Multilingual Content Processing

Gregor Thurmair

linguatec
Gottfried Keller Str. 12
D 81245 Munich
g.thurmair@linguatec.de

Abstract

This contribution describes the consequences of a multilingual set-up, as used in internet information gathering, search and content processing, for the architecture and the different components of such a system. First the scenario is briefly outlined; then the existing technology components are reviewed, with the focus of the effects of multilingual content to such components. Finally, the linguistic resources are discussed which form the backbone of such a multifunctional and multilingual system. It is shown that adding multilinguality to an information system has massive consequences for the design of all system components and resources.

1 The Crosslingual Scenario

Information Acquisition has become a major challenge in the internet age. The amount of information to be monitored has grown significantly, but time and resources to perform this task are still constrained. Tools like personalised electronic news clipping, automatic knowledge mining, or internet monitoring try to match the new requirements. As a result, a scenario needs to be designed which requires massive natural language support to meet its goals:

- automatic **classification** of incoming texts is used as a first-level filter, to get rid of irrelevant material
- key word and **index term** extraction allows for a finer-graded determination of the topic of a document. Links to some classification or ontology scheme take the analysis beyond pure text search into the field of knowledge mining and the semantic web.
- Named entity recognition and **fact extraction** take care of the fact that people do not search strings but semantically meaningful objects, i.e. content. Profiles of interest look for content, not just string comparison.

Such information systems (developed e.g. in the EC-SENSUS project and described in (Thurmair, 2004)) face an additional challenge now, as **multilinguality** is a major obstacle for efficient information access. In globalised contexts, relevant information is available in a variety of languages, many of them not spoken by the information recipient. Queries, interest profile definitions etc. are stated in the interface language (i.e. the users' native language) but the documents and the information sources to be looked up are in different languages, the document languages.

One could claim now that the extension consists only of adding two additional translation steps, one on the query side, and one on the answering side (to re-translate hits e.g. in Korean into the interface languages, or somehow make the content available for further use), and perhaps a language identification tool to determine the document language.

However, this is not quite the case; indeed there are massive consequences on the single components and technologies used, if multilingual aspects should be considered. This refers to the design of such components themselves (section 2 below) as well as the design of the linguistic resources required (section 3).

It is known that some technologies are restricted to, or give better results in, specialised domains. This fact corresponds to a difference in the targeted markets, i.e. there is a difference between the corporate market (large internationalised companies), and the general internet market. While the former can often be constrained to specific applications (like: automotive, banking), and therefore produce higher quality in linguistic support tools, the internet applications must cope with any request. Experience in Machine Translation but also Information Extraction shows that results can be better, and systems are more likely to be accepted by users, if they can be tailored to a given application or context; general purpose tools are much more difficult to design.

2 Challenges in crosslingual Scenarios

A closer look into the scenario just described shows that the requirement of multilinguality influences the architecture and the design of all components of an information system; this section discusses some of them.

2.1 Techniques for Query Analysis and Expansion

This section discusses the query part of a crosslingual scenario, and design adaptations on the search components.

Definition of Search terms

The standard practice of analysing a search request is to remove the stop words, stem the content words, create some link between them (e.g. Boolean operators) and send the query to the search backend.

This is not simply feasible in a crosslingual environment.

- conventional techniques to increase recall (like truncation or **stemming**) are not applicable because the resulting strings often cannot be translated: If *integration* is stemmed to *integr-* this is a non-translatable string. Therefore they cannot be searched in the document languages, and so they decrease the overall. system quality.
- It is known that most terms in special languages are **multiword terms**. Translating the single words very often leads to useless document language queries, and to completely irrelevant search results: Searching for en *nuclear power plants* (de *Atomkraftwerk*) fails if *nuclear* (->*Kern-*), *power* (->*Macht*) and *plant* (-> *Pflanze*) are translated in isolation.

- Removing **stop words** is an inadequate strategy if such stopwords are part of a multiword term (like en *House of Commons* -> de *Unterhaus*) as the remaining fraction of the term cannot be properly translated any more.

So in query analysis, it has to be kept in mind that the target terms of query analysis must be units which can be translated. Lemmatisation and multiword recognition would be appropriate technologies in this case.

It should also be noted that translation may change the shape of the query significantly: Single word terms in one language (de *Kernkraftwerk*) may become multiword terms in the other language (en *nuclear power plant*), and operations based on word sequence or adjacency may become invalid after translation: en *money laundering* -> es *lavado del dinero* changes both word sequence and distance, and de *Geldwäsche* is even a single word (German compound). The only stable fact in these structures is an identical **head-modifier relationship**, which requires significant linguistic efforts to analyse.

Translation of Search terms

Good translation of the search terms is essential for good retrieval results. Inaccurate or vague translations decrease results significantly, as can be concluded from previous experiences:

- General purpose termbanks often lack the specific terms which are relevant for a concrete application, esp. in specialised domains. This also holds for generalised resources like WordNet or EuroWordNet (Verdejo et al., 2000).
- Standard machine translation tools' dictionaries are also rather generic, and in addition they face the problem that queries usually are pretty short, and lack the context needed by these systems for disambiguation. Even combining several MT systems did not improve the search results significantly. (Braschler et al., 2000).
- Queries built from bilingual parallel text do not represent the standard situation; in fact, once parallel text is available, crosslingual search is not needed any more.

Best results can be achieved using a **specialised terminological resource** which is based on the corpus documents that should be searched. This requires some extra effort, namely to build specific multilingual term banks for a given application domain.

Expansion of Search terms

Research e.g. in TREC has shown that search results can be improved significantly if appropriate query expansion techniques are applied (Voorhees/Harmann, 2000).

Query Expansion usually follow two lines: either the search documents themselves are analysed, and the closest matching search terms are used for expansion, or some linguistic resource (e.g. an ontology) is used for expansion (synonyms, narrower terms).

The first approach is not easily feasible in a multilingual environment as it stays language-specific. It could be done for each language separately but this presupposes properly translated search terms.

The second approach would simply enrich the existing multilingual resource by conceptual relations, the most important of which are the thesaurus-type ones (synonym, broader/narrower term).

The result is a **multilingual concept net**, modelling hierarchical (and other) relations between concepts, and offering multilingual information on its concept nodes. Such resources have been proposed several times (Volk et al., 2002); and it seems as if a common conceptual backbone for different languages provides more use for a crosslingual environment than a EuroWordNet type approach whereby each language defines its own conceptual hierarchies.

Query Expansion will consist in selecting synonyms and narrower terms in the target language; other conceptual relations may be added depending on the query.

A final point in query expansion is morphological expansion, i.e. the generation of all inflected forms for a translation of a term, to improve recall. This requires knowledge about the inflectional behaviour of terms (single words, heads of multiwords).

Proper Names, information objects

A special problem are proper names: In a monolingual string-based search, there is no need to distinguish between terms and proper names, but in a cross-lingual scenario there is. Not every string which cannot be translated is a proper name, it may be just a term which is missing in a particular resource. While untranslated terms in the interface language will drop the search results significantly (no hits in the document language), this does not hold for proper names, which still may return valid search results So proper name recognition may be required to determine if a unknown word should be sent to the foreign language search or not. It should be noted that proper names are inflected in some languages (like Russian).

In a multilingual context, an additional problem consists in the fact that names sometimes need to be translated (en *Milan* it *Milano* de *Mailand*), and transliterated from foreign writing systems (fr *Abou Moussa* de *Abu Mussa*). As there is a common referent, the different names will be transformed into each other.

2.2 Multilingual Search

A multilingual scenario modifies the architecture of a search system not just on the query frontend but also on the search system backend.

Classification

Text classification tools, if based on statistical techniques, stay monolingual; as a result, the training data for a given classification schema must be duplicated. In case the classification is based on domain terminology, a multilingual resource, if properly equipped, can be used for multilingual classification.

System index

The system index which is searched must also be adapted to a multilingual setting.

The index should be **language-specific**. If the index does not reflect the language of the documents, two consequences will have to be taken into account:

- The system will produce false hits in non-intended languages (de *post kind Seine* conflict with English or French terms), which increase noise in the output

- The use of stoplists will produce unwanted side-effects in other languages: Many stopwords in one language are meaningful terms in other languages.

One index per language will avoid such problems; however a language identifier is required to select the proper index for a document to be attached to.

The organisation of an index usually is based on **single words**; and multiword terms are built by postcoordination. If an index should reflect terms as semantic units, precoordination would be an appropriate strategy for the treatment of multiword terms, with massive consequences on the rest of the system (e.g. computing statistical relations between document terms). Automatic indexing with precoordinated terms has not been explored yet, the problem being the disambiguation of alternative segmentations; postcoordination however needs more intelligence than simple adjacency information, to avoid noise in the document result set.

Semantic annotations

Adding more intelligence to the index by storing **semantic information** also affects the index structure. Searches for *Roma / ethical group* as opposed to searches for *Roma / city* needs additional intelligence in the index to cope with this distinction; the alternative to run a filter on the search results at query time would slow down the query process significantly. Experiments using document sections to solve the problem have been made in (Thurmair, 2004); however searching for relations in addition to objects would probably use some alternative search and display strategy (structured search or graphical search).

2.3 Result Presentation

Assume a search returns 56 documents, 23 being in English, 12 in Korean, 16 in Arabic, others in Russian, French, and German. Two basic problems are immediately obvious: Ranking, and result understanding.

Ranking of result documents

Current approaches to ranking are based on a comparison of term occurrences. This is a monolingual view, and it is assumed that the terms to be compared form a common universe. However, such an assumption does not hold in a crosslingual setting, as the term sets there are completely disjoint. How could a French result document be ranked against a Russian result document if there is no common basis of comparison?

While many systems simply offer different result sets for different languages, probably ranked within the respective sets, overall ranking can only be attempted if there is some similarity between the documents which can be exploited:

One alternative is to use bilingual texts / corpora. In this case, the text itself offers the means to translate it into a state in which it would be comparable to other texts. Such a strategy, however, is somewhat artificial as in practice parallel texts are not really frequent.

Another alternative would again use a bilingual resource, and map some of the document language's terms into a kind of pivot representation. Such a ranking procedure depends crucially on the quality of such a resource (and would have to solve problems like word sense disambiguation etc.). But ranking would be based more on content than in previous approaches.

Translation

Translation is the only serious available means to transport the meaning of a found document back from the document language into the interface language. Translation must be fast, instantly available, and as accurate as possible to give a correct understanding of the content.

The only solution for such a situation is machine translation. While in document production and localisation, there is the alternative of human translation, this is not available in the crosslingual scenario; here the alternative is machine translation or no translation. This fact is a strong argument in favour of continuing MT research.

Two main alternatives are available, one being key term translation, the other one full MT.

In **key term translation** (also called term substitution), the most important terms of a text are re-translated into the interface language. This technique requires three main components:

- determining the relevancy of a term for a given text; this is to avoid a situation where many terms are translated but they are not relevant; this produces an improper text understanding of the user.
- basic morphosyntactic analysis of a text to bring the text words into some canonical representation which can be looked up in the dictionary
- bilingual dictionary resource which links terms in the interface and in the document language.

It should be noticed that the key component, the bilingual resource, would be available in a crosslingual scenario because it has already been used to support query term translation. If this resource is designed in a bidirectional way then key term translation is a valid first-level support option for understanding the content of the target text.

This type of translation can work as a filter, to determine if a result document is relevant for a given profile or not (this would rule out about 30 to 50% of a search result, however, and users appreciate this 'abstracting' effect). A better level of understanding can be provided by full machine translation.

Full machine translation in current settings has still room for improvement, the main challenge being translation quality. Research to improve translation quality, however, is not really wide-spread today.

Statistical machine translation (Knight et al., 2003) has not proven to reach superior quality compared to trained rule-based systems; so its contribution on quality can only be an add-on to the existing rule-based technology.

Quality improvements for machine translation can be seen in several areas:

- **Extending syntactic coverage** is a slow but necessary task; and finding good fail-soft mechanisms for parse failures will always be necessary.
- **Dictionary coverage** is the most straight-forward approach; there is still a significant potential for system improvement (Weber, 2003). However, to make use of large dictionaries requires to improve the transfer selection strategies of current systems; otherwise there is no improvement, and the system just accumulates alternative translations of a term.
- **Transfer selection** in situation of 1:many transfer options is a common source of errors in current MT systems. While statistical approaches often do not even address this issue and work on a 1:1 translation

assumption, rule-based systems often fall short when applying pure linguistic heuristics in situations where the context does not provide the required linguistic clues.

Therefore, rule-based systems must be enriched by contextual knowledge, to improve the selection strategy. Subject area selection, clustering, automatic topic detection and other techniques are available to support such an idea.

In addition, when comparing human readable and machine dictionaries (Beryl et al., 1995), it turns out that the former have much finer-grained entries, and use much more semantic information for disambiguation; esp. the notion of a 'typical representative' is used, which triggers a certain transfer. Such a system behaviour can only be modelled by using more semantic and ontological information than most current systems do; and there has been progress in this area to exploit such knowledge sources (Beale et al., 1995; Bel et al., 2000).

- **Tuning** is a key element to MT success; and in fact well-tuned systems can show a high rate of acceptance. Tuning is not just a lexical but also a syntactic matter: Different subject areas use certain syntactic constructions with different frequency, and the analyser should be able to adjust to this fact.

  To do this, corpus-specific statistical data need to be collected, which can provide the statistical data for weighting the respective rules of a grammar.
- Increasing the intelligence of the text understanding by applying means of **information extraction** will also contribute to better translation quality. Being able to assign types of names to an unknown string can not just improve parsing results (Babych/Hartley, 2003) but also the interpretation of related elements (like prepositions, pronouns etc.)

As a result, machine translation is a condition sine qua non in the multilingual information scenario, as there is no alternative technology, and there is still significant improvement potential.

Abstracting by Extraction

Another option in coping with foreign language material is abstracting, but not on the basis of selecting the 'most meaningful' sentences but on the basis of information extraction.

The approach consists of two parts:

- Extraction of the relevant objects and relations, based on a profile of interest, using information extraction technology, and creating a formal representation of the extraction result, e.g. in the paradigm of Topic Maps (ISO 13250)
- Generation of abstracts from the resulting fact representation; this generation can produce several monolingual variants.

This approach, described e.g. in (Ritzke,2000), is a kind of 'interlingual' view to the translation problem, and uses the representation of the extraction step as an interlingual structure. It is also a way of linking the text content to formal kinds or representations in the context of the Semantic Web, by relating the text content so some kind of conceptual hierarchy / ontology.

However, only a fraction (albeit the most interesting one) of the text content can be represented (and translated) this way, and there is still a significant part of a text which would get lost if this would be the only representation of the text content. A combination of abstracting by extraction and some machine translation of relevant text parts would probably the best offer for a user to generate.

3 The Role of Resources

As can be seen, many of the components of a state-of-the-art information systems need to be adapted if multilingual aspects want to be taken into account.

A key element in all these adaptations are the linguistic resources. They need to be used in different facets and complexity by the different system components, so an effort what exactly would be required, and how such resources should be structured, seems to be appropriate.

3.1 Requirements

The requirements for such a resource can be derived from the architectural descriptions as follows:

- The resource must be multilingual, and cover as many languages as possible, given the fact that it cannot be determined a priori in which language relevant information would be found.
- The resource must be bi-directional, and support translation from native into foreign language (for query translation) as well as from foreign into native language.
- The resource must match the document base / corpus which is supposed to be searched. Relevant foreign language terms that are in the corpus but not in the resource will lead to poor search results.
- The resource must be consistent. If query translation uses different translations, or has different coverage, than text re-translation then users will become confused, and the overall system acceptance will drop.
- The resource must contain conceptual structures and links, to provide a basis for query expansion. Ontological structures must be included, leading to the result of a multilingual concept net, whereby the nodes have multilingual denominations.
- In order to support linguistic processing (like shallow parsing for information extraction, full MT parsing, inflection generation for query expansion etc.), the resource needs to have formal linguistic annotations for each term, like part of speech, gender, inflection classes, verb argument structures, and the like; the scope of this is defined by the MT and IE analysis requirements.
- Beyond monolingual annotations, the resource also needs bidirectional information as used for MT, like transfer conditions and actions. This requirement makes it a challenging combination of a multilingual / nondirected and bilingual/directed resource paradigm.

Technically, as not all system components need all information items, the preferred approach would perhaps be to have one central repository where the whole multilingual concept net is stored and maintained, and compile the necessary information items into the dictionaries of the respective components (like MT lexicons, extraction gazetteers, expansion ontologies etc.), without loosing consistency. The data exchange involved would require to review existing exchange formats, like

OLIF (Thurmair/Lieske, 2002), or MILE (Calzolari et al., 2002).

3.2 Tools for resource maintenance

Such a centralised multilingual concept net needs a specialised administration interface for maintenance, and several off-line components to make it work.

Tools to build up resources

Given the fact that the multilingual concept net should match a specific document set, or corpus, it should be built (or at least enriched) by the terms of this particular domain.

- Term Extraction would be one of the key tools for such a scenario. It is supposed to extract term candidates from a text base. Note that the candidates should be in base form instead of text form, and should contain multiword as well as single word candidates, in all languages to be supported; this includes a certain amount of linguistic intelligence (Thurmair, 2003).
- If bilingual texts are available (sometimes the translation departments have translation memories), tools could try to extract translation equivalents from such corpora; they would be preferable to standard dictionary translations because they would match the terminology really used by this particular application.

The result would be a list of (possibly bilingual) term candidates, to be merged with the existing resources, to match the terminology of a particular domain.

As this task is specific for each application, the tools must try to be cost-effective and fast; otherwise it would be too expensive for an end-user to set up such a resource.

Tools to annotate resources

Annotation of the extracted resources is the next task. There are several levels of complexity in this task:

- **Basic linguistic annotations** (part of speech, inflectional behaviour etc.) can be derived by linguistic defaulting techniques. For instance, most of existing MT systems provide such tools for coding support.
- **Complex linguistic annotations** like verb frames, semantic annotations etc. have also been tried in literature, with good results (Grishman/Sterling, 1992); there is a trade-off, however, between the effort to provide the corpus material for training, and the effort of hand-coding.
- **Conceptual annotations** are not straightforward to extract, and while there is research to find certain links between concepts (Maedche/Staab, 2001), most of the key links (the hierarchy itself) would have to be created by hand. A strategy which starts from a top level ontology (like EuroWordnet) and adds to the lower level nodes, linking concepts e.g. by adding subject area nodes (Magnini/Cavaglià, 2000) seems to be the most promising approach, and also allow a certain amount of exchangeability.

  It should be mentioned that the most useful conceptual links are the most concrete, detailed, and application-specific ones. Using mainly higher-level WordNet-type ontological links does not necessarily improve search results as it introduces noise in the search terms.
- There are also approaches to automatically find MT **transfer tests** and actions (Meyers et al., 2000); however, as this depends of the context in which such entries are used, some manual coding will always remain to be done.

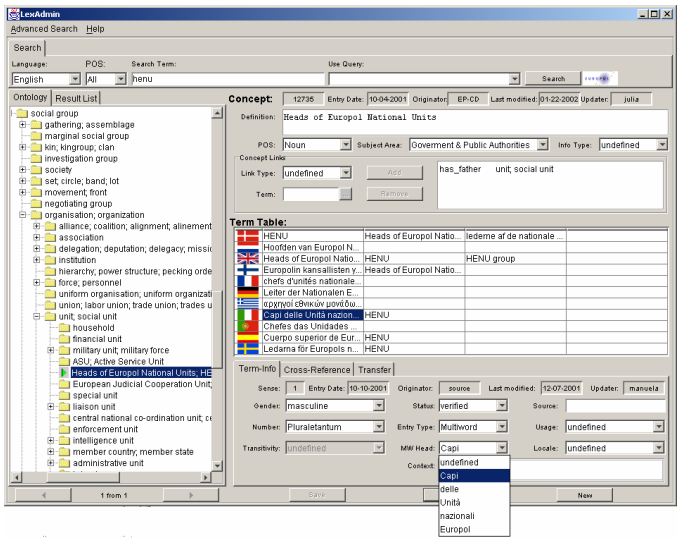

Fig. 1: GUI of administration tool

Tools to maintain the database

All the annotations need to be kept consistent, and maintained in the central resource. This resource therefore needs a powerful administration tool. It includes aspects of ontology management tools like (Stojanovic et al., 2003), but extends them for linguistic requirements.

Fig. 1 shows the basic interface of such an application. It was built to support a multilingual concept net for a multilingual information system, the application was law enforcement. About 16000 concepts were collected, with term descriptions in 11 languages, overall about 250 K entries.

The left part of the screen gives the **ontology**; it starts from the EuroWordNet top level ontology and goes down to a fine-grained representation of the terms in the law enforcement domain. With a right mouse click, users can add new nodes to this ontology, or move and copy nodes and branches. New imported terms go into a 'unlinked' branch, and can be moved to the proper place in the ontology by drag&drop operations.

The **concept description** is given in the right upper corner. A concept is described by a definition, a part of speech, and a subject area code. The concept can be linked to the hierarchy as visualised in the ontology tree, but can have additional links to other concepts (like *part_of*, *made_of* etc.; a subset of the EuroWordNet link types is used here).

In the right middle field, the **terms** for the concept are given, in the respective languages. Several terms for one language can be entered, forming synonyms. The terms were created by using the term extraction tools mentioned, running on the corpus material of the application.

Each term has its **linguistic annotations**, as given in the right lower corner. These annotations are used in formal linguistic analysis components (inflection class, gender, multiword structure, and others). They were chosen by investigating existing standards like OLIF which determine what features are common to most current MT

systems. Information on transfers (like: which is the preferred translation) or other term-specific links (like: abbreviation, head of a multiword term) are given here as well.
A more detailed description of the database can be found in (Jackson et al., 2002).

Linguistic Compilers
The linguistic resource maintained with the interface just shown needs to be used in different parts of the system with specific fractions of information: Query translation needs multilingual information, information extraction gazetteers mainly need monolingual morphosyntactic annotations. Instead of forcing the respective tools to access the central resource, an approach was chosen to compile the needed information for the different tools.
For this purpose, compilers were implemented using OLIF as exchange format. Compilers into the query analysis and translation tool, as well as the MT lexicon, use this interchange format, and convert the DB resources into the proprietary format of these tools, optimised for fast and compact access.

4 Outlook

Internal tests, as well as a survey of the literature, have shown that the quality of the multilingual linguistic resource is decisive for the quality of a multilingual scenario. With a good resource, there was nearly no deterioration of multilingual search as compared to a monolingual one. Poor, non-matching translation resources turn a multilingual search into a garbage producing device.
Based on such a resource, all components of an information system need to be adapted to a crosslingual scenario; it has been shown that this requires significantly more linguistic intelligence than in conventional systems, and drives the system a bit further to the direction of content understanding.

References

Babych, B., Hartley, A. (2003). Improving Machine Translation Quality with Automatic named Entity Recognition. Proc. EACL-EAMT, Budapest.

Beale, St., Nirenburg, S., Mahesh, K. (1995). Semantic Analysis in The Microcosmos Machine Translation Project. Proc. SNLP, 1995, Bangkok

Bel, N., Busa, F., Calzolari, N., Gola, E., Lenci, A., Monacchini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. Proc. LREC 2000, Athens

Beryl, T., Atkins, S., Lewes, Sx., Fillmore, Ch.., Heid, U., Christ, O. (1995): Lexicographical relevance in Corpus Evidence. DELIS Report, 1995

Braschler, M., Kamn, M., Schäuble, P., Klavans, J.(2000): The Eurospider Retrieval System and the TREC-8 Cross-Language Track. Proc. TREC 8

Calzolari, N., Bertagna, F., Lenci, A., Monachini, M., ed., (2002). Standards and best practice for multilingual computational Lexicons and MILE (the Multilingual ISLE Lexical Entry). ISLE-Report 2002

Calzolari, N., Grishman, R., Palmer, M., (2002). Standards & best practice for multilingual computational lexicons: ISLE MILE and more. Proc. LREC 2002, Gran Canaria

Chakrabarti, S., (2003). Mining the Web. Discovering Knowledge from Hypertext Data. (Morgan Kaufmann).

Eberle, K. (2001). FUDR-based MT, Head Switching and the Lexicon. Proc. MT-Summit 8, 2001

Grishman, R., Sterling, J. (1992). Acquisition of Selectional Patterns. Proc 14th COLING, Nantes

Jackson, A., Lewandowski, M., Thurmair, Gr., Zwickl, J. (2002): Concept Manager, Pflege multilingualer Ontologien im crosslingualen Retrieval. Proc. ISI, Regensburg.

Jackson, P., Moulinier, I., (2002). Natural Language Processing for Online Applications; Text Retrieval, Extraction and Categorization. (John Benjamins)

Knight, K., Koehn, Ph., 2003: Introduction to Statistical Machine Translation. Tutorial MT Summitt 2003, New Orleans

Maedche, A., Staab, St. (2001): Ontology Learning for the semantic web. IEEE Intell.Systems 16

Magnini, B., Cavaglià, G. (2000). Integrating Subject Field Codes into WordNet. Proc. LREC Athens

Meyers, A., Kosaka, M., Grishman, R., 2000: Chart-Based Transfer Rule Application in Machine Translation ,Proc. COLING 2000

Moreno, A., Pérez, Ch. (2000). Reusing the Microcosmos Ontology for Concept-based Multilingual Terminology Databases. Proc. LREC 2000, Athens.

Piperidis, St., Boutsis, S., Demiros, J. (1997). Automatic Translation Lexicon Generation from Multilingual texts.. Proc. AAAI 1997.

Richardson, St., Dolan, W., Menezes, A., Pinkham, J.: Achieving Commercial-quality Translation with Example-based Methods.

Ritzke, J. (2000). SEN-DNL-Gen: Dynamic Natural Language generation within the SENSUS system environment. Sensus Report.

Stojanovic, N., Hartmann, J., Gonzalez, J. (2003): OntoManager – a system for usage-based ontology management. Proc. FGML Workshop, GI.

Thurmair, G. (2003). Making Term Extraction Tools Usable. Proc EAMT-CLAW Dublin.

Thurmair, G. (2004): Comparing rule-based and statistical MT output. Proc. LREC Workshop.

Thurmair, G. (2004): Sprachtechnoogie in einem Informationssystem. To appear.

Thurmair, G., Lieske, C. (2002): Lexical Exchange Formats – DXLT and OLIF, Proc. LRC Workshop on Standards in Localisation, Dublin, 2002

Verdejo, F., Gonzalo, J., Peñas, J., López, F., Fernández, D. (2000). Evaluating wordnets in Cross-Language Information Retrieval: the ITEM search engine. Proc. LREC Athens

Volk, M., Ripplinger, B., Vintar, S., Buitelaar, P., Raileanu, D., Sacaleanu, B., 2002: Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval. in: International Journal of Medical Informatics, Volume 67:1-3, December 2002.

Voorhees, E., Harmann, D. (2000): Overview of the Eight Text REtrieval Conference (TREC-9). Proc. TREC 8

Weber, N.: MÜ-Lexikografie. Proc. GLDV, Köthen, 2003