On lexical objects and their properties

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

Paper presented at the workshop on
Web-Based Language Documentation and Description
12-15 December 2000, Philadelphia, USA.

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 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.

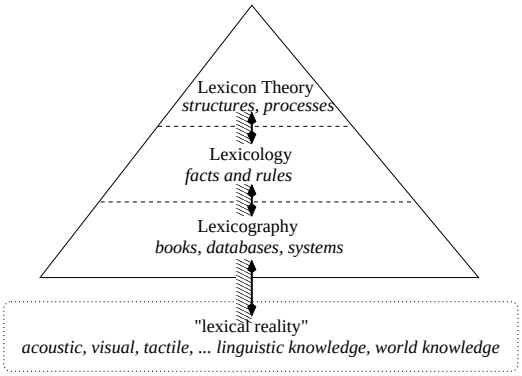

Figure 1: Relations between the lexicon sciences.

Underlying the approach presented in the present contribution is the idea that the complexity of lexicographic documentation — whether Web–based or not — has become so

complex that all the lexicon sciences need to provide input to the development process if the unproductive variety 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.

A lexicon is already a form of 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.

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 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 approaches to lexicon metadata characterisation and standardisation, from the accepted traditional norms used in typological linguistics (cf. [Coward & Grimes 1995]). to the technology oriented work of the EAGLES projct series and the MARTIF ISO terminological standards.

Today the focus in lexicography is 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 2 is concerned with characterising large and small lexical objects; in Section 3, a model for characterising types of lexical information is proposed, based on contemporary linguistic and media theory; Section 4 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 5 focusses on the status of hyperlexicon realisations of lexical documents in the context of the semiotic model, and, finally, Section 6 summarises the approach.

2 Macrostructure: 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.

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.

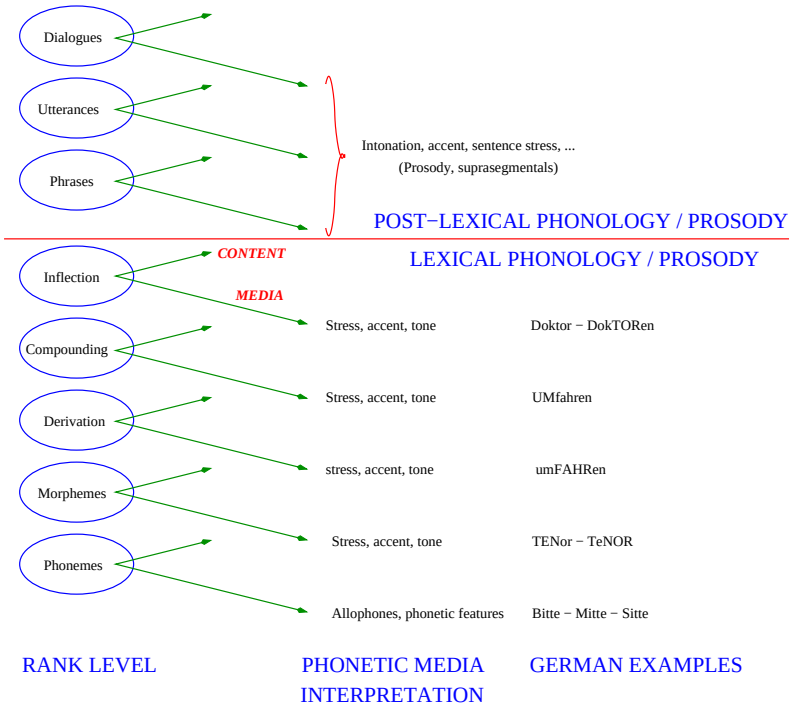

Figure 2: The rank hierarchy of lexical objects and their content and media semantics.

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 betwen 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.

3 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.

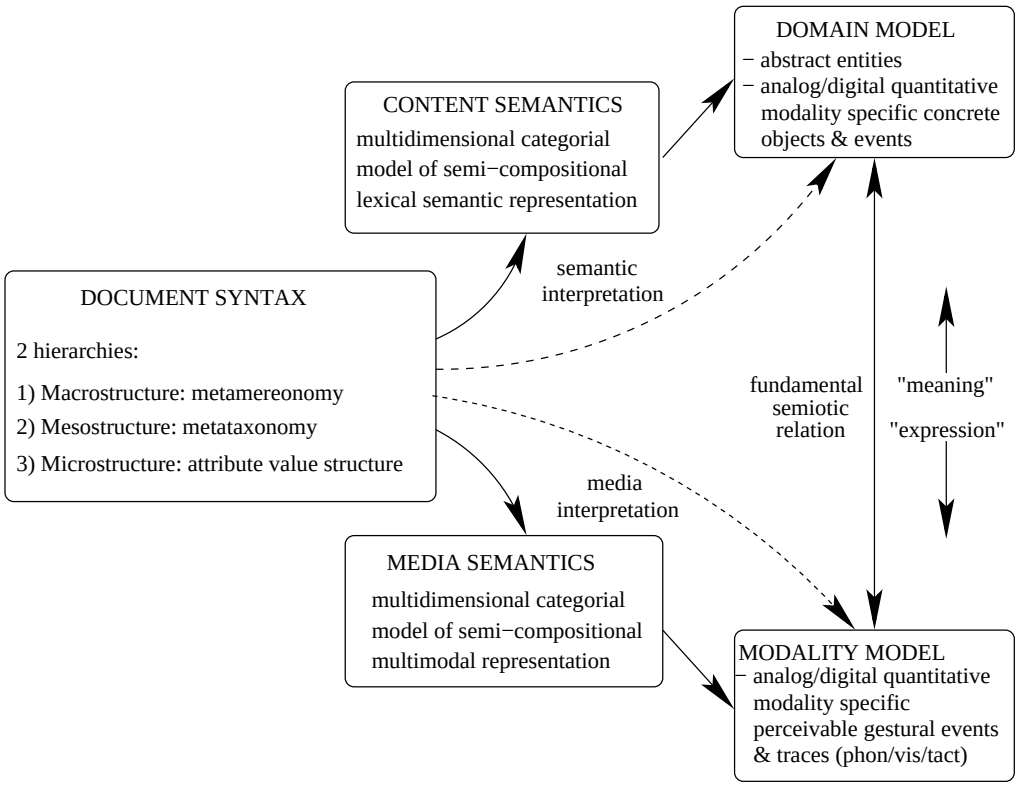

Figure 3: A semiotic model of document structure with content and media semantics.

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);
- 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 lager) 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 implmementation, the annotation lattice approach of [Bird & Liberman 1999] is clearly relevant as a partial solution to this problem.

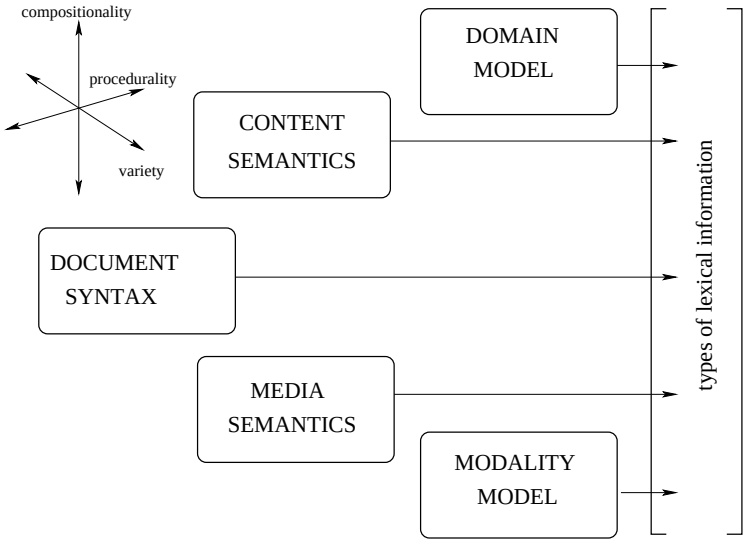

Figure 4: Projection of semiotic model into a lexicon microstructure vector.

Whatever formalism, abstract data structure or concrete format is selected, a mapping of the data model into a microstructure vector, visualised in simplified form in Figure 4, provides additional layers of structure for theoretical, heuristic and database implementation purposes.

4 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 multiplied in order to cope with these additional types; traditional microstructures have an ad hoc combination of these.

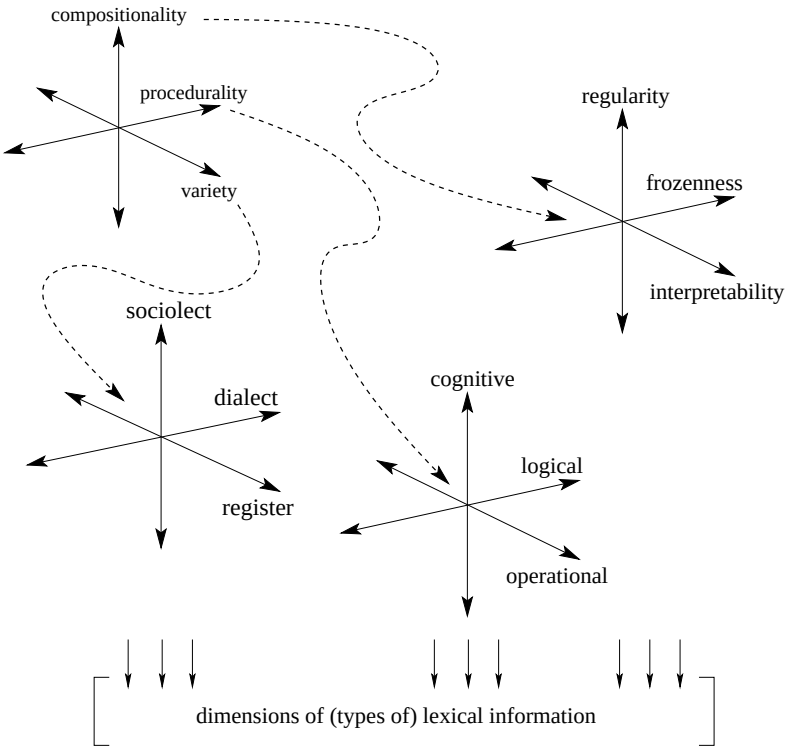

Figure 5: Increasing dimensionality of lexical information.

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. Nor does the fractioning process stop here.

5 On classifying hypertext lexica

The concept of hypertext, and thus also 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 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). 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 as a database integrity–preserving technique (cf. [Gibbon & Lüngen 2000]). 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.

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. The VerbMobilHyprLex (1995-1996) can be visited at http://coral.lili.uni-bielefeld.de/VM-HyprLex/. Further applications of the hyperlexicon principle are to be found at http://coral.lili.uni-bielefeld.de/HyprLex/. 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.

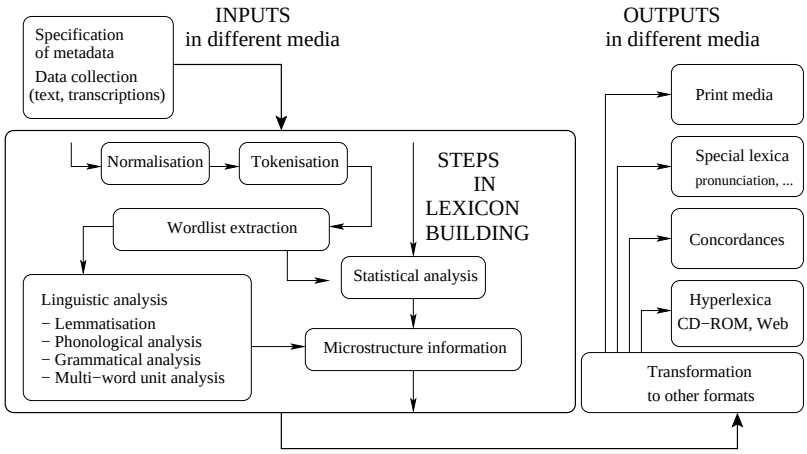

Figure 6: Generalised VM-HyprLex lexicographic logistics.

Figure 6 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.

In language documentation, the most well–known hyperlexicon is Bird's HyprLex (1998), to be visited right here at http://morph.ldc.upenn.edu/hyperlex/. Bird's HyperLex bears some similarities to Gibbon's VM-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.

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

6 Steps towards lexicon standardisation

So, in conclusion, why all this background discussion of the principles of lexical organisation when this workshop is concerned with web–based documentation only?

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

1. 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).

2. 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.
3. 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.
4. 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.
5. 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 I advocate a 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.

References

[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-Dictioary 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 hyper-lexicon 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.