Computational lexicography: a training programme for language documentation in West Africa

Dafydd Gibbon[1] *and Nadine Borchardt*[2]

[1]*Universität Bielefeld,* [2]*Humboldt-Universität Berlin*

1 Training lexicographers for language documentation

The goal of the present contribution is to provide an overview and a practical foundation for teaching the lexicographic aspects of language documentation for local languages. Perhaps paradoxically, most of the contribution will be somewhat abstract: for reasons of space, the choice is either to be highly selective with lots of examples (the approach usually taken), or to be comprehensive and highly structured (though intermediate approaches would also have been possible). Since systematic overviews do not exist yet for modern lexicography, we have opted for the second approach.

The starting point for the contribution was the need to provide teaching materials for conducting practical training courses on lexicography in language documentation at the Université de Cocody, Abidjan, Côte d'Ivoire, and the University of Uyo, Akwa Ibom State, Nigeria.[1] In each case, students have B.A. level training in all areas of linguistics, and consequently, the courses can concentrate on the lexicographic subdiscipline of lexicography in language documentation.

The lexicographic application domains include lexicography for field linguistics, language teaching and speech technology, three fields which seem worlds apart, but which experience shows to haveat least some similar basic lexicographic requirements. The specific task which we address is the training of local lexicographers for local language documentation activities.

Based on experience of lexicography in field linguistics, language teaching and speech technology (particularly text-to-speech synthesis), we clarify our basic terminology as follows:

1. We take *lexicography* to be the scientific and technological discipline concerned with the theory and practice of *constructing* user-oriented documents such as dictionaries, lexicons, encyclopaedias, glossaries, terminology manuals and lexical databases for manual and computational consultation.
2. We take *lexicology* to be the theory and practice of selection and description of the *content* of any of these document types, consequently to be at the same time a foundational discipline for and a component of lexicography. Lexicology is often restricted to "lexis", i.e. lexical semantics and pragmatics. Our usage differs from this, for reasons of consistency, and includes all types of lexical information which figure in the microstructure of lexical entries, including syntactic, morphological and phonological information; from this point of view, phonology is just as much a subdomain of lexicology as is lexical semantics.

In this sense, lexicology isa part of linguistics which is both a foundational science for and a part of lexicography. We presuppose knowledge of lexicology in the course, and concentrate on operational aspects of lexicography.

Our definitions of lexicography and lexicology are deliberately broad. On the one hand, the breadth of definition is chosen in order to avoid unfruitful lapses into terminological conflicts within linguistic lexicology, and arguments on delimiting, say, dictionaries from encyclopaedias, lexicons from terminological dictionaries or lexical databases from speech technology system components. On the other hand, the definition is also kept broad in order to underline the common practical, operational criteria and procedures which are required by all of these activities.

From the point of view of course material construction, existing introductory or overview information on modern computationally supported lexicography is less than satisfactory, both in

1 We are grateful to the *Deutscher Akademischer Austauschdienst* (DAAD), the German Academic Exchange Service, for funding the project "M.A. curriculum for computational langauge documentation", 2002-2006.

terms of content and in terms of availability. There are many studies of the structure of the lexicon from a theoretical viewpoint (mainly in theoretical and computational linguistics) and on the content of the lexicon (in theoretical and descriptive lexicology, morphology and phonology), but there are no corresponding modern studies on lexicography as understood here: although the lexicon - in the various senses of the term - is a central component in many modern linguistic theories, as well as in speech technology and in the practical work of a field linguist, oddly enough, there is no modern comprehensive introduction to lexicography. What theoretical and practical information there is on lexicography tends to be strongly oriented towards lexicology of specific languages, and to be scattered around in hard to find project reports, software user guides (e.g. for the widely used *Shoebox* or *Toolbox* software), and unpublished articles on the internet.

Older standard publications, such as Zgusta (1971) or Landau (1984) refer to the "art" and the "craft" of lexicography in the context of classical publishing oriented lexicography of written texts, but not to the "science" or the "technology" of lexicography. Neither do they consider the intricacies of the lexicography of spoken language and of languages with young orthographies, both of which are relevant for field linguistics, language teaching and speech technology, and range from the practicalities of the annotation of speech recordings to the use of sustainable formats and standard fonts.

Other more recent studies tend to be exclusively lexicologically oriented, that is, oriented towards the content of a lexicon, rather than towards its design, production and use. The informal essay by Haviland (2006) in the context of the documentation of endangered languages presents an interesting but idiosyncratic selection of aspects of cognitive semantics and their representation; it deals with a very small part of lexicology, and consequently only with an extremely small part of lexicographic documentation. A more practical study, by Mosel (2004), is also still more lexicologically than lexicographically oriented, in the sense in which we have introduced the term. General introductions to computational lexicography of spoken language for speech technology are provided by Gibbon (2000) and (2004), but lack the broader context of lexicographic methodology.

On the operational side of lexicography, the most comprehensive linguistically oriented database tool is the *Toolbox* package of SIL International (2006), which is provided with extensive tutorial and illustrative material. The package is oriented towards field linguistics and the production of print media dictionaries but can be used for other purposes. The tutorial materials provided with the package are well-constructed and quite extensive, but still presuppose considerable lexicographic background knowledge, as well as computational abilities, for effective use: creation of *Toolbox* projects is not a trivial matter.

2 Course design considerations

2.1 Development strategy

Examination of the background sketched in the preceding section has made it apparent that existing lexicographic instruction materials are either one-sided or elementary, often neglecting very basic components of modern lexicographic procedures such as text and transcription corpus handling, automatic concordancing, word sketch analysis or lexical database design. This situation made it necessary to construct lexicographic teaching materials from scratch. The course design procedure follows the following pattern:

1. Course objectives and content.
2. Design of course structure.
3. Implementation of course materials: The materials consist of slides prepared with OpenOffice (exported for convenience of re-use and distribution to PowerPoint and PDF formats) together with practical exercises using local languages, and construction of a basic lexical databases using spreadsheet techniques and the *Toolbox* software.

The following sections deal in detail with these points.

2.2 Course objectives and content

The knowledge and skills required of participants at the end of the course have to be kept within

realistic bounds. The basic philosophy is to try to achieve a multiplier effect: to ensure that local teaching staff are able to teach the same materials and to extend these, and to ensure that students have the basic intellectual and practical equipment to enable them to acquire further knowledge and skills independently, and to apply these to the creation of lexicons for their chosen languages.
The content of the course covers lexicological elements (selection of domain and types of lexical information), lexicographic elements such as the design of lexicon structures, questions of the implementation of lexicons in different media, as well as practical issues of project design and logistics.
At the end of the course, participants should therefore have knowledge and skills which should enable them to:

- revise and extend basic knowledge of lexicography, architecture of a dictionary;
- understand and acquire skills in using procedures for creating lexical databases and dictionaries, in particular with Toolbox;
- understand the dissemination formats of electronic and print dictionaries;
- independently acquire further knowledge and skills in these areas.

2.3 Course structure

The course was designed to offer a maximum of structured introduction in a compact form which can be taken up by local linguists and used effectively to extend the knowledge and skills of students in local coursework. The course consequently has both a theoretical and a practical component. These components are divided into four phases:

1. Project design: The first point to convey to the students is the need to abstract away from theoretical and descriptive linguistic concerns, and to focus on systematic project design in terms of *lexicographic tasks* and the assignment of *material resources*, *human resources*, and *time resources* to these tasks using practical planning techniques taken from professional project management. Without infrastructural skills of this kind, even the most worthwhile project goals often turn out to be impracticable.
2. Basic lexicographic concepts:
    1. The distinction between
        1. the content of a lexicon (e.g. as a corpus lexicon, a thematic lexicon, a terminological lexicon, a multilingual lexicon),
        2. the underlying structure of a lexicon,
        3. the realisations of a lexicon as implementations in different media: database, print, hyperlexicon.
    2. The relation between text or transcription corpora and lexicons.
    3. Lexicons at different levels of abstraction (wordlist, concordance, tabular lexicon, nested lexicon, generalised lexicon).
3. Elementary computational methods: this is an essential point in an introduction to modern lexicography. The area chosen for this is concordancing, a technique which is neglected in other introductory materials (though concordancing techniques are possible with the Toolbox software).
4. Practical work: The practical part of the course starts with demonstrations, followed by construction of a simple lexicon database and then exercises with basic components of the *Toolbox* software package within the context of a “micro-project”.

The four phases were distributed over the following units, to suit local constraints:

Unit 1: Introduction to language documentation.
Unit 2: Introduction to lexicographic project work.
Unit 3: Basic concepts in lexicography.
Unit 4: Phases in the lexicographic workflow.
Unit 5: Elements of computational lexicography: concordances.
Unit 6: Practical work.
Unit 7: Test and course evaluation.

2.4 Course implementation: slides, exercises, test

The presentation format of the course consists of slides presented from a laptop with a data projector; in addition to the slides, practical software demonstrations are included. All materials are made available to the participants. Participants also have access to a few desktop or laptop computers and are therefore able to utilise the materials directly.
The material prerequisites for the course are:

1. PC with Windows XP for presenter and participants; data projector for presenter and exercise presentations by participants.
2. OpenOffice (text, presentation, spreadsheet).
3. Toolbox (lexicon database management system).
4. Perl (scripting language; used for concordancing in computational lexicography).
5. Traditional classroom materials.

The main didactic methods used in the course are as follows:

1. Lecture (theoretical background; illustration and demonstrations).
2. Student tasks (individual and group tasks, in class and homework):
    1. Prepare a portfolio (learner diary) containing
        1. notes on lecture, group work & discussion, results
        2. additional information from other sources
        3. glossary of technical terms
    2. Microprojects
3. Discussion of results:
    1. Summary of results (short presentations)
    2. Quizzes on results of classes
4. Final test:
    1. describe the lexicographic work flow cycle,
    2. define selected technical terms,
    3. describe lexicon structures.

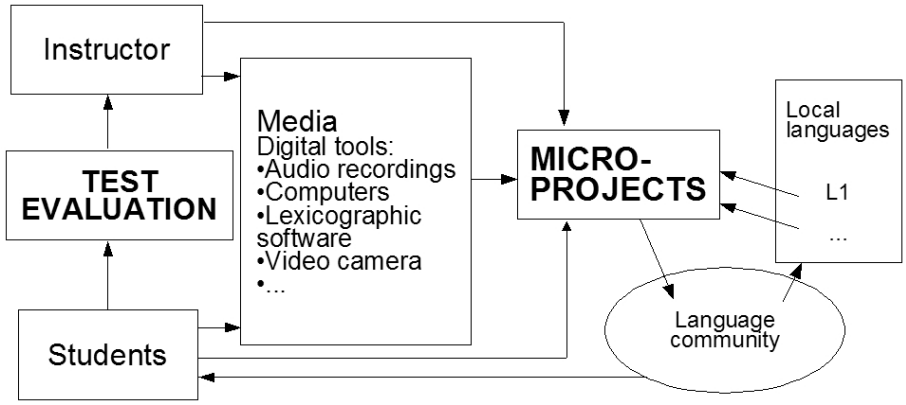

Figure 1: Teaching model.

An overview of the teaching model used is shown in Figure 1. Infrastructural problems must also be taken into consideration in course planning. The greatest infrastructural problem which arose during the course (and which will predictably arise in other contexts) was power failure, but this was handled by switching to a standalone generator.

3 Course Units

3.1 Unit 1: Introduction to language documentation

We distinguish between *language documentation* and *language description*, following Himmelmann (1998). Sometimes the term "language documentation" is restricted to the activities of linguistic fieldworkers, particularly when concerned with endangered languages, but this view does not do justice to the content and methodology of the discipline of language documentation as a

whole. In contemporary linguistics, language documentation is often seen as an antidote to subjective data creation methods based on introspection by native speaker linguists, the standard method in generative, post-generative and formal linguistics. Language documentation is given a more modern interpretation here than those usually cited: the scientific discipline concerned with the provision of high quality recorded empirical data for linguistic work, for education, for technological applications, and for the preservation of the heritage of language communities. The emphasis on high quality recorded empirical data is central: for some linguistic purposes, such as the preliminary identification of problem areas and initial theory formation, introspective data is useful; however, the introspective method inevitably leads to the neglect of large areas of both written and spoken language. The functionality of language documentation is illustrated in Figure 2.

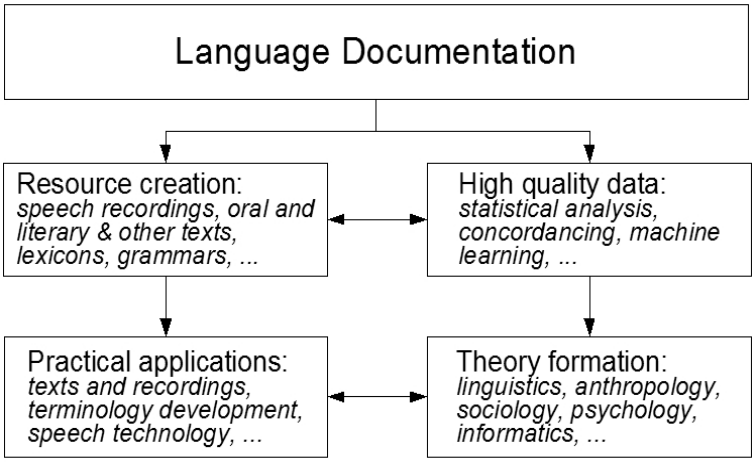

Figure 2: Functionalities of Language Documentation.

As a general guideline for language documentation we follow the WELD model ("Workable Efficient Language Documentation", Gibbon 2004), which specifies the following five criteria, the "FACES":

1. *Fairness*: language documentation, particularly of local languages, should feed into local communities and not constitute yet another knowledge drain for technology in affluent societies.
2. *Affordability*: the language documentation technologies to be used, particularly in local language contexts, should be affordable in the sense that it should not require the latest in broadband high speed internet connections, the latest versions of software and computer operating systems, the latest hardware, but should operate with conventional widespread office-level equipment.
3. *Comprehensiveness*: language documentation should be comprehensive in relation to the requirements - total documentation is not possible, but size and relevance of text and speech corpora are important criteria for effective documentation which can really be used.
4. *Efficient*: regardless of whether the latest software and hardware can be used, computational tools should be used wherever possible for speech and text annotation.
5. *State-of-the-art*: regardless of whether the latest software and hardware can be used, the latest developments in intellectual and operational tools should be used, for example text linguistics in document theory, text technology for analysis and automatic formatting of documents, machine learning as an aid to grammar and lexicon development.

Language documentation has historical roots in at least the following three disciplines, roughly ordered in terms of the times when these disciplines emerged:

1. *Descriptive and anthropological field linguistics*, as practised mainly by structuralist and functionalist linguists in describing unwritten languages, often as background for the translation of religious texts, with the aim of providing writing systems; accounts of language typology also emerged. The origins of this contributory linguistic discipline lie in

the first half of the 20th century (with older roots in the philologies), with developments in typology in the second half. The discipline has been re-focussed in the context of endangered language documentation since the last decade of the 20th century. The mutual relevance of field linguistics and language documentation in general lies in the provision of empirical text and speech data of high quality for work in applications such as language and culture heritage conservation or education, but also for theoretical and descriptive linguistics.

2. *Natural Language Processing* (NLP), and *Computational Linguistics*, from which more recently the discipline of Text Technology has developed. NLP is, broadly, concerned with the computational analysis and generation of texts in natural languages; application areas include document generation, information extraction, internet search. This work started in the mid-20th century, with text statistics (character and word distributions) and concordance construction (the Bible, Shakespeare, legal texts); the discipline is now one of the foundational disciplines of the internet and has come full circle with modern versions of text statistics and concordance construction in internet search portals such as Google. The relevance of NLP to language documentation lies in the techniques developed for the formal description and machine learning of document structure, from the character level to the text network level, as a basis for automatic document analysis, classification and production (for example in the cases of printed and hypertext dictionaries).
3. *Speech technology*, which is concerned with speech input and output devices for computational systems, including automatic speech recognition, speech synthesis and dictation software. Speech technology, an engineering discipline which started in experimental phonetics in the 1960s, has also continued to develop rapidly. In the language documentation context, the area of speech synthesis has been most prominent, with the results of empirical fieldwork being used together with Natural Language Processing methods in order to create the language models used in practical speech synthesis systems for local languages in many parts of the world. Speech technology and NLP meet in hybrid computer systems such as dictation software or tools for analysing and producing multimedia documents.

These source disciplines for language documentation are concerned with two main domains of language: the domain of written text data (NLP) and the domain of speech data (field linguistics and speech technology). The two domains overlap in some areas. For example, handwritten text poses similar physical pattern recognition problems to speech, text spatial and speech temporal. Another case is when speech is transcribed, transcriptions also being text. But in general, text documentation and speech documentation have very different empirical and technological foundations, particularly where speech is understood as multimodal (acoustic and visual) communication.

In each of these areas, language documentation has three main kinds of methodology:

1. *Data acquisition*. The design or selection and acquisition of high quality linguistic data, either by planned experiments, by interviews, or by collecting planned or unplanned corpora of texts or speech. These methods involve human interactions of a variety of types, with or without computer support.
2. *Data processing*. Initial decisions in language data processing are based on categorisations and procedures which have both intuitive and formal foundations. Analysis of text corpora has led to significant changes in linguistic subdisciplines such as computational corpus linguistics, i.e. computational text corpus analysis, which is now conducted on large corpora with computational tools like concordancers for text or stochastic modelling for speech.
3. *Data storage and access*. The criteria of sustainability, interpretability and reusability are central for language documentation: the goal of language documentation is to provide a stable and reliable empirical basis for analysis and application of data.

3.2 Unit 2: Introduction to lexicographic project work

Unlike conventional introductions to descriptive linguistic and other language modelling work, we consider it necessary to have a good working understanding of what it means to organise a research

and development project. This understanding is essential for efficient work and effective use of resources, whether the project is a a large funded activity or a Ph.D. or M.A. thesis.
The first step is to introduce a basic structure for producing specified outcomes. A useful model for project design is the following, adapted from a traditional approach to software development:

1. *Definition of needs, objectives and outcomes*: in the present case, the need is to provide lexicographic material for linguistic analysis, educational applications, and for use in speech synthesis systems; the objectives are to create lexicographic infrastructure; the outcome is the product in the form a lexicons, concordance, etc.
2. *Architecture*: in general terms, the architecture of the product consists of a specification of the *modules* of a system and of the *interfaces* between them; in the case of lexicography, the architecture concerns the components of the lexicon (front matter, main body, lexical entries, cross-references, back matter, ...), their interrelationships, and their processing requirements.
3. *Implementation*: a lexicographic product can be produced as a conventional print medium book, or as a database which is made accessible to the user in the form of a hypertext such as a hyperlexicon or hyperconcordance on the internet or on CD. A very small-scale lexicon production process may use word processing techniques, though these are highly inefficient for lexicon production, and inevitably lead to inflexibility and inconsistency, which are hindrances in scaling-up lexicon size. The state-of-the-art approach is to create a lexical database, and then generate the correctly formatted product, whether book or hypertext, automatically from the database information.
4. *Evaluation*: the outcome requires evaluation, i.e. quality control, before being regarded as a finished product. In the case of lexicography, the criteria are based mainly on content (coverage of a given domain, detail of lexical information) and on usability (ease of access of the information) by the intended user community.

A typical workflow cycle during the implementation and evaluation phases is shown in Figure 3.

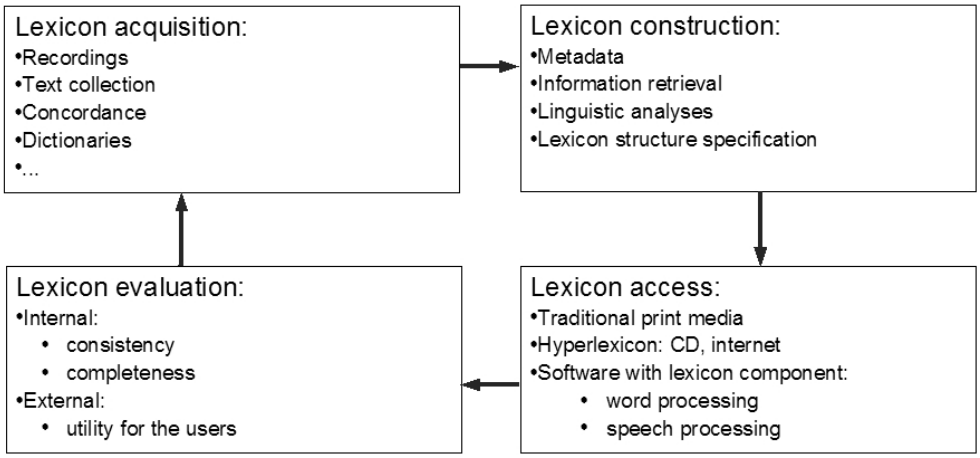

Figure 3: Lexicographic workflow cycle.

The main phases are as follows:

1. *Lexical data acquisition*. Lexical data may be acquired from many sources, the most important of which are
    1. intuitions of the lexicographer,
    2. existing lexicons,
    3. systematic interviews with native speakers based on standard wordlists or topic areas such as body parts, domestic tasks, agriculture,
    4. corpus linguistics, i.e. text and transcription corpus mining by means of automatically produced wordlists, concordances, and statistical analyses.
2. *Lexicon construction*. Construction of the lexicon involves decisions about the components

of the lexicon such as

1. *megastructure*: front matter with metadata about the lexicon, grammatical information, main body, back matter with metadata about the lexicon,
2. *macrostructure*: overall organisation of the lexical entries,
3. *microstructure*: the information contained in the lexical entries,
4. *mesostructure*: cross-references between the lexical entries and each other as well as to other information sources such as source texts, sketch grammar, abbreviations.

3. *Lexicon access*. The realisation of the lexicon in different media - as database, hyperlexicon or as a book - depends very much on the requirements of the potential user. Each medium is appropriate for different kinds of use and access.
4. *Lexicon evaluation*. The evaluation process covers
    1. *diagnostic evaluation*, for errors, for extrinsic coverage in terms of numbers of lexical entries, for internal coverage in terms of detail of lexical information, in terms of
        1. *correctness*,
        2. *completeness*,
        3. *consistency*.
    2. *user evaluation*, (sometimes referred to as *field evaluation*) for ergonomic aspects such as relevance and ease of use.

3.3 Unit 3: Basic concepts in lexicography

Lexicographic products are extremely varied. There are innumerable varieties of print dictionaries, ranging from simple glossaries, rhyming dictionaries and concordances through terminological dictionaries and bilingual dictionaries, to highly structured thesauri and complex etymological dictionaries with very detailed lexical information and extensive citation of sources of examples. Dictionaries are extremely complex documents with a highly differentiated "semantics":

1. *Publication context*: the megastructure of a dictionary consists of the front matter, the main body, and the back matter. The front and back matter contain both metadata about the dictionary (e.g. title, editors, publisher, date, place of publication and other information in prefaces, forwards and table of contents), and mesostructural information with generalisations about orthography, spelling-to-sound rules, word structure, prosody, morphology and phrasal syntax. The main body contains the lexical entries with selected types of lexical information, to which the further extensional and intensional criteria apply.
2. *Content organisation*:
    1. *Extensional*:
        1. *Macrostructure*: the overall *organisation* of the lexical entries as a list, a table, a hierarchy, a network.
        2. *Extensional coverage*: the *selection* of lexical entries from a particular topic domain, the *number* of lexical entries.
        3. *Functionality*: the use cases from the point of view of someone consulting a dictionary:
            1. *semasiological* (decoding or reader's dictionaries), organised by word forms (e.g. alphabetically), with meanings as the information searched for.
            2. *onomasiological* (encoding or writer's dictionaries), organised by meaning-based search keys, e.g. as a thesaurus or synonym dictionary, with the word form as the information searched for.
    2. *Intensional*:
        1. *Microstructure*: The *organisation* of types of lexical entry, for instance as a "flat list" of types of lexical information about form (orthography, pronunciation, syllabification, accentuation), structure (morphology, i.e. internal word structure, syntactic category and subcategory), and function (definition, semantic components, semantic relations such as synonyms, antonyms, hyponyms, hyperonyms), pragmatic usage properties, examples, and (in lexical databases) metadata with date of editing

and identity of the lexicographer.

2. *Intensional coverage*: The *selection* of types of lexical information associated with each entry and the *number* of items of lexical information.

3. *Informational*:
    1. *Mesostructure*: A *network* of cross-references between lexical entries and to additional sources of lexical information.
    2. *General coverage*: Cross-references constitute generalisations in the lexicon; the fact that the lexicon contains generalisations contradicts the widespread and false assumption that the lexicon only contains idiosyncratic information about lexical items. A lexical entry in a conventional dictionary, in whatever medium, can only be understood if certain assumptions are made about its general properties:
        1. *conventions* about orthography, pronunciation transcription, part of speech, morphological structure, are abstracted out of the individual entry and expressed by explicit or implicit cross-references to a grammar sketch and abbreviations in the front or back matter,
        2. *definitions* using other terms (lexical items) from the dictionary to specify the meaning of the entry,
        3. *cross-references* to semantically related synonyms and antonyms (in this respect, a *thesaurus* or synonym dictionary is organised in terms of an explicit metastructure),
        4. *examples* or cross-references to sources of examples in texts.

The interrelations between different components of lexicon structure are visualised in Figure 4. Megastructure contains all other components, in particular metadata and elements of mesostructure, as well as the macrostructure, which in turn contains the lexical entries and their microstructure organised, in principle, as a table.

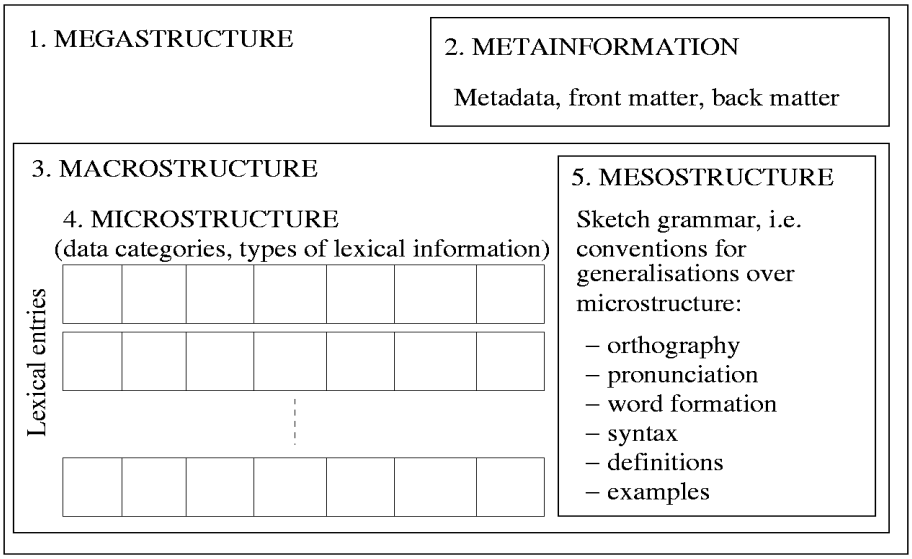

Figure 4: Components of lexicon structure.

The fundamental insight is about the content, structure and realisation of a lexicon is that a lexicon is a semiotic product, that is, a product concerned with signs, and that its design is dependent on the awareness that signs have four main properties, in addition to their context of use, which can be organised into two dimensions of structure and interpretation:

1. *Structure*:
    1. *Internal structure*: for words, this is their morphological structure (word formation by derivation and compounding, and inflection); for idioms this is their phrasal syntax.
    2. *External structure*: for words, this is the phrasal context in which words occur, summarised in their part of speech, i.e. grammatical category (e.g. verb) and subcategory

(e.g. transitive), as well as the context of use; for idioms it is essentially the context of use.

2. *Interpretation*:
    1. *Semantic interpretation*: for all lexical items the semantic interpretation is the assignment of a meaning, for example by definition, by reference to semantic fields of synonyms and antonyms, or by example.
    2. *Media interpretation* (appearance): for all lexical items the media interpretation is twofold, either visual (assignment of an orthography) or acoustic (the assignment of a pronunciation by means of a transcription); in principle, many conversational gestures are also a kind of lexical item, realised visually, often together with specific conventional lexical items or idioms.

Lexical items are signs, but the entire lexicon is also a complex sign in the sense that it is a special kind of text, a semiotic product: it also has a semantic interpretation in terms of its coverage, and a complex structure (megastructure, macrostructure, microstructure, mesostructure). Two basic terms for lexical signs are:

1. *Lemma*: an item in a dictionary, usually represented by an inflected form of a lexeme, described by a *lexicon article*. In European languages usually by the nominative singular for nouns, and infinitive for verbs, but in other languages the inflectional typology plays a role; for instance, with Niger-Congo languages with nominal classification prefixes, decisions on useful lemmatisations (with or without prefix) and on alphabetic orderings have to be made.
2. *Lexeme*: a word occurring in a corpus, stripped of its inflectional affixes, representing the basic lexical stem and the lexical meaning of the word.

The procedure of affix-stripping in computational linguistics is usually known as lemmatisation; strictly speaking the use of the term "lemma" for this stage of the lexicon acquisition procedure is premature since the item is not (yet) an entry in a dictionary. A term "lexematisation" would be more appropriate, but it would be pedantic to insist on this.

3.4 Unit 4: Phases in the lexicographic workflow

The four phases in a standard form of lexicographic workflow are, as already noted, *lexicon acquisition*, *lexicon construction*, *lexicon access* and *lexicon evaluation*. This unit focusses mainly on the acquisition of lexical information.

3.4.1 Acquisition of lexical information

The acquisition of lexical information is heavily dependent on linguistic theory; in many dictionaries the underlying theoretical foundation is made explicit in the grammar sketch contained in the front or back matter, in general it is left somewhat inexplicit, or represented by an informal set of claims about the language.

A dictionary in general does not contain information about the lexicographic procedures involved in acquiring the lexical information which it contains. This information may come from many sources, including the following:

1. intuitions of the lexicographer,
2. existing lexicons,
3. systematic interviews with native speakers based on standard wordlists or topic areas (such as body parts, domestic tasks or agriculture),
4. corpus liguistics, i.e. text and transcription corpus mining by means of automatically produced wordlists, concordances, and statistical analyses.

The last of these areas, corpus linguistics, is now recognised to be the most important source of lexical information. In Figure 5 an overview of relations between corpus sources and different levels of abstraction of lexical information is given.

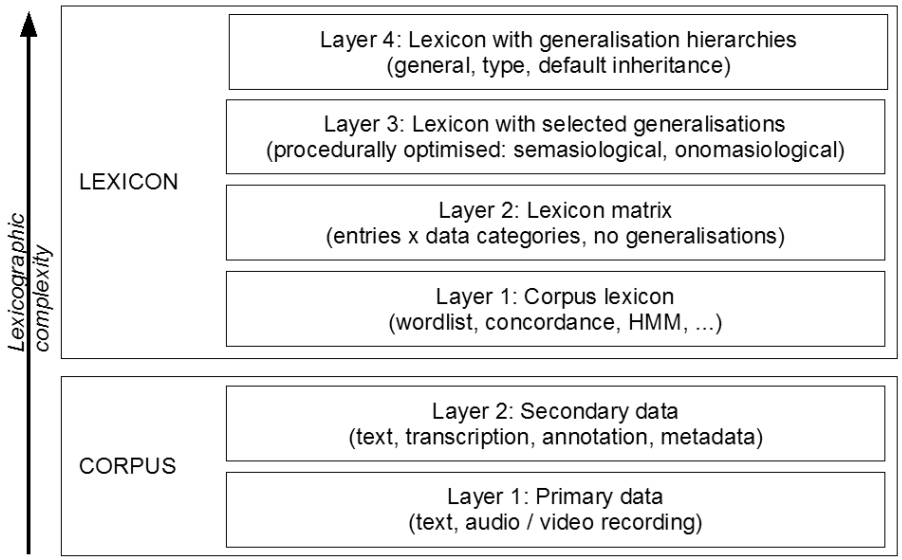

Figure 5: Abstraction hierarchy of lexicon types.

The *primary data* for the acquisition of lexical information come from the collection of texts or speech recordings, or both, which constitute the lexicographic corpus; cf. the lower block in Figure 5. The corpus may consist of authentic texts and speech recordings, where "authentic" in this context means "not created for the purpose of lexicography". In the case of spoken language, speech recordings are in general purpose-built, i.e. created systematically from a particular domain. In conversation analysis, recordings are of everyday or institutional interactions; in speech technology, recordings are generally of such special scenarios as such as appointment scheduling, travel booking, car navigation systems, etc. Authentic speech recordings are often taken from mass media broadcasts; authentic speech obtained by surreptitious clandestine recording is ethically deprecated and is in general illegal.

On the basis of the primary data, the *secondary data* of annotations, which constitute a first level of generalisation and linguistic analysis, are created:

1. *Text annotation*: texts are consistently formatted and provided with structural markup or tags (part of speech category or phrasal category); where sentence structure is marked up, the secondary data are referred to as *tree bank*. Each language requires its own *tagset* of categories, based on a linguistic analysis of the language.
2. *Speech annotation*: speech recordings are provided with transcriptions, which are used to provide time-aligned labels or indices of the recordings; each item in the transcription is provided with a time-stamp (a temporal point or interval in the speech signal) which indicates to which part of the recording the transcription item applies. The items may be whole words or phrases in orthography or phonemic transcription, or syllables, or individual phonemes in phonemic or phonetic transcription. Insofar as transcriptions are also written texts, they may also be provided with the same kind of structural markup as regular written texts.

The first lexicographic step, shown in the upper block of Figure 5, is to create a *corpus lexicon* consisting of *wordlists*, and *concordances* based on these wordlists. A concordance is basically a list of words with the contexts in which they occur in the corpus, i.e. an elementary semasiological lexicon with a very simple macrostructure and a microstructure in which only contextual definitions by example are given. However, in conventional lexicography, concordances have the status of intermediate stages in lexicon construction, rather than as a lexicon in its own right. The most well-known kind of concordance is known as the KWIC concordance,where the abbreviation stands for "KeyWord In Context". Concordances may be enhanced with additional kinds of lexical information if the corpus is tagged. A further level of concordance abstraction is the *word sketch* (Kilgarriff & Tugwell 2002). Concordances are dealt with in more detail in the section on computational lexicography.

The second lexicographic step is the creation of a lexicon matrix or table (or a set of matrices or tables) as a lexical database containing all the lexical entries, each associated with the required types of lexical information. This table can be extremely large, with tens of thousands of entries (or more) as rows in the table, and five or more kinds of lexical information in columns in the table. In practice, a table of this kind is organised in a database system.

The third lexicographic step is to create a *lexicon based on selective generalisations* which are dependent on procedural decisions about how the lexical information is to be accessed: semasiologically, onomasiologically, or according to other criteria. Additionally, decisions have to be made about introducing hierarchical structures into both the macrostructure (for instance in a thesaurus) and the microstructure (to combine polysemous readings under the same lemma form). This is the organisational or sorting and structuring step which immediately precedes production of a practical dictionary.

The fourth lexicographic step, the creation of a *lexicon with generalisation hierarchies* shown in Figure 5 is determined by lexicon theory, and is not usually part of conventional lexicography. It involves abstracting hierarchies of generalisations from the lexical entries, forming classes and subclasses of entries containing types of lexical information ranging from the specific to the general. Effectively, the basic hierarchical structures in this kind of lexicon are taxonomies, but they refer not just to semantic taxonomies but also to morphological taxonomies (e.g. conjugation classes and subclasses) and phonological taxonomies (e.g. syllable classes and subclasses). Lexicons of this kind are often referred to as inheritance lexicons (Gibbon 2002). In an inheritance lexicon, the lexical entry itself is maximally underspecified, that is, it contains only idiosyncratic information which serves to identify it; the same applies to all the subclasses and their superclasses in the hierarchy. Entries, and the classes to which they belong, are fully specified by inheriting as much generalisable information as possible from their superclasses.

3.4.2 Lexicon construction

The decisions to be taken in constructing an actual lexicon are practical ones concerned with deciding on the structural components of the outcome of a lexicographic project. Many of these aspects have been introduced in preceding sections, so the main concern in the course at this point is to fill in the abstract theoretical framework with practical examples of different kinds of lexicon.

An important component of this phase is the introduction of *lexicon databases* as a first step in computational lexicography. There are many kinds of database, but the most suitable, and the most widespread kind for lexicographic work is the relational database, in which objects (here lexical entries) are related to their properties (values of attributes), and to other objects, in the form of tables. In each table, the objects are represented by the rows of a table, the attributes by the columns of the table, and the properties are the values of the attributes in the cells of the table.

It is possible to start with creating simple dictionary tables in a word processing system such as OpenOffice Writer, or MS-Word (provided that genuine table object formats are used, and not simply table lookalikes where columns are spaced with blanks or tabs).

A better way of starting out with database construction is to use spreadsheet software such as OpenOffice Calc or MS-Excel, a technique which many linguists use. Spreadsheet software permits tables of arbitrary size (not limited by page width) to be created, and also permits flexible sorting, and linking of cells.

The optimal technique, after an introduction to “tabular thinking”, is to use a professional database management system (DBMS), however. The most popular DBMS among linguist is the *Toolbox* DBMS with a wide range of features required by linguists, permitting use of language-specific fonts and attributes as well as links to a tagged (interlinearised) corpus, and the production of an appropriately formatted print medium dictionary with ready-made software tools such as *MDF* (Multi-Dictionary Formatter), or *Lexique Pro*, or with custom-made tools. For large commercial lexical databases, conventional DBMS applications such as *MS-Access*, *Fox-Pro* or *mSQL* are used.

3.4.3 Lexicon access

The major decisions to be taken in ensuring convenient lexical access based on user needs are

concerned with

1. the organisation of the macrostructure of the lexicon (e.g. semasiological, onomasiological),
2. the choice of media interpretation for the realisation of the lexicon in print, on the World Wide Web, on CD, etc.

The selection of a particular macrostructure for a lexicon determines the ease of accessibility of lexical information like the selection of a particular view for a database, as with address database which may be viewed according to different selection and sorting criteria, e.g. names, or towns, or telephone numbers. In the case of a bilingual or multilingual lexicon, each language determines a possible organisation of the lexicon (e.g. English-Ibibio, Ibibio-English), with different degrees of relevance to different users.

The selection of a particular medium determines accessibility to lexical information in a different way. Under many circumstances, a printed dictionary is the optimal product. However, for more general and flexible search purposes a hyperlexicon on CD or on the internet may be preferred; the utility of such a lexicon depends on the availability of computational tools for viewing it (and of course for creating it).

3.4.4 Lexicon evaluation

The evaluation of a lexicon is a complex matter, though the basic principle is straightforward. As with any evaluation process, the evaluation of a lexicon consists in the comparison of the properties of the finished system (or subsystem) with its requirements specification and its design specification. The empirical comparison procedure itself uses the same methodologies as other comparison procedures; cf. Gibbon (1997, 2000).

During lexicon construction, diagnostic evaluation processes provide for checking the *correctness* and *completeness* (i.e. the *extensional* and *intensional* coverage) and *consistency* of lexical entries, and for the correction of any errors of these three kinds. The completed lexicon needs to be evaluated in field trials in terms of *usability* and *relevance* and other dimensions of immediate relevance to the user and commercial or other distribution.

The field of lexicon evaluation is rapidly expanding, and cannot be dealt with in a brief overview (cf. contributions to Corréard 2002). Furthermore, the initial priority within the framework of the present course in lexicography is the creation process, and in the present context more than a minimum of information about lexicon evaluation procedures is not relevant.

3.5 Unit 5: Elements of computational lexicography: concordances

An overview of elements of computational lexicography is provided in contributions to van Eynde & Gibbon (2000), with an introductory overview by Gibbon (2000). Much of the infrastructural value of computing in lexicography lies two areas:

1. the quantity of data which can be processed in a given time,
2. the consistency of performance, given an initial validation.

Reformatting even a very large dictionary for use as a hyperlexicon takes minutes rather than months if computational methods are used and the principles outlined in the present overview are followed. Modern computational methods also support the FACES (fairness, affordability, comprehensiveness, efficiency, state of the art) criteria for workable effective language documentation, as outlined in the first unit.

A simple and extremely useful application of computational corpus linguistics lies in providing *word frequency lists* and *rank lists*, and in calculating the *type/token* ratio (ratio of word forms to the number of occurrences of these word forms) for texts. Information of this kind was used in the first attempts to categorise the formal styles of authors: such rank lists and ratios tend to stay approximately constant for a given author. This kind of information is also used in evaluating dictionaries.

One of the most useful computational tools for lexicon analysis of words in context is the concordancer, a software tool for producing concordances.. The simplest and most well-known kind of concordance is the KWIC concordance, or KeyWord in Context concordance. A KWIC concordance is fundamentally a list of all the words in a corpus, together with the contexts in which

each word occurs. For example, the following is an extract of just 10 occurrences of the word "cat" at different places in an automatically generated concordance from a machine-readable version of T. S. Eliot's *Old Possum's Book of Practical Cats*, for the keyword "cat" (with frequency information):

CAT

```
(01) When I tell you, a cat must have THREE DIFFERENT NAMES.
(02) But I tell you, a cat needs a name that's particular,
(03) Names that never belong to more than one cat.
(04) But THE CAT HIMSELF KNOWS, and will never confess.
(05) When you notice a cat in profound meditation,
(06) Growltiger was a Bravo Cat, who travelled on a barge:
(07) In fact he was the roughtest cat that ever roamed at large.
(08) And woe to any Cat with whom Growltiger came to grips!
(09) The Rum Tum Tug Cat is a Curious Cat:
(10) Yes the Rum Tum Tugger is a Curious Cat
```

The following example is of a concordance in the standard keyword-centred format, for a key phrase "eine Leiste" (a bar) in a corpus of dialogue transcriptions made for the German VerbMobil speech-to-speech machine translation project:

```
HyprLex: Plain substring KWIC concordance for SFB-A3
coral.lili.uni-bielefeld.de, Sat Dec 30 17:37:08 MET 2006
22 x eine Leiste
dia_12_I:          jetzt wird "ah die kl eine Leiste  mit der Schraube drin auf den
dia_12_I:              jetzt wird die kl eine Leiste  rechtwinkelig unter den roten
dia_12_I: ge Leiste wird "ahm auf die kl eine Leiste  dr"ubergelegt , so da"s "ah d
dia_12_I:              "ahm es wird noch  eine Leiste  mit drei L"ochern ben"otigt .
dia_12_I:                bitte legen Sie  eine Leiste
dia_12_I:                bitte legen Sie  eine Leiste  mit sieben L"ochern rechtwink
dia_15_I: Leiste mit drei L"ochern mu"s  eine Leiste  mit f"unf L"ochern mit zwei S
dia_15_I: in die letzten beiden L"ocher  eine Leiste  mit f"unf L"ochern angeschrau
dia_15_I:       in das zweite Loch mu"s  eine Leiste  mit "ah sieben L"ochern senkr
dia_15_I: eiste mit f"unf L"ochern mu"s  eine Leiste  mit drei L"ochern , so da"s e
```

From the lexicographic point of view, a KWIC concordance is therefore essentially an elementary corpus-based semasiological lexicon, with a very simple macrostructure, and a microstructure in which only contextual definitions by example are given. However, in conventional lexicography, concordances are not regarded as lexicon in their own right, but have the status of intermediate stages in lexicon construction. Concordances may be enhanced with additional kinds of lexical information if the corpus is tagged. A further level of concordance abstraction is the *word sketch* (Kilgarriff & Tugwell 2002).

Until the 1950s, concordances of texts such as the Bible, Shakespeare's plays, and legal texts, were made by hand, filling huge volumes and consuming an enormous amount of time and energy. The earliest computational concordances were consequently also made mainly by those theologians and literary scholars who were interested in close text analysis and in the statistical analysis of texts. In fact, a full-text keyword search machine on the internet, such as Google, is also a variety of concordance, and indeed is frequently used for corpus linguistic and lexicographic research.

The basic procedure for making a KWIC concordance is as follows:

1. Collect a corpus of texts in electronic format
2. Pre-process each text (tokenisation):
    1. process punctuation marks, numbers, abbreviations
    2. break the text into context units (lines/sentences)
3. Extract a keywordlist with all the words from the text
4. Locate keywords in context:

```
for each keyword in the keywordlist:
      for each text in the corpus
            for each context unit in the text
```

```
if the word occurs in the context unit
then mark keyword in context and store
else continue
```

5. Format the stored output, e.g. with the keyword in the centre of the context.

The project planning steps for making a KWIC concordance are the same as for project planning in general (the terms used here are used in software development):

1. Requirements specification: who, for whom, what, how, where, when; i.e. resources (people, software, hardware), timetable, e.g. topic area, computational or print medium use.
2. Design: architecture: modules and interfaces; user interface (perhaps: graphical user interface, GUI); see below.
3. Implementation: selection of programming language, coding; in academic projects, the most frequently used programming language is Perl, a standard language for computational text processing.
4. Evaluation: operation (internal system diagnostic and performance evaluation); utility (external fitness-for-use evaluation).

The modules (boxes) in the concordance construction procedure, and their interfaces (lines), are visualised in Figure 6.

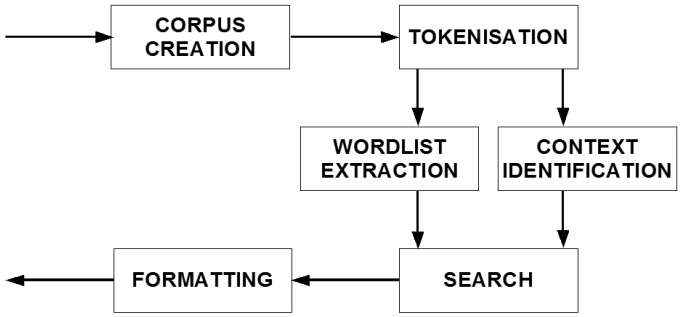

Figure 6: Concordance creation project architecture.

For teaching purposes, the software modules were implemented in Perl.

The code is reproduced here without detailed explanation; use of the code requires elementary knowledge of programming in cooperation with a computational linguist or computer scientist.

Preprocessing:

```
$wordlist = "" ;
while(<>) {
      chomp;
      s/e\.g\./EG/ ;
      s/M\.A\./MA/ ;
      tr/[.,;:„"-)(]/ / ;
      tr/[A-Z]/[a-z]/ ;
      tr/\t/ / ;
      s/  */ /g ;
      $wordlist = $wordlist . $_ ;
      }
```

Context identification:

```
$contextlength = 5 ;
@contextlist = () ;
for ($i=0; $i<(@wordlist - $contextlength); $i++) {
      print OUTPUT $wordlist[$i] ;
      $contextlist[$i] = $wordlist[$i] ;
      for ($j=1; $j<$contextlength; $j++) {
```

```
            print OUTPUT " " . $wordlist[$i+$j] ;
            $contextlist[$i] = $contextlist[$i] . " " . $wordlist[$i+$j] ;
            }
      print OUTPUT "\n" ;
      }
```

Keyword list:

```
@wordlist = split(/ /,$wordlist) ;
@sortedwordlist = sort { $a cmp $b } @wordlist ;
$prev = "" ;
$count = 0;
@uniquewordlist = () ;
for ( $i=0; $i<@sortedwordlist; $i++ ) {
      $a = $sortedwordlist[$i] ;
      if ( $a ne $prev ) {
            $prev = $a ;
            print OUTPUT $a . "\n" ;
            $uniquewordlist[$count] = $a ;
            $count++ ;
            }
      }
```

Search:

```
for ($i=0; $i<@uniquewordlist; $i++) {
      $a = $uniquewordlist[$i] ;
      for ($j=0; $j<@contextlist; $j++) {
            @context = split(/ /,$contextlist[$j]) ;
            if ($a eq $context[2]) {
                  $context = $context[0] . " " . $context[1] . " " .
$context[2] . $context[3] . " " . $context[4] ;
                  print OUTPUT $context ;
                  }
            }
      }
```

Finally, a formatting procedure such as those previously illustrated must be selected, with appropriate titles and headings for the concordance table, and implemented for printing or as a hypertext for the internet.

3.6 Unit 6: Practical work (microproject)

The practical work consists of a microproject on specific aspects of the preceding units, involving recordings, annotations, construction of wordlists and concordances as well as a basic lexicon spreadsheet table for selected local languages. As the basis for a “toy” corpus and lexicon, recipes were selected: the ingredients list is effectively a selective corpus lexicon, and the method is effectively a text corpus which contains words from this lexicon.
As a final stage, the lexicon table is re-implemented as a *Toolbox* lexicon database, a procedure which would require a lengthy further description.
This section is left to the imagination of the experienced course teacher.

4 Evaluation of the course, conclusion, recommendation

Two conventional levels of course evaluation were included:

1. *Student evaluation by teacher*, by a final test at the end of the course. The test is highly selective, as there are alway numerous possibilities for constructing a detailed test, with different types of question. In view of the experimental nature of the course, a two-hour open question scheme was selected, with the following questions:

1. *Explain the lexicographic workflow cycle, and give details for each of the four parts*.
2. *Define the KWIC type of concordance and define the six steps involved in concordance construction*.
3. *Describe the four main components of the architecture of the lexicon, giving examples to illustrate each component*.

2. *Teacher evaluation by the student*, by course and teaching evaluation. The course and teaching evaluation was held, and used by the host university for internal purposes. Feedback was in general positive, with special emphasis on the high level of informativity of the course, and the introduction of computational methods. However, the lack of equipment was commented on, and the difficulty of understanding several parts without adequate means of practising.

A French version of the course was first presented by the authors in March 2006 at the Université de Cocody, Abidjan, to Maîtrise and Doctorat students. An English version of the course was later presented in August 2006 at the University of Uyo as an official component of the M.A. course in "Computational Documentation of Local Languages". In both cases, the courses filled a gap in local expertise in the area of computational lexicography. The courses were attended by local staff, and were designed and modified in cooperation with them.

The tentative conclusion is that in future the computational language documentation course can be adopted and presented by local linguistics and computer science staff, and the general recommendation which can be made (in agreement with the local staff) is that the course can be re-used and, judging by interest at UNESCO and other international conferences (Ekpenyong & al. 2006; Urua & al. 2006), it is also likely to be usable at other universities with similar needs.

5 References

Corréard, Marie-Hélène, ed. (2002). Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins. EURALEX. Distributed by www.ims.uni-stuttgart.de/euralex/ .

Ekpenyong, Moses, Nnamso Umoh, Mfon Udoinyang, Golden Ibiang, Eno-Abasi Urua, Dafydd Gibbon (2006). Infrastructure to Empowerment: An OSWA+GIS Model for Documenting Local Languages. E-MELD: http://linguistlist.org/cfdocs/emeld/workshop/2006/papers/ekpenyong.html.

Gibbon, Dafydd (2000). Computational Lexicography. In Frank van Eynde and Dafydd Gibbon, editors, *Lexicon Development for Speech and Language Processing.* Dordrecht: Kluwer Academic Publishers. pp. 1-42.

Gibbon, Dafydd (2002). Compositionality in the Inheritance Lexicon: English Nouns. In: Behrens, Leila & Dietmar Zaefferer, eds. *The Lexicon in Focus: Competition and Convergence in Current Lexicology*. Frankfurt etc., Peter Lang.

Gibbon, Dafydd (2005). Spoken language lexicography: an integrative framework. In: Zybatow, Lew, ed. *Translatologie - neue Ideen und Ansätze.* Innsbrucker Ringvorlesungen zur Translationswissenschaft IV. Forum Translationswissenschaft Band 5, pp. 247-289.

Gibbon, Dafydd, Roger Moore & Richard Winski (1997). *Handboook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Dafydd Gibbon, Inge Mertins, and Roger Moore, eds. (2000). *Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology and Product Evaluation*. Doordrecht, Kluwer Academic Publishers.

Gippert, Jost, Nikolaus P. Himmelmann, Ulrike Mosel (2006). *Essentials of Language Documentation*. Berlin: Mouton de Gruyter.

Himmelmann, Nikolaus P. (1998). Documentary and descriptive linguistics. *Linguistics 36*: 161-195.

Haviland, John B. (2006). Documenting lexical knowledge. In: Gippert & al., eds., *Essentials of Language Documentation*. Berlin: Mouton de Gruyter.

Kilgarriff, Adam & David Tugwell) Sketching words. In: Marie-Hélène Corréard (ed.), *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins.* EURALEX: 125-137.

Landau, Sidney I. (1984). *Dictionaries: The Art and Craft of Lexicography*. New York: Scribner.

Mosel, Ulrike (2004) Dictionary making in endangered speech communities. In: Peter Austin (ed.)

Language documentation and description. Vol. 2, *Endangered Languages Project.* London: School of Oriental and African Studies. pp. 39-54. (cf. also http://www.mpi.nl/lrec/2002/papers/lrec-pap-07-Dictionary_Endangered_SpComm.pdf)

Urua, Eno-Abasi Essien, Dafydd Gibbon, Firmin Ahoua, Moses Ekpenyong and Eddi Gbery (2006). The WALA Initiative: an overview. A paper presented at the UNESCO/ACALAN meeting, Bamako, Mali March 23-25, 2006.

van Eynde, Frank & Dafydd Gibbon, eds. (2000). *Lexicon Development for Speech and Language Processing.* Dordrecht: Kluwer Academic Publishers.

Zgusta, Ladislav (1971). *Manual of Lexicography*. The Hague: Mouton.