Multilingual Cataloguing of Product Information of Specific Domains: Case Mkbeem System

Aarno Lehtola, Jarno Tenni and Tuula Käpylä
VTT Information Technology
POB 1201, FIN-02044 VTT
Finland
Aarno.Lehtola@vtt.fi

Abstract

This paper describes the Cataloguing Tool of the Mkbeem multilingual eCommerce mediation system. The Cataloguing Tool is used by product suppliers to provide necessary product information in a form required by the mediation system. The relevant information includes product articles that are maintained in a pivot language and automatically translated to other supported languages. From the supplier viewpoint the Cataloguing Tool implements "write-once-publish-many" paradigm. Other functionalities of the tool include automatic extraction of product properties from text articles along to an ontological product model and automatic classification of products based on their properties and ontology models. This paper describes the Cataloguing Tool and discusses in more detail about the use of ontologies in automatic interpretation of the semantics of product articles and user queries.

1 Introduction

In the year 2000 native English speakers became outnumbered by native users of other languages in the Internet population. Since then the trend has continued and in September 2002 the others comprised already over 63 % of the population [GlobalReach 2002]. The linguistic diversity is huge among the population. Even within Europe there can easily be counted over 60 languages among which even the smaller ones, like the over 100000 Icelandic speakers, comprise potential customers groups for international eShops. There is a remarkable need for cost effective IT solutions for enabling multilinguality in consumer Internet trading. This is the market where the Mkbeem mediation system has been positioned.

The Mkbeem mediation system (Multilingual Knowledge-Based European Electronic Marketplace, outcome of the EC project IST-1999-10589) adapts the language and the trading conditions of an Internet sales point according to its international customers [Leger & al. 2001, Mkbeem 2000-2002]. The system supports three use cases (shown by the big arrows in Figure 1). Each use case has its own tool. The multilingual Cataloguing Tool is for content and service providers to describe their products by means of text articles in a write-once-publish-many manner so that the information is maintained monolingually. There is a tool for the providers to define the contract conditions of their goods within the context of a consumer sales related legislation model, which enables adaptation of the contracts depending on how the transactions cross national borders. Finally, for consumers there is a system that implements cross-lingual IR from the product databases based on combining NL queries, graphical navigation in localised product hierarchies and use of search forms – the user chooses the modality. The language and contract adaptation is based on ontological models covering product models, legislation and related generic knowledge like time, materials and colours. The ontologies function for narrowing the scope of NL processing and for mediating in between different languages [Leger & al. 2000, Gomez-Perez & al. 2001].

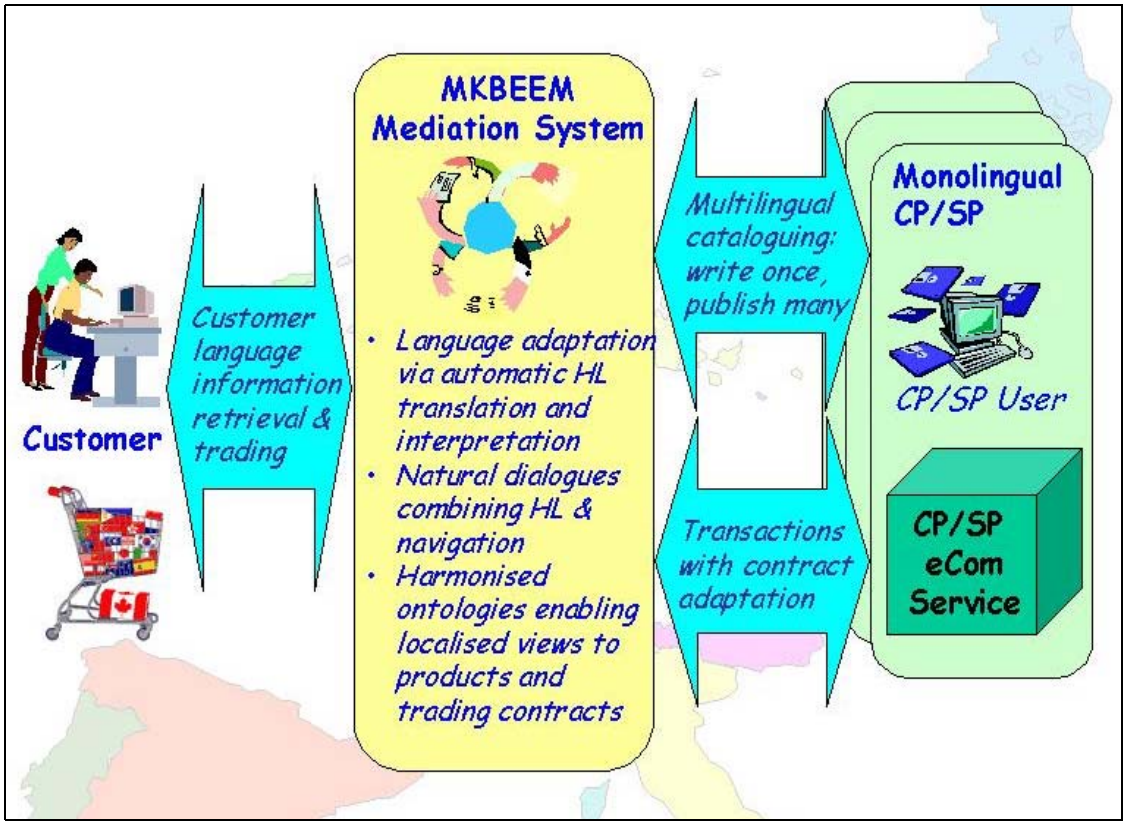

Figure 1: The operating context of the Mkbeem mediation system.

Next this paper describes the multilingual Cataloguing Tool. Its very central functionality, the meaning extraction from product articles and NL queries is presented in more detail. In the end there is a short summary of experiences from the use of the Cataloguing Tool.

2 Multilingual Cataloguing Tool

Multilingual Cataloguing Tool is used to publish monolingual product information in multiple languages in a write-once-publish-many manner. Cataloguing of a new product goes through multiple steps, taking into account linguistic, product model and culture-specific issues. These steps include:

1. Text checking
2. Property extraction
3. Categorisation
4. Machine translation
5. NL query processing

Text checking functionality is used to verify that the description of the product conforms to the language model in order to guarantee the quality of the automatic processing. This means that the terms used must be found in the lexicon and the sentence syntax must comply with the language model. In a case of unknown or erroneous word or sentence, the checking tool suggests possible corrections. If needed, it also allows a qualified user to edit language model (terminology, language rules and ontology definitions).

Property extraction functionality finds product properties from the textual product descriptions along the provided ontological product model, which describes the parts and their properties of product types in terms of concepts and their attributes. The extracted properties present information about a product in a language-independent way and they are stored in a relational database for further uses in inference, product classification and end-user information request processing. Inference means that new information is concluded based on ontologies. E.g., for a given cloth qualitative facts are inferred based on a material ontology and the known material composition of the cloth. We can also infer origin (e.g. are they natural) of materials.

Based on a colour ontology, colour similarities and harmonies can be inferred.

Categorisation functionality provides culture and market-specific categorisation of products into product hierarchies in an eCommerce site. For instance, this means that for a wind-proof jacket the system suggests that this product should be found both in outdoor clothing and in jacket categories of the eCommerce site. The categorisation is based on the extracted properties and simple rules. Culture and market-specific issues may arise. For instance, the concept of a *winter cloth* differs in Finland and in Greece.

Machine translation in the Cataloguing Tool relies on the Webtran machine translation system. A checked and accepted product description text is passed to Webtran to be translated to multiple target languages.

Natural language query processing is used for testing the newly added products so that they are found easily and from the correct places in the catalogue. The queries are analysed and the extracted properties are matched against saved properties of the products in the database.

After the processing steps the product information is stored into the product database. This includes the translated product articles, the extracted properties, and the results of inferences based on the properties and the market specific categories where the product belongs.

The Cataloguing Tool supports access profiles for four classes of users. The *Proof-readers* have rights to correct existing product information, the *cataloguers* can add new products to and remove old ones from the catalogue, the *language modellers* can add new terms to the lexica and the *content provider manager* has all the rights, including editing the ontologies and the language models. The user interface is adapted to the profile of the current user so that only relevant and accessible parts are shown on the screen.

From the language processing point of view the cataloguing involves two central processes: automatic sublanguage translation, and meaning extraction from product articles and NL information requests. The text checking and the machine translation are directly based on the Webtran MT software of VTT [Lehtola & al. 1998, Lehtola & al. 1999a & 1999b, Tenni 1999]. Section 3 concentrates on describing the meaning extraction process that associates the input texts to ontological domain models. It finds for input texts a language-independent semantic representation in terms of the description logic language CARIN [Levy & Rousset 1998] and based on the provided domain ontologies and the associated language models. The analysed text inputs include product description articles and NL queries. Meaning extraction is central when product properties are recognised from textual product articles, when products are classified, and when NL queries are processed. Webtran system has been modified to assist in the meaning extraction, as well.

The Cataloguing Tool is implemented using Enterprise Java Beans. The overall system consists of two parts: the core server and the end-user interface. The core server provides natural language processing services and an inference engine for using the ontologies. The end-user interface is a Java Applet running in a WWW browser. The tool interacts with the core server using internet protocols. The architecture needs just one installation of the core server and then the Cataloguing Tool can be used from multiple places without basically any extra installations, e.g., through company intranet at different subsidiaries or through extranet by subproviders. This makes also maintenance of the system relatively easy.

3 Meaning Extraction Process

Webtran MT system has its own formalism for describing domain-specific sublanguages. This Augmented Lexical Entries (ALE) formalism provides multidirectional rules denoting equal, non-directed natural language excerpts on the desired linguistic abstraction levels. An entry can describe linguistic information in one, two or more languages. In an entry, each language is represented in its own section. Entries can also be understood as partial dependency parse trees. For detailed technical description of the ALE formalism, see [Lehtola et al. 1999]. Below is an example of an ALE:

```
[cloth.material.composition
    [fi  ^(A){clothProd} tag_percentage(X)
         (B){textileMaterial ptv}]
    [fr ^(A){clothProd}  en tag_percentage(X)
        (B){textileMaterial}]
    [en ^(A){clothProd} of tag_percentage(X)
         (B){textileMaterial}]
    [se ^(A){clothProd} av tag_percentage(X)
         (B){textileMaterial}]
```

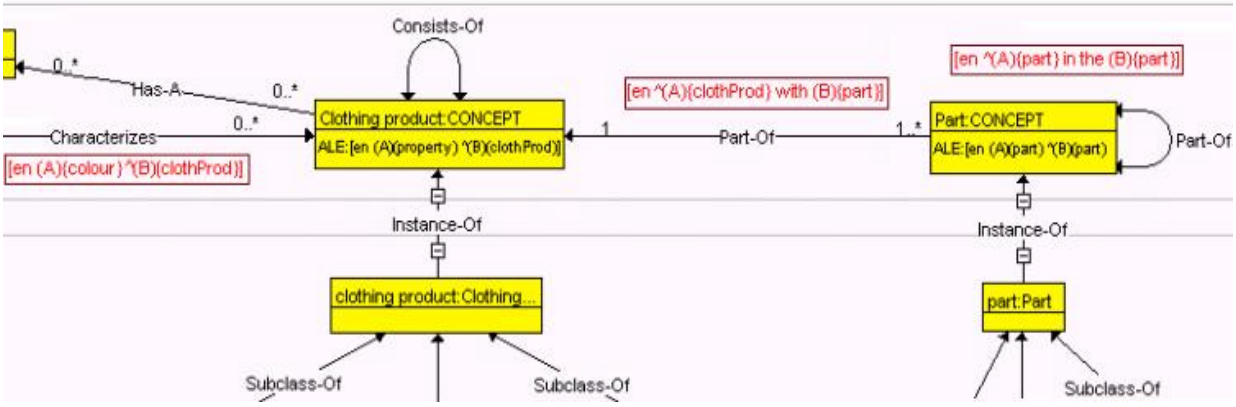

Figure 2: An excerpt from an ontology with ALE rules associated.

The previous ALE concerns a phrase with product name and the material that the product is made of with the material percentage. For example: 'housut 100% puuvillaa' (fi) translates into 'pantalon en 100% coton' (fr), 'trousers of 100% cotton' (en) and 'byxa av 100% bomull' (se). It also marks the product name to be the head of the implied partial dependency tree.

ALEs are used also to describe the linguistic constituents for concept matching in meaning extraction. Concept matching ALEs can be included into the concept property and relations descriptions in domain ontologies, e.g., the product models of Mkbeem. The idea is schematically illustrated in Figure 2. When corresponding language construct is recognised the system automatically associates it to the ontology concept. The given ALEs control how concepts and relations can be recognised from constructs of human language, as well as, how a human language paraphrase can be generated from an ontological expression. This reverse function is not currently used.

The meaning extraction process includes five phases:

1. Lexical analysis
2. Dependence analysis
3. Concept matching and verification
4. Refining semantics in particular themes
5. Syntactic translation into CARIN

Figure 3 illustrates the process and names the intermediate results. The phases 1-3 make up syntactico-semantic analysis.

The **lexical analysis** includes tokenising of the input and incorporating morpho-lexical information to each token. After this phase we know for each token (word, number, abbreviation etc.) what is known from it based on the information that is available from the lexicon and without considering any context information.

The **dependence analysis** involves finding syntactical relationships between the constituents of the sentence. It produces a set of syntax trees.

The **concept matching and verification** involves finding the conceptual bindings to the domain ontology that the user input embedded. After this phase we have a set of syntax trees with the relationships of its subparts to the domain ontology concepts explicitly marked. In fact, we have a dependence syntax tree that is extended with the relevant parts of the domain ontology through these relationships. We call this presentation briefly a *semantic graph*.

After the set of semantic graphs has been derived, there follows ontological inference of the CARIN formulas in the phases 4 and 5.

The **refining semantics in particular themes** is based on additional generic ontologies like ontologies of colours, materials, distances etc. The refining takes a semantic graph and deduces and makes explicit in it additional knowledge concerning the particular themes. The deduction process with colours, materials and expressions of time is described in more detail in [Lehtola & al. 2003].

The analysis results are translated into CARIN language in a format that is standard for all further processing in the Mkbeem system. The translation involves, e.g., removing of the linguistic informa-

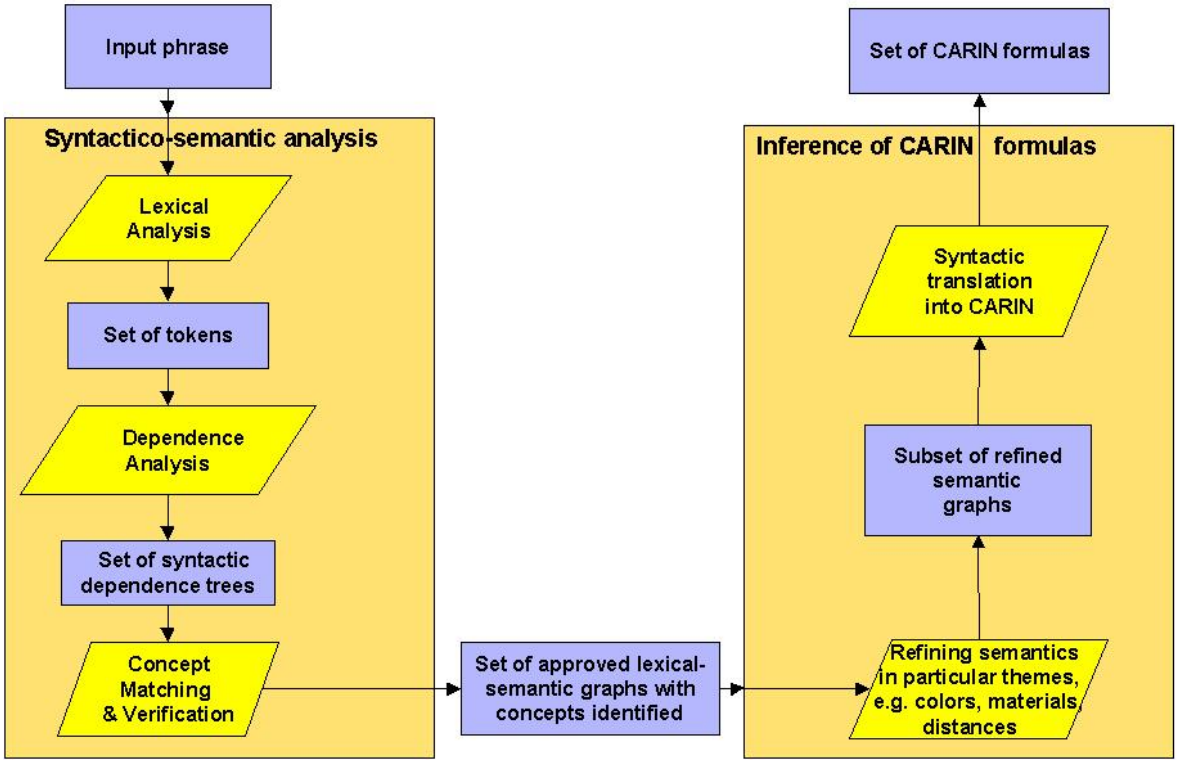

Figure 3: The meaning extraction process in the Cataloguing Tool.

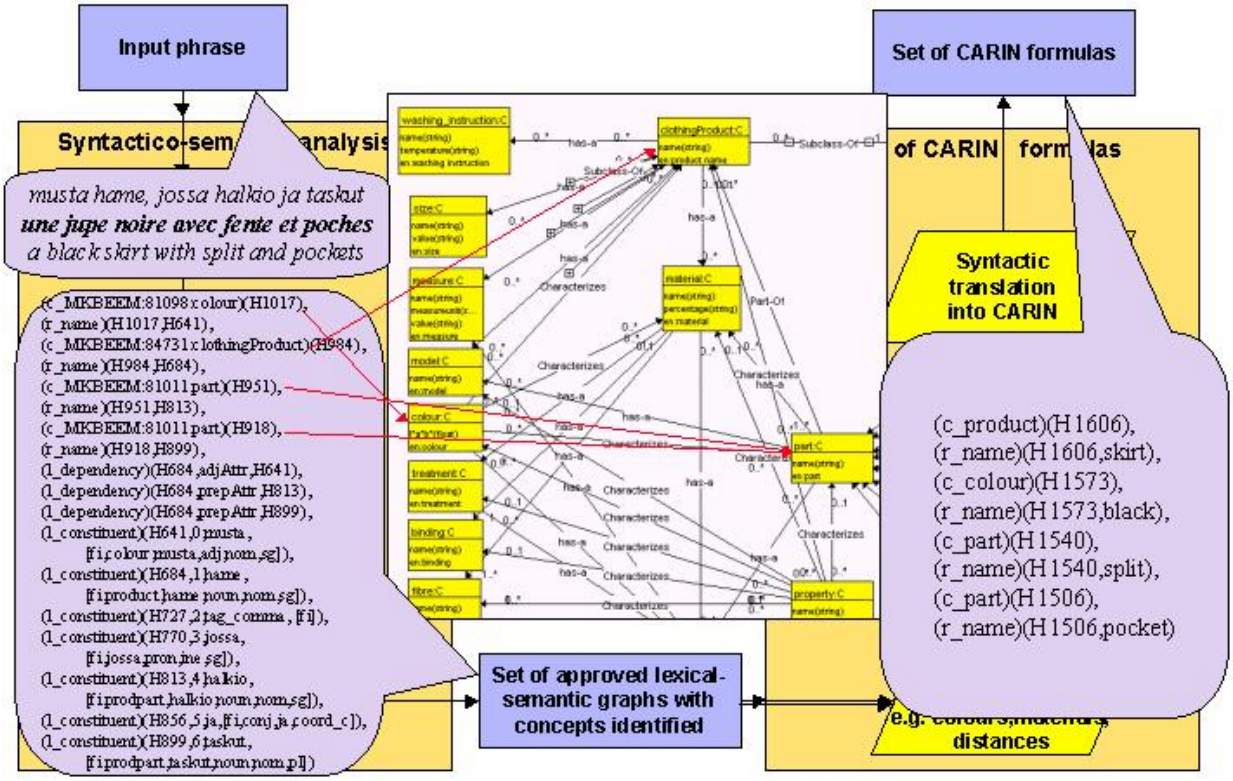

Figure 4: An example of the data flow when extracting meaning of a NL query.

	Finnish	French	English
Domain specific lexicon (morphological entries)	4500	1700	1500
Domain specific lexicon (number of entries)	2800	1300	1400
ALE Rules: - total number	965		
- of which bound to ontologies	150		
Cataloguing Rules	96		
Ontology concepts/attribute values	307/1050		

Table 1: The sizes of the Mkbeem specific linguistic knowledge bases.

tion that has been retained this far in the intermediate data structures.

Figure 4 contains an example about how the meaning extraction process goes with the NL query *"musta hame, jossa halkio ja taskut"* ("a black skirt with split and pockets"). The corresponding lexical semantic graph is shown. The graph includes both linguistic analysis results and references of the constituents to the recognised concepts in a domain ontology, which describes properties of clothing products.

Table 1 summarises the sizes of the linguistic knowledge bases that were implemented for the cataloguing of descriptions of clothing products from Finnish to French and to English in the field trial tests of the Cataloguing Tool.

4 Test User Experiences

The testing of the Cataloguing Tool was carried out by the mail-order company Ellos Postimyynti Oy with their sales articles being women clothes. The first tests were carried out in the middle of the project in September 2001 in order to guide the development work of the following second phase. The test results presented here concern the second phase trials. The field tests were done during September 2002. The idea of the testing was first to test the concept of the Cataloguing Tool, i.e. the maintenance of the multilingual catalogue using a single tool. Secondly, the tests concerned the usability of the Cataloguing Tool in a real working environment. The test group consisted of 3 cataloguing professionals. Testers were interviewed twice. Before the tests they were asked about their background, experiences and expectations. After the tests they were made an interview that focused systematically to every function of the cataloguing tool. Testing period was one month during which they were using the system from their own machines.

Test results were very positive. The cataloguing process as a whole was seen as an easy and efficient way of producing and classifying product information. Also the tool itself got good remarks: it was considered to be a useful tool for the production of multilingual product information and each of the main features (see Section 2) was considered as good. Besides the very important possibility of semi-automatic translation into target languages, test-users named functionalities like property extraction and inference with colours and materials to be important in bringing the customers new possibilities to find complementary information from goods deduced from additional sources.

One important advantage in an integrated cataloguing environment is that it helps in producing consistent and uniform information as the whole cataloguing process is based on joint language and product models that conform to the company knowledge of the domain. Moreover, the test users anticipated that the use of the Cataloguing Tool can make the working process faster and it reduces the amount of manual, repeated routine procedures. Also the knowledge base maintenance tools were considered to suit to their task well.

The MT component Webtran of the Cataloguing Tool has been in production use at Ellos since the year 2000. The EUROMAP case study by CSC Inc. [Loimaranta 2000] reports savings of over 30% in translation time having been reached after a relatively short use of the MT tool.

5 Conclusions

This paper described the Cataloguing Tool of the Mkbeem multilingual eCommerce mediation system. The Cataloguing Tool is the software for product suppliers to author multilingual product information. This paper describes briefly the main functionalities (text checking, property extraction, product categorisation, machine translation, natural language query processing) of the Cataloguing Tool. The central linguistic function, called meaning extraction, was presented in more detail. Meaning extraction associates linguistic expressions to the domain ontology concepts and relations and provides a language neutral semantic representation for input texts.

The ALE formalism for describing domain-specific sublanguage was briefly explained. There was also outlined how ALE rules can be associated into concepts and relations of an ontology model in order to user them for analysing meanings of NL inputs. The results are expressed in specific ontological formulas using CARIN language. The ALE formalism includes required elements to describe the rules for language checking, machine translation, and matching NL inputs to the concepts and relations of an ontology model.

Finally the user-test settings and results were summarised. Experienced catalogue maintenance professionals carried out tests and the overall impression was positive. Both the concept of a "cataloguing tool" and the overall software were seen as very useful. The cataloguing process as a whole was seen as an easy and efficient way of producing and classifying product information. The tool itself was considered to be a useful tool for production of multilingual product information and each of the main features was considered important. The results give us a good reason to continue this work into the future and to bring this technology into everyday use and to adapt it to new domains of goods and new languages. Of the companies involved in the Mkbeem project, Ellos as well as SNCF and France Telecom are planning to utilise the technology in their business operations.

Acknowledgements

The authors would like to thank all their colleagues in the Mkbeem consortium (France Telecom, SNCF, Fidal, SchlumbergerSema, UPM, NTUA, CNRS and VTT) for the excellent co-operation and the European Commission for supporting the reported work.

References

GlobalReach (2002): Statistics on the website of Global Reach Inc, September 2002, URL http://www.glreach.com/

Gomez-Perez, Asun; Corcho Carcia, Oscar; Fernandez Lopez, M.; Lehtola, Aarno; Taveter, Kuldar; Sorva, Juha; Käpylä, Tuula; Toumani, Farouk; Soualmia, L.; Barboux, Cecile; Castro, E.; Sallatin, Jean; Arbant, Geraldine; Bonnaric, Annabelle (2001): Requirement, Choice of a Knowledge Representation and Tools. Public Report of MKBEEM project (EC IST-1999-10589), Version 2.0, available e.g. from www.mkbeem.com, 2001, 93 p.

Jaaranen, Kristiina, Lehtola, Aarno, Tenni, Jarno, Bounsaythip, Catherine (2000): Webtran tools for in-company language support. In: *Language Technologies for Dynamic Business in the Age of the Media*. Köln, 23 - 25 Nov. 2000. Vereinigung für Sprache und Wirtschaft. Köln (2000), pp. 145 - 155.

Lehtola, A., Bounsaythip, C., and Tenni, J. (1998): Controlled Language Technology in Multilingual User Interfaces. In: *Proceedings of the 4th ERCIM Workshop on User Interfaces for All* (UI4ALL´98), Stockholm, 1998, pp. 73-78.

Lehtola, A., Heinecke, J., Bounsaythip, C (2003): Intelligent Human Language Query Processing in Mkbeem. In: *Proceedings of HCI International/UAHCI 2003*, Crete, June 22-27, in print.

Lehtola, A., Tenni, J., Bounsaythip, C., and Jaaranen, K. (1999a): Controlled Languages as the Basis for Multilingual Catalogues on the WWW. In: Jean-Yves Roger, Brian Stanford-Smith and Paul T. Kidd (Eds.). In: *Business and Work in the Information Society: New Technologies and Applications*. IOS-Press, Amsterdam, pp. 207-213.

Lehtola A., Tenni J., Bounsaythip C., and Jaaranen K. (1999b): WEBTRAN: A Controlled Language Machine Translation System for Building Multilingual Services on Internet. In: *Proceedings of: Machine Translation Summit VII `99* (MT Summit 99), September 13-17, 1999, Singapore, pp. 487 - 495.

Levy, A.; Rousset, M.C. (1998): CARIN: A Representation Language Combining Horn rules and Description Logics. *Artificial Intelligence Journal*, vol 104. September 1998.

Leger, Alain; Michel, Geraldine; Barrett, Peter; Gitton, Sylvain; Gomez-Pere, Asuncion; Lehtola, Aarno; Mokkila, Kristiina; Rodrigez, Santiago; Sallantin, Jean; Varvarigou, Theodora; Vinesse, Jerome. Ontology domain modeling support for multi-lingual services in E-Commerce: MKBEEM 14th European *Conference on Artificial Intelligence ECAI'00, Workshop on Applications of Ontologies and Problem-Solving Methods*. Berlin, DE, 20 - 25 August 2000. Berlin (2000), 4 p. URL http://delicias.dia.fi.upm.es/WORKSHOP/ECAI00/19.pdf

Leger, Alain; Lehtola, Aarno, Villagra, Victor (2000): MKBEEM – Developing Multilingual Knowledge-Based Marketplace. *ERCIM News*, July 2001, pp. 50-52. Reprinted in Research News of VTT Information Technology, December 2001, pp. 1 – 3.

Loimaranta, Outi (2000): EUROMAP HLT Case Study: Webtran – a controlled language machine translation system for building multilingual services on Internet. December 2000, http://www.hltcentral.org/usr_docs/case_studies/euromap/FIN_webtrans.doc

Mkbeem (2000-2002). The Mkbeem project. URL http://www.mkbeem.com/