TERMINOLOGY PRINCIPLES AND SUPPORT FOR SPOKEN LANGUAGE SYSTEM DEVELOPMENT

Dafydd Gibbon, Silke Kölsch, Inge Mertins, Michaela Schulte, Thorsten Trippel

ABSTRACT

Spoken language (SL) system development is an increasingly interdisciplinary effort. Speech-to-speech system development, for example, involves speech engineers, software engineers, phoneticians, and a variety of computational linguistic subdisciplines from morphology, syntax and lexicology through semantics and pragmatics, each with their own historically motivated terminology. In our experience this 'terminology barrier' makes communication between the disciplines unnecessarily difficult. As a contribution to reducing the terminology barrier we propose a set of new speech specific terminological principles and a prototype term bank with an Internet interface for this specific purpose. The result is one of the outputs of the spoken language working group of the LE EAGLES Phase II project (LE3-4244 10484/0).

1. INTRODUCTION

In order to take the Scylla of heterogeneity in the field of SL terminology [2] into account, and avoid, on the other hand, the Charybdis of a completely *ad hoc* hybrid description, a new approach was developed in the EAGLES Phase II project which

- combines the traditional semasiological and onomasiological approaches to terminology characterisation,
- re-uses existing computer readable terminological documentation and relevant text,
- develops a notion of a terminological hypergraph model and applies this in the construction of a terminological hyperlexicon.

2. A HYPERGRAPH BASED APPROACH

With this goal in mind, the traditional device of *conceptual graphs* in the onomasiological characterisation of terminology is replaced by an explicitly defined macrostructure with substructures which are designed to be realised as a *terminological hypergraph*, which in turn serves as a specification for the design of a *hyperlexicon* for implementation in CD-ROM and World Wide Web contexts (EAGLET HyperLexicon). With the advent of computers, lexicology (the scientific study of lexical information) and lexicography (the principles and techniques of lexicon construction) have been converging. One may distinguish between *latent hyperstructures* and *manifest hyperstructures* in texts. On this basis the structures used in traditional dictionaries can be used as a starting point for defining latent lexical hyperstructures underlying dictionaries, and strategies can be defined for developing manifest lexical hyperstructures as a hypertext. A hyperlexicon is then defined as a hypertext based on a latent or manifest lexical hyperstructure. In the EAGLET HyperLexicon the 'leaves' of the hypergraph are the terms; terms and their vector of defining properties is used to specify the data categories and records of the terminological relational database. An individual entry has to be structured according to a set of types of lexical information, or data categories: the *microstructure*. The EAGLET Term Database is operational with provisional functionality, and EAGLET HyperLexicon will remain for the medium term future as a specification.

3. CONCEPTUAL PARTS

Architectural model: For EAGLET, a single relation, the microstructure, is defined. The architectural model is the specification for the implementation in database software.
Database engine: In EAGLET development, the engine used is that of the software package mSQL. To provide a maximum of platform independence and to prevent inconsistencies and controversies resulting from use of different versions the database is stored on one machine as a single token, and accessed via the World Wide Web.
Front end tools: In the EAGLET implementation, JavaScript menu control is used. This script language is implemented in most modern WWW browsers and is platform independent. A text interface for PDA browsers and the like is being prepared.
Normalisation rules: An HTML form input mask is used. Only the categories and data fields that are implemented in the form can be entered and displayed.

4. INFORMATION STORAGE

The following three main types of field are currently envisaged for the EAGLET relation:
Static: the entity (here: term attribute value) is stored in an ASCII coded format (e.g. simple text, SAMPA notation, LaTeX code).
Hyperlink: term relations that depend on the query's context are built up as URL*s*.
Media event: data structures are coded (if necessary in an appropriate data format *'on-the-fly'*). They are stored outside the DB and are referenced by URLs.

5. SYSTEM COMPONENTS

Database server: Currently an SQL Database Server is in use (*Hughes Technologies mSQL* Version 2.0.3) .

Filters to import external data: UNIX Scripts for Solaris 2.5.1 and Linux are available to manipulate external data into a suitable format, building import functions via the mSQL script language *lite* - Version 2.0.3.
Database query language: the mSQL CGI interpreter and script language *lite* - Version 2.0.3 .
Interface application: any HTML 3.2 and JavaScript 1.1 enabled browser.
Interface programming language: HTML 3.2.
Interface validation language: JavaScript 1.1.

6. STRUCTURE

The overall structure of the EAGLET database is shown in Figure 1.

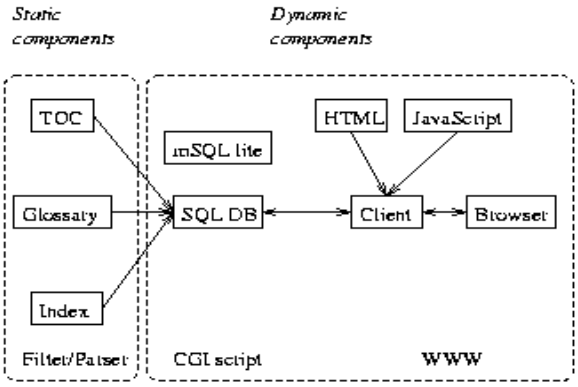

Figure 1: Structural overview of EAGLET

The database with its structured entries (based on the table of contents, glossary and index of the *Handbook of Standards and Resources of Spoken Language Systems* [3] is of a fairly static nature, with facilities for evaluation and manipulation via query and format filters. The SQL database machine serves as the interface via the scripting language *mSQL lite* to the dynamic, interactive components of the client. Queries are submitted via an HTML form (validated by a JavaScript applet) displayed on the client browser. The form is generated dynamically: every interface page is generated with the up to date data of the database, and queries can be reduced to the user's needs.

7. EAGLET MACROSTRUCTURE

In view of the complexities involved – and the very large number of degrees of freedom – a pragmatic approach has been taken in the development of the EAGLET concept. The approach involved developing a macromodel for spoken language terminology based on the macrostructure of the Parts and Chapters of the *Handbook*. For this purpose, the following textual components will be incorporated into a hypergraph design for a terminological hyperlexicon:

1. *Table of contents (TOC):* The TOC represents a possible onomasiological structure for the content, and provides an elementary variety of onomasiological indices into the text.
2. *Body of text*: The body of the *Handbook* provides expert-developed contexts in which terminology is authentically attested; the body of text in the chapters therefore defines an authentic corpus of attested forms in context.
3. *Glossary*: The Glossary is effectively a semasiological dictionary with headword and definitions, usually of the *genus proximum et differentia specifica* type.
4. *Index*: The Index indirectly provides a semasiological concordance, with headword and pointers into the corpus of attested forms in context.

7.1 Sub-taxonomies

The text source for the terminology hyperlexicon provides taxonomies with a greater degree of granularity than that outlined so far, based on a subtree of the table of contents of the *Handbook*. For convenience in representation, the taxonomy is divided into the sub-taxonomies:

- system design,
- corpus design (see Figure 2),
- lexicon development,
- language models,
- physical characterisation,
- assessment methodology,
- recogniser assessment,
- speaker verification assessment,
- synthesis assessment, and
- interactive dialogue system assessment.

The structure is modified from the basic text organisation of the *Handbook*, and is intended to represent, in each case, a first starting point for a pragmatic applications orientated basic system design taxonomy.
Taken together, the sub-taxonomies constitute a comprehensive taxonomic hierarchy of fine granularity; the sub-taxonomies have been curtailed at a coarse-grained level, but as the textual structure of the *Handbook* shows, much finer grain is available. In later versions of the work on spoken language terminology, this will be used for graphically oriented access to term definitions.

7.2 Graphical representations of sub-taxonomies

In Figure 2, the sub-taxonomy for '*corpus design'* is given. The sub-taxonomies for different areas show very different kinds of structure, as to be expected. The differences encompass the following topological and semantic features of the graphs which are, with few exceptions, tree graphs:

- number of nodes,
- depth of branching,
- breadth of branching,
- differences in node interpretation, e.g. in terms of *formalism* (notation, terminology, nomenclature), *empirical method*, or *sub-domain* (field, subject),
- differences in edge interpretation, e.g. as *ISA* or strict taxonomic interpretation, vs. *PARTOF* or mereonomic interpretation.

However, the explicit graphical representation of the taxonomies provides a useful basis for future work, in which similarities between the different sub-taxonomies can be examined in more detail and, in some cases, merged.

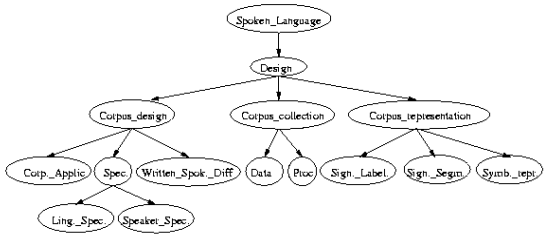

Figure 2: A basic corpus design taxonomy

8. EAGLET MICROSTRUCTURE FOR SPOKEN LANGUAGE TERMINOLOGY

Microstructures of hyperlexica differ from terminology to terminology. A detailed discussion of relevant data categories can be found in [4]. The currently implemented EAGLET microstructure contains the following data categories as fields in the database:

1. *Form/Orthography:* Standard British English orthography.
2. *Form/Pronunciation*: Phonemic transcription in SAMPA notation; cf. the revision in [3].
3. *Form/POS*: The structure of compounds is given in an attribute-value notation. Example: The term 'text-to-speech system' is analysed as '[N: [N: text][PREP: to][N: speech][N: system]]'.
4. *Form/Inflections*: As nearly all terms in EAGLET are nouns, this category basically indicates the plural form(s) of terms. The possible values are: -s ('badger' - 'badgers'), -es ('search' - 'searches'), -0 ('sheep' - 'sheep'), none ('Bayesian decision theory'); for non-regular forms and the '-ies' plural in words like 'frequencies' the plural form is given in full.
5. *Semantic Domain*: 'Domain' refers here to the individual chapter of the *Handbook* [3] the term can be assigned to. The default value 'Spoken Language Technology' has been entered for all terms, and, where possible, the more specific subject field such as 'physical characterisation', 'corpora', 'lexicon', 'interactive dialogue systems' is added. Example: For 'Hidden Markov Model' the value is 'Spoken Language Technology: language modelling'. Many terms are difficult to place because they are very general, for example 'orthographic transcription', a term that occurs in nearly all *Handbook* chapters and, like many others, is not restricted to the domain of spoken language technology.
6. *Semantics/Hyperonyms*: The data category 'hyperonyms' corresponds to the classical genera proxima in terminological theory. A *hyperonym* is the verbal representation of the superordinate concept of a term in a taxonomy. Examples: *morph* is a hyperonym of *bound morph* because '*A bound morph is a type of morph'* is acceptable.
7. *Semantics/Hyponyms*: A *hyponym* is the verbal representation of the subordinate concept of the term in question. Examples: A *bound morph* is a hyponym of *morph* because *A bound morph is a kind of morph* is an acceptable sentence.
8. *Semantics/Synonyms*: A synonym is a term that represents the same concept as the main entry term in a term entry. In EAGLET, no distinction is made between genuine synonyms and quasi synonyms. Quasi synonyms are terms that represent the same concept in the same language, but for which interchangeability is limited to some contexts and inapplicable in others. Example: *wolf* is a synonym of *skilled impostor*.
9. *Semantics/Antonyms*: This data category covers terms denoting all types of lexical opposite. Complementaries, i.e. terms that "divide some conceptual domain into two mutually exclusive compartments" [1], p. 198, are treated as a subset of antonyms. Example: *cardioid microphone* and *hypercardioid microphone* are antonyms of *supercardioid microphone*.
10. *Semantics/Definitions*: As in most standard general dictionaries, EAGLET not only contains analytical definitions, i.e. definitions which give a noun phrase providing the meaning of the term in question [5], but also definitions that contain nonessential characteristics and information that would be classified as `world knowledge'. In many cases also the source of the definition is given. Example: The unidirectional type of microphone is most sensitive to sound arriving from one direction and more or less attenu-

ates incident sound from other directions. Thus, unidirectional microphones will suppress intended sound when pointed at the wanted sound source, i.e. the speaker. [3], p. 303

11. *Semantics/ Meronymic superordinates*: Terms that are superordinates in a PARTOF hierarchy. Example: *syllable* is a meronymic superordinate of *onset* because *The/An onset is part of a syllable* is an acceptable sentence.
12. *Semantics*: Meronymic subordinates. Terms that are subordinates in a PARTOF hierarchy. Example: *onset* is a meronym of *syllable*, because *An onset is a part of a syllable.* is an acceptable sentence.
    In EAGLET no distinction is made between facultative and non-facultative parts, and no information is given as to whether constituents occur in a certain order or not: for example, the order onset-nucleus-coda (= parts of a syllable) is not expressed in EAGLET. Note that two or more meronymic hierarchies may co-exist depending on the classificatory criterion.
13. *Context/Examples*: A term and its definition is exemplified. Example: 'un' and 'able' in 'unbearable' are affixes.
14. *Context/Graphic models*: This data category is reserved for visual representations of a concept.
15. *Context/Audio models*: This data category is reserved for auditory representations of a concept.
16. *Context/Formulas*: Here formulas are given that might replace a textual definition.
17. *Context/References*: Here the occurrences of a term in the WWW edition of [3] is given. At the moment this information is not accessible.
18. *Author/Date*: This administrational category shows the date of the last change of the record.
19. *Author/Author*: The administrators who performed the changes to the record are given.

9. CONCLUSION

The approach taken here differs from standard approaches in terminology science [6] for a number of reasons. First, terminology science is concerned with providing tools for the unambiguous treatment of linguistic, phonetic and human language technology terms by non-specialists with respect to language. This is a crucial difference: in the spoken language technology field the addressees are very different. The users themselves have highly sophisticated approaches to language and are often specialists in terminological and lexical matters themselves. Second, from a linguistic point of view there are two grave errors which are often made in lexical work, which reduce to the same general point: the confusion of *unique index* with an *information element* of the microstructure (field of the database). The first, and commonest, example in lexicography is the dual use of *orthographic information* as a unique ordering index. The second is the comparable dual use of *conceptual information* in terminologies as a unique ordering index. This use is avoided in our approach, which is neutral with respect to onomasiological (semantically ordered) and semasiological (form ordered) lexica. Remaining problems of cross-reference will be solved by the use of hyperlinks and, ultimately, by the use of object-oriented databases and other inheritance formalisms.

REFERENCES

[1] D. Cruse (1986). *Lexical semantics*. CUP, Cambridge.

[2] H. Felber, Gerhard Budin (1989), Terminologie in Theorie und Praxis. Gunter Narr Verlag, Tübingen.

[3] D. Gibbon, R. Moore, R. Winski (eds.) (1997). *Handbook of standards and resources for spoken language systems*. Mouton de Gruyter, Berlin.

[4] A. K. Melby, S. E. Wright (1998), The CLS Framework (draft version), http://www.ttt.org/clsframe/index.html.

[5] J. Sager M.-C. L'Homme (1994). A model for the definition of concepts: Rules for analytical definitions in terminological databases. *Terminology* 1(2): 351-374.

[6] K.-D. Schmitz, G. Budin, C. Galinski (1994), Empfehlung für Planung und Aufbau von Terminologiedatenbanken. Gesellschaft für Terminologie und Wissenstransfer e. V.