KNOWLEDGE BASED MACHINE AIDED TRANSLATION

Walther v. Hahn

Abstract

The paper presents and demonstrates the system DB-MAT, a machine aided translation prototype system.

Central aim: support of human translation of industrial technical texts by allowing for clarification questions about the domain of the text. Additionally, pictures from the domain are included in retrieval and in the lexicon of the system.

Functionality: The translator can select chunks of the source text (or the target text or even from answers to previous queries) and chooses from a nested query menu. The system will derive answers from an internal language independent knowledge base and will present the answer in coherent natural language (at the moment German and Russian). The demo domain is oil/water pollution texts in German and Bulgarian.

Prof. Dr. Walther v. Hahn

Prof. Dr. Walther v. Hahn is professor of computer science at University of Hamburg and director of its "Natural Language Systems" division. He conducted numerous projects in applied computational linguistics and is member of the national VERBMOBIL project (member of steering committee).

Computer Science Department of the University of Hamburg

Computer Science Department of the University of Hamburg is one of the biggest CS faculties in Germany and houses 5 institutions dealing with knowledge based systems / cognition.

Prof. Dr. Walther v. Hahn
Computer Science Department
Natural Language Systems Division
Vogt-Koelln-St. 30
D - 22527 HAMBURG
Telephone: +49 - 5494-2434
Fax +49-5494-2515
E-mail: vhahn@informatik.uni-hamburg.de
Internet: http://www.informatik.uni-hamburg.de/NATS/staff/vhahn.html

PROVIDING MULTILINGUAL TERM EXPLANATIONS IN MACHINE AIDED TRANSLATION[1]

Walther von Hahn[2]

University of Hamburg, Germany,
Natural Language Systems Division

0. Abstract

This paper describes the system DBR-MAT which supports translators by allowing for domain specific queries from an abstract knowledge base instead of giving monolingual canned text explanations or term definitions. The interaction of the lexicon, the knowledge base, the graphical objects and the generator are explained in more detail. The second part of the paper describes the user tools for lexicon acquisition and browsing the knowledge base. Further information and all publications may be obtained from the DBR-MAT homepage[3].

1. Project Objectives:

DBR-MAT (Deutsch-Bulgarisch-Rumänisches MAT) is a Machine Aided Translation (MAT) project. Its central aim is to support translators' work by providing domain knowledge, i.e. allowing for clarification questions and integrating pictures.

The methods applied in this project are an interconnection between the lexicon and a language independent knowledge base of Conceptual Graphs well as a stock of fully indexed pictures.

The pilot system has been tested in the domain of oil/water pollution and corresponding separation technology.

[1] Research is funded by Volkswagen Foundation (1992 - 1997)

[2] This paper includes scientific results achieved by Dr. Galja Angelova, Kalina Bontcheva (Sofia) and Heike Petermann (Hamburg). See also v.Hahn/Angelova 1996

[3] For further information:
http: //www.informatik.uni-hamburg.de/Arbeitsbereiche/NATS/projects/db-mat.html

2. Motivation

To day translation support mainly means linguistic data and translation memory. If there is any domain knowledge included at all, it is a selection of domain texts. Such explanatory texts may appear as product descriptions by the customer, term definitions in the term bank, or encyclopedic texts on the domain as a whole. These texts, however, are either monolingual or must be translated again to several other „standard" languages to be useful as background knowledge for translators in more than one language pair.

Even common sense will claim that understanding the domain knowledge included in a text is crucial for the quality of translation. Additionally, questionnaires from translators[4] verified that about 40% of the overall translation time is spent for contents clarifications. The aim of the project is to reduce these 40% by a flexible and user friendly information facility for a translation tool.

The design principle derived from these facts is to support flexible, language independent, global and integrated access to domain knowledge.

3. The DBR-MAT paradigm

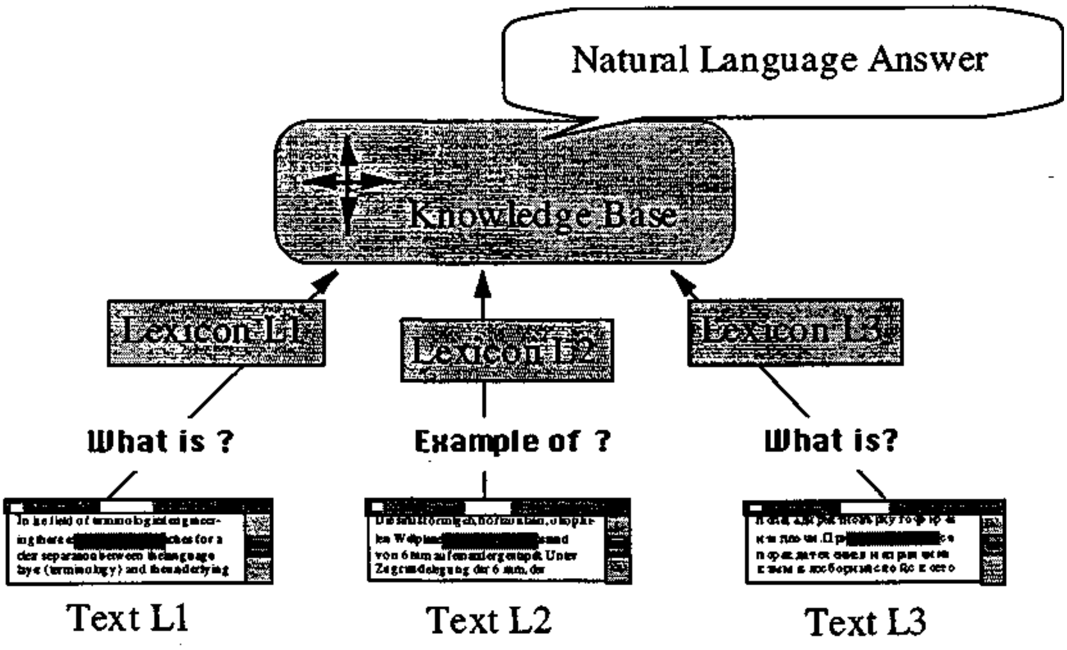


figure 1: from selecting a term to natural language answers

[4] Kieselbach /Winschiers 1990

The central functionality provided by DBR-MAT consists of the following steps:

The user

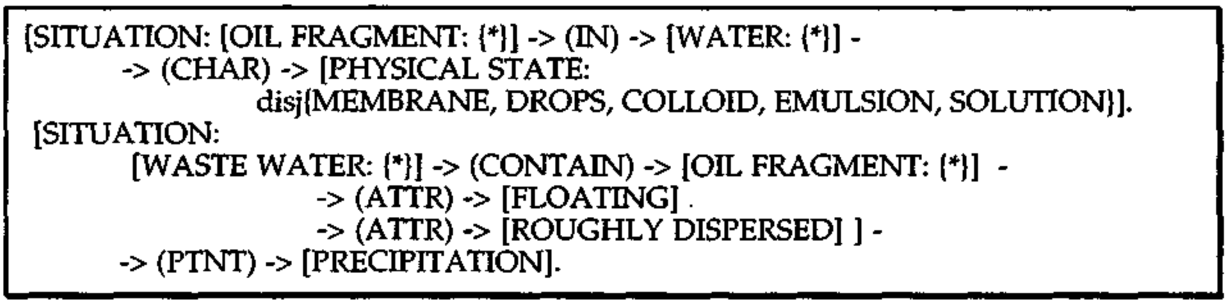
- translates in two windows (source and target text),
- selects a piece of text either in the target language or the source language or the previous explanation (text 1,2,3)
- chooses a question type from the "Information" menu (e.g., *What is?)*
- receives a natural language answer for the selected text (if it corresponds to a technical term) from an abstract knowledge base.
- may ask for corresponding graphics.

4. The Knowledge Base

The objects of the KB are primarily concepts such as [DEVICE], but not lexical units, they are secondarily linked to lexical entries. Every object has an internal arbitrary object name (a formal designator) like [OIL_SEPARATOR_1]. Concepts are connected by conceptual relations.

In the representation language "Conceptual Graphs" (Sowal984), concept types and relation types are organized in a type hierarchy.

An example of two conceptual graphs (contexts) in "linear notation":



```
[SITUATION: [OIL FRAGMENT: {*}] -> (IN) -> [WATER: {*}] -
        -> (CHAR) -> [PHYSICAL STATE:
                disj{MEMBRANE, DROPS, COLLOID, EMULSION, SOLUTION}].
 [SITUATION:
        [WASTE WATER: {*}] -> (CONTAIN) -> [OIL FRAGMENT: {*}]  -
                -> (ATTR) -> [FLOATING] .
                -> (ATTR) -> [ROUGHLY DISPERSED] ] -
        -> (PTNT) -> [PRECIPITATION].
```

table 1: A situation in Conceptual Graphs

5. How does the system organize its knowledge?

One single coherent domain model (the Knowledge base) supports all meanings of terminological entries in all languages. In contrast to similar approaches the answers to clarification questions are not direct quotes from the knowledge base (in the formal representation), but natural language answers. They are composed from the result of the query by applying graph operations and a natural language generator. The knowledge base (the conceptual structure) is attached to lexicon entries as their "meaning".

A meaning in DBR-MAT is one (or more) pointer to "starting nodes" in the knowledge base (KB). From this starting node the search procedure selects the information specified by the selection from the query menu (What is? exampie, characteristics, differences, ...). Due to this technique the borders of word meanings are fuzzy because a user can start querying iteratively out of the system's answers.

The knowledge base (KB) represents:

KB objects	*Example*
concepts	[OIL SEPARATOR]
individuals	the separator C334
conceptual relations	part_of
contexts (situations)	[SITUATION: [WASTE WATER: {*}] -> ...
a type hierarchy	[SEPARATOR] ↓ [OIL SEPARATOR]

table 2: objects of the knowledge base

The contents of the KB is acquired from a textbook of the domain rather than from lexical material to make it independent of the test material.

The representation formalism of Conceptual Graphs is well suited for such types of tasks, because it can represent typical properties of language-near objects (terms) and allows for grouping objects ("contexts"). This is an interesting method to represent the fact that words in different languages have a different coverage of a conceptual array:

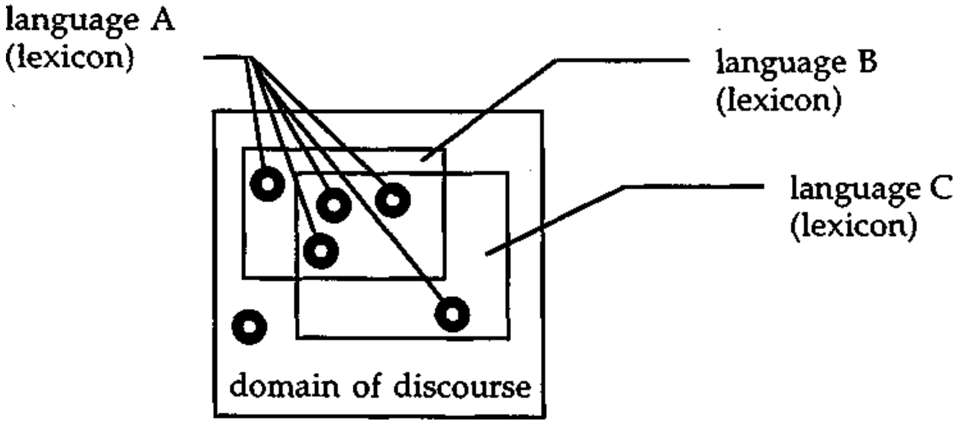


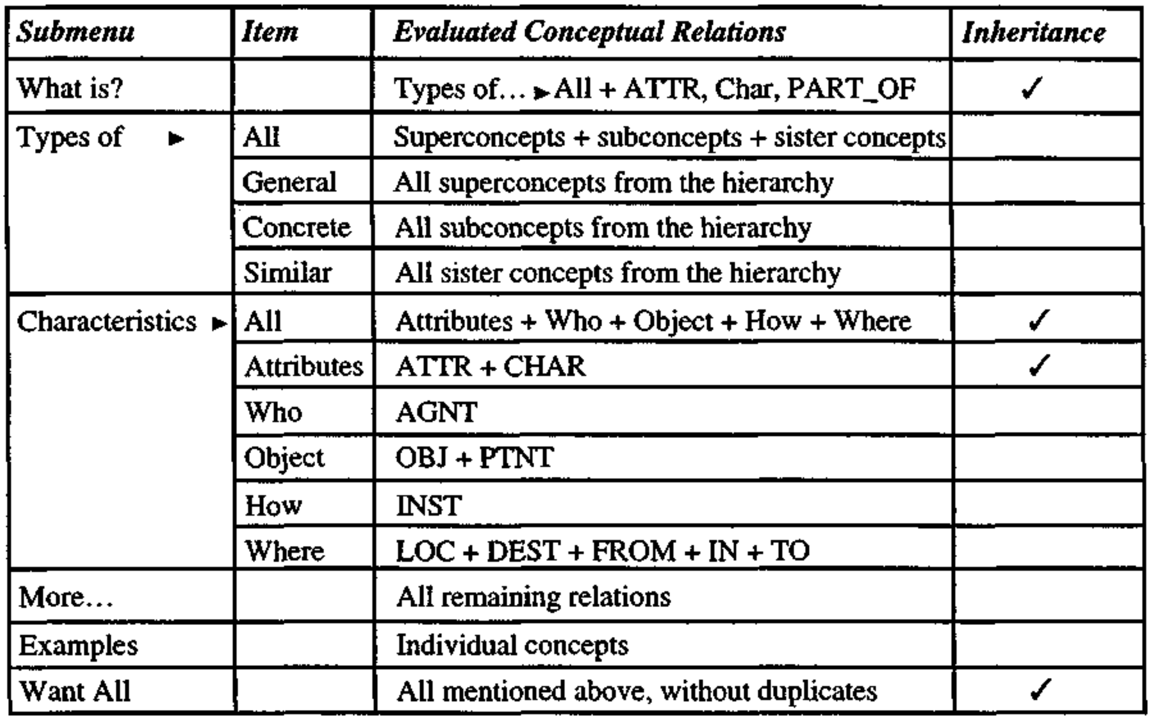
figure 2: conceptualization in different languages

In a given technical domain language A has terms for all objects in this small sample domain, whereas language B and language C depict the concept area with less granularity (or a different conceptualization) each.

All information sources are interconnected. Lexical items point systematically to conceptual items and to facts in the domain. Figures are accessible from the lexicon as well as from knowledge items. Complex terms *("Very Large Scale Integration", e.g.)* are handled correctly.

6. Traversing Rules of the Query Mapper

When answering clarification questions the system follows specific paths in the knowledge base (KB) according to the type of question chosen by the translator. In the following table you find in the first two columns the query types and their subtypes, in the third column the evaluated relations.



Submenu	*Item*	*Evaluated Conceptual Relations*	*Inheritance*
What is?		Types of... ►All + ATTR, Char, PART_OF	✓
Types of ►	All	Superconcepts + subconcepts + sister concepts	
	General	All superconcepts from the hierarchy	
	Concrete	All subconcepts from the hierarchy	
	Similar	All sister concepts from the hierarchy	
Characteristics ►	All	Attributes + Who + Object + How + Where	✓
	Attributes	ATTR + CHAR	✓
	Who	AGNT	
	Object	OBJ + PTNT	
	How	INST	
	Where	LOC + DEST + FROM + IN + TO	
More...		All remaining relations	
Examples		Individual concepts	
Want All		All mentioned above, without duplicates	✓

table 3: DBR-MAT's query mapper

Example: A user highlighted the word "Wellplatte" in the source text and selected the query **What is?** in the menu. The word "Wellplatte" is looked up in the lexicon, there the system will find an Id of a KB object, say „2245". Thus the enter point in the KB is defined. Starting from this point the first line of table of the query mapper is executed:

1. find all relations given under **Types of ... —► All** which means
   - Superconcepts
   - Subconcepts and
   - Sister concepts

2. Find all “Attr” relations,
3. Find all “Char” relations,
4. Find all “part_of” relations.
5. Apply inheritance (if “More” is selected under *“Details”,* see paragraph 9).

7. **Generating natural language** (cf Bontcheva[5] 95)

The result of applying these search rules is already the answer to the query. However, the Conceptual Graphs formalism is not readable by a naive user. Moreover, the result set contains duplicates and trivialities and is not interconnected.

Therefore a generator (EGEN) evaluates the result set coming from the knowledge base (avoiding duplicates, connecting subparts etc.) and produces a natural language answer. The generation is kept language independent until the last steps (e.g. consulting the KB, traversing the type hierarchy, extracting the relevant knowledge items, determining “what-to-say” is language independent). This facilitates the development of new generators of other languages. DBR-MAT presently can run generators for German and Russian.

8. The DBR-MAT Lexicon

The lexicon of DB-MAT consists of several sublexicons and has the following modular structure:

{<LexEntry>, <LexMorpho>, <LexSyntax>, <LexText>, <LexTrans>}

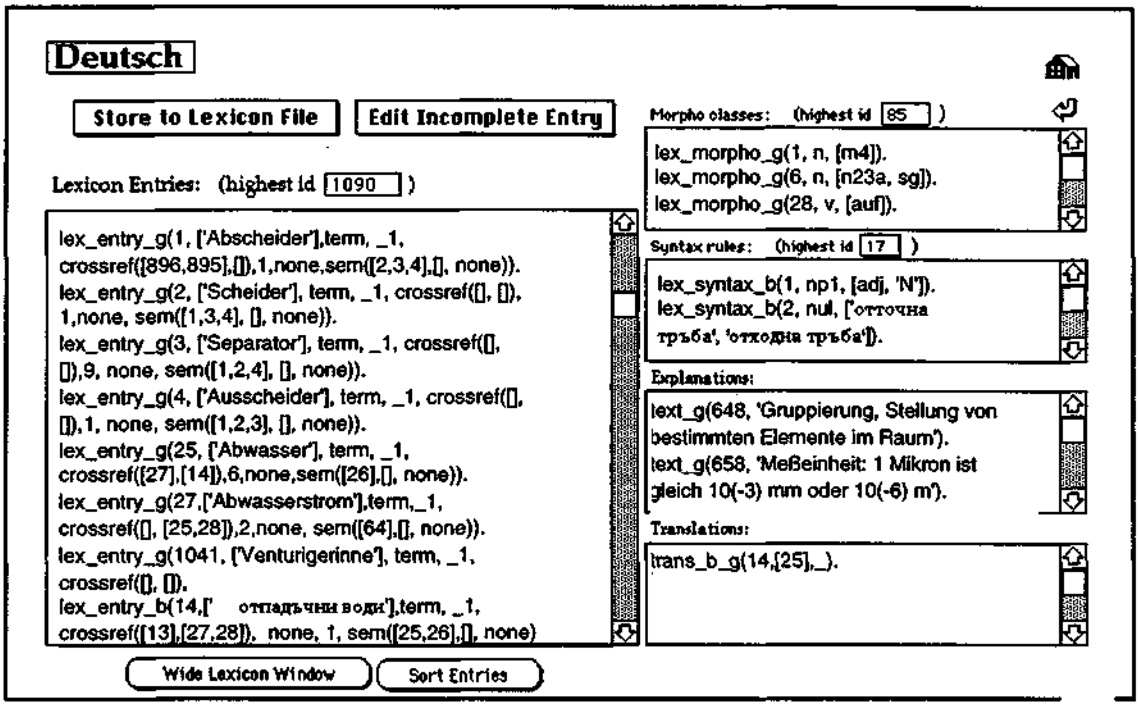
where LexEntry is the main list of entries:

```
<LexEntry> := lex_entry_<x>( <Id>, <Entry>, <Type>, <Annotation>,
                 [<CrossRefGroup>, <MorphoGroupId>, <SyntaxGroupId>,
                 SemGroup] ).
```

<CrossRefGroup> is a set of Ids referring to those lexicon entries which are contained in the entry at hand or which contain this entry. The arguments <MorphoGroupId> and <SyntaxGroupId> refer to the corresponding rule modules of the lexicon, in which their structure is described. <LexSyntax> is only relevant for complex terms and describes their phrase structure.

In figure 3 a fragment of the lexicon is displayed as it is acquired by the tool HyperLAT (see chapter 14).

[5] Bontcheva, Kalina: Generation of Multilingual Explanations from Conceptual Graphs. In: Processings of Recent Advances in Natural Language Processing, 14-16. Sept. 1995, Tzigov Chark/Bulgaria. S. 184-190.



Lexicon Entries: (highest id 1090)

```
lex_entry_g(1, ['Abscheider'],term, _1,
crossref([896,895],[]),1,none,sem([2,3,4],[], none)).
lex_entry_g(2, ['Scheider'], term, _1, crossref([], []),
1,none, sem([1,3,4], [], none)).
lex_entry_g(3, ['Separator'], term, _1, crossref([],
[]),9, none, sem([1,2,4], [], none)).
lex_entry_g(4, ['Ausscheider'], term, _1, crossref([],
[]),1, none, sem([1,2,3], [], none)).
lex_entry_g(25, ['Abwasser'], term, _1,
crossref([27],[14]),6,none,sem([26],[], none)).
lex_entry_g(27,['Abwasserstrom'],term,_1,
crossref([], [25,28]),2,none, sem([64],[], none)).
lex_entry_g(1041, ['Venturigerinne'], term, _1,
crossref([], []).
lex_entry_b(14,['    отпадъчни води'],term, _1,
crossref([13],[27,28]),  none, 1, sem([25,26],[], none)
```

Morpho classes: (highest id 85)

```
lex_morpho_g(1, n, [m4]).
lex_morpho_g(6, n, [n23a, sg]).
lex_morpho_g(28, v, [auf]).
```

Syntax rules: (highest id 17)

```
lex_syntax_b(1, np1, [adj, 'N']).
lex_syntax_b(2, nul, ['отточна
тръба', 'отходна тръба']).
```

Explanations:

```
text_g(648, 'Gruppierung, Stellung von
bestimmten Elemente im Raum').
text_g(658, 'Meßeinheit: 1 Mikron ist
gleich 10(-3) mm oder 10(-6) m').
```

Translations:

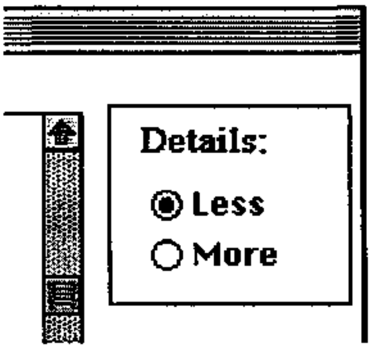
```
trans_b_g(14,[25],_).
```

figure 3: DBR-MAT lexicon

9. Linking Lexical Material

The links from the lexicon entries to the knowledge units are bi-directional:

Lexicon to KB:
starting from a lexicon entry (the one selected in the text) the semantics of a term can be calculated by traversing the KB, or
KB to lexicon:
the generator can search for appropriate terms for KB units
The KB can be evaluated to a defined depth depending on the specification of the user (figure 4)



Details:
● Less
○ More

figure 4. choice of detail

less: the local environment of a concept is consulted only,
more: the concept hierarchy is evaluated to include inherited features of more general concept types.

This modular and interconnected representation method has the following advantages:

- you see the tacit assumptions included in terms (inherited from the whole terminological system),

- you can traverse the whole terminological material in a coherent conceptual system, and
- you can inspect the terminological environment of lexical gaps.

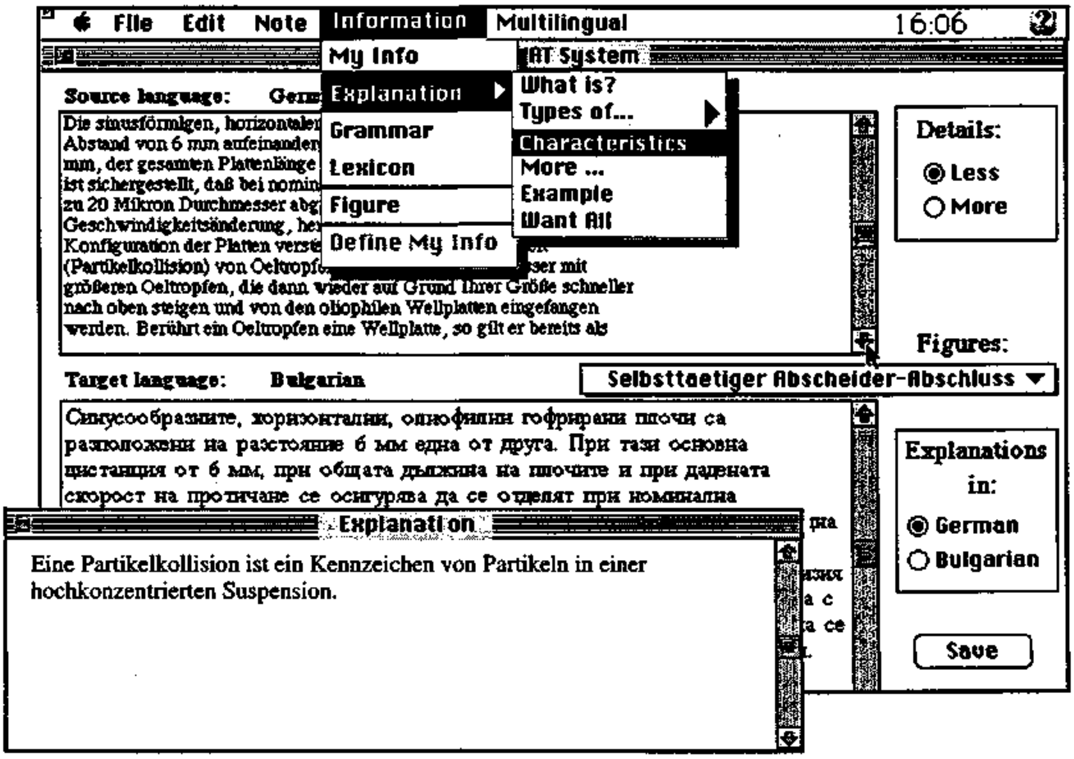
10. Costs and benefits for the translator

The DBR-MAT paradigm requires the preparation of additional data. The effort to do so, however, must be compared to the effort to prepare and maintain a term bank. It is well known that maintenance of term banks concerning consistency and homogeneity is rather expensive. In DBR-MAT, tools support the terminologist and lexicographer to a high degree, they reduce the amount of work and guarantee consistency and formal correctness.

The benefit of the knowledge based approach are

- new terms can immediately be verbalized in all languages, for which a generator already exists,
- no translation is necessary in term banks,
- term definitions in DBR-MAT are interconnected,
- the user will see not only definitions but further explanations.
- the modularity of components, esp. of the generator, make them reusable for other tasks

11. DBR-MAT User Interface



Source language: Germ

Die sinusförmigen, horizontaler
Abstand von 6 mm aufeinander
mm, der gesamten Plattenlänge
ist sichergestellt, daß bei nomin
zu 20 Mikron Durchmesser abg
Geschwindigkeitsänderung, he
Konfiguration der Platten verst
(Partikelkollision) von Oeltropfe ser mit
größeren Oeltropfen, die dann wieder auf Grund ihrer Größe schneller
nach oben steigen und von den oliophilen Wellplatten eingefangen
werden. Berührt ein Oeltropfen eine Wellplatte, so gilt er bereits als

Target language: Bulgarian

Selbsttaetiger Abscheider-Abschluss ▼

Синусообразните, хоризонтални, олиофилни гофрирани плочи са
разположени на разстояние 6 мм една от друга. При тази основна
дистанция от 6 мм, при общата дължина на плочите и при дадената
скорост на протичане се осигурява да се отделят при номинална

Eine Partikelkollision ist ein Kennzeichen von Partikeln in einer hochkonzentrierten Suspension.

figure 5: DBR-MAT user interface

The user interface contains two text windows and a variety of menus to access the lexicons and the knowledge base. The menu at the moment contains (besides usual items like **File** and **Edit):**

Note to insert notes and flags to the text, which (in a previous version of DBR-MAT) were included in a translation document, where the translators can see all the flags in context,
Information with the facilities to ask questions about the domain, and linguistics,
Multilingual for language correspondences.

12. The Architecture of DBR-MAT

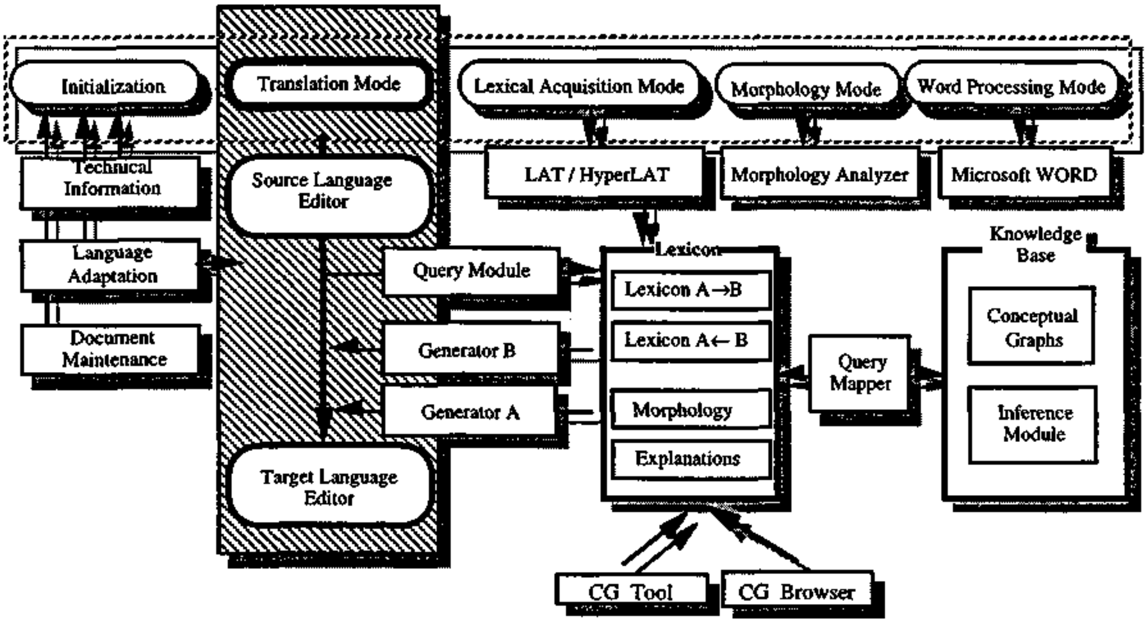


figure 6: System architecture of DBR-MAT

DBR-MAT is designed to have 5 modes (the vertical columns of figure 6) , three of which are implemented yet. In the center of the graphic the flow of information from queries to the lexicon, the query mapper and the KB in both directions is sketched.

13. No System without User Tools

In a realistic environment complex systems without additional tools will need as much time for maintenance as they save by their usage. Therefore the maintenance and the acquisition of data must be supported by powerful and easy tools.

In DBR-MAT the acquisition and maintenance of the lexicon and the knowledge base can be done (in near future) by a normal user, otherwise the costs for these activities will exceed the benefits of DBR-MAT. The following tools are available or under work (gray) for the languages given in the balloons:

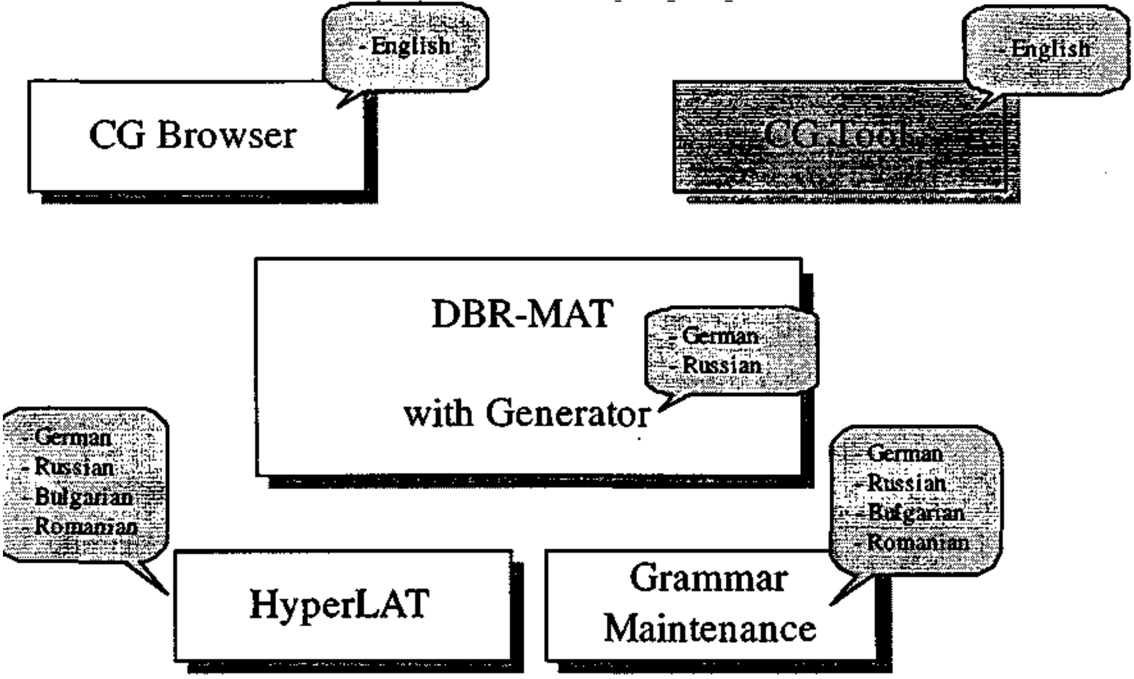
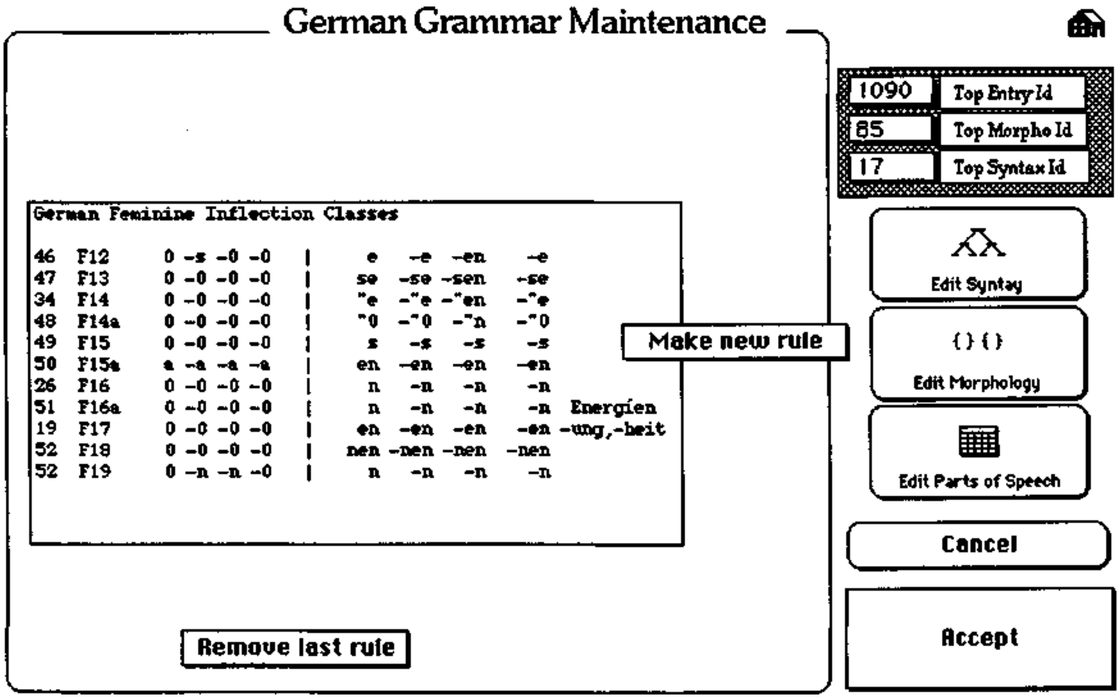


figure 7: tools around DBR-MAT

14. HyperLAT Lexicon Tool



```
German Feminine Inflection Classes

46  F12    0 -s -0 -0   |   e   -e  -en   -e
47  F13    0 -0 -0 -0   |  se  -se -sen  -se
34  F14    0 -0 -0 -0   |  "e  -"e -"en  -"e
48  F14a   0 -0 -0 -0   |  "0  -"0  -"n  -"0
49  F15    0 -0 -0 -0   |   s   -s   -s   -s
50  F15a   a -a -a -a   |  en  -en  -en  -en
26  F16    0 -0 -0 -0   |   n   -n   -n   -n
51  F16a   0 -0 -0 -0   |   n   -n   -n   -n Energien
19  F17    0 -0 -0 -0   |  en  -en  -en  -en -ung,-heit
52  F18    0 -0 -0 -0   | nen -nen -nen -nen
52  F19    0 -n -n -0   |   n   -n   -n   -n
```

figure 8: Grammar definition tool of HyperLAT

Only a few lines about the lexicon acquisition tool HyperLAT:

This tool relies (except entering the entry itself) only on clicking values from specification tables. This principle rules out any typing errors and allows only reasonable values for a given linguistic item. Endless code lists on the lexicologists table are unnecessary, because everything is displayed on the screen and only when linguistically appropriate. On figure 8 a user is just about defining the encoding table for feminine nouns in German. To make the choice unambiguous the table shows all endings of each classes.

15. CG Browser

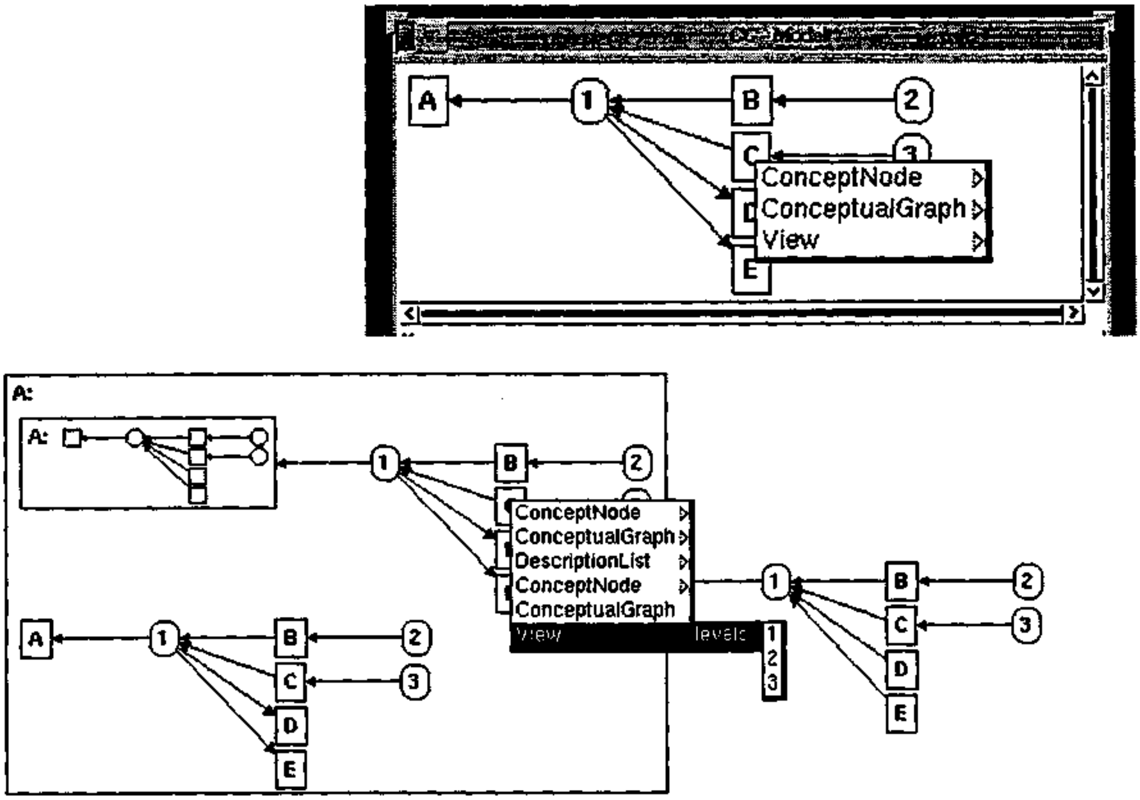


figure 9: Visual CG browser

16. Status of Implementation

The current laboratory version of DBR-MAT illustrates all important design principles by fully running components (implemented in LPA Prolog). Components in German, Bulgarian, Romanian and Russian are implemented.

17. References:

ANGELOVA, G. / BONTCHEVA K. (1996): DB-MAT: Knowledge Acquisition, Processing and NL Generation using Conceptual Graphs. To appear in Proc. ICCS-96, August, Lecture Notes of Artificial Intelligence.

BONTCHEVA, K. (1995): Generation of Multilingual Explanations from Conceptual Graphs. In: Processings of Recent Advances in Natural Language Processing (RANLP), 14-16. Sept. 1995, Tzigov Chark/Bulgaria. S. 184-190.

BOGURAEV, B./ PUSTEJOVSKY, J. (1990): Lexical Ambiguity and the Role of Knowledge Representation in Lexicon Design. In: COLING-90.

BOWKER,L./ MEYER, I. (1993): Beyond “Textbook” Concept Systems: Handling Multidimensionality in a New Generation of Term Banks. In: Proceedings of TKE'93: Terminology and Knowledge Engineering.

FULFORD, H./ HOEGE, M./ AHMAD, K.(1990): Translator’s Workbench Project. User Requirements Study. Report, March.

v.HAHN, W. (1992): Innovative Concepts for Machine Aided Translation. In: Proceedings VAKKI, Vaasa, Finland, pp. 13-25.

v.HAHN, W. / ANGELOVA, G.: (1996): Combining Terminology, Lexical Semantics and Knowledge Representation in Machine Aided Translation. In Galinski, Chr. and Schmitz, K.-D. (eds.): Proceedings of TKE'96: Terminology and Knowledge Engineering. Frankfurt/M. 304 - 314.

KIESELBACH, C./WINSCHIERS, H. (1990): Studie zur Anforderungsspezifikation einer computergestützten Übersetzerumgebung. Studienarbeit, Universität Hamburg.

MEYER, I. (1994): Helping Terminologist Do Knowledge Engineering: Some Linguistic Strategies and Computer Aids. In: Actualite Terminologique, December.

MEYER, I./ SKUCE, D. / BOWKER,L. /ECK, K. (1992): Towards a New Generation of Terminological Resources: An Experiment in Building a Terminological Knowledge Base. In: COLING-92.

SKUCE, D. / LETHBRIDGE, T. (1996): CODE4: A Unified System for Managing Conceptual Knowledge. To appear in International Journal of Human-Computer Studies and Knowledge Acquisition.

SOWA, J. (1984): Conceptual Structures: Information Processing in Mind and Machine. Addison Wesley.