[From: Wolfgang Menzel & Cristina Vertan (eds.) *Natural Language Processing between Linguistic Inquiry and System Engineering*. Iaşi: Editura Universităţii "Alexandru Oan Cuza", 2003.]

Structuring Terminology: between Lexicons and Domain Knowledge Representation

Galia Angelova
Bulgarian Academy of Sciences, Sofia, Bulgaria

galia@lml.bas.bg

Abstract

This article presents a revisited view on the project DB-MAT and its approach to model multilingual terminology using a single knowledge base.

1 German-Bulgarian terminology in DB-MAT

The project DB-MAT (1992-1995)[1] aimed at the design and implementation of a translators' workbench providing linguistic and domain explanations to human translators, within the paradigm of the knowledge-based Machine Aided Translation (MAT) [1]. Most generally, the innovative idea is to integrate in the MAT workbench a domain model (a knowledge base of conceptual graphs) and to generate on the fly explanations, when the translator highlights unknown terms in the source text to be translated. The project had to deal with German and Bulgarian languages, which opened the question how to link the corresponding entries of the bilingual lexicon to the Knowledge Base (KB) entities. Figure 1 presents an early model of the pointers between the lexicon and the KB items.

In principle, keeping phrasal lexicons is an acceptable strategy to support multilingual terminology (in the 80ies, several projects and prototypes of the so-called "knowledge-based term banks" seemed to approach the issue in a similar way). However, especially in DB-MAT, it became clear that the picture shown at Figure 1 has two potential "defects". First, it contains repeating information in the lexicons (see for instance all Bulgarian noun phrases including the word *court).* Second, the Bulgarian noun phrases are to be declined during the surface verbalisation according to complicated grammar rules (since the articles in Bulgarian are augmented at the end of the noun or the preceding adjective); so it turned out that the generation grammar for Bulgarian would work more easily with non-phrasal lexicons (since declination rules are to be supported anyway). In this way, to avoid repeating information in the (phrasal) lexicons and to provide more uniform treatment and even some elegance in the process of multilingual

[1] between Hamburg University (NATS) and the Linguistic Modelling Department, Central Laboratory for Parallel Processing. Bulgarian Academy of Sciences, funded by Volkswagen Foundation (Germany)

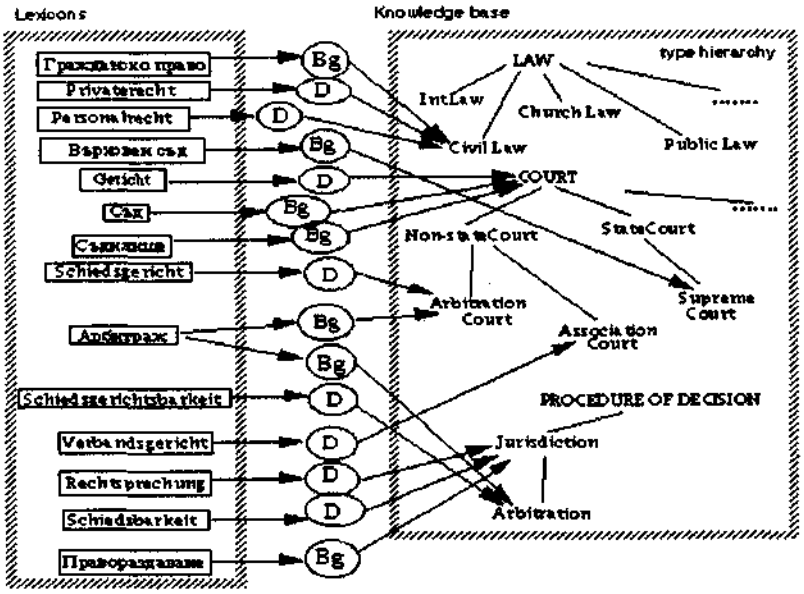

Figure 1. Full-phrase lexicons linked to the KB type hierarchy in DB-MAT (1993)

NL generation, DB-MAT adopted a more sophisticated view to the conceptualisation of bilingual terminology: German compounds were linked to *contexts* in the KB of conceptual graphs (and this motivated their "phrasal" behavior) while the corresponding Bulgarian translations were generated on the fly from the concepts "denoted by" (i.e. linked to) single words in the lexicon (see Figure 2).

Unfortunately, the idea illustrated at Figure 2 has another "defect": it is almost impossibl and somehow makes no sense to model very complex terminological corres-

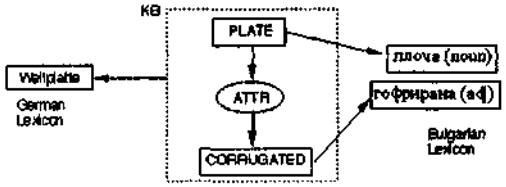

Figure 2. Supporting different granularity of lexicon items in DB-MAT (1994-1995)

pondences in this way. For instance, a KB context conceptualising the Bulgarian translation of the German term *Eindringharteprufung* turned to be too complicated to provide correct NL generation of the corresponding Bulgarian term and therefore useless as a KB content (additionally, it requires quite complicated inference procedures). Due to all these difficulties, DB-MAT system supported on the one hand phrasal lexicons (storing explicitly all terms which were too complicated to be modelled formally in the KB of conceptual graphs) and, on the other hand, KB contexts defining as conceptual graphs the semantics of German compounds which are translated into Bulgarian as relatively simple noun phrases. In this way, one important question was open: what is the reasonable balance between explicit declaration of phrases in the lexicon and forma! semantic definitions of the phrasal meaning in the knowledge base? In order to evaluate our knowledge-based model of multilingual terminology, in the next section we briefly overview an advanced approach to machine translation using the MikroKosmos ontology.

2. MikroKosmos as a backbone for formalising translation equivalents

MikroKosmos [2] is a formal ontology developed several years ago, to be applied primarily in advanced Machine Translation (MT) (see [3]). A useful comparison between MikroKosmos and other ontologies is presented in [4]. MikroKosmos is an important artifact in the field of knowledge-intensive NLP due to several reasons:

- it is a manually acquired, relatively large semantic network with more than 5000 concepts, and was one of the biggest knowledge resources used in industrially-oriented NLP prototypes;
- it offers systematically-elaborated solutions for treatment of meaning ambiguities in multilingual cases, which is an important achievement in MT (and differs from the monolingual approaches to formal semantics in NL understanding).

Table 1 presents samples of MikroKosmos nodes and labeled arcs. It is easily readable since MikroKosmos contains a canned text definition for each node. Encoded in a (Lisp-compatible) text format, MikroKosmos is a typical semantic network. The whole ontology of 5000 concepts is an inter-linked, spaghetti-like set of nodes and labeled arcs. The labels express the semantic relations between the concepts they connect. MikroKosmos is designed to support MT, i.e. the main task is to find the proper translation of every word, sentence, paragraph etc. in the source text which often requires resolution of semantic ambiguities among word meanings in the multilingual case. Having in mind the experience collected in DB-MAT, it is interesting to ask the following questions: *How much* knowledge MikroKosmos contains, and *why* its designers decided that especially this knowledge is worth to be acquired and encoded? *What* is MikroKosmos *used for* in ontology-based MT, and *how* it is used? Answering all these questions help us understand why MikroKosmos ontology looks as shown in Table 1.

(**ACADEMIC-BUILDING**
(IS-A (VALUE (**INSTITUTIONAL-BUILDING**)))
(SUBCLASSES (VALUE (**LIBRARY**)))
(PART-OF (SEM (**UNIVERSITY**)))
(DEFINITION (VALUE ("a building in which learning takes place, usually part of a school or university")))
(LOCATION-OF (SEM (**LECTURE PRINTED-MEDIA RECEPTION TECHNICAL-DEMONSTRATION**)))
(INSTRUMENT-OF (SEM (**EVENT**))))

(BELIEVE
(IS-A (VALUE (**ACTIVE-COGNITIVE-EVENT**)))
(DEFINITION (VALUE ("to take (an idea, thought, etc) as true, real, etc.")))
(THEME (SEM (**OBJECT EVENT**))))

(**BEHIND**
(IS-A (VALUE (**INVERSE-SPATIAL-RELATION**)))
(DEFINITION (VALUE ("self-explanatory--one object is behind another")))
(INVERSE (VALUE (**IN-FRONT-OF**)))
(DOMAIN (SEM (**OBJECT**))))

Table *1*. Fragments of MikroKosmos ontology (1998): definition of physical object (ACADEMIC-BUILDING), event (BELIEVE) and spatial relation (BEHIND). For convenience, concept nodes are bolded and semantic relations are shown in italics.

MT systems analyse the source language text, often create an internal representation (interlingua) and generate from there text translation in the target language. So the MT goal is to resolve the ambiguities in the source text providing the correct transitions

Source language text —> Interlingua —> Target language text

MikroKosmos was developed to support the resolution of ambiguities related to word meanings in the multilingual case.

Since words in different languages have different meanings, it often happens that one word has to be translated in another language by a phrase explanation keeping the exact word sense. For instance, the German verb *fressen* is translated into English as *eat* with *non-human agent.* MT systems have to deal with these meanings and that is why it is very important to decide where and how to store the words and the encoding of their meanings (i.e. the representation of their lexical semantics) within the MT system. The standard is to enumerate words in lexicons but there is (still) no widely accepted standard how to encode their semantics. There might be several approaches:

• to adopt a word-sense view to the internal ontology, i.e. to create an internal **concept-per-word representation** (this is the so called ontological promiscuity [5]). This idea might be feasible in the monolingual case (although too verbose [5]), but in MT it is

useless because the translation system still has to decide how source meanings are translated to target meanings, i.e. the "meaning correspondence" problem would be only shifted from word-level to internal conceptual level without any solution;

• to invent a small restricted set of semantic primitives, hoping that each word meaning in the two languages can be decomposed to the basic primitives. This is the so called decompositional approach, which proved to be unsuitable for larger amount of words in open domains;

• to combine the two approaches, defining an open set of ontological primitives, and to encode the lexical semantics of every lexicon word using the ontological concepts. MikroKosmos was especially designed to provide this combination of the two approaches, as a trade-off between Lexical Semantics and Ontology in machine translation. Every word meaning is defined in the lexicon as a *"possibly augmented instantiations of ontological concepts"*. Phrase meanings are obtained from word meanings by a *"combination operator"* [6].

Let us consider a simple example [6]. MikroKosmos contains a node ***INGEST*** with corresponding definitions of *agents, theme,* etc. The concept ***INGEST*** is an ontological primitive and the meanings of all semantically related verbs is explained in the corresponding lexicons using "possibly augmented instantiations of ***INGEST"***. Instantiations are in fact specialisations of the neighbor concepts, i.e. concepts directly linked to ***INGEST*** through conceptual relations. In the German lexicon, for instance, the semantics of the verb *fressen* is defined as ***INGEST*** with non-human agent, while in the English lexicon *eat* has meaning ***INGEST*** with animate agent. In this way MikroKosmos was manually acquired according to the following knowledge acquisition guidelines:

• two words W1 in language L1 and W2 in language L2 are "translations" if their meanings in the corresponding languages overlap; so there should be a conceptual similarity between W1 and W2. Knowledge Acquisition aims at the acquisition of an ontological node C1 encoding the common meaning of W1 and W2 in the proper way;

• concept C1 has the necessary number of conceptual relations and links to other concepts, to assure the proper ontological framework for instantiation and encoding of the word meanings W1 and W2, and

• the ontology is an open set of semantic primitives, i.e. the designers can always add a new concept providing correct translation of new words in new languages. Having in mind the current amount of MikroKosmos concepts, we empirically decide that the set of conceptual relations is more or less stable, i.e. a critical mass of semantic relations should have been already identified and acquired. Certainly these relations are not universal, but they provide the translation task in the chosen domain of knowledge-intensive MT.

3. Discussion and Conclusion

DB-MAT was a translator's workbench, generating explanations for separate terms selected in the source text, and provided neither analysis of the source language text, nor generation of target language fragments. In this way DB-MAT main task was quite different from the MT problems and solutions. That is why DB-MAT focused mostly on the conceptual representation of the terminology (noun phrases) and approached it from an NLG perspective, dealing in fact with knowledge-based modelling of noun phrases only. DB-MAT prototype was delivered with a relatively small knowledge base of about 300 concepts. In contrast, MikroKosmos contains semantic primitives for modelling the meaning of common lexica and helps resolving translation ambiguities in MT. The case form figure 2 *(Wellplatte <-> corrugated plate)* is not interesting for knowledge-based MT, since it is predefined and unambiguous.

However, on an abstract level, we find some similarity in the abovementioned approaches, which aim at the formalisation of the semantics of bilingual lexicon entries. Both projects attempted to use a single underlying knowledge base (a kind of semantic network) and developed rather similar techniques for expressing lexical meanings corresponding to the granularity of one KB concept node. Both projects used 20-30 semantic relations in the knowledge base. The successful examples, reported in the literature, illustrate flexible solutions concerning single words in the bilingual lexicon (i.e. the verbs *fressen, eat)* or simple noun phrases (see Figure 2). Complex translation equivalents are kept in the bilingual lexicons as explicitly defined phrases, since the knowledge-based approaches to lexical semantics have well-known limitations (the author discovered them progressively with the evolution of the DB-MAT project). DB-MAT tried to model the semantics of German compounds, wrt their translation in Bulgarian, and the obtained KB is practically dependent on the language pair. Adding a new language would require revision. Similarly, MikroKosmos was advertised with successful examples of single words expressed via semantic primitives but obviously a new language pair would require revision of the ontological primitives to provide correct definition of the meaning into the new language.

Today it is clear that knowledge-intensive NLP is extremely expensive and most often ends up with the acquisition of task-dependent conceptual resources. Advancement is to be expected in the closed worlds of well-defined domains. At the same time it is obvious that the challenging mono- and multilingual new fields (for instance the Semantic Web) deal again and again with (multilingual) ontologies structuring the terminology as a conceptual representation [7]. So we can expect that in the foreseeable future all abovementioned problems of how to link multilingual lexicons to the entities of a single KB will be treated again and again in numerous projects and hopefully, some better solutions might be found.

Acknowledgements: Many Bulgarian M.Sc. students who worked in DB-MAT had the chance to start their scientific carrier from this project:

• *Kalina Bontcheva,* who implemented the German generator in 1995 (and later became a PhD student at Sheffield University, UK; her PhD thesis in NLG was delivered in 2001);

• *Kristina Toutanova,* who implemented the Bulgarian generator in 1998 (and is currently a PhD student at Stanford University, USA);

• *Ani Nenkova,* who implemented the user-modelling module in 1998 (and is currently a PhD student at Columbia University, USA);

• *Ognian Kalaydjiev,* who implemented a term-translation checker in 1998 (and is currently a PhD student at CLPP, Bulgarian Academy of Sciences).

For all of them, DB-MAT and its continuation DBR-MAT were fascinating and challenging windows to the NLP world, moreover their early scientific maturity and related papers helped them to receive full PhD support abroad. The author is also grateful to Prof. Dr. Walther von Hahn (this was her first international project) for the successful and fruitful co-operation in DB-MAT and DBR-MAT.

References

[1] v. Hahn, W. and Angelova, G. *Providing Factual Information in MAT.* In: Proceedings of the International Conference *"Machine Translation - Ten Years on",* Cranfield, United Kingdom, November 1994, pp. 11/1 -11/16.

[2] Mahesh, K. and Sergei Nirenburg. *A situated ontology for practical NLP.* In Proceedings of IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, August 1995. See also http://crl.nmsu.edu/Research/Projects/mikro/

[3] Computing Research Laboratory at New Mexico State University, http: //crl.nmsu.edu/

[4] Nirenburg, S., K. Mahesh, and S. Beale. *Measuring Semantic Coverage.* In Proc. COLING-96, Copenhagen, Denmark 1996, Volume 1, pp. 83 - 88.

[5] Hobbs, J. *Ontological Promiscuity.* Proc. 23rd Annual Meeting of ACL, Chicago, IL, July 1985, pp. 61-69.

[6] Onyshkevych, B. and S. Nirenburg. *A lexicon for Knowledge-Based MT.* Machine Translation 10, 1995, pp. 5-57.

[7] Guarino, N. *Foundational Ontologies and Lexical Resources for the Semantic Web.* OntoWeb Workshop, December 2000, see http://www.ontoweb.org