (semasiological organisation)
    - Search for lexical forms by looking up meanings (onomasiological organisation, thesaurus)
    - other lookup criteria (roots, morphemes, orthography, syntactic category)
    - other response criteria (any microstructure items)
    - concordance (search for occurrences in corpus by looking up lexical forms)
    - other
2. Electronic lexica:
    - search for meanings by looking up forms (semasiological organisation)
    - search for forms by looking up meanings (onomasiological organisation, thesaurus)
    - other lookup criteria (roots, morphemes, orthography, syntactic category)
    - other response criteria (any microstructure items)
    - concordance (search for occurrences in corpus by looking up lexical forms) with textual, audio, graphic, video output, ...
    - other
3. What other search tasks would you like to be able to perform?
4. Comments

8 Lexical support tools

1. Which lexicon or lexical database formats do you currently have to convert?
2. Which lexicon or lexical database formats would you like to be able to convert?
3. Do you currently have corpus or lexicon statistics in your lexicon?
4. Would find corpus or lexicon statistics in your lexicon useful?
5. Would you find a concordance tool useful (cf. [Gibbon & al. 2001])?
6. Would you find a hyperlexicon production tool useful? I.e. generation of a lexicon in hypertext format for quick computer lookup of cross-references.
7. Would you find a tool useful which automatically generates additional corpus annotation tiers from information in the lexicon about lexical entries?
8. Comments

9 General comments and recommendations

Are there any further kinds of specification which you have not found in the checklist formulated in this document?

What would you recommend as a minimum but flexible specification for a lexicon for documenting endangered languages? Please bear in mind that this should include lexical metadata, and that some kinds of information cannot be reconstructed after the language has died out.

What would be a minimum lexical toolset for making and accessing a lexicon in the context of documenting endangered languages?

References

[Dwyer 2001] Dwyer, Arienne (2001). DOBES linguistic markup scheme: Towards a Minimal Annotation Standard for Encoding Linguistic Information. Universität Mainz: DOBES Technical Report *3*

[Gibbon 2001] Gibbon, Dafydd (2001). On lexical objects and their properties. A contribution to the 'MetaLex' requirements specification for spoken language lexicon documentation. Universität Bielefeld: DOBES Technical Report *2*

[Gibbon & al. 2001] Gibbon, Dafydd (2001). Preliminary Specification, Design and Proof–of–Concept Implementation of a Portable Audio Concordance (PAC). Universität Bielefeld: DOBES Technical Report *4*

[Lieb & Drude 2001] Lieb, Hans–Heinrich & Sebastian Drude (2001). Advanced Glossing. Freie Universität Berlin: DOBES Technical Report *1*