A multi-view hyperlexicon resource for speech and language system development

Dafydd Gibbon, Thorsten Trippel

Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Postfach 100 131, D–33501 Bielefeld, Germany
{gibbon, ttrippel}@spectrum.uni-bielefeld.de

Abstract

New generations of integrated multimodal speech and language systems with dictation, readback or talking face facilities require multiple sources of lexical information for development and evaluation. Recent developments in hyperlexicon development offer new perspectives for the development of such resources which are at the same time practically useful, computationally feasible, and theoretically well–founded. We describe the specification, three–level lexical document design principles, and implementation of a MARTIF document structure and several presentation structures for a terminological lexicon, including both on demand access and full hypertext lexicon compilation. The underlying resource is a relational lexical database with SQL querying and access via a CGI internet interface. This resource is mapped on to the hypergraph structure which defines the macrostructure of the hyperlexicon.

1. Overview

New generations of integrated multimodal speech and language systems with dictation, readback or talking face facilities require multiple sources of lexical information for development and evaluation. Recent developments in hyperlexicon development offer new perspectives for the development of such resources which are at the same time practically useful, computationally feasible, and theoretically well founded. We describe an interactive hyperlexicon for on demand parametrised lexical access and full hypertext lexicon compilation with multiple views of lexical resources. The underlying resource is a relational lexical database with SQL querying and access via a CGI internet interface. This resource is mapped on to the hypergraph structure which defines the macrostructure of the hyperlexicon. Fuller discussion of lexical and terminological resources for spoken language systems can be found in (Gibbon et al., 1997a) and (Gibbon et al., 1999 forthcoming).

2. Specification

In development environments for the new generation of multimodal systems there is a growing need for processing many simultaneous and heterogeneous sources of lexical knowledge representing acoustic and visual signal information, references to marked up corpora, and symbolic information at many levels of linguistic representation. Advanced techniques for processing data with multiple tiers of annotation, often in high-level markup languages such as XML, are currently under development at a number of centres.

Here we address the complementary problem of user access to complex information of this kind, and develop a hyperlexicon interface for such resources as a first step towards creating a standardisable lexicographic resource. We contend that the notion of a document is too complex to be dealt with using purely application oriented decisions, and consequently we base our approach on a theoretical definition of documents as complex signs, using the term 'sign' in the same sense in which it is used for smaller linguistic units such as *word* and *sentence* in current linguistic theory. Advances in computational techniques have led to a rapprochement between lexicon theory in computational linguistics and large–scale corpus–based computational lexicography, minimising possible theory–application conflicts in the present case. In the present approach, a complete *document representation* contains

1. a *text syntax* or document structure DS, specifying (a) the category or context of occurrence of the document and (b) the parts (immediate constituents or daughters) of a document;
2. a *text interpretation* or interpretation structure IS, specifying on semiotic grounds an interpretation of DS in terms of a Content Structure CS and a Presentation Structure PS (analogous to the semantic and phonetic interpretation pair of mainstream linguistic theories).

These distinctions are not made simply on theoretical grounds, however, but in order to provide a specification for practical lexicographic implementation. In the case of a lexicon,

1. DS specifies the structure of a lexical database, e.g. relational, object–oriented, SGML/XML, or as a lexical knowledge base in DATR or another LKRL (lexicon knowledge representation language), its context (e.g. in relation to a corpus, a grammar, other lexica), its parts, and their contexts.
2. IS specifies

   (a) CS in terms of category definitions for the fields of the lexicon (e.g. 'types of lexical information'),
   (b) PS in terms of a superimposed set of relations over document constituents defining 'surface realisations' as

      • a database view with appropriate front end,

- a hyperlexicon in hypertext format in an on-line help environment, on CD–ROM, on the web, in clipped version for WAP or a PDA,
- a printed book,
- ...

The existing resource on which the present work is built is a relational database, with lexical microstructure at DS level as a database record structure, and fields representing underlying types of lexical information which are relevant for IS (CS and PS) interpretation (Gibbon et al., 1999). We do not address the issue of CS semantic representations but take these to be implicit in the user's understanding of his or her domain. The database concerned has been in regular use on the web with a JavaScript/HTML form interface for SQL queries some two years.

The domain selected for this task is spoken language and multimodal system terminology, as required by an intelligent software agent for potential developers and users. The domain was selected because it also defines an an important resource type for HLT (human language technologies) system development, the technical terminology of the field.

Although we adhere to ISO specifications for terminology, practical experience in HLT and recent developments in lexicology theory show that there are fundamentally untenable constraints involved in current ISO definitions, e.g.

1. The inherently procedural onomasiological view of terminological lexicography, i.e. a direct mapping from concepts to terms, is an obstacle in the way of flexible lexicographic views (and, though adhered to in theory for pure terminologies, generally flaunted in practical termbanks thanks to the quasi–isomorphic relation postulated between concepts and wordforms).

2. The identification of keys with concepts is just as much a confusion of categories in onomasiological dictionaries as the identification of keys with spelling in conventional semasiological dictionaries. Since this $key = concept$ assumption only holds within one homogeneous discipline and domain, in the multidisciplinary domain of HLT in which concepts evolve dynamically, approaches change, and conceptual structures are hybrid, it fails seriously.

We resolve these conflicts by applying the distinction between underlying DS, conceptual CS and surface PS representations (Gibbon et al., 2000, forthcoming), taking a declarative stance in regarding DS as neutral between CS and PS.

3. Hyperlexicon design and implementation

Following the requirements outlined above, we introduce more detail than previous approaches to hypertext document characterisation, and informally define (cf. also Figure 2):

Document Representation (DR): $< DS, IS >$

Document Structure (DS):
$< CATEGORY, PARTS >$

Interpretation Structure (IS):
$< CONTENT, PRESENTATION >$

CATEGORY: the location of a document or document part in a larger structure, e.g. in a library, on a web site, or in some other archive;

PARTS: a part of a document, such as a table of contents, chapter, index, lexical entry;

CONTENT: a real–world semantic domain into which Document Structure is mapped in document planning and understanding (in the case of a terminological lexicon, for instance, the set of objects, relations, states and events in the technical domain concerned);

PRESENTATION: a visual and/or acoustic real–world domain into which Document Structure is mapped in document production and perception (in the case of a terminological lexicon, for instance, as a database front–end, a hypertextual web site, a conventional paper dictionary).
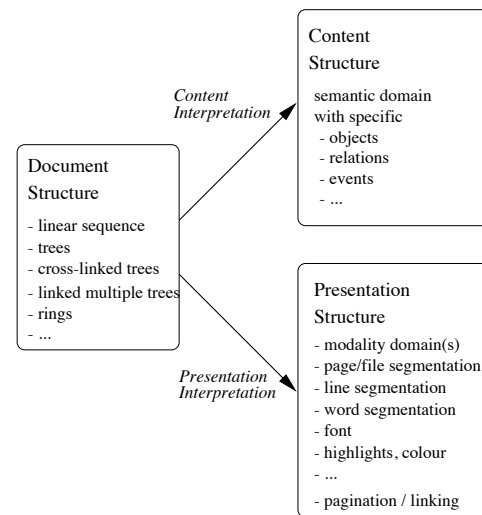

Figure 1: General document representation schema.

Constructor functions (i.e. a text grammar) are associated with DS, and a pair of interpretation functions CI (Content Interpretation) and PI (Presentation Interpretation) is associated with IS (cf. also Figure 2):

$$CI : \text{DS} \rightarrow CS$$
$$PI : \text{DS} \rightarrow PS$$

It is PI, in the form of converters for print make–up and hypertextualisation with which we are concerned here. PI is in fact a family of functions, each of which may be quite complex and involve multimodal visual (text, graphic, animated) and acoustic mappings rather than straightforward textual mappings.

The main differences between this basis for design and much current wisdom and folklore about hypertexts are (a) that the present approach is deliberately based on linguistic theory; (b) it is fundamentally triadic, $< DS, CS, PS >$,

in contrast to the simpler 'logical structure / presentation' approach of many current textbooks; (c) applied to lexica, it specifies the meta–semiotic structure of a lexical document as a complex sign which in turn describes object–semiotic signs, i.e. the lexical entries.

4. Specifying DS, PS and PI for a hyperlexicon

Hyperlexicon PS. A hyperlexicon is a lexicon with hypertextual PS, in general based on the 'latent hypertext' structure formalised in DS. We take a set–theoretic view of the structure of hypertexts such as hyperlexica and other hyperdocuments, and define hypertexts as relations between text constituents, where the relations represent hyperlinks. We depart from conventional wisdom in regarding conventional linear textual structure simply as a special case of hypertextual structure, and a simple document as the elementary hypertext.

We note that PI is determined by procedural considerations of lexical lookup, and in addition to the traditional onomasiological–semasiological distinction we also distinguish between two main procedural notions:

1. *generator hypertext*, i.e. the complete structure designed by the author or constructed by software,
2. *receiver hypertext*, i.e. the structure actually navigated and retrieved by the reader (or perhaps a hypertext spider agent), possibly fragmentary in relation to the generated hypertext, and with different relations over the parts.

The specification of receiver PI may well be pretty nearly identical for different PS — after all, a dictionary user may simply look up an entry, find a synonym, look up the synonym and check out its definition. This path is not too different in paper or electronic books, though the details of PS may be different.

In principle, the relations which may be specified *ad hoc* for receiver PI will be arbitrary members of the powerset of the Cartesian product $E \times E$, where E is the set of linked documents in PS. The members are completely arbitrary, because in principle a user may make arbitrary on the fly jumps. However, more straightforward cases can be taken as the norm, in which receiver PI is simply a subset of generator PI. We therefore concentrate on generator PI.

The traditional onomasiological and semasiological criteria for organising lexica are procedural: an onomasiological lexicon favours concept–based lookup in search of appropriate forms, while a semasiological lexicon favours form–based lookup in search of definitions. The typical semasiological lexicon has a top node, the lexicon, second level nodes for lexical entries, and leaf nodes representing types of lexical information; some lexica have lemma–sublemma structures, thus involving an additional level in the tree structure. However, the tree structure does not capture a fundamental feature of semasological lexica, since each pre–leaf node has, in general, the same number of leaves, in the same order, and information types; this suggests a tabular structure as being more suitable, though a user may not actually find this a convenient format. An onomasiological lexicon, often a thesaurus or synonym lexicon, is more naturally represented as a tree structure; this is a presentation structure which would in theory be suited to conventional termbanks, though these are in practice presented as onomasiologically oriented structures.

Tree–structured and tabular DS. The Document Structure of a lexical document could be organised on the procedurally biassed lines; however, we propose adhering to as strictly declarative a DS as possible, with procedural criteria being used to define different PS. But a declarative DS is not simple. We propose a multi–level DS:

1. An underlying n–place relation L, where n is the number of types of lexical information (the *intensional coverage*), and each n–tuple represents a lexical entry. A tabular or matrix structure of this kind does not on its own yield natural representations of multiple mappings between types of lexical information however (the injective onomasiological mapping of polysemy or homonymy etc., including homophony, homography, or the injective semasiological relations of synonymy). In a tabular representation, such relations have to be spelled out, leading to a massive inflation of table structure.
2. A set of more complex relations superimposed on the elements of the relation L, which can be represented graphically as a tree (e.g. for the hyperonym–hyponym relation), a ring (e.g. for partial synonyms), a fully connected graph (e.g. for co–hyponyms, antonyms), a linear graph (e.g. for predicate–argument type relations).

The basic matrix concept DS is supplemented in current theoretically well–founded lexica by generalisation (type or default inheritance) hierarchies coupled with attribute–value structures, which permit underspecified ('sparse') representations of lexical entry vectors, and reduction of overly inflated tabular structures to more interesting and tractable tree structures. We do not discuss these methods here, however.

Traditional problems with the matrix representation structure are (cf. also (Gibbon et al., 1997a), Chapter 6, and (Gibbon et al., 1999 forthcoming), Chapter 4):

1. polysemy, i.e. closely related meanings associated with one lexical form (does each meaning require a separate entry?);
2. homonymy (homophony, homography), i.e. the more straightforward case in which one lexical form has completely disparate meanings, with variants of multiple spellings for one pronunciation or multiple pronunciations for one spelling;
3. (partial) synonymy, i.e. a lexical form shares at least one meaning with another (the extreme case being abbreviations and their full forms).

The simplest solution, conceptually (though in general intractable) is to spell out all these additional entries and define a superimposed superstructure of more complex relations which may then be used to reduce the tabular structure

systematically as a tractable PS, with the maximal table preserved by associating the rows or records with explicit superimposed relations or 'rules'. The rules are then available as *hyperlinks* between relevant types of lexical information, e.g. 'pointing' to synonyms, alternative spellings. More abstractly: the 'rules' are available as specifications of a type or default hierarchy over attribute–value structures in a higly–structured and theoretically motivated lexicon.

In summary: the basic, procedurally neutral lexicographic DS is tabular with well–defined superimposed tree structures representing paradigmatic similarity relations within the lexicon. It is this concept of DS for a flexible hyperlexicon which we suggest as a suitable design. In a sense, this is no surprise, because numerous lexical databases adhere to this structure.

5. Implementation of the DS graph

The standard text resource annotation languages SGML (Standard Generalised Markup Language) (ISO 8879:1986, 1986) and its derivative XML (eXtensible Markup Language) are not entirely adequate for lexicon markup as they are primarily tree–based, and therefore not immediately suited to formulating relation tables or non–tree graph structures in the most expressive manner. They are adequate for many purposes, however, as tree structures are very common at all levels of lexicon representation. In particular they can easily be construed as typed attribute–value languages, and therefore isomorphic mappings can conveniently be constructed between SGML/XML markup and the conventional attribute formalisms of current linguistic theory. Practical advantages of SGML/XML are their standardised status, which promotes reusabillity of resources, and the availability of Document Type Definitions and tools such as parsers and structure editors. This technique was adopted for mapping the the EAGLES-Termbank (EAGLET, 1997-99) database relation to additional tree structures, as described above, using the MARTIF DTD (ISO 12200, 1998) with the lexical categories described in (Gibbon et al., 1997b) based on (ISO 12620, 1998). The EAGLES Termbank covers the domain of terminology for spoken language and multimodal systems.

This procedure is illustrated here with the lexicographic term *abstract lemma*.

Encoding the database relation table. The basic lexical information type vector is encoded with the attribute `type` of the element `<termNote type=...>` Different terms are included in the lexicon relation, representing different types of lexical information. A subject field, which permits a hierarchical domain classification, is defined:

```
<descripGrp>
<descrip type="subjectField">
Spoken Language Technology: lexicon
</descrip>
</descripGrp>
```

Unconventionally for terminological lexica, which are in theory concept–based, detailed word form information is included (the pronunciation is given in standard SAMPA notation):

```
<ntig lang=en>
<termGrp>
<term id="abstractlemma">
abstract lemma
</term>
<termNote type="termType">
international scientific term
</termNote>
<termNote type="partOfSpeech">
[N: [AJ: abstract][N: lemma]]
</termNote>
<termNote type="PLU">
[plural: abstract lemmata]
</termNote>
<termNote type="PRON">
/'\{bstr\{kt 'lem@/
</termNote>
</termGrp>
```

Encoding superimposed tree relations. The database relation is fully specified, with disjunction–free fields. In order to define relations between entries (records), additional relations are superimposed. First, for each term an identifier *id* is defined to enable linking from other lexcial entries to this term. Relations with elements of a set of related terms such as *lemma* and *lexical lemma* are included, defined as *links* to these other terms:

```
<ref type="relatedTerm" target="lemma">
lemma
</ref>
```

```
<ref type="relatedTerm"
target="lexicallemma">
lexical lemma
</ref>
```

Both the definition and the concordance–like reference are also included; however, the reference itself is given in another place of the document with all other references and it is refered to by a link again:

```
<descripGrp>
<descrip type="definition">
An abstract lemma is an access key
which may have any convenient unique
name or number (or indeed be labelled
by the spelling of the canonical
inflected form); all properties have
equal status, so that the abstract
lemma is neutral with respect to
different types of lexical access,
through spelling, pronunciation,
semantics, etc.
</descrip>
<ref type="sourceIdentifier"
target="Gibbonetal1997">
p. 200
</ref>
</descripGrp>
```

Meta–meta–information on the lexicographer(s), is also available, which can evidently be linked to information on the lexicographer and other contextual information:

```
<date type="modification">3.3.99</date>
<ptr type="originator" target="MS">
<adminGrp>
<admin type="responsibility">
EAGLES 1999
</admin>
</adminGrp>
```

In this notation, the DS can be checked by a conventional SGML/XML parser for consistency.
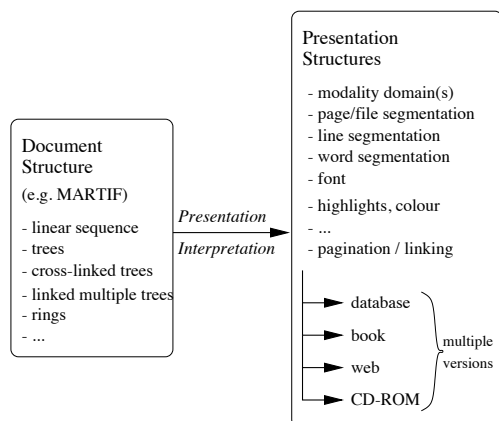
6. Implementation of PS graphs

Figure 2: Multiple Presentation Structures.

The original relational database structure, with the additional relations described above, and the SGML/XML format can be seen as equivalent alternative DS representations. The different database indexings and views, with different user interfaces, provided by an appropriate database management system are entirely analogous to the PS structures already discussed. From both the database relations and the SGML/XML DS representation it is relatively straightforward to derive database views, print media formats, as well as hypermedia formats.
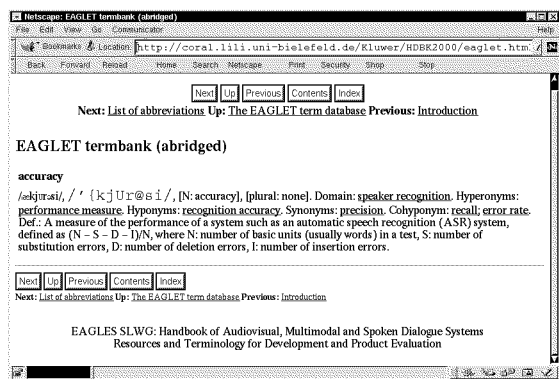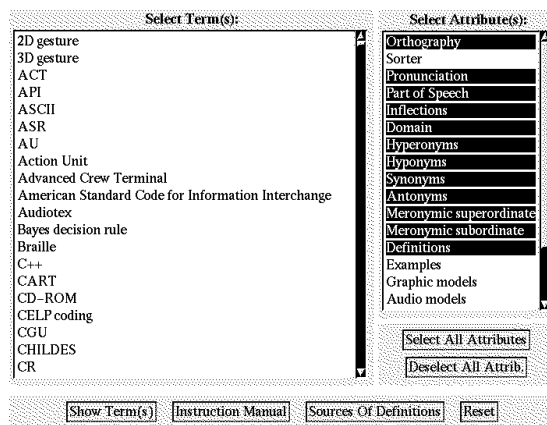

Figure 5: Hyperlexicon PS prototype.

The current relational database DS has the following properties (the DBMS is *mSQL*):

- DS: a matrix with efficient internal representation, and exported for interchange purposes in ASCII CSV

Figure 6: Web database interface PS.

(character separated value) format with further field coding where necessary. Acoustic visual signal information as well as concordance references to source text corpora are stored as local or network resource names (filenames, URLs).

- User and lexicographer PS interfaces: JavaScript driven HTML forms.
- Relational DB query PS output: filing card style browser appearance.
- Compiled and on-the-fly hyperlexicon PS output: hyperlinked pages representing unifications of information-type-specific tree structures, with reentrancy for shared subtrees, yielding a hypergraph structure.

The current implementation of the hyperlexicon constructor uses UNIX scripting, and is in the process of being ported for use with servers in different operating system environments.

7. Prototype and further development

We have demonstrated the feasibility of specifying a PI mapping of the lexical DS relation expressed either as a relational database or in SGML/XML format, and converting

```
ORT: accuracy
KEY: accuracy
PRO: /'{kjUr@si/
POS: [N: accuracy]
INF: [plural: none]
DOM: Spoken Language Technology: speaker recognition
SUP: performance measure
SUB: recognition accuracy
SYN: precision
ANT: recall; error rate
DEF: A measure of the performance of a system such as an automatic speech
recognition (ASR) system, defined as (N - S - D - I)/N, where N: number
of basic units(usually words) in a test, S: number of substitution errors,
D: number of deletion errors, I: number of insertion errors.
DAT: 25.11.98, 20.1.99, 16.2.99, 23.2.99
AUT: ms, ms, ms, IMO
```

Figure 3: Database *DS* format.

accuracy
/'ækjuːrəsi/, /'{kjUr@si/, [N: accuracy], [plural: none]. Domain: speaker recognition. Hyperonyms: performance measure. Hyponyms: recognition accuracy. Synonyms: precision. Cohyponym: recall; error rate. Def.: A measure of the performance of a system such as an automatic speech recognition (ASR) system, defined as (N - S - D - I)/N, where N: number of basic units (usually words) in a test, S: number of substitution errors, D: number of deletion errors, I: number of insertion errors.

Figure 4: Book lexicon *PS*.

this into several practical hypergraph (hyperlexicon) *PS* macrostructures. In this way, fast, reliable and consistent maintenance of a lexical resource for a range of different user types can be maintained. The present approach differs from recent commercial 'hypertextualisation' systems in being based on a consistent, formally expressed linguistic model for textual signs.

The termbank currently has an extensional coverage (record count) of about 1300 terms, and an intensional coverage (field count) of 20, yielding 26k information items (cells). It is planned to re–implement the *PS* conversion with DSSSL or XSL in order to achieve maximal reusability in the future; cf. also (Witt et al., 2000).

8. References

EAGLET, 1997-99. Eagles-termbase (eaglet). URL coral.lili.uni-bielefeld.de/EAGLES/. Termbank for Spoken Language Systems defined in: LE EAGLES Phase II project (LE3-4244 10484/0) for the European Commission.

Gibbon, Dafydd, Silke Kölsch, Inge Mertins, Michaela Schulte, and Thorsten Trippel, 1999. Terminology principles and support for spoken language system development. In *Proceedings of EUROSPEECH '99*. Budapest: EUROSPEECH.

Gibbon, Dafydd, Inge Mertins, and Roger Moore (eds.), 1999 forthcoming. *Handbook of Audiovisual, Multimodal and Spoken Dialogue Systems Resources and Terminology for Development and Product Evaluation*. Dordrecht: Kluwer.

Gibbon, Dafydd, Inge Mertins, and Roger Moore (eds.), 2000, forthcoming. *Handbook of Audiovisual, Multimodal and Spoken Dialogue Systems Resources and Terminology for Development and Product Evaluation*. Doordrecht, New York: Kluwer. Final Report of LE EAGLES Phase II project (LE3-4244 10484/0) for the European Commission.

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

Gibbon, Dafydd, Holger Nord, and Thorsten Trippel, 1997b. Eaglet: Eagles termbank for spoken language systems. DRAFT EAGLES Interim Deliverable, University of Bielefeld.

ISO 12200, 1998. *Computer applications in terminology – Machine-readable terminology interchange format (MARTIF) –Negociated interchange ISO/DIS 12200*. International Organization for Standardization (ISO), Genéva.

ISO 12620, 1998. *Computer applications in terminology – Data categories ISO/DIS 12620*. International Organization for Standardization (ISO), Genéva.

ISO 8879:1986, 1986. *Information processing – Text and office systems – Standard Generalized Markup Language (SGML)*. International Organization for Standardization (ISO), Genéve.

Witt, Andreas, Harald Lüngen, and Gibbon Dafydd, 2000. Enhancing speech corpus resources with multiple lexical tag layers. In *LREC 2000 Proceedings*. Athens.