

Providing factual information in MAT

Walther v. Hahn* & Galia Angelova♦

*University of Hamburg, Germany
vhahn@nats2.informatik.uni-hamburg.de

♦Bulgarian Academy of Sciences, Bulgaria
galja@bgearn.bitnet

Abstract

Most translations are needed for technical documents in specific domains and often the domain knowledge available to the translator is crucial for the efficiency and quality of the translation task. Our project¹ aims at the investigation of a MAT-paradigm where the human user is supported by linguistic as well as by subject information ([vHa90], [vHAN92]).

The basic hypotheses of the approach are:

- domain knowledge is not encoded in the lexicon entries, i.e. we clearly distinguish between the language layer and the conceptual layer;
- the representation of domain knowledge is language independent and replaces most of the semantic entries in a traditional semantic lexicon of MT/MAT-systems;
- the user accesses domain information by highlighting a sequence in the source text and specifying the type of query;
- factual explanations to the user should be simple and transparent although the underlying formalisms for knowledge representation and processing might be very complex;
- as a language for knowledge representation, conceptual graphs (CGs) of Sowa [Sow84] were chosen.

In providing connections between the terms (lexical entries) and the knowledge base our approach will be compared to terminological knowledge bases (TKBs) which are hybrid systems between concept-oriented term banks and knowledge bases.

¹for a German-Bulgarian KB MAT system ("DB-MAT"), funded by Volkswagen Foundation for three years (7/1992-6/1995).

This paper presents:

- a contrastive view to knowledge based techniques in MAT,
- mechanisms for mapping the "ordinary" linguistic lexicon and the terminological lexicon of two languages onto one knowledge base,
- methods to access the domain knowledge in a flexible way without allowing completely free linguistic dialogues,
- techniques to present the result of queries to the translator in restricted natural language, and
- use of domain knowledge to solve specific translation difficulties.

Keywords:

Machine Aided Translation, German, Bulgarian, Knowledge Based Methods, Domain Knowledge

1 Translating and domain knowledge

It is nearly trivial to say that domain knowledge is necessary to produce at least an acceptable translation [PSc86]. But it is always presupposed that translators have the corresponding domain knowledge or can acquire it easily in terms of time and resources.

Example 1:

to translate the term "Notar" from German to American English requires very detailed knowledge about the legal system in the U.S.A., or even in some of its states. The translation might be a non-terminological description of a legal role.

According to our questionnaire based inquiries [KiWi92] and other research [FHAh90] the time effort, however, to provide factual information about some details of the source text must be estimated at 30% - 50% of the translation time.

The traditional answer of the translators was to include domain knowledge in the dictionaries. Then in the last century specific domain dictionaries appeared and in our century terminological standardisation was established as an activity with its own methods and documentation. Nowadays the terminological system ideally should contain the domain knowledge by virtue of an unambiguous mapping of concepts and facts onto terms.

2 The proper place of domain knowledge in MAT

Domain knowledge is used by translators not only to understand the source text and to identify the correct terms. Translators must also find verbal descriptions, if a term is missing in the target language, or even must coin new terms.

In principle, the customer of the translator could give him/her some machine readable texts describing the subject area and the translator can find information by using the FIND option in the menu. But texts are monolingual and difficult to maintain. Moreover, the translator will always get the same text, which cannot be adapted to the specific context.

Better methods to support the translator, by giving him/her domain knowledge, are partly discussed above: It can be kept in a dictionary or in a terminological/conceptual data base. But all these sources of information are still unconnected from the translation task as are paper dictionaries or data bases. Moreover, there is no direct interconnection between lexicon entries, terminology and domain knowledge

Unquestionably, the approach described here is not the ultimate solution for all translation difficulties, because we do not assume that in such an MAT system huge amounts of conceptual material, especially in the area of everyday knowledge, can be accumulated and maintained. But the domain knowledge in our approach is language independent; it is linked to the terms and the lexicon in general, and the answers to queries are user-oriented, more specific, and language dependent.

3 Knowledge based approaches

In MAT, links which relate terms (lexicon entries) and underlying concepts ("knowledge") have been discussed for many years. We distinguish the following approaches:

- concept-oriented term banks (TBs), where a conceptual skeleton usually specifies hierarchical (thesaurus) relations among terms,
- and terminological knowledge bases (TKBs), where the meaning of a term is modelled by corresponding fragments of a conceptual representation.

In both cases the domain knowledge is organised around the lexical entries, thus being artificially disconnected and cut into pieces fitting to the semantics of the specified entry. The approach described in section 4 aims to give the knowledge the necessary central role. This is supported by its contingent representation, consistent entries, and proper algorithms for formal processing.

3.1 Term banks

The relations between terms and concepts are always critical. Two positions in concept-oriented TBs may be sketched here:

- in [CoAm93] the term is the concept label as well as the lexicon item itself. The relations among terms and concepts can be many-to-many. Bilingualism is not discussed.
- IpsiLex [Fis93] can treat bilingual material. Each natural language has its own set of conceptual structures. The conceptual representations each may differ in structure but not in their attributes and relations. A "translation relation" is a "homomorphic mapping relation" from the concepts of the source language onto the concepts of the target language. The authors do not discuss complicated translation problems for example lexical gaps.

A major weakness of the TBs is that the conceptual information included in the definitions cannot be processed computationally. The definitions are entered by different writers in different natural languages and thus the entries are inconsistent, redundant and unstructured.

3.2 Terminological knowledge bases

A terminological knowledge base is proposed as a further step in the development of knowledge based methods in term banks (for references see e.g. [Sch93]). While TBs are term-to-concept oriented, TKBs are concept-to-term oriented. This may be highly important for information retrieval: in TKBs a user can enter the characteristics of a concept asking for the correct corresponding term ("What is the correct term for a device with the function X?"[MeA192]). In translation tasks, however, this facility does not occur, because the translator is always concentrated on the two equivalent texts, not on conceptual structures (see section 6). Moreover, TBs and TKBs are not integrated in the text manipulation and the translation task.

Two approaches, which address similar problem areas as the system described below in section 4, are:

COGNITERM

Cogniterm is a bilingual TKB with terms in English and French [MeA192], constructed at the University of Ottawa. Each concept is represented by a frame-like structure with much weaker inference mechanisms. Conceptual knowledge can be visualised by semantic networks. Knowledge is accessed only by names of concepts or names of their characteristics, which is not suitable for translation tasks. The network is presented to the user directly in a conceptual (i.e. hierarchical) or alphabetical order. In [SkMe90] there are some remarks on how to treat bilingualism: introducing a second language, the authors define a second knowledge base for the target-language "based on the translation equivalents provided for concepts in the source-language knowledge base". It is not clear how complex translation phenomena can be treated.

TWB Term Bank

A TKB based on the formalism of Conceptual Graphs (CGs) has been developed for the ESPRIT-project TWB [HHAh92, HoAh92] at the University of Surrey. A corresponding conceptual graph editor has been implemented. This tool is used for the interactive graphical formation of fragments of CGs for terms in an already existing term bank. In the last available report published in 1992, however, the knowledge base contained only about 30 partially connected concepts, each related to several terms via synonyms and foreign language equivalents. There is no strategy on how to build up a domain model using the fragmentary descriptions of term meanings. The CGs representing the terminological data are networks whose nodes are designated directly by term names in one of the natural languages, which would cause difficulties with complex bilingual mappings.

4 The DB-MAT project

The project has the aim of designing and implementing an integrated MAT system, where

- language independent domain knowledge is represented by a set of Conceptual Graphs,
- there is only one integrated domain representation (the knowledge base), * the knowledge base being an autonomous domain description derived from domain material (manuals, text books, introductions, ..),
- access to the domain knowledge is always gained by natural language (source text items and generated explanations),

- links between this knowledge base and the (terminological) entries in the lexicon (or several lexica) serve as "name" relations in the graphs,
- menu-based question types together with a highlighted sequence of the text start a query,
- corresponding mapping strategies of the question onto the graph selection will return the necessary information,
- a generator will verbalise the result set for each of the different languages,
- more linguistic information can be used especially in morphology,
- Latin and Cyrillic texts can be processed flexibly on the same screen, and
- the surface language of the system (explanations, designators, etc.) can be selected.

Bilingual translations (German/Bulgarian) of the following application areas have been checked: banking, car repair, arbitration and chemical devices for oil separation.

5 Conceptual Graphs

The domain knowledge is represented by CGs. The set of CGs is called the Knowledge Base (KB). The KB consists of a type hierarchy (a lattice) and canonical graphs (basic knowledge statements). Additionally, conceptual graphs (the details of the domain) can be derived from canonical graphs according to well defined formation rules.

Example 2: ([concept A] -> (relation) -> [concept B])

(Example 2 uses the linear notation; Figure 16 - 1 contains graphical notations of CGs.)

CGs overcome some disadvantages of other formalisms applied in knowledge representation (e.g. semantic networks or frames) as they are logically adequate, distinguish elegantly between instances and classes of objects and provide a powerful inheritance mechanism.

Especially for an integrated view of concepts, terminology and other lexical items, CGs provide convenient mechanisms:

- conceptual and lexical knowledge (antonyms, synonyms,...) can be represented in the same formalism if necessary;
- the relations between terms and concepts can be represented like conceptual relations [Sow93];
- declarative statements which are typical of term definitions ("X is a part of Y") can easily be represented and generated;
- formation rules can build graphs so the translator can browse through terms by means of the relationships.

6 General Functionality

The lower right part of Figure 16 - 1 gives an impression of the surface² and the menus, the nested items of which are listed in the next section. The two windows for the source and target text are scrolling in parallel. Additional small windows will appear when necessary for requests or answers. Requests/queries are triggered by highlighted sequences of the text and menu items. Most answers are given by a generator in the specified system language (see below).

Under **File**, the user can create/select documents and inspect their records. The record contains

- practical information about the task (customer, languages, deadlines etc.),
- all flags (see above),
- all notes, and
- the customised search sequences.

Here the translator can specify the "language of the system", i.e. the natural language of the menus, the linguistic information, and the explanations.

The rest of the items under **File** and **Edit** are the usual options of the operating system and of text processing.

Under **Note** the translator can set flags (e.g. "end of edited text", "to be checked later") or insert notes, which are put into the record of the file.

Under **Info** there is the option **Explanations** with all its submenus described in section 8. **MyInfo** is a "favourite" search strategy to be customised by the user. The second option is **Grammar**, where parsing information and pre-calculated morphological tables are planned. The third option is **Lexicon**, where lexical information can be obtained. All answers of the system belonging to one marked sequence accumulate in a temporal window.

For these facilities the system has at its disposal four interconnected sources: the domain knowledge (represented in Conceptual Graphs), the (bilingual) lexicon, the morphological data (not yet implemented) and the grammar data for parsing (not yet implemented).

Under **Translation** the system presents direct correspondences or examples (from previous text). If more than one word is highlighted, either an idiom-word-term is given or the procedures of **Explanations** are triggered.

² Most of this work has been done by Kiril Simov

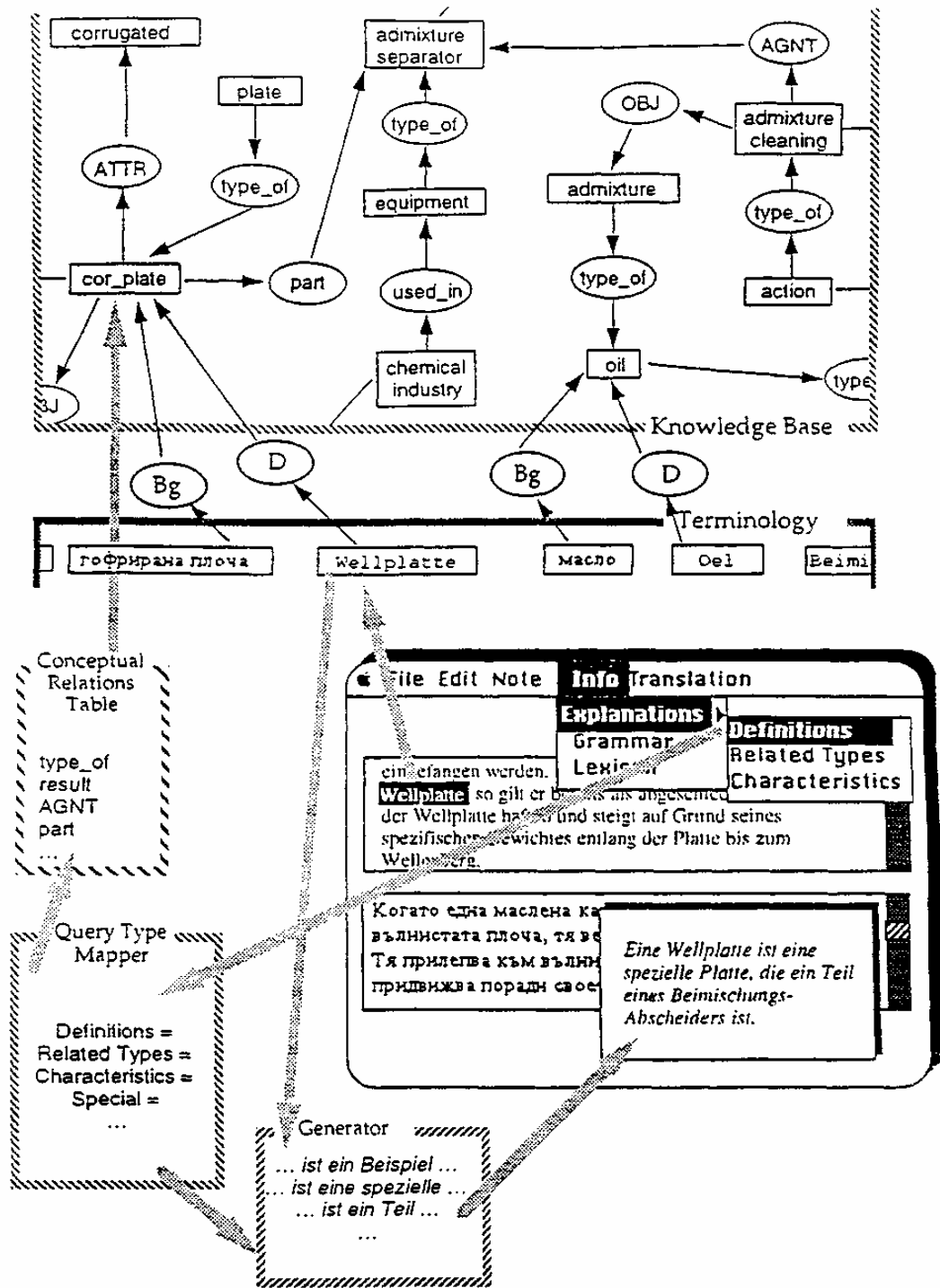


Figure 16 - 1 An explanation request and its answer

7 Accessing Domain Knowledge

Three basic assumptions guide the design of the system:

1. The internal organisation of the knowledge is different from its presentation to the user. Developers or other trained staff can build up or edit the knowledge base and the lexicon entries. Acquisition tools are planned. Translators, in contrast, will not accept plain formal expressions but expect natural language explanations which can be read more easily.
2. On the other hand the user must have a guideline on how to use the facilities that the system provides. Nested menus (instead of free text input) seem to be the appropriate way for quick and precise information.
3. Each translator may have a different view about the domain and about what is an appropriate answer. Therefore the generation of the surface answer can be customised according to the selection of graphs (the CG-Mapper) and the degree of compression.

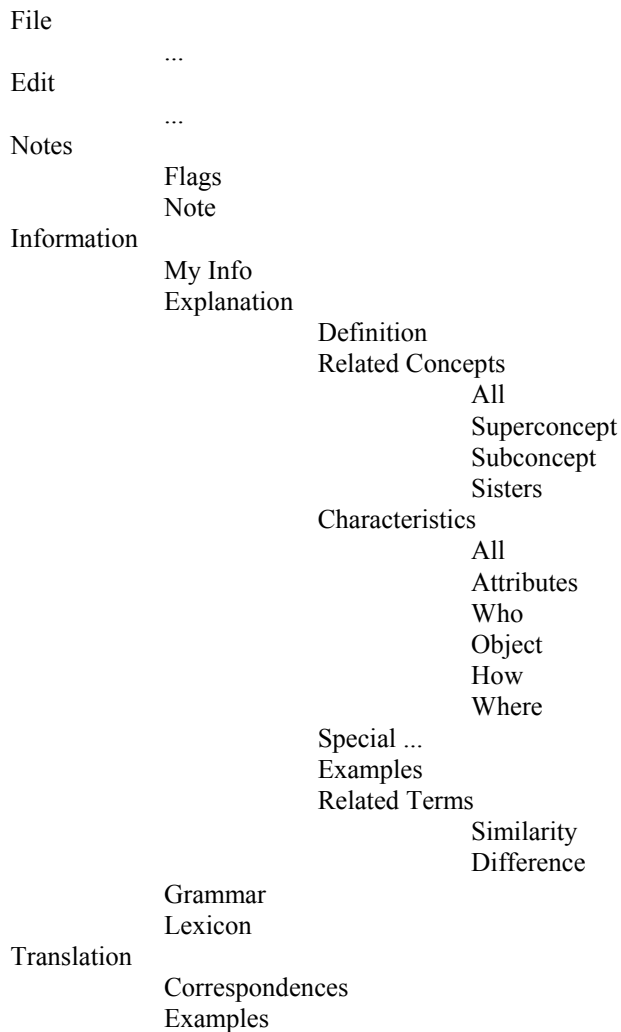
A consequence of the second assumption is the need to avoid too much natural language processing of the system and thus to organise the options for clarification in nested menus. Figure 16 - 2 shows the whole menu.

All options under *Explanation* and under *Translation* will produce results if one or more elements in the source text are highlighted (see Figure 16 - 1). In the standard case of *Explanation* there will exist a link from the corresponding lexicon entry to an item in the knowledge base. This will be verbalised according to the specific search strategy defined in the CG-MAPPER list. Non-standard cases are discussed in section 8.

Example 3: **Definition** is mapped

- onto first order arcs of the node (corresponding to the highlighted word in the text)
- for the following relation types:
 - Superconcept(s),
 - Subconcepts,
 - Sisters,
 - Attributes,
 - Function,
 - Instances.

The translator can edit this list. This is very suitable for experienced users who want to customise the reactions of the system. The menu item *My Info* is a customised free selection of functions in all submenus. Moreover the system keeps a memory of previous sequences of questions and the user can decide to repeat this pattern. Thus an implicit customisation is possible.

**Figure 16 - 2**

A feature derived from the questionnaires is "iterative clarification": The user can repeat asking about knowledge by highlighting a word in the generated answer (instead of the source text). Again he/she has the choice of the whole menu.

8 Solving Translation Difficulties by Knowledge and Multilingual Access

As pointed out in section 3, domain knowledge is mainly required when term equivalencies will not allow a direct ("blind") translation. Thus, translation problems especially arise

- when it is necessary to replace the source expression by a paraphrase in the target text (terminological gap),
- when the source expression must be replaced by a new term,
- when the meaning of the source phrase is ambiguous or not clear at all ; so the translator needs information about which interpretation is consistent to the domain knowledge (e.g. by checking attributes, parts or functions).

In the DB-MAT project we try to find solutions for these complex ("non-standard", see previous section) translation situations. The following examples may demonstrate the leading ideas:

8.1 Equivalence gaps between lexicon entries

Under *Correspondences* the lexical gap problem can be solved: If no direct translation of a term exists the user gets the *lexicalised environment*, i.e. the nearest superconcept which is lexicalised in both languages and a list of the mutually lexicalised sister concepts. [AnWi93] discusses the corresponding knowledge processing using an example from the domain of law (arbitration): Assume the user highlights "Verbandsgericht" (association court) in the source text. Accordingly, starting from the German lexicon entry the concept [association court] in the knowledge base is found. But there is no corresponding Bulgarian lexical entry reachable from the concept. To fill the lexical gap the type hierarchy of concepts and the canonical graphs are extracted; eventually, the user gets a rough idea about the term's meaning from the following output:

association court is a non-state court,
association court operates on rules,
another non-state court is the arbitration court,
 etc. [see AnWi93]

8.2 Phrase explanations

Under Explanations, more complex phenomena can be resolved: When the translator highlights a phrase of the source text he/she presupposes the existence of a complex term. At present, there are two types of phrases where the system will support the user to formulate a correct translation (or a paraphrase):

a) Adjective + Noun
 (existing in the lexicon) (existing in the lexicon, with link to the KB)

According to classical linguistic semantics, we suppose that a positive adjective is a (domain-specific) property of the noun restricting its semantics, omitting other interpretations for the moment. So, any solution of the problem can only be found in the subtree of the concept. The system thus verbalises the lexicalised conceptual environment in the target language only, for the attribute-relations (ATTR) connected to the concept of the noun itself and those of all its sub-concepts, respectively. Iterative requests will be resolved within the subtree, too.

b) Adjective + Noun
 (not in the lexicon) (existing in the lexicon, with link to the KB)

If the adjective is not contained in the lexicon we suppose it is not a trivial, every-day adjective but has some domain-specific meaning (the translator would not highlight e.g. "new court").

The system presents the lexicalised conceptual environment in both source and target languages, for the ATTR-relations connected to the concept of the noun itself and those of all its sub-concepts, respectively.

In this way for phrase explanations in cases a) and b) the translator can see the domain structure and interpret it, finding either an adequate synonym or another noun/term (linked to a subconcept) for a collocation.

8.3 Processing similarity and difference

Under *Related Terms* the translator can obtain information about the interrelation of two terms (one term highlighted in the text and one entered in a special window, see [AnWi93]). This support is helpful in two different situations:

- The translator asks about the relation of a term he knows to an unknown term to learn its meaning according to the domain knowledge in the KB;
- The translator asks about the relation among two unknown terms of the source text in order to get the context and the translation.

In both cases, similarity and difference, the system follows the links of the highlighted terms to their concepts in the KB and searches in the type hierarchy for the first common superconcept of the chosen concepts. Since the type hierarchy is a lattice there always exists at least one common superconcept.

a) Similarity

Likewise, the system searches for similarity by retrieving canonical graphs, which differ only in one of the chosen concepts. Asking about the similarity of [Arbitration court] and [Association court] would produce the following result from the type hierarchy and the canonical graphs:

[Arbitration court]->(type_of)->[Court]<-(type_of)<-[Association court]

[Arbitration court]->(attribute)->[non state]

[Association court]->(attribute)->[non state],

which may be expressed by the generator as:

*"Arbitration court is a type of court "
 "Association court is a type of court "
 "Arbitration court is a non-state court " and
 "Association court is a non-state court ".*

b) Difference

The difference between two chosen concepts is found by checking the type hierarchy for the lowest common superconcept and then displaying the differences between the daughters of that superconcept. Given the following type hierarchy asking about the difference between arbitration court and supreme court

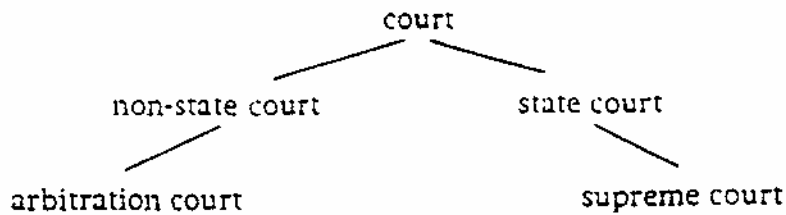


Figure 16 - 3

would produce the following answer:

"Arbitration court is a non-state court " "Supreme court is a state court ".

9 Layers and links in the lexicon

The general organisation of the lexicon follows the principles derived from the questionnaires for the translators. There exists only one lexicon in one general format containing

- the basic vocabulary, which often is called "normal" or "ordinary" (everyday words, functions words etc.), and
- the terminology.

with a stem entry each. An elementary morphology, which is necessary anyway for the morphological information and the generator option will reduce selected text sequences.

This unified lexicon structure has been chosen because from domain to domain words can change their membership in a split lexicon structure. Moreover, first, there is no difference between the entries of ordinary words and terms and, second, the user will sometimes not be aware of the term character of a word.

Entries of ordinary words consist of the usual information (part of speech, gender, inflection class, multilingual equivalencies, abbreviations, etc.) in a formalised and ordered way. A first sketch of an acquisition tool exists already.

Entries of terms and other domain specific entries differ from ordinary words only in having a link to the knowledge base. Complex terms are represented by pointers among lexical entries (see Figure 16 - 4). This structure allows us to recognise complex terms contained in complex terms.

Some details are still to be defined: The combinatorics of complex/compound words or terms and simple/complex/compound multilingual equivalents require a rather sophisticated structure. (see [vHAn94])

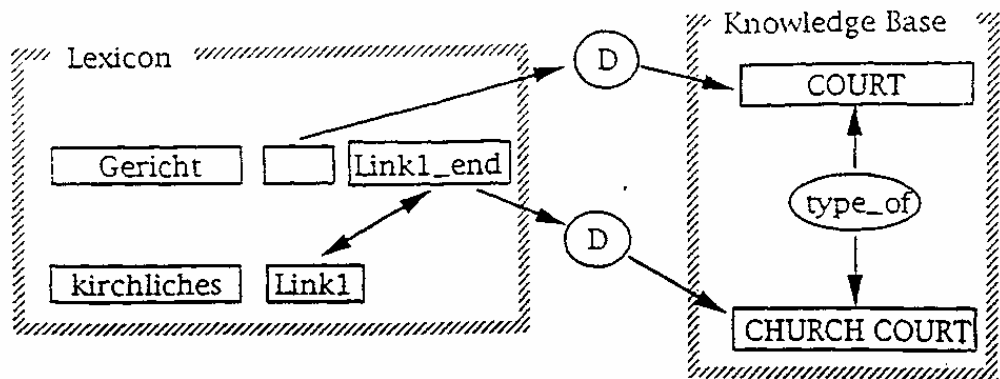


Figure 16 - 4

10 Generation

According to the application in technical translations of manuals and instructions, the generation component must be simple enough to be duplicated for other languages. Therefore a very pragmatic compromise between full grammatical generation and simple template generation has been chosen. Templates and compressing mechanisms (quantifiers, conjunction) according to the morphology will be sufficient to make the output readable and allows for easy maintenance.

11 Discussion

Some technical features of the system are still under design and some general characteristics of the system might be subject to criticism, among them:

1. Every knowledge representation is only a small fragment of what an expert might have in mind when speaking about his/her field. Therefore we do not claim that all possible factual ambiguities or translation problems can be solved by the system described in this paper. We only encode some important concepts of the field. If these fields are small enough the set of represented concepts is sufficient for considerable clarification. Moreover the user can move around the field by asking consequent queries. Furthermore the envisaged application of such a system will be in technical areas where for example manuals or instructions will be translated for companies into less used languages and in frequent iterations. The objects "behind" the texts can be represented under a very restricted scope and without any elaborated upper model. The system shares this problem with all dictionaries and term banks.

2. It might be seen as a practical problem that formal representations in real life applications are not available by default. But such representations are needed more and more for other system types, for example in in-house information and documentation systems. Additionally, with growing markets in the future, translations into less used languages will be necessary even for middle size enterprises. For some languages of the Far East or the Pacific area only general translators will be available, who do not have expertise in specific domains. The DB-MAT activities can help to produce better results.

3. Formal representations are more difficult to acquire in comparison to dictionary or data base entries. But the development of knowledge based tools (using knowledge about the representation language and its consistency rules for example) with intelligent prompts and examples will make acquisition much easier. Furthermore, building and maintaining MT-lexica does not seem to be simpler. System specific lexica, however, remain system-dependent, whereas knowledge bases can be used in more than one embedding system.

4. The DB-MAT approach can be used only in restricted domains (in [vHAn92] described as finite). In terms of operational characteristics such finite domains are technical domains where most referents (objects) denoted by a term or expression exist physically and are independent of a specific linguistic/cultural environment. On the other hand, conceptual models of several domains will be accumulated in time and raise the power of the system linearly.

5. Some decisions in the lexicon structure are still open.

12 Implementation

Since summer 1994 there exists a complete implementation of the user interface (menus, windows with Latin and Cyrillic characters, parallel scrolling) and the basic linguistic customising techniques as well as a first sketch of the conceptual retrieval system with a German generator. The programs are running on Macintosh and LPAProlog. Other components are being implemented.

References:

[AnWi93] Angelova, G. and Winschiers, H. Knowledge based explanations in MAT. Project "DB-MAT", Rep.4-93, Uni Hamburg, Dec. 1993.

[CoAm93] Condamines A. and Amsili P. Terminology between language and knowledge: an example of termin. knowledge base. In [Sch93].

[FHAh90] Fulford, H., Hoege, M., Ahmad,K.: TWB Project. User Requirements Study. Rep. 3/1990.

[Fis93] Fischer D. Consistency rules and triggers for multilingual terminology. In [Sch93].

[FrHe90] Freibott G. and Heid U. Terminological and lexical knowledge for computer-aided translation and technical writing. 2nd Int. Congress on Terminology and Knowledge Engineering TKE, 1990.

[Hei93] Heid,U.On the Representation of Collocational Phenomena in Sublanguage Lexicons. In [Sch93].

[HHAh92] Holmes-Higgin, P. and Ahmad, K. The Machine Assisted Terminology Elicitation Environment: Text and Data Processing and Management in Prolog. Rep. 8, Uni Surrey, 1992.

- [HoAh92] Hook, S. Ahmad, K.: Conceptual Graphs and Term Elaboration: Explicating (Terminological) Knowledge. TWB: ESPRIT II No.2315, TR 10, Univ. of Surrey, July 1992.
- [KiWi90] Kieselbach, C. and H. Winschiers. Studie zur Anforderungsspezifikation einer computergestützten Übersetzerumgebung. Studienarbeit, Uni Hamburg, 1990.
- [KiWi92] Kieselbach, C. and H. Winschiers. Benutzungsschnittstelle und Komponenteninteraktion einer wissensbasierten, integrierten, computergestützten Übersetzerumgebung. Dipl. Thesis, Uni Hamburg, 1992.
- [Kuk90] Kukulska-Hulme A. Speed understanding of an unfamiliar Domain. 2nd Int. Congress on TKE. Indeks Verlag, 1990.
- [MeAl92] Meyer I., Skuce D., Bowker L. and Eck K. Towards a new generation of terminological resources: an experiment in building a TKB. COLING'92.
- [NiGo92] Goodman, K. and S. Nirenburg (eds.) The KBMT project: A Case Study in Knowledge-Based MT. Morgan Kaufmann Publ., Inc., 1991.
- [PSc86] Schmitt, P. Die "Eindeutigkeit" von Fachtexten: Bemerkungen zu einer Fiktion. In: M. Snell-Hornby (ed.) Übersetzungswissenschaft - eine Neuorientierung, Francke Verlag Tübingen, 1986, pp. 253-282.
- [Sch93] Schmitz, K. (ed.) Terminology and Knowledge Engineering. Pr. 3rd Int. Congr., 1993, FRG.
- [Schin92] Schindel, Petra. Die Schiedsgerichtsbarkeit - eine Methode der Konfliktregelung im internationalen Wirtschaftsverkehr. 1992.
- [Sim93] Simov, K. The Role of the Morphological Component in MAT. Project "DB-MAT", Rep. 2-93, Uni Hamburg, July 1993.
- [SkMe90] Skuce D. and Meyer I. Concept Analysis and Terminology: A Knowledge-Based Approach to Documentation. COLING'90, pp. 56-58.
- [Sow84] Sowa, J. Conceptual Structures: Information Processing in Mind and Machine. 1984.
- [Sow93] Sowa, J. Lexical Structure and Conceptual Structures. In: Pustejovsky, J. (ed.) Semantics in the Lexicon.. Kluwer Academic Publishers, 1993.
- [vHa90] v.Hahn, W.: Considerations for the design of a translator's workbench. 10th Ann. Conference LSP and Theory of Translation. Vaasa, Finland, 1990.
- [vHAn92] v.Hahn, W. and G. Angelova. Knowledge Based MAT. (to appear Computers and AI, Bratislava)
- [vHAn94] v. Hahn, Walther and G. Angelova. System Architecture and Some System-Specific Components in Knowledge Based MAT. Project DB-MAT, Technical Report 1-94, University of Hamburg, July 1994.