On lexical objects and their properties

A contribution to the 'MetaLex' requirements specification for spoken language lexicon documentation

Dafydd Gibbon
Universität Bielefeld
`gibbon@spectrum.uni-bielefeld.de`

DOBES Technical Report *n* (Ega)
(Status: RFC draft, January 2000 — printed January 5, 2001)
Revised version of a paper presented at the workshop on
Web-Based Language Documentation and Description
12-15 December 2000, Philadelphia, USA.

Contents

Abstract

The implementation of complex multimedia lexica in hypertext formats is potentially a task of extremely high complexity. It is suggested that as a preliminary step towards designing electronic lexica of various kinds, including Web lexica, a requirements specification in terms of first principles of the lexicon sciences is needed. On this basis a data model for generic lexical databases can be designed, and multimedia hypertext can be systematically derived as differently optimised views on this database model. The aspects discussed include the notions of semasiological and onomasiological macrostructures as procedural views on the same lexicon, different ranks of lexical objects, and a contemporary semiotic model for defining the core of a lexicon microstructure as types of lexical information. Additional dimensions of lexical complexity which apply to all lexical objects are also discussed. A new concept of mesostructure is introduced, to capture the partial regularities which may be abstracted out of invididual lexical entries, and constitutes an important part of lexical metadata. The principles described have been applied to terminological work in the EAGLES project and are currently in use in the DOBES consortium within the project "Ega: a documentation model for an endangered Ivorian language."

1 Preliminaries

This paper is a contribution to the lexicographic component of language documentation and metadata specification, whereby the relation between language documentation and linguistic description (and metadescription) is understood as a continuum, not a sharp divide. We will term this specification the 'MetaLex' specification. Furthermore, since documents are linguistic objects, they can themselves be the subject of linguistic descriptions; it would therefore seem to be rather foolish to ignore linguistic criteria in this area, and particularly foolish for the linguist to do so (as, alas, many linguists do).

But rather than plunging into the midst of detailed discussions of lexicon architecture, data structures, formats, acquisition and database tool implementation, types of lexical information for specific lexica, and so on, this contribution describes an attempt to step back for a moment from the grip of design and implementation issues in lexicographic development and to specify the linguistic foundations of a requirements specification for lexical development before going into further practical application details.

The term 'requirements specification' is meant literally in the sense of software engineering: in this case, it means a specification of the requirements which lexical documentation should fulfil, derived from very general requirements such as the intensional and extensional coverage of lexical information, and the reusability, interoperability, portability and ergonomic utility of lexical software. There are, of course, other technical and logistic requirements which are not covered here.

This is not to say that the issues of lexicon design, implementation, evaluation, acquisition and application logistics are unimportant. On the contrary. But we claim that they must be derived from general specifications of requirements if they are not to risk being ignored by others. In discussing these points, this document draws to different extents on criteria from linguistic theory, descriptive lexicology, lexicography, terminography, computational linguistics and software engineering, and on extensive experience in the lexicography and terminography of spoken language in the EAGLES and Verbmobil language engineering projects, and the Bielefeld–Abidjan "Encyclopédie des language de Côte d'Ivoire" documentation design project.

2 The lexicon *is* metadata

A lexicon is already a form of *corpus metadata* in the sense that it contains more or less generalised descriptive facts about a corpus or introspected data, and it was treated as such in [Gibbon & al. 1997], i.e. as "linguistic characterisation" of corpora.[1]

But lexicographers often speak of "lexical data" in the sense of the information in the lexicon itself. In this sense, a lexicon itself needs description in terms of a higher level of *lexical metadata*, designed

1. to distinguish between types of conventional lexicon, electronic hyperlexicon, terminological database, encyclopaedia;
2. to characterise lexical macrostructure (e.g. onomasiological vs. semasiological, word rank vs. idiom rank etc.);
3. to characterise lexical microstructure (i.e. types and dimensions of lexical information);
4. to characterise lexical mesostructure (i.e. generalisations about partial regularities in the lexicon);
5. to characterise development and application history; ...

There are numerous approaches to lexicon metadata characterisation and standardisation, from the accepted traditional norms used in typological linguistics (cf. [Coward & Grimes 1995]). to the technology oriented *de facto* standards work of the EAGLES project series[2] and the industry standard MARTIF (ISO 12200) for terminological databases, based on standard markup (ISO 8879 SGML). A modern XML version is under development (cf. ISO FDIS 12620); in this context, it is interesting to note that the traditionally automous practical engineering discpline of terminography is gradually accepting more general linguistically based lexicographic standards, though MARTIF extensions to general lexicography are not available.

Today the main foci in lexicography are often more on the standardisation of markup and implementation techniques than on conceptual harmonisation. But the more complex the issues — and in lexicography they are very complex — the harder it becomes even to think of documentation standards without looking at the broader picture of the other lexical sciences and the conceptual support they can provide to lexicography.

The present contribution takes a broader view of the position of the lexicon in this unsettled scene from the point of view of some small lexical objects and their properties. Section 3

---

[1] The ideas presented here owe a considerable debt to colleagues and students in many contexts, particularly to Bruce Connell and Firmin Ahoua, co–partners in the DOBES Ega project, for their ideas on lexicography for endangered languages. I am indebted to Stephen Bird for discussion over many years on hyperlexica and computational issues; to David Weber and Nicholas Oster for detailed discussion in a memorable locale after the Exploration 2000 conference on web–based language documentation and description, particularly on the dimensionality of lexical information; to many discussions on the lexicon with my EAGLES project partners, in particular Nicoletta Calzolari and John McNaught; to the inexorable insistence and hard–headed operational lexicographic requirements of my VerbMobil project partners; to the critical and well–informed particpants in the 1997 ELSNET summer school on computational lexicography in Leuven (cf. [van Eynde & Gibbon2000]); above all, to intensive discussions with many colleagues and students in Bielefeld and Abidjan on the formalisation of lexicography.

[2] URL: `http://www.ilc.pi.cnr.it/EAGLES/browse.htmlff`

Section 4 is concerned with characterising large and small lexical objects; in Section 5, a model for characterising types of lexical information is proposed, based on contemporary linguistic and media theory; Section 6 is concerned with the complexity of lexical information and additional, particularly pragmatic and operational dimensions of lexical information to be accounted for in metadata. Section 7 focusses on the status of hyperlexicon realisations of lexical documents in the context of the semiotic model, and, finally, Section 8 summarises the approach.

A number of aspects are excluded from consideration in this document, not because they are unimportant, but because they deserve separate, detailed treatment.

One aspect of standardisation which is only touched on in passing is the procedural or operational side of lexicography. There are two main aspects: first, the manual and/or computer–supported acquisition of lexical information (for example by conscious reflection on lexical objects or by corpus analysis), and second, access to lexical information in paper and electronic media, particularly in hypertext format. The question of acquisition of lexical data and lexical knowledge Section 7).

A further aspect which is excluded from present discussion is the ethics of lexicography, i.e. how lexical information is gathered, which culturally significant lexical information should be included, which lexical information should be disseminated publicly, how the commercial value of lexical information should feed back to the originators of the lexical information, intellectual property rights (IRP) on lexical information, including different forms of the same document.

3 Lexicon sciences and lexicon standards

Lexicon theory, descriptive lexicology and operational lexicography — the lexicon sciences — are old sciences and technologies, and use of computational modelling and large–scale corpus processing is rapidly leading to a convergence of these three areas. A general outline of the interrelations between these disciplines is shown in Figure 1.

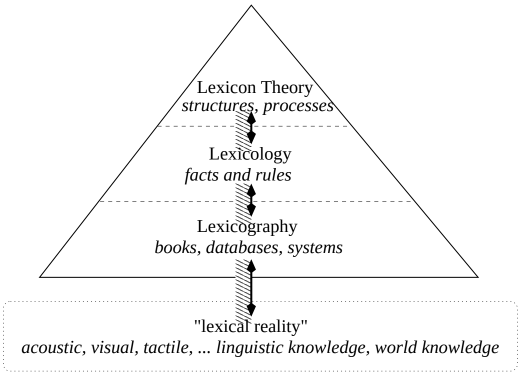

Lexicon Theory
structures, processes
Lexicology
facts and rules
Lexicography
books, databases, systems
"lexical reality"
acoustic, visual, tactile, ... linguistic knowledge, world knowledge

Figure 1: Relations between the lexicon sciences.

Underlying the approach presented in the present contribution is the idea that the complexity of lexicographic documentation — whether paper or electronic — has become so complex that all the lexicon sciences need to provide input to the development process if an unproductive kind of chaos is not to ensue. A number of approaches in the area of the human language technologies where this principle is practised are discussed in the contributions to [van Eynde & Gibbon2000].

It is a truism to state that archiving and documentation are inconceivable without standardisation — standardisation at levels which do not prejudice creative scientific and technological innovation. De facto standards have arisen over the past 10 years with the development of the PC in the context of the World Wide Web into a mass Information and Communication Technology product. Some 'standards' come and go, or develop too quickly to be regarded as standards except for a transitory period; examples of these are hardware configurations and software norms for media, text and multimedia documents. Other standards are more lasting, in particular those to do with design and quality control of archives, documents and systems (cf. [Gibbon & al. 1997], [Gibbon & al. 2000]). It is standards of this kind which fall into the area of *metadata* as opposed to being artefacts of specific archives or implementations.

4 Large and small lexical objects

Any inventarised form which may be abstracted from tokens of speech, inscriptions of text, or gestural events, including iconic and indexical signs as well as the conventional symbols, is a lexical object. Because of its generality, this is not a very useful definition as it stands, except to distinguish lexical objects from completely compositional, transparently interpretable complex signs. The definition encompasses a vast spectrum of objects, from the regular phonetic realisations of phonemes and prosodies to the constituents of handwriting and printed or electronic text, through morphemes, words (simplex or complex), phrasal idioms to entire anthologised texts. And there are weird lexical objects, too, such as hums and haws, coughs and tut–tuts, as well as a wide range of conventional, stylised and codified visual gesture systems, all of which have communicative functions which are closely related to the more central aspects of language.

Before proceeding, four central structural concepts for lexicon design will be introduced. Two of these are traditional, though modified for present purposes; the third is new, the fourth is currently topical in the area of language resources in general.

Lexicon macrostructure: The macrostructure of a lexicon is its overall structure or architecture, defined in terms of the arrangement of lexical objects, i.e. lexical entries. It may encompass different entry types, e.g. words vs. idioms, or different procedurally motivated structural optimisations, e.g. a function from form to meaning, semasiological macrostructure, or from meaning to form, onomasiological macrostructure (though the traditional semasiological–onomasiological distinction is inadequate in view of the complexity of lexical information as understood today).

Lexicon microstructure: The microstructure of a lexicon is the structure of the properties of the invidual lexical objects, i.e. the structure of the information associated with the lexical entries, from media information (pronunciation, orthography, gesture) through morphological and syntactic information to semantic, pragmatic information, information (via a concordance) on occurrence contexts in corpora, and record–level housekeeping metadata.

Mesostructure: The mesostructure of a lexicon is a set of generalisations about microstructures and macrostructures. In traditional lexica this consists of definitions of parts of speech, rules for spelling and pronunciation, etc., which are common to all entries, or at least to large classes of entries. In contemporary formal lexica it consists of a type or default hierarchy or other systems of implication relations.

Lexicon metadata: Lexicon metadata consist of (a) the lexicon mesostructure; (b) sources, such as examples and media (text, audio, graphic, video) data; (c) authoring data such as identity of lexicographers, dates of creation and modification; (d) specification of markup conventions and their interpretation.

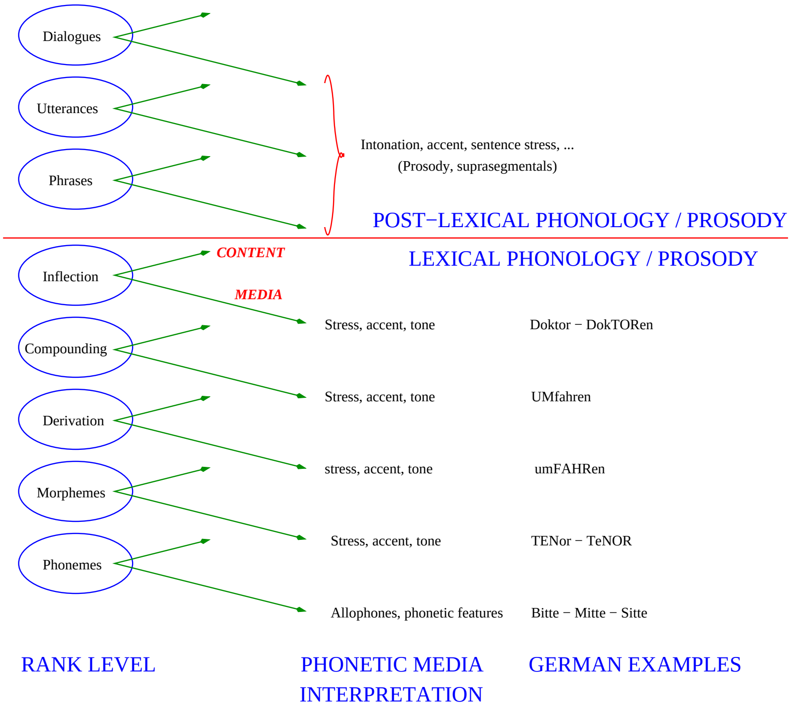


Figure 2: The rank hierarchy of lexical objects and their content and media semantics.

The first (declarative) aspect of macrostructure classifies lexical objects according to two main criteria:

Sorts of lexical object: Sorts of lexical object pertain to the level of abstraction involved: lemma (perhaps with some canonical headword representation) vs. stem vs. fully inflected form vs. conceptual category vs. transfer unit (in a multilingual lexicon) vs. ...

Ranks of lexical object: Inventarised lexical objects — in the generalised sense used here — differ in size, from the phoneme, the syllable, consonant clusters (*glare*, *gleam*, *glitter*, *glisten*, *glossy*, *glow*, ...) through the morpheme, lexeme, derived and compound stem at word level to phrasal, sentential idioms, fixed and ritual texts and conventionalised routine or liturgical dialogue. At the textual level, the distinction between text and lexical object becomes fuzzy: is an anthology of poetry a lexicon? It certainly has much in common with one.

Figure 2 illustrates the core *rank hierarchy* of conventional lexical objects, with which other lexical objects may be related via notions such as the prosodic hierarchy in speech, layout hierarchies in printed matter, and gestural hierarchies. Conventional lexical objects thus vary in rank from the very small (e.g. font characters and their parts) to the very large (e.g. a standard religious text).

The model shown in Figure 2 is too general to be very helpful when it comes to describing types of lexical information, but it is a useful start. In particular, in the world of multimedia documentation, the idea that a lexicon is basically concerned with *words* needs to be scotched once and for all. A phraseological unit, for instance, is a lexical object in its own right, at its own rank, and not only by virtue of the words it contains; listing idioms by words is a matter of procedural convenience, not of conceptual clarity, and has led to much confusion in linguistics over the past 40 years. Likewise, an image or a sound may be a lexical object.

Lexicon macrostructure is determined not only by the rank hierarchy of large and small lexical objects and their interpretation, but also, and traditionally more typically, in terms of procedural orderings of lexical microstructure.

We will not discuss macrostructure or mesostructure further in this document; We stipulate that

- data models for lexical databases must permit the transformation (indexing, automatic re–structuring) of the database for different views corresponding to different macrostructures, and that
- mesostructures are best dealt with separately in the context of linguistic descriptions.

It will be sufficient at this stage to propose providing lexical metadata at two levels (cf. also [Coward & Grimes 1995] for a practical approach):

1. the macrostructure level: acquisition of the entire lexicon (e.g. source, date, coordinating lexicographer, modification history);
2. the microstructure level: specific lexical entries (also source, data, lexicographer, modification history).

5 Microstructure: types of lexical information

The conventional view of types of lexical information was formulated in a classic article ([Fillmore 1971]). Types of lexical information in this sense underlie the *microstructure* of a lexicon.

Figure 3 visualises a contemporary semiotic model of relations between levels of abstraction for the description of signs. Lexicon microstructures are typically represented in some kind of vector format, for example:

- a record structure in a relational database (most lexical database).
- a list of linked objects such as paragraphs or files (hypertext lexicon);
- a list, perhaps numbered, perhaps mildly hierarchical with a lemma or headword and polysemous, homophonous, homographic or categorial variants (traditional dictionary);
- a feature vector (traditional linguistic theory);
- an attribute–value structure, possibly hierarchical (contemporary linguistics theory);

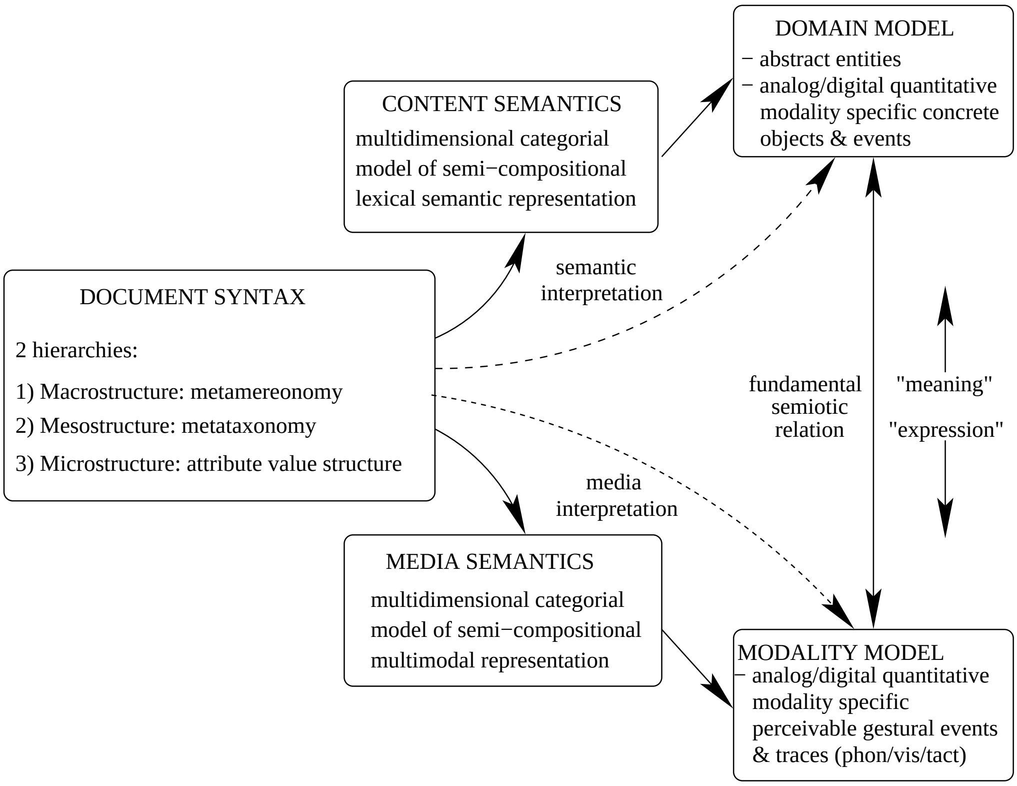

DOMAIN MODEL
− abstract entities
− analog/digital quantitative modality specific concrete objects & events
CONTENT SEMANTICS
multidimensional categorial model of semi−compositional lexical semantic representation
DOCUMENT SYNTAX
2 hierarchies:
1) Macrostructure: metamereonomy
2) Mesostructure: metataxonomy
3) Microstructure: attribute value structure
semantic interpretation
media interpretation
fundamental semiotic relation
"meaning"
"expression"
MEDIA SEMANTICS
multidimensional categorial model of semi−compositional multimodal representation
MODALITY MODEL
− analog/digital quantitative modality specific perceivable gestural events & traces (phon/vis/tact)

Figure 3: A semiotic model of document structure with content and media semantics.

- a generalised attribute–value structure in a type or default hierarchy (inheritance or object–oriented lexicon).

There are also other important issues to do with lexicon microstructure. One of these is the representation of *lattice–structured* or *multilinear* information which receives a media interpretation of *simultaneity* rather than *sequentiality*. This issue applies immediately to the representation of

- word–level stress, pitch accent, lexical tone and other prosodies;
- phrasal (and larger) size lexical prosodies, as in greetings;
- accompanying gestural behaviour;
- autonomous gestural symbols, as in waving or in sign languages.

A thorough discussion of lexical information which is interpreted as simultaneity relations at different levels is given in [Carson–Berndsen 1998], based on some principles of Event Phonology, as first formulated in [Bird & Klein 1989], and on Prosodic Time Types formulated in [Gibbon 1992]. At the level of resource implementation, the annotation lattice approach of [Bird & Liberman 1999] is clearly relevant as a partial solution to this problem.

The basic semiotic model has a theoretically sound basis, but also a heuristic value as a structured ‘checklist’ for types of lexical information. On this view, the components of the semiotic model

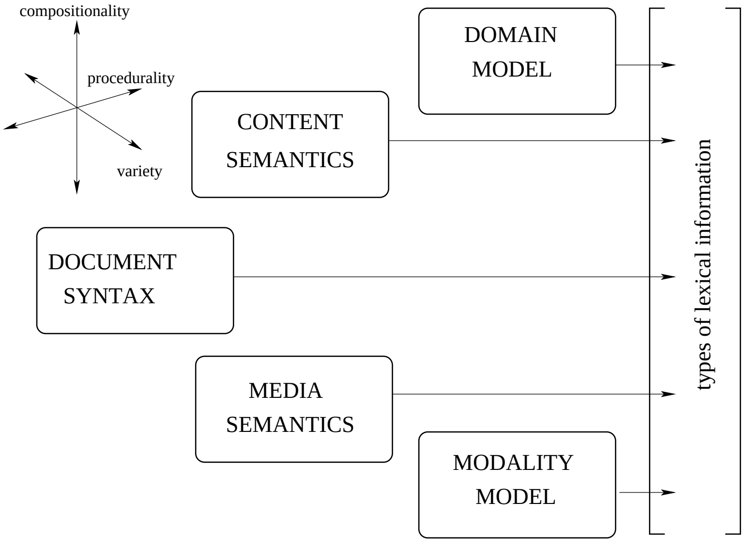


Figure 4: Projection of semiotic model into a lexicon microstructure vector.

are projected on to a microstructure of types of lexical information, whether in a the lexicological context of a linguistic description, or in the lexicographic context of a paper or electronic lexicon. This microstructural vector contains media–oriented types of information on orthography and pronunciation (perhaps also on other gestural properties) as well as other kinds of operational information concerned with lexical acquisition and lexical access keys.

Whatever formalism, abstract data structure or concrete format is selected, this idea of mapping a semiotic model into a microstructure vector, visualised in simplified form in Figure 4, is an essential defining characteristic of a well–defined lexicon, illustrating the theoretical linguistic basis for lexical metadata.

6 Further dimensions of lexical information

The conventional kinds of lexicon microstructure, even modelled at different rank levels as discussed above in connection with lexicon microstructure, are only sufficient for creating lexical resources of a standard language — the type of lexicon suited to current standard language oriented speech technology, or, in more jocular terms, to the Scrabble player.

Embedded in Figure 4 is a small diagramme showing three additional dimensions to which the main types of lexical information need to be generalised: compositionality, variety, and procedurality. In principle, the types of lexical information need to be multiplied in order to cope with these additional types; traditional microstructures have an ad hoc combination of these.

Figure 4 elaborates on the theme of dimensionality: each of the dimensions described so far can be further analysed, fractal–like, into subdimensions; three dimensions are chosen to represent the higher dimensionality in each case more on associative than on principled grounds:

1. Variety: a central question in lexicography is its domain in terms of language varieties defined in terms of situational or contextual consitutive factors. There are many dimensions

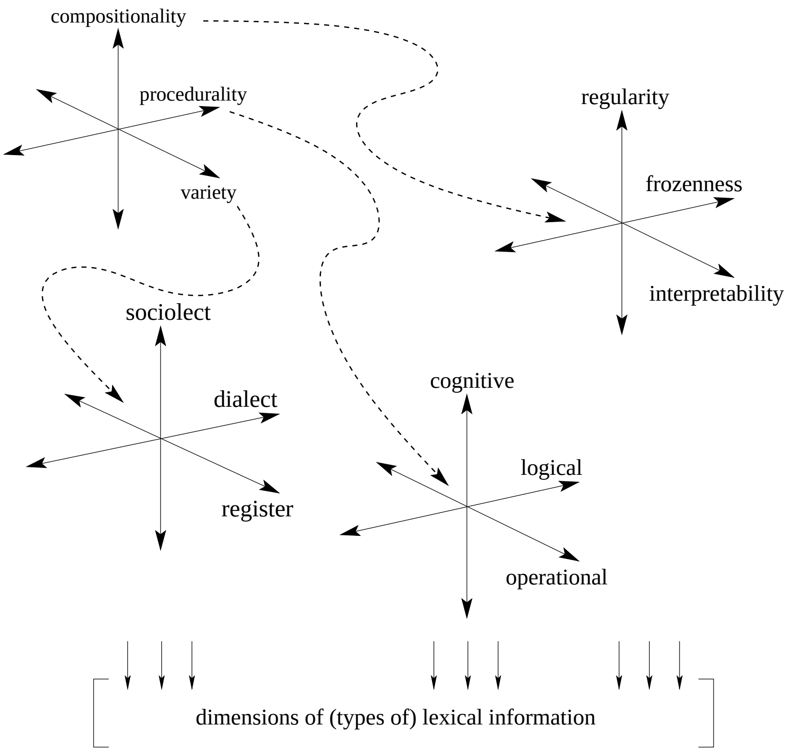


Figure 5: Increasing dimensionality of lexical information.

of variation, both historical and synchronic, and the terminology for language variation is extensive, ranging from well–known terms such as *dialect* and *sociolect* through *style* and *genre*, and to more technical terms such as *register* and *sublanguage*. The dimension of variety fractions out into three main dimensions shown in Figure 4:

(a) Dialect, defined by regional parameters, with transitions between neighbouring dialects often parametrically defined in terms of different (and not always co–extensive) isoglosses.

(b) Sociolect, defined by biological and social parameters such as sex or gender, age or maturity class, position in conventional political, administrative, economic or religious hierarchies.

(c) Register, defined by functional parameters of language use in different situated activities, different situations, different channels.

Terms such as 'genre' and 'style' are sometimes used to cover special aspects of register, sometimes to cover both the register and sociolect dimensions.

The historical dimension, which classifies genetically related diachronic forms of a language, is defined in terms of projections of these three dimensions onto a long–term temporal axis of language change, under the influence of internal change and external influence of other languages and varieties.

2. Compositionality: the central structural dimension of language, defining an interplay between inventories of basic units of various sizes and the complex structures within which they relate to each other. The three main subdimensions of compositionality are determined by hierarchicality, idiomaticity, and semantic interpretation:

(a) Regularity of form, defined in terms of different sized hierarchical 'ranks' or 'layers' of structure, such as dialogue, text, sentence, word, and in partially hierarchical constituent structures at each of these ranks.

(b) Frozenness of lexical items, defined in terms of finite vocabularies of lexicalised items at each rank, from dialogue rituals through fixed textual and sentential expressions (idioms) to complex and simplex words.

(c) Interpretability of meanings, defined in terms of semantic transparency and opacity, influenced by factors such as metaphor, ellipsis, idiomaticity, and conventionality or rituality of use.

It is the inverse of compositionality which defines the lexicon: the lexicon is a finite repository of items graded in terms of hierarchies of relative irregularity, relative frozenness, and relative opacity.

3. Procedurality: constraints on the processes of acquisition of lexical data and lexical information, and on the access to lexical data and information from different epistemological points of view:

   (a) Cognitive: defined in terms of the 'mental lexicon', with neurologically plausible data structures which can be constructed and accessed by procedures which have cognitively plausible temporal and storage properties. The cognitive dimension is also heuristically important in lexicon documentation, for example in perceptual tests or tests of mutual intelligibility to validate lexical information.

   (b) Logical: defined in terms of lexical information as axioms, theorems, and inference procedures involving lexical mesostructures, from the minimally procedural 'declarative lexicon' involving only one inference rule, such as *modus ponens* to the more 'procedural lexicon' with complex 'lexical rules' which have a variety of logical properties.

   (c) Operational: defined in terms of computable abstract data structures and appropriate generalisation or specialisation (learning) and search algorithms, leading to format and tool specifications for appropriate tool specifications.

   The first two of these three sub–dimensions are concerned with the static properties of lexica and lexical items, the third is concerned with their dynamic properties in the context of a 'learning' or 'remembering' system.

$$\begin{bmatrix} \text{Variety:} & \begin{bmatrix}\text{Dialect: ...}\\ \text{Sociolect: ...}\\ \text{Register: ...}\end{bmatrix}\\ \text{Compositionality:} & \begin{bmatrix}\text{Regularity: ...}\\ \text{Frozenness: ...}\\ \text{Interpretability: ...}\end{bmatrix}\\ \text{Procedurality:} & \begin{bmatrix}\text{Cognitive: ...}\\ \text{Logical: ...}\\ \text{Operational: ...}\end{bmatrix} \end{bmatrix} \Rightarrow \begin{bmatrix}\text{microstructure field}_1\\ \text{...}\\ \text{...}\\ \text{...}\\ \text{...}\\ \text{...}\\ \text{...}\\ \text{...}\\ \text{microstructure field}_n\end{bmatrix}$$

Figure 6: Lexical properties as AVS mapped to flat microstructure.

The fractioning of lexical dimensions does not stop here; lexical macrostructures may involve more complexity than just varietal mappings, lexical microstructures may involve more complexity than structural information (but including any or all of the information types involved in the parameters discussed above), and the specification of an operational lexicon is a complex task indeed.

For any given task in lexical documentation not all of these dimensions and sub–dimensions will be relevant. However, from the point of view of the systematic definition of granularity levels for lexical metadata, the fractal–like characterisations given here, abstract though they may seem at first glance, provide a useful starting point.

In a formal specification for document technology purposes, the dimensions of lexical information may be represented as a recursive attribute–value structure (AVS), as shown in Figure 6.

7 Aspects of lexical documentation as hypertext

In this section, one type of document organisation which has consequences for operational lexica is discussed: hypertext (and related notions such as hypermedia, hyperdocument).

The concept of hypertext, and thus also of hyperlexicon, is a presentation level concept, derivable from a more fundamental lexical document structure by means of a media interpretation function.

The file split and hyperlinking functions of a hypertext are comparable to the procedure used for printers' make–up (pagination, line and page breaks, index and table of contents page references, footnoting and endnoting). Hypertexts are sometimes defined as 'non–linear' in structure, in contrast to conventional texts; rarely, however, is the level of definition specified. Many texts are non–linear (hierarchical, tree or graph–structured) at the *document syntax* level and no doubt (if the meaning were sufficiently precisely specified) also st the *document semantics* level. Linearity here means only the relatively trivial property of a full sequential ordering of printed pages at the presentation (media interpretation) level. But we all know that books, especially reference books, are frequently neither written nor read sequentially. Conversely, there is nothing to prevent a hypertext from being organised sequentially, with each page having only one link to a successor page. And again, there is nothing to stop us from producing or accessing such a hypertext non–linearly. The semiotically based five–component document model is more complex than the usual 'logical structure' vs. 'rendering' dichotomy of document specifications in much of the hypertext literature, and helps us to avoid such simplistic characterisations.

Summarising: the five component semiotic model introduced in the present contribution locates hypertext at the level of MEDIA SEMANTICS; the distinction between hypertext description and graphical or textual hypertext rendering is captured by the additional MODALITY MODEL component.

In 1995, the concept of a hyperlexicon on the web was explicitly introduced by the author as a database integrity–preserving technique and further developed during the following years (cf. [Gibbon & Lüngen 2000]).[3] The Verbmobil VM–HyprLex website was one of the first very large–scale CGI database applications on the World–Wide Web, and provided a single–token, simultaneous multiple–access shared database for the 30 or so laboratories around the world who were members of the VerbMobil consortium. The lexicographic task was to standardise and integrate approximately 25 types of lexical information which were made available by the partners in a variety of non–standardised formats.

---

[3] http://coral.lili.uni-bielefeld.de/VM-HyprLex/

The extensional coverage (number of entries) is 10000, and the intensional coverage (number of types of lexical information) is 25 (varying with different applications); a number of different search strategies, including regular expressions (restricted to prevent overloading the download channel) and formatting types. Further applications of the hyperlexicon principle have been developed outside the Verbmobil project.[4] The HyprLex approach is multi–level:

1. content structure is defined in a database;
2. document structure is defined with an inheritance network formalism;
3. the hypertext lexicon, with integrated online help and dynamically generated on the fly sublexica for the extraction of phonetically similar subsets, concordance, etc., was generated automatically in HTML format from the document structure; in later versions, an intermediate formatting language (IKE) was used.

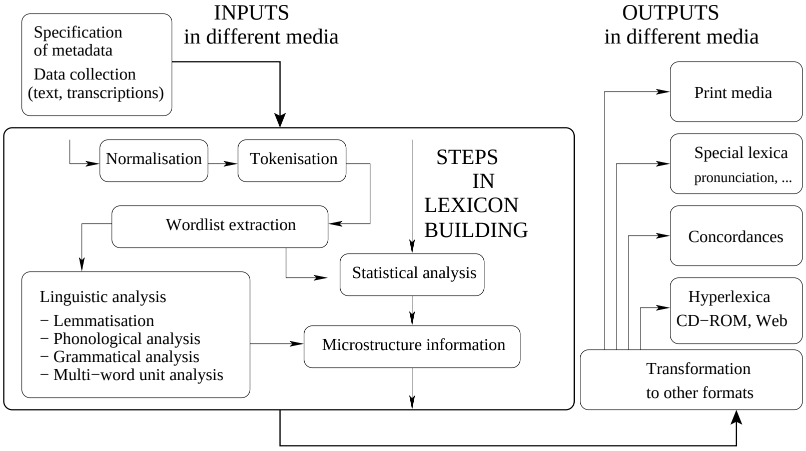


Figure 7: Generalised VM-HyprLex lexicographic logistics.

Figure 7 shows a generalised perspective on the logistics of the VM–HyprLex lexicographic task, which is applicable to a wide range of lexicographic tasks in language documentation. The VM–HyprLex source code and databases are available via the BAS and ELRA dissemination agencies.

In language documentation, the most well–known hyperlexicon is Bird's Hyperlex (cf. [Bird 1997]).[5] Hyperlex has been applied to a number of languages, in the area of endangered languages notably by Connell to Mambila and Amith to Nahuatl.[6]

Hyperlex bears some similarities to HyprLex, in that it has a CGI–based search concept, and integrates other on the fly calculations into the lexicon via CGI routines; the degree of integration of these add-ons is higher than with the VM–HyprLex application. Both hyperlexicon interfaces offer search and display filters over the microstructure entries of their lexical databases.

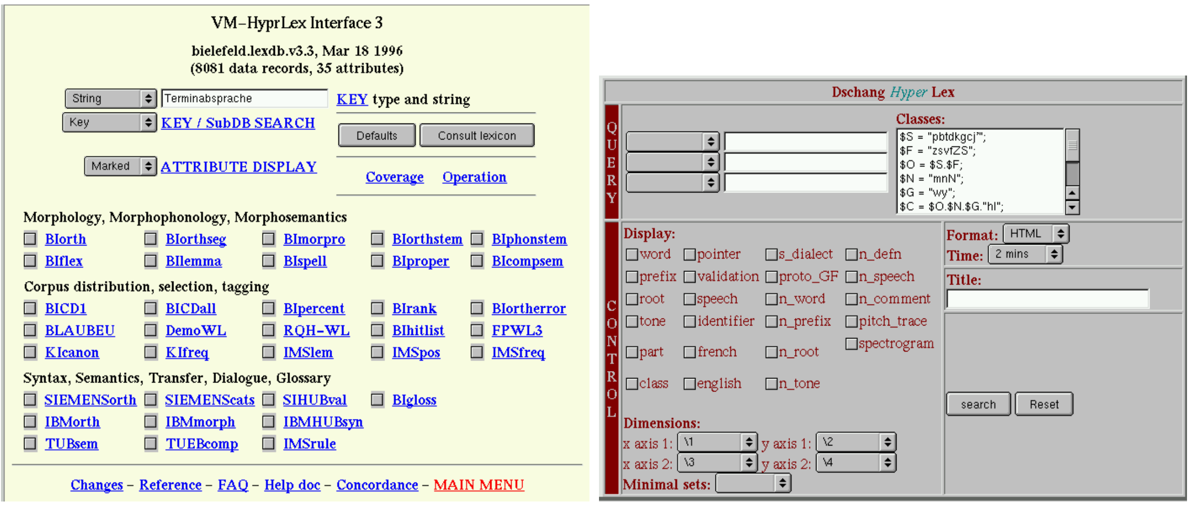
The HyprLex approach was further developed by Gibbon & Trippel in [Gibbon & Trippel 2000] in the domain of terminological lexicography, using a textual database for generating a variety of

---

[4] URL: http://coral.lili.uni-bielefeld.de/HyprLex/

[5] URL: http://morph.ldc.upenn.edu/hyperlex/

[6] The Mambila and Nahuatl hyperlexica (among others) are at the Hyperlex URL.



VM–HyprLex Interface 3

bielefeld.lexdb.v3.3, Mar 18 1996
(8081 data records, 35 attributes)

String: Terminabsprache

Key

Marked

Classes:

```
$S = "pbtdkgcj'";
$F = "zsvfZS";
$O = $S.$F;
$N = "mnN";
$G = "wy";
$C = $O.$N.$G."hl";
```

Format: HTML

Time: 2 mins

x axis 1: \1

y axis 1: \2

x axis 2: \3

y axis 2: \4

KEY type and string
KEY / SubDB SEARCH
Defaults
Consult lexicon
ATTRIBUTE DISPLAY
Coverage
Operation
Morphology, Morphophonology, Morphosemantics
☐ BIorth ☐ BIorthseg ☐ BImorpro ☐ BIorthstem ☐ BIphonstem
☐ BIflex ☐ BIlemma ☐ BIspell ☐ BIproper ☐ BIcompsem
Corpus distribution, selection, tagging
☐ BICD1 ☐ BICDall ☐ BIpercent ☐ BIrank ☐ BIortherror
☐ BLAUBEU ☐ DemoWL ☐ RQH–WL ☐ BIhitlist ☐ FPWL3
☐ KIcanon ☐ KIfreq ☐ IMSlem ☐ IMSpos ☐ IMSfreq
Syntax, Semantics, Transfer, Dialogue, Glossary
☐ SIEMENSorth ☐ SIEMENScats ☐ SIHUBval ☐ BIgloss
☐ IBMorth ☐ IBMmorph ☐ IBMHUBsyn
☐ TUBsem ☐ TUEBcomp ☐ IMSrule
Changes – Reference – FAQ – Help doc – Concordance – MAIN MENU
Dschang *Hyper* Lex
QUERY
CONTROL
Display:
☐word ☐pointer ☐s_dialect ☐n_defn
☐prefix ☐validation ☐proto_GF ☐n_speech
☐root ☐speech ☐n_word ☐n_comment
☐tone ☐identifier ☐n_prefix ☐pitch_trace
☐spectrogram
☐part ☐french ☐n_root
☐class ☐english ☐n_tone
Title:
search
Reset
Dimensions:
Minimal sets:

Figure 8: VM-HyprLex (Gibbon) and Dschang-Hyperlex (Bird) interfaces.

media interpretations. The same approach was used in generating the different media involved in the publication of [Gibbon & al. 1997] and [Gibbon & al. 2000].

8 Steps towards lexicon standardisation

So, in conclusion, why all this background discussion of the principles of lexical organisation?

The answer can be stated in terms of a few basic principles:

1. Lexicography today is in the intersection between of software engineering, computational linguistics, linguistic lexicography, and industrial terminography; consequently the requirements specification, design, implementation, evaluation, documentation and maintenance standards of these disciplines need to be met.

2. In very basic declarative terms, lexical structure is conveniently seen as two dimensional: the macrostructure as a rank hierarchy and the microstructure as a vector of atomic information (nothing implied here about the ontology of these atoms).

3. In procedural terms, alternative operational macrostructures can be defined as a *semasiological* mapping from media information to content information (the conventional dictionary or encyclopaedia), as an *onomasiological mapping* from content information to media information (as in the conventional thesaurus), or indeed in any other function from some combination of lexical properties to sets of lexical objects, or to other lexical properties.

4. From the database engineering point of view, any of the alternative lexical macrostructures and lexical microstructures can be seen as the foundation for the design of a *database view*, implemented systematically as specific optimal indexings of one and the same database, with specific output filters and formatting.

5. The Web is a special case of a database with

   (a) an associative data model, i.e. a more general data model than the model which underlies current relational or object–oriented databases;
   (b) simultaneous orthogonal views of the database;
   (c) arbitrary cross–linking not only between lexical microstructure elements and lexical macrostructure elements, but also between views.

6. Lexical mesostructure can be included as on–line help, as in VM–HyprLex, or as on the fly generalisations over lexical information, as in both VM–HyprLex and in Bird's HyperLex.

Faced with this plethora of possibilities we advocate an explicit return to a semiotic model of the lexicon which can be incrementally extended according to fundamental linguistic and operational principles until coherent design strategies for lexical database views can be clearly defined, and hypermedia lexica can be derived automatically. A start may be made by defining a rank hierarchy of lexical objects, and a procedurally neutral microstructure on accepted linguistic typological principles, with definitions of generic metadata as a well–structured mesostructure, in addition to traditional forms of metadata.

We suggest that the next steps are along two dimensions:

The first dimension is to formulate a system design document (including the human factors of lexicographers in the field, the office and the lab, and users of various kinds), an implementation document (with open source after the alpha evaluation stages), an evaluation document, and a maintenance strategy (also including the human factors), based on an overall documentation policy.

The second is to filter out requirements specifications for specific branches of lexicography (for example in the contexts of the scientific documentation of endangered languages, as well as its applications in terminography, language teaching, and other information services), and to derive specific designs from these specifications and the general design document, specific implementations and evaluation procedures.

References

[Bird 1997] Bird, Steven (1997). A lexical database tool for quantitative phonological research. Computation and Language, abstract cmp-lg/9707011.

[Bird & Klein 1989] Bird, Steven & Ewan Klein (1990). Phonological Events. In: *Journal of Linguistics* 26: 33–56.

[Bird & Liberman 1999] Bird, Steven & Mark Liberman (1999). A formal framework for linguistic annotation. Technical Report MS-CIS-99-01, Computer and Information Science, University of Pennsylvania.

[Carson–Berndsen 1998] Carson–Berndsen, Julie (1998). *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition.* Dordrecht & Boston: Kluwer Academic Publishers.

[Coward & Grimes 1995] Coward, David F. & Charles E. Grimes (1995). *Making Dictionaries: A guide to lexicography and the Multi-Dictionary Formatter.* Waxhaw, NC: Summer Institute of Linguistics.

[van Eynde & Gibbon2000] van Eynde, Frank & Dafydd Gibbon eds. (2000). *Lexicon Development for Speech and Language.* Boston & Dordrecht: Kluwer Academic Publishers.

[Fillmore 1971] Fillmore, Charles (1971). Types of lexical information. In: Danny D. Steinberg & Leon A. Jakobovits, eds. *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology.* Cambridge: Cambridge University Press.

[Gibbon 1992] Gibbon, Dafydd (1992). Prosody, time types and linguistic design factors in spoken language system architectures. In: G. Görz, ed., *KONVENS 92, 1. Konferenz "Verarbeitung natrlicher Sprache"*, Berlin: Springer–Verlag.

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

[Gibbon & Lüngen 2000] Gibbon, Dafydd & Harald Lüngen (2000). Speech lexica and consistent multilingual vocabularies. In: Wolfgang Wahlster, ed., *Verbmobil: Foundations of Speech–to–Speech Translation.* Berlin: Springer Verlag.

[Gibbon & Trippel 2000] Gibbon, Dafydd & Thorsten Trippel (1999). A multi-view hyperlexicon resource for speech and language system development. In *LREC Proceedings 2000.* Athens, Greece.

[Gibbon & al. 2000] Gibbon, Dafydd, Inge Mertins & Roger Moore, eds. (2000). *Handbook of Multimodal and Spoken Language Systems: Resources, Terminology and Product Evaluation.* Boston & Dordrecht: Kluwer Academic Publishers.