A REUSABLE LEXICAL DATABASE TOOL FOR MACHINE TRANSLATION

BRIGITTE BLÄSER
IBM Germany
Institute for Knowledge
Based Systems
P.O. Box 10 30 68
W-6900 Heidelberg
Email: alschwee
at dhdibm1.bitnet

ULRIKE SCHWALL
IBM Germany
Institute for Knowledge
Based Systems
P.O. Box 10 30 68
W-6900 Heidelberg
schwall
at dhdibm1.bitnet

ANGELIKA STORRER
University of Tübingen
Seminar für
natürlichsprachl. Systeme
Wilhelmstr. 113
W-7400 Tübingen
storrer at arbuckle.sns.
neuphilologie.uni-tuebingen.de

0. ABSTRACT

This paper describes the lexical database tool LOLA (Linguistic-Oriented Lexical database Approach) which has been developed for the construction and maintenance of lexicons for the machine translation system LMT. First, the requirements such a tool should meet are discussed, then LMT and the lexical information it requires, and some issues concerning vocabulary acquisition are presented. Afterwards the architecture and the components of the LOLA system are described and it is shown how we tried to meet the requirements worked out earlier. Although LOLA originally has been designed and implemented for the German-English LMT prototype, it aimed from the beginning at a representation of lexical data that can be reused for other LMT or MT prototypes or even other NLP applications. A special point of discussion will therefore be the adaptability of the tool and its components as well as the reusability of the lexical data stored in the database for the lexicon development for LMT or for other applications.

1. Introduction

The availability of large-scale lexical information has widely been recognized as a bottleneck in the construction of Natural Language Processing (NLP) systems. The lexical database LOLA has been developed in connection with the Logic-programming-based Machine Translation (LMT) system and shall be presented here. This work is part of the objectives of the project TransLexis launched in 1991 at the Institute of Knowledge Based Systems of the IBM Germany Scientific Center. TransLexis aims at the theoretically and empirically well motivated lexical description and the management of the lexical information of LMT in a database. It is conceived as a first step towards a reusable lexical knowledge base.

1.1. Requirements for convenient construction and maintenance of Lexicons

Based on our experience and existing literature, a tool for the construction and maintenance of large NLP lexicons with a complex entry structure should meet the following requirements:

- Adequate expressive power of the representation formalism: the expressive power must be sufficient to cover the facts of lexical description.
- Methodology for the description of lexical information: criteria and guidelines relevant for encoding should be developed and documented.
- Orientation towards lexicographic procedure: the design of the tool should take the logical course of the lexicographic work procedure into consideration and support it during all its steps and phases. The lexicographer should be enabled to concentrate on the lexicographic description of lexical units while the tool itself automatically takes care of the remaining tasks in lexicon development.
- Consistency and integrity checking of the lexical data: when entries are added or updated, the system should reject invalid values for particular features and check if the input leads to inconsistency of the database.
- Data independence: An extreme dependency between the structuring of lexical data stored in the database and the structure of the lexical entries in a given application system should be avoided. In this way the lexical data will remain resistant to modifications in the NLP/MT-systems that make use of these data.
- Reusability/Reversability of the data (cf. Calzolari 1989, Heid 1991): lexical data should be represented in such a way that it can — apart from its transfer specific components — be reused for other MT-prototypes with the same source or target language, or with the reverse language pair (e.g. German-English and English-German). Ideally, the lexical data should be independent to such a degree that they are also reusable for other NLP-applications.
- Multi-user access: it should be possible for several users to work on the lexicon simultaneously.
- Help facilities: the criteria and guidelines for lexical description should be easily accessible. The availability of monolingual and bilingual dictionaries are to support the lexicographer's linguistic competence.

1.2 LMT

LMT, developed by Michael McCord, is in basic design a source-based transfer system in which the source analysis is done with Slot Grammar (cf. McCord 1989, 1990, forthcoming). Two main characteristics of LMT should be emphasized:

1. the lexicalism, arising from Slot Grammar source analysis;

2. a large language-general X-to-Y-translation shell.

Both features facilitate the development of prototypes for new language pairs[1] . Versions of LMT (in various stages) exist currently for nine language pairs.
LMT currently requires the following types of information to be specified for lexical units (LU):

- part of speech;
- word senses;
- morphological properties;
- agreement features;
- the valency, i.e. the frame of optional/obligatory complement slots;
- the specification of the fillers (NPs, subordinate clauses) for each slot;
- semantic compatibility constraints and collocations;
- characterization of multiword lexemes;
- subject area;
- translation relations;
- lexical transformations.

In McCord (forthcoming), an external lexical format (ELF) is presented which allows the representation of the above information. Until now, however, the lexical data has been kept in sequential files and updating has been done with a text editor. Thus most of the above-mentioned requirements could not be met.

1.3 Vocabulary Acquisition

The hand-coding of dictionaries is a laborious and time-consuming task. Therefore a number of attempts have been made to exploit corpora and/or machine readable dictionaries (MRDs) for the build-up of NLP-lexicons (cf. 3.5)[2] . In many cases, however, the lexical information in MRD's is neither complete nor sufficiently explicit for NLP/MT purposes and has to be revised by lexicographers. Ideally, the demands on a lexicographer should only be of linguistic nature. For this reason a sophisticated tool is needed to guide and support the NLP/MT-lexicographer in revising entries automatically converted from machine readable sources as well as in building up new vocabulary.

2. LOLA - architecture and components

The lexical data base tool LOLA aims at meeting the above mentioned requirements. Its design and development are based on work achieved in the LEX-project and the COLEX-project[3] . LOLA makes use of automatic consistency and integrity checks as well as of the support of multi-user access provided as standard facilities by the relational DBMS SQL/DS. Updates are made with the help of a user interface that supports the lexicographer during the encoding process. The representation of the lexical data has been worked out to be as independent as possible of the format of a specific application lexicon, thus increasing the degree of reusability of the lexical data. In addition, a catalogue of criteria and guidelines for lexical description is being elaborated and will be integrated into the tool.

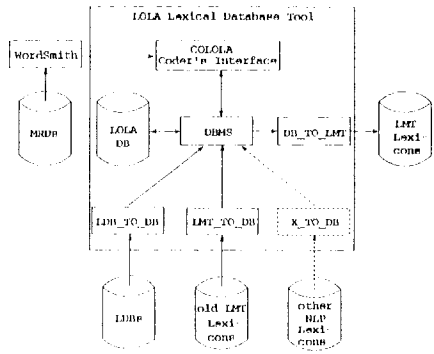

Figure 1. LOLA - architecture

The main components of the architecture of the LOLA system are the following (cf. Figure 1):

1. LOLA-DB: the database itself.
2. COLOLA (COder's interface to LOLA): Interface for hand-coding and modification of the lexical data, stored in LOLA-DB.
3. DB_TO_LMT: program that generates LMT lexicon entries from the lexical data stored in LOLA-DB.
4. LMT_TO_DB: program that loads already existing LMT lexicons into LOLA-DB.
5. LDB_TO_DB: program that converts data from MRD's into LOLA-DB.

In the following we give a brief description of these components.

2.1. The Database

The database was designed in two steps: development of the **conceptual scheme** and development of the **database scheme**.
In the conceptual design phase, the lexical objects, their properties, and their interrelations were represented in an entity-relationship diagram (cf.

[1] LMT is the technical basis of an international project at IBM with cooperation between IBM Research, the IBM Science Centers in Heidelberg, Madrid, Paris, Haifa, and Cairo, and IBM European Language Services in Copenhagen (cf. Rimon et al. 1991).

[2] Cf. Byrd et al. 1987; for an overview of related activities within the LMT-project, cf. Rimon et al. 1991, pp. 14-15.

[3] Cf. Barnett et al. 1986; Blumenthal et al. 1988; Storrer 1990.

Chen 1976). Although the ER-model does not have the expressive power to cover all aspects of lexical description, especially complex constraints, it has been chosen here as a compromise between a complete lexical representation and the realization in a traditional database system.
The resulting ER-diagram for the German-English lexicon is shown in Figure 2[4].
The conceptual scheme is still independent of the choice of a specific DBMS and of other implementation aspects. The basic principles of the conceptual design of our database will be sketched out in the following.

Orientation towards linguistic structure, not towards the structure of the application lexicon.
The diagram reflects, in the first place, the structure of the linguistic objects, their properties and and their interrelations, and it is influenced to a smaller degree by the structure of the application lexicon. As a consequence, the data is quite resistant to structural changes in the format of the application lexicon. The abstraction from the structures of the application lexicon has a positive side effect with regard to the exploitation of machine readable lexical resources: on one hand, we can handle cases, in which not all information required by LMT is provided in the entries of MRD's. The information acquired can be stored as entries to be completed and revised later. On the other hand, we are free to store types of lexical information that are of relevance for NLP applications and can be acquired from MRD's or other NLP lexicons but are not processed in a current LMT-version. We can save them in the database as coding aids for the lexicographers, for future prototype versions, or other NLP applications.

Analogous structure for source and target language wherever possible.
The lower part of the ER-diagram represents the German source, the upper part the English target language. For both languages, an entity of the type **entry** can have one or more **homonyms**, each of which can have one or more **senses**. The senses themselves can open one or more sense-specific **slots** (one-to-many relations). A sense-specific slot can be filled by several types of **fillers** and the same type of filler can fill several sense-specific slots (many-to-many relation). The basic types of entities and relations, which are the same for all languages, are described by their characteristic features represented as attributes. The number of attributes as well as their values may differ according to language-specific peculiarities[5].

Many-to-many relations between the lexical objects of both languages.
We represent the relation of lexical equivalence between source and target senses as a many-to-many relation (one source sense can have multiple target equivalents and vice versa). This breaks with the traditional hierarchical entry structure of bilingual dictionaries (Calzolari et al. 1990), but it avoids redundant description and storage of one target sense that is lexically equivalent to different source senses. Another relation holds for the sense-specific slots of two senses that are regarded as lexically equivalent. We decided to establish this relation between slots and not between slot frames. This way we can elegantly describe lexically equivalent senses with non-corresponding slotframes[6]. In this way the relations between the two languages may be used to a great extent bidirectionally for the XY- as well as for the YX-language pair.

The conceptual scheme captured in the ER-diagram was then mapped into a **database scheme** and implemented in the relational DBMS SQL/DS. We chose a relational DBMS, because – for the maintenance of the large LMT-GE lexicon (about 50,000 entries) – we were in need of a stable DBMS which supports multi-user access, has facilities for automatic checking of consistency and integrity of the lexical data and allows for the specification of multiple user-specific views on the data. To avoid redundancy and update anomalies we tried to normalize our relations as far it was useful with respect to our approach. In total, 32 tables are implemented: 25 tables describe lexical objects and relations by means of attributes with associated values, 7 tables serve to store "knowledge about the lexical knowledge", e.g. the admitted values for attributes such as *semantic type*, *filler-type*, *slot-type* for both languages.

2.2. COLOLA: the user interface to LOLA

COLOLA is the user interface to LOLA-DB that looks up the lexical data of a given search word and displays it on sequentially connected menus. The design of the menus as well as their sequential order was guided by the manner in which lexicographers describe lexical entries. The following operations can be performed:

---

[4] The boxes represent types of entities, the diamonds represent types of relations between entities, the ellipses represent attributes which characterize types of entities or relations. The labels of the connection lines indicate whether the relation in question is a one-to-one, one-to-many, many-to-one or many-to-many relation.
The ER-diagram is a simplified version of the actual conceptual model. For the purpose of this paper, several entity types, attributes, and relations have been left out.

[5] E.g.: in German a preposition like "*auf*" can govern either an accusative NP ("*warten auf*") or a dative NP ("*lasten auf*") depending on the verb that takes the prepositional phrase with the respective preposition as a complement. Therefore "case" is a feature, relevant for the description of German slot fillers filling a prepositional complement slot.

[6] E.g. cases like "*like*" and "*gefallen*" where the subject of the English verb corresponds to the dative object of the German verb; or cases like "*geigen*" and "*play the violin*" where the English direct object filler "*the violin*" is incorporated in the semantics of the German verb "*geigen*".

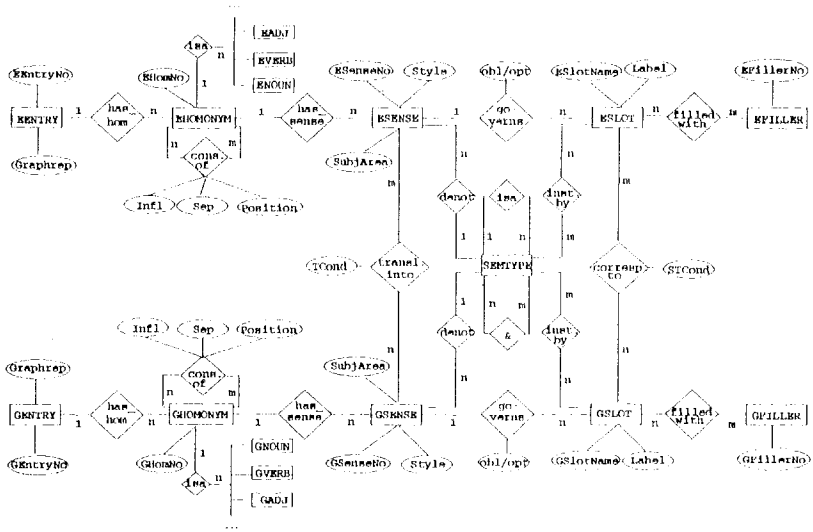

Figure 2. Entity Relationship Diagram for German-English

- addition of new source or target entries
- deletion of existing entries
- change of existing or addition of new features to existing entries
- deletion of features from existing entries
- assignment of new or deletion of existing translation equivalents for a given source sense
- update, insertion or deletion of transfer information for each pair of translation equivalents.

For each part of speech, a specific sequence of menus is defined. There are menus for homonyms and senses of source and target entries; the "linking" of the source senses and target senses regarded to be lexically equivalent is done via transfer menus. This allows lexicographers to specialize on specific parts of speech or on specific features which can be locally updated.

COLOLA controls multi-user access to the LOLA-DB so that several lexicographers can update the lexical database simultaneously. The logical unit of work is the source or target homonym: when a lexicographer requests to update a homonym, this homonym, together with its senses, is locked for other users.

If a new entry contains blanks, a multiword menu is called where the multiword is split up into its components. For each component, the following lexical information is gathered: the part of speech, whether the word may inflect within the multiword, and whether a phrase can be inserted between one multiword component and the previous one without doing away with idiomaticity.

If new homonyms or senses are inserted on the multiword menu as well as on other menus, default values for features are displayed. They can either be accepted or rejected and overwritten by the lexicographer[7] . The assumptions on default values for attributes of lexical information may differ according to different grammars and systems. We therefore decided to store the complete lexical information and use default values as proposals in the user interface. With this approach we allow for two advantages: on one hand, the data in the database can be used for different applications having distinct theory specific assumptions on defaults. On the other hand, the user of COLOLA can benefit from the economic advantages of default assumptions.

COLOLA does extensive consistency checking of the values entered by the lexicographers. Illegal values are rejected and warning messages are displayed in situations where errors might easily occur. Although much of the consistency checking is supported by the database management system, some extensions were necessary.

Further support for the lexicographers is provided by an interface to the WordSmith on-line dictionary system (cf. Byrd/Neff 1987). Several machine readable dictionaries are available e.g. Collins German-English, English-German, Longman's Dictionary of Contemporary English, and Webster 7th Collegiate dictionary. The lexicographer can look up entries in these dictionaries during the encoding process. Furthermore, help menus are provided in which the valid values for specific features can be looked up.

---

[7] Default values are provided, for instance, for slot fillers. German direct object slots get an accusative noun phrase as the default filler. The lexicographers may accept this, add other fillers or write over it with another filler.

2.3. DB_TO_LMT

A conversion program DB_TO_LMT has been developed which extracts lexical information stored in the relations of LOLA-DB and converts it into LMT-format. DB_TO_LMT consists of two components:

- a **database extractor** and
- a **conversion program**

The **database extractor** selects the source entries and the corresponding target entries and stores them in database format. This format can be regarded as an intermediate representation between database scheme and LMT-format. It consists of a set of Prolog predicates which correspond to the relations of the database scheme. There are, for instance, entry, homonym, sense, and slot predicates which correspond to the entry, homonym, sense, and slot relations in the database. The **conversion program** finally converts the database format into the LMT-format. It has to be adapted according to the changes or extensions of the LMT-format.

2.4. LMT_TO_DB

Before and during LOLA design and development, LMT lexicons in ELF were already created and updated in files. Since these lexicons still need updating and since this is much better supported by LOLA, a conversion program LMT_TO_DB was needed which converts ELF entries of lexicon files into the database format and loads them into LOLA-DB. LMT_TO_DB consists of three components:

- the **lexicon compiler** of LMT,
- a **conversion component**, and
- a **database loader**.

The **lexicon compiler** is the component of the LMT system which converts the ELF into the internal LMT format[8]. In the internal format all abbreviation conventions and default assumptions are already interpreted and expanded accordingly so that the complete lexical information is represented explicitly. The **conversion component** then converts the internal LMT-format to database format. The **database loader** generates the SQL-statements and updates the database. It has to check first whether the homonym or sense to be inserted is identical with an homonym or sense stored in the database. If all the features of two homonyms or senses can be unified, they are regarded to be identical and the already existing entry is merged with the converted entry. In all other cases the homonym or sense is inserted into the database and merging has to be done by the lexicographers with COLOLA.

2.5. LDB_TO_DB

To supplement the lexical coverage of the LMT system, a dictionary access module has been developed which allows real-time access (cf. Neff/McCord 1990) to Collins bilingual dictionaries available as lexical data bases (LDBs)[9]. The module includes a language pair independent shell component COLLXY and language-specific components and converts the lexical data of the LDB into the LMT-format. LDB_TO_DB is based on these programs. It consists of

- a **pattern matching component**,
- a **restructuring component**,
- a **conversion component**, and
- the **database loader** of LMT_TO_DB.

With the **pattern matching component**, those features (sub-trees) that are to be converted are selected from the dictionary entries. In printed dictionaries, features common to more than one sub-tree are often factored out in order to save space. With the **restructuring component**, those features can be moved to the sub-trees they logically belong to. The **conversion component** converts the restructured dictionary entry to database format. The **database loader** of LMT_TO_DB merges the entry with a possibly already existing one in LOLA-DB and generates the SQL-statements to update the database. The converted entries can be revised by the lexicographers with COLOLA.

3. Reusability of the LOLA system

3.1 Reusability of the tool components

The first LOLA prototype was developed to support lexicon development for the language pair German-English. In the meantime, work has been started to make the tool usable for lexicon development of the English-Danish and English-Spanish LMT systems. As a positive result of the design principles described in section 3.1., the database scheme had to be modified only slightly with regard to prototype-specific differences[10]. The values for language-specific attributes such as types of slots and fillers will be defined for the "new" languages Spanish and Danish and will be stored in the database. They can then be used for consistency checking (only defined values can be updated in the database). In COLOLA we had to take into account the homonym level on the target side, where

[8] In the morpho-lexical processing and compiling phase, ELF entries are converted into an internal format (cf. McCord (forthcoming): sect. 2) which represents the initial source and transfer analysis of an individual input word string.

[9] An LDB provides a tree representation of the hierarchical structure of the dictionary entries. The nodes of the tree are labeled with attributes having specific values for each individual entry. The LDB can be queried with the specialized query language LQL (cf. Neff/Byrd/Rizk 1988).

[10] English-Danish and English-Spanish use lexicon driven morphology for the target languages Spanish (cf. Rimon et al. 1991) and Danish, whereas German-English uses a rule-based target morphology for English (cf. McCord/Wolff 1988).

the features of Spanish and Danish morphology have to be specified. The programs that convert the database entries into the format of the application lexicons and vice versa (DB_TO_LMT and LMT_TO_DB) need generalization in order to achieve an abstraction from prototype-specific features of LMT.

3.2 Reusability of the lexical data

In order to meet the requirement of data independence, the representation of lexical entries in the database is highly independent of that in the application lexicon. In the database, the description of linguistic entities and their interrelations is given in a set of tables where specific values are stored for the characteristic attributes of each individual entity. On these tables, different views can be defined for different types of users. Different programs (like DB_TO_LMT) can extract exactly the attribute values needed for their respective application and convert them into each given format. This way, from one and the same data base several lexicons can be generated, in which the same 'linguistic world' is structured differently or represented in a completely different way. The possibilities of reusability are naturally defined and limited by the number of the registered types of lexical information in the original data base. As far as the LOLA database is concerned, the very detailed description of slot frames as well as the information about multiwords and the properties of their components may be reused for other NLP applications with one of the languages involved. The reusability of the transfer information (specified in the transfer relations between the languages of a given language pair) for other MT systems depends highly on the respective MT approach. As to the question of reusabilty of the data in the LMT system "family", three different cases have to be distinguished:

1. lexical-data description given for a source language X is reused for another language pair having X as source language,
2. lexical-data description given for a source language X is reused for another language pair having X as target language,
3. lexical-data description given for a target language Y is reused for another language pair having Y as source language.

In the first two cases, reusability of the lexical data of language X is very high. In the third case, the description of Y as source language may have to be more detailed in order to achieve an adequate syntactic analysis[11] . New attributes or even new types of entities or relationships may be needed and the database scheme will have to be enhanced accordingly.

4. Outlook

Our long-term goal is a multilingual database, in which the lexical knowledge for each language involved in the LMT project is represented only once. Application lexicons for LMT prototypes with different language pairs are generated by extracting the required information from the database and by converting it into the respective LMT-format. Furthermore, the tool is to be extended in such a way that it is not restricted to the construction of MT lexicons, but can also be used as a terminology workbench and thus support the construction and maintenance of terminology. An integrated MT and terminology database would have the advantage that the lexical knowledge encoded by terminologists and translators can be used by the translation system as well. For refinement and completion of the description of the German language, it is planned to integrate further information from available German NLP lexicons into the LOLA-DB. A basic problem concerning this undertaking will be to identify and to match the basic categories "entries", "homonyms", "senses", which are defined in various lexical resources according to different criteria, only some of which being transparent. With this effort, we hope to gain further knowledge on the limits and possibilities concerning the reusability of lexical data.

5. References

Barnett, B., H. Lehmann, M. Zoeppritz (1986): "A Word Database for Natural Language Processing", *Proceedings 11th International Conference on Computational Linguistics COLING86 August 25th to 29th, 1986, Bonn, Federal Republic of Germany*, pp. 435-440.

Blumenthal et al. (1988): "Was ist eigentlich ein Verweis? - Konzeptuelle Datenmodellierung als Voraussetzung computergestützter Verweisbehandlung." , Harras,G.(ed.): *Das Wörterbuch. Artikel und Verweisstrukturen.* Düsseldorf 1988.

Byrd, R., N. Calzolari , M. Chodorow, J. Klavans, M. S. Neff, O. Rizk(1987): "Tools and Methods for Computational Lexicography", *Computational Linguistics*,13, 3-4.

Byrd, R. J., M. S. Neff (1987): *WordSmith User's Guide*, Research Report, IBM Research Division, Yorktown Heights, NY 10598.

Chen, Peter P.-S. (1976): "The Entity-Relationship Model - Towards a Unified View of Data" , *ACM Transactions on Database Systems 1*, pp. 9-36.

Calzolari, N. (1989): "The Development of Large Mono- and Bilingual Lexical Databases", *Contribution to the IBM Europe Institute "Computer based*

[11] E.g. in a source-based translation system like LMT, the information on whether a target slot is obligatory is not directly encoded in the LMT transfer-lexicon; the system controls target slot assignment by the presence of a corresponding source slot in a given input sentence and the mapping relation specified within the transfer lexicon entry. On the source side, however, the feature of slot obligatoriness is used for purposes of analysis disambiguation.

Translation of Natural Language", Garmisch-Partenkirchen.

Calzolari et al. (1990): "Computational Model of the Dictionary Entry - Preliminary Report", *Project AQUILEX*, Pisa.

Heid, U. (1991): *A short report on the EUROTRA-7 Study* , Univ. of Stuttgart 1991.

McCord, M. C. (1989) "A New Version of the Machine Translation System LMT", *Literary and Linguistic Computing*, 4, pp. 218-229.

McCord, M. C. (1990): "Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars", In R. Studer (ed.), *Natural Language and Logic: International Scientific Symposium*, Lecture Notes in Computer Science, Springer Verlag, Berlin, pp. 118-145.

McCord, M. C. (forthcoming): "The Slot Grammar System", In J. Wedekind and Ch. Rohrer (ed.), *Unification in Grammar*, to appear in MIT Press.

McCord, M. C., Wolff, S. (1988): *The Lexicon and Morphology for LMT, a Prolog-based MT system*, Research Report RC 13403, IBM Research Division, Yorktown Heights, NY 10598.

Neff/Byrd/Rizk (1988) M. S. Neff, R. J. Byrd, O. A. Rizk: "Creating and Querying Hierarchical Lexical Data Bases", *Proceedings of the 2nd ACL Conference on Applied NLP*, 1988.

Neff, M. S., M. C. McCord (1990): "Acquiring Lexical Data From Machine-readable Dictionary Resources for Machine Translation", *Proceedings of the 3rd Int. Conf. on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, pp. 87-92. Linguistics Research Center, Univ. of Texas, Austin.

Storrer, A. (1990): "Überlegungen zur Repräsentation der Verbsyntax in einer multifunktional-polytheoretischen lexikalischen Datenbank" , Schaeder,B./Rieger,B. (ed.): *Lexikon und Lexikographie: maschinell - maschinell gestützt. Grundlagen - Entwicklungen - Produkte*, Hildesheim, pp. 120-133.

Rimon, M., McCord, M. C., Schwall, U., Martínez, P. (1991): "Advances in Machine Translation Research in IBM," *Proceedings of MT Summit III*, pp. 11-18, Washington D.C.