[*Terminologie et Traduction* 1.1998, pp.140-166]

POUL ANDERSEN

Translation Tools for the CEEC Candidates for EU Membership - an Overview

Introduction

With the expected enlargement of the EU following the accession of up to ten Central and Eastern European countries (referred to as CEECs, which is the usual EU abbreviation), the translation complexity takes a quantum leap. The current EU languages (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish) can be translated in 110 language combinations, as each of the 11 languages can be translated into 10 other languages.

With the addition of 10 new languages (Estonian, Latvian, Lithuanian, Polish, Czech, Slovak, Hungarian, Slovenian, Romanian and Bulgarian) the complexity goes up to 21 x 20 = 420 language combinations, but there is no obvious political or linguistic justification for changing the European Union's official policy of supporting multilingualism, which finds its expression in the MLIS programme[1], among others.

Given this background, technological means to facilitate the translation process become even more important, particularly in view of the fact that most of the new languages are little known or understood outside their respective countries.

The present overview will be only of limited interest to the professional translator who is looking for tools to facilitate his work. The objective is rather to reach two different target groups:

1 Access to information for persons who during the course of their work are confronted with documents written in a language they do not understand, and who do not have access to "real" translation. In the contacts between EU and CEECs, it might be useful to distinguish between two typical situations:

1.1 Persons in CEECs confronted with documents written in major EU languages, such as English, French and German. These persons will often have a certain basic knowledge of the language, or will be able to find a colleague with such a basic knowledge. They may be helped with access to an on-line dictionary, which would allow them for instance to click on unknown words and get a translation in a window on their screen. Such users will essentially understand the rest of the text, and may not even need a proper translation.

1.2 Persons in EU countries confronted with documents written in CEEC languages. These persons will often not have any knowledge at all of the foreign language. Most CEEC languages belong to a different language family, and the reader cannot even "guess" the approximate contents through similar words (except perhaps for a Finnish reader confronted with a document in Estonian).

In this case, an on-line dictionary will not be of much help, whereas a simple translation system, which provides a rough translation, may give the reader an approximate idea about the contents of the document, until a proper translation can be provided.

2 Researchers, commercial companies and funding bodies (national authorities, EU services) who are interested or can be made interested in setting up projects for development of Machine Translation or more restricted Translation Tools between CEEC languages and EU languages.

Information sources

An important source of information on translation tools is the *Language Engineering Directory*[2] (LED). It lists 46 Machine Translation products and 24 Computer-Aided Translation products, but only two of these 70 products include one of the 10 CEEC languages, in both cases Polish (and see below under *Polish* for the negative results of our inquiries). Two other products claim that they cover "all languages", which at least must presuppose some preparatory work: as one of the providers mentions, his system "requires the compilation of specialised user dictionaries".

A more targeted source for CEECs is the *ELSNET survey*[3]. A search with the keywords "Machine Translation" gives 9 different hits among the 110 research institutes and companies from the 10 candidate CEECs listed in the survey. These hits concern the Czech Republic, Poland, Slovenia, Latvia and Hungary, but include hits on expressions such as "bilingual *morphological* dictionaries which may be used in machine-aided translation", and hits which mention in passing that the institution gives courses in Machine Translation. The search also gives a few hits for other

CEECs/NIS, such as Ukraine, Belarus, Uzbekistan and Georgia, which are outside the scope of this article.

A third source, which gives an estimate of the political importance attached to these issues in CEECs, is the *Panel Report* on the Action Plan for IS in the CEECs[4]. This report mentions the development of electronic language resources, including development of translation and localisation tools, as a priority domain for the promotion of the languages of CEECs (cf "idea" N° 13 "Multilingual Support for the Information Society" in the report), and it includes presentations of Information Society activities for 9 of the 10 candidate CEECs (Bulgaria is missing) and the FYR of Macedonia. The presentations were prepared by CEEC government representatives in the Panel.

In the following, we will list for each language the information that can be extracted from these three publicly available sources. However, we have relied to a large degree on individual communications from contacts in each of the CEECs. The main network of contacts has been the participants in the TELRI Concerted Action[5], but other persons have also contributed.

In many cases the information is provided by the persons involved in the projects. This is a first overview, which we are sure is neither balanced nor complete, and no quality assessment of the products and projects mentioned has been carried out. We have in some cases mentioned URLs, e-mail addresses or other contact information, but we shall be happy to provide more information on an individual basis and we appreciate all kind of comments and supplementary information - see contact address at the end of the article.

Scope of the overview

As is evident from LED and from the ELSNET survey, developments for CEEC languages lag far behind those for international languages such as English, German, French, Russian and Japanese, but they are also very modest compared with other smaller national languages such as the Scandinavian languages.

In order not to end up with an empty inventory for most of the languages, we have extended our search in two directions :

- we include information on tools such as *electronic bilingual dictionaries and morphological parsers,* which can either facilitate the translator's work

right now or form the basis of more sophisticated or automatic translation aids; however we do not consider printed dictionaries;
- we include information about ongoing *projects,* plans for future developments and commercial and other actors who have shown interest in the development of translation tools, and already have some kind of experience in the area of NLP.

The overview is mainly based on indigenous developments in the countries concerned. Politically, it is preferable that the countries concerned take care of their own languages, and in practice we have not come across many development activities or plans for these languages abroad, except in joint collaboration schemes such as those between CEECs and EU partners.

As an exception to this indigenous approach, we should mention that SYSTRAN Software Inc (SSI, a subsidiary of the French company SYSTRAN SA based in La Jolla, California) will start to develop a Polish-English MT system in 1998 (they are already developing Serbo-Croatian into English, but Serbo-Croatian is not among the candidate CEEC languages). In the medium to long term (2000-2005) SSI has plans for other CEEC languages, such as Czech, Slovenian and possibly others. At present, SSI is doing this development work on contract for the US National Air Intelligence Centre.

The SSI developments are based on a minimum vocabulary of 50 000 words, and include a general dictionary and specialised dictionaries for subjects such as computer science, electronics etc. As far as we are informed, developments for CEEC languages have until now been done in the USA, but it is possible that partners from CEECs will become more involved in the future. SYSTRAN'S CEEC related activities are not yet presented on the Internet.

We would also like to mention the activities of Xerox Research Centre in Grenoble, France, which co-operates closely with institutions and individual researchers from CEECs. Xerox does not produce fully automatic MT systems, but an important basis for future developments are morphological analysers and part-of-speech taggers, which have been developed or are under construction for Hungarian, Czech, Polish and Romanian. Other CEEC languages may be added later.

These linguistic tools are integrated into various applications including online translation aids, such as Compass and TANS, translation memory, bilingual terminology management or cross-language information retrieval. Xerox technology has been exploited in the EU-funded CALL project GLOSSER, with CEEC partners from Estonia, Hungary and Bulgaria, and is currently exploited in the project STEEL, which is co-ordinated by Xerox, with CEEC partners in the Czech Republic and Poland.

For more information on the company's activities, see: http://www.xrce.xerox.com/research/, and in particular the link to Multilingual Theory and Technology. The contact person at Xerox for CEEC-related activities is JEAN-PIERRE CHANOD: Jean.Pierre.Chanod@xrce .xerox.com.

The 10 languages of the CEEC candidates

In the following sections, the information we have been able to find is presented for each language, going from North to South.

Estonian

Estonia has ambitious plans on informatisation of the country, in the form of a computerisation programme for schools called "Tiger's Leap". The country also has an active national language policy. Estonia's university, which is located in the second largest town, Tartu, has participated in several European co-operation projects, together with the Institute of Estonian Language. There are also close links between the university and a commercial company, Filosoft, at whose home page one can find a list of:

- Programs for use via Internet, ia:
- Speller for HTML-documents in Estonian
- Morphological analyzer for Estonian
- Lemmatizer for Estonian
- Hyphenator for Estonian
- Thesaurus for Estonian
- Speller for HTML-documents in Latvian
- Speller for HTML-documents in English
- English-Estonian word-to-word translation aid

- Filosoft software Products:
- Speller for Estonian
- Hyphenator for Estonian
- Thesaurus for Estonian
- Speller for Latvian
- Hyphenator for Latvian

This is printed from the English version of the home page: http://www.filosoft.ee/index_en.html. Most of the links are in Estonian.

We are not aware of any proper translation tools. In spite of the country's recent history, and of the large Russian-speaking population, English is the most "interesting" foreign language, as can be seen from the services provided by Filosoft.

An important translation centre is Estonian Translation and Legislative Support Centre (ETLSC), which translates European legislation into Estonian and Estonian legislation into English, and has about 50 employees: translators, revisers, terminologists, computer specialists, etc. They are looking for software for terminology database and full-text database, and have tested some commercial products from abroad, without having found any currently available versions which satisfy their needs.

The following is a quote from a comprehensive Evaluation Report from ETLSC :

> *Continuous development of the Centre's information technology resources is essential. The development and purchase of software to manage full-text databases and terminological data in a way that is cost-efficient, employs resources efficiently and effectively, and focuses on their utilisation for the broadest impact and benefit possible will lead to an increase in the speed and efficiency of translation by facilitating the standardisation of terminology and avoiding the re-translation of repeated text. (It has been estimated that up to 1/3 of new European Community law is a repetition of sections from previously adopted European Community law.) This in turn allows for the better use of the Centre's human resources and increased productivity in general.*

Estonian is closely related to Finnish, and one might consider adapting existing translation tools for Finnish to Estonian, eg "recycle" a Finnish-English system into an Estonian-English system. I have presented this idea to Finnish and Estonian linguists, and both parties have confirmed that such an approach is reasonable.

Latvian

Like Estonia, Latvia has an active national language policy. In the Panel Report (see note 4), the Latvian representative mentions "Multilingual Support for the Information Society" as the first among seven Latvian priorities recommended by Latvian information experts.

He further mentions "software development and computer linguistics" as one of the four main fields of research. On the subject of participation in European Information Society projects, he mentions that Latvia took part in regional Awareness Seminar on Language in Baltic States (which was organised in November 1994 in Latvia's capital Riga, for all three Baltic states - PA) and that the Institute of Mathematics and Computer Science of the University of Latvia has been researching the PC-based processing of Latvian language dictionaries and texts since 1988.

Since 1993 the Artificial Intelligence Laboratory of this Institute deals with the development of a Machine Translation (MT) model for Latvian. This work is supported by the Latvian Council of Sciences through two projects: "Limited Model of Automated MT System for Latvian" (1993-1996) and "Development of Probabilistic Methods for Automated Disambiguation of Natural Language Texts and Applications for MT" (1997-1999).

At the initial stage different multilingual MT systems were studied, and the approach of MT system SWETRA (Lund University, Sweden) was chosen as the basis for further development of the Latvian model.

SWETRA is a multilingual interlingua-based MT system for automated translation of stock-market texts between English, Swedish, Russian, Polish and German. The interlingua (functional representation) in the system provides an unambiguous and common language-independent description of the sentence.

The head of the Artificial Intelligence Laboratory is Andrejs Spektors, and the person most actively involved in MT developments is Inguna Greitane. The MT system under development is called LATRA, and I have received the following status report on their activities (February 1998) from Ms Greitane:

> *For the moment LATRA has been tested on stock market texts - some samples are from the SWETRA demo (see* above - PA) others are from our *business newspaper* Dienas Bizness (The Daily Business). *During my stay in Sweden I did several tests between SWETRA languages and Latvian: Swedish-Latvian, Latvian-English and Russian-Latvian. Because LATRA is a scientific project, tests are done only at sentence level.*
>
> *Now we have plans to make the LATRA module more complex by adding preference rules. We also hope to get some results in translation of EU documentation. For this purpose we started studies of "White Paper" (on preparation of the associated CEECs for integration into the internal market of EU - PA), ie we work on statistical analysis of text corpora to have a more detailed description of the language and to develop preference rules.*

Since last year we have a good co-operation with Translation and Terminology Centre (TTC) - the official translator of EC and Latvian documents. During the next 3-4 months we will work on the development of a terminology database for TTC.

An example of a translation and the intermediate representation:

Gada sākumu iezimēja ari struktúrkapitāla palielināšanas maratons, kuru ir spiestas uzsākt komercbanku lielākā dala.

```
[subj
   (s(s
     (s(m(statutarycap,sg),[]),m(increase,sg),[]),
     m(marathon,sg),
     [subj(s(s(m(def,_11675),m(commercialbank,pl),[]),m(most,sg),[])),
     pred(m(m(begin,nonf),m(must,pres))),
     obj(m(marathon,sg)),obj([]),advl([]),advl([]),advl([]),advl([])]
)),
pred(m(m(feature,past),[])),
obj(s(s(m(def,_8757),m(year,sg),[]),m(beginning,sg),[])),
obj([]),advl(D),advl(D),advl(m(also,_9206)),advl([]),
co
   (s(s
     (s(m(statutarycap,sg),[]),m(increase,sg),[]),
     m(marathon,sg),
     [subj(s(s(m(def,_11675),m(commercialbank,pl),[]),m(most,sg),[])),
        pred(m(m(begin,nonf),m(must,pres))),
        obj(m(marathon,sg)),obj([]),advl([]),advl([]),advl([]),advl([])]
),[.])]
```

The marathon of the increase of statuary capital which the most of the commercials bank must begin characterised the beginning of the year also.

Lithuanian

Lithuania participates actively in several co-operative projects with partners in EU in the area of Language Engineering and Multilingual issues, and hosted a very successful seminar on "Language Applications for a Multilingual Europe" in Kaunas in April 1997, but we have no information concerning Translation Tools or MT for Lithuanian.

The ELSNET survey lists no less than 16 institutions in Lithuania, including three commercial companies, which might be possible contact points for development work:

1 *Sekasoft Ltd*

During past years Sekasoft has taken part in the standardisation of character sets in the ISO INSTA/IT subcommittee. Sekasoft develops and distributes computer fonts and tools for Lithuanian and Russian language support for various operating systems (MS DOS, Windows, OS/2 and UNIX), text converters between different text coding formats, etc. Sekasoft has won the tender and completed Lithuanian Government contract for the lithuanisation of UNIX systems (SCO, IBM, HP, SUN). We also take part in the Computer terminology and legislation committees in Lithuania. Major directions of work: groupware system implementations, Lithuanian language support for Lotus Notes and workflow automation based on person-to-person interactions (ActionWorkflow methodology).

2 *Fotonija Ltd*

The main activities of the company are directed towards software engineering for PCs. We use Windows NT and Windows 95 in our company. The programming is done using C++, Fortran, Assembly language and rapid application development tools such as Borland Delphi. Fotonija produces various Lithuanian language support packages for Windows 3.1x, Windows NT and Windows 95. We support other local languages as well, Polish and Russian included.

3 *Publishing House TEV*

Publishing House TEV is a joint Lithuanian-Russian-Dutch company created in 1991. The main fields of activity are:

1 Typesetting services rendered for different scientific publishers around the world: (SIAM, Plenum (USA), VSP, Balkema, Brill (The Netherlands), Omsha (japan) etc;

2 Publishing for domestic market (textbooks in mathematics and physics, bilingual and specialised multilingual dictionaries);

3 Electronic dictionaries for MS DOS, Windows 3.XX, Windows 95, OS/2. We consider an electronic dictionary as consisting of a computer program (viewer) together with the contents of a dictionary. The viewer allows the user to search and retrieve information from a dictionary contents file. A dictionary contents file is characterised by the fact that to each unique word (root entry) corresponds one article (explanation). The explanation could contain text, links to other entries and multi media elements (sound). [...] The source code, used for the

paper version, is converted to SGML. The dictionaries are compiled to the internal viewer format.

Polish

In the Panel Report (note 4) the longest contribution (6 pages) is from Poland, but the contribution strikes a somewhat pessimistic general note on IS issues: "One of the consequences of a general underfinancing of research is that interest in taking part in European initiatives on Information Society is very scarce. Another consequence of a relative scarcity of researchers in Poland is a very sceptical attitude of opinion-making intellectuals to the issues of information civilization". The only mention of language-related activities also ends on a pessimistic note: "Awareness Seminars on Language: A seminar on Language and Technology was organized back in April 1995 in Poznan, and further activities are being carried out, including an Internet discussion list NPL-L@uci.agh.edu.pl. However, there is no interest from industrial or commercial partners".

As mentioned above, Polish is the only language mentioned in the LED in the sectors Machine Translation and Computer-Assisted Translation, with the two products:

NeuroTran. Machine Translation: NeuroTran carries out sentence to sentence translation using artificial neural network technology to remember grammar rules and to generalise from stored knowledge. However, an inquiry into the declared supplier resulted in an angry answer, stating that they *"have nothing whatsoever to do with Language Translation software ...".*

LexiTools. Computer-Aided Translation: Performs rough translation of whole sentences and texts from English to Polish. It includes a program for compiling and supplementing user glossaries and dictionaries. It also includes a word processor but can be used with other word processing programs provided they can handle plain text. It comprises a 5 000 entry glossary of computer technology, and a 22 000 entry glossary on tourism and everyday life. Optional: 16 000 entry glossary of general business terms. It is marketed by the firm LexiLAB, Wadowice, Poland, but we have not been able to contact this firm.

An experimental Russian-Polish MT system, SCANLAN, was presented at the abovementioned Awareness Seminar in Poznan in 1995[7]. The project was started in 1986 by Prof HIPPE in Rzeszow, and the main executor was Dr ANDRZEJ KACZMAREK, Politechnika Rzeszowska, Katedra

Informatyki Chemicznej. The system is oriented to translations from Russian into Polish with special stress on technical, chemical texts.

In a recent e-mail Dr KACZMAREK writes that "SCANLAN is restricted to translate text in one direction because the analysis and synthesis module are only adequately implemented for Russian and Polish. In order to expand SCANLAN functionality to translate other pairs of languages the proper analysis, synthesis and transformation modules should be developed. [...] Because of my limited access to a Web server, SCANLAN is not presented in the WWW. I am currently preparing a new implementation of SCANLAN using Borland Delphi as a programming environment. I hope that because of its Windows compatibility, the new version will be more hardware-independent".

A group of researchers at Adam Mickiewicz University and the Technical University of Poznan is interested in starting up projects in the domain of translation technology. In Poznan, there already is a project on *Automatic Spoken Language Translation: Polish-English,* headed by WIKTOR JASSEM.

The contact person for MT is ZYGMUNT VETULANI, who is in close contact with WIKTOR JASSEM and his son, KRZYSZTOF JASSEM. VETULANI's research group has participated in several EU cooperation projects, dealing with standards for electronic dictionaries, and is willing to build a consortium for a medium size MT project involving Polish, English and, possibly, French.

The MT related experience in Poznan covers the following aspects :

- parsing of Polish sentences (using Prolog),
- non-modular Polish-to-English machine translation,
- technology of bilingual electronic dictionaries with quick access time,
- linking the system with text-to-speech voice synthesisers,
- large morphological dictionaries of Polish,
- tools for language engineering (lemmatisers, taggers, concordance generators),
- finite automata for compressing and accessing lexical data,
- understanding of text by a computer (a question answering system POLINT for dialoguing with a computer has been constructed),
- NL communication with autonomous robots involving reasoning about space (for future developments).

The main achievement to date is the prototype of a Polish-to-English system written in SWI-Prolog, based on a dictionary of 2 000 lexemes (selected on the basis of a corpus of text in computer sciences). This system is the object of current research and development.

Cooperation could include a small commercial company in Poznan, LEX SC, with language engineering as its main field of activity.

Czech

The Institute of Applied and Formal Linguistics within the Faculty of Mathematics and Physics at Charles University in Prague is a centre of international renown. It is headed by EVA HAJICOVÁ. Also at Charles University is the Institute of Theoretical and Computational Linguistics under the Faculty of Philosophy, headed by VLADIMIR PETKEVICZ, and the Czech National Corpus, headed by FRANTISEK CERMAK.

R&D in Machine Translation at Charles University goes back to 1961, with a very restricted English-Czech system. In recent years, the Institute of Applied and Formal Linguistics has been engaged in two series of MT projects:

- English-Czech, since 1974; there were several experiments with the system APAC (acronym for "Automaticky preklad z anglictiny do cestiny"); attention focused on abstracts from the domain of electronics; some theoretical issues had some success, such as the fail-soft measures[8] and the use of dependency grammar for the analysis of English; the project was "frozen" because of lack of money on the developmental side. The system runs in batch-mode (not inter-active). Contact person at Charles University is ALEXANDR ROSEN (rosen@ff.cuni.cz).
- Czech-Russian, since 1985; based on a more or less "direct" method, using experience with the English-Czech system (see reference in note 8). Also frozen in 1990, due to lack of interest (it was supported by an Institute which specialised in software development, and needed translations of software manuals from Czech to Russian, which was then obligatory in the COMECON countries);

Contact persons at Charles University are EVA HAJICOVÁ (hajicova@ufal.mff.cuni.cz) or VLADISLAV KUBON (vk@ufal.mff.cuni.cz).

Another academic institution with MT-related interests is the Department of Information Technologies, Faculty of Informatics, Masaryk University, Brno, (contact person is KAREL PALA). They cooperate closely with Charles University and will soon join the EU Telematics project EuroWordNet 2, where they will produce a Czech lexical database, ie Czech WordNet, which will be linked to WordNets in English, German, French, Dutch, Italian, Spanish and Estonian. The ultimate result then will be a strong multilingual tool, eg for browsing and information retrieval within the WWW.

The market of commercial MT systems in the Czech Republic currently contains several competing systems. Most of these systems are aimed at individual users with a very limited knowledge of foreign languages, who would like to have access to information contained in documents written in English, German, Spanish or French. In general, both the price (100-250 DM) and the quality of these systems are rather low, and they often give misleading results, but they may be adequate for accessing information. Quantitative tests of their performance have shown that they are able to translate approximately 60% of the source text, with the best of them, TRANSEN, able to translate about 65%.

There are no Czech systems aimed at professional translators. The large translating companies or companies engaged in software localisation are using large translation memory based systems such as TRADOS or IBM Translation Manager.

Independent sources report that the best quality among these systems is offered by TRANSEN, from the company POSY s r o. TRANSEN contains a knowledge base describing syntactic relations among words. The system is able to identify the subject and object of a sentence, and it contains a rich system of morphemic classes which are useful in the generation of morphologically correct text (other systems sometimes issue an incorrect word form). The system is also to a certain extent able to learn from sentences which it was not able to analyse.

TRANSEN was originally built for English-Slovak translations and later adapted for Czech.

Other simple MT systems for Czech:

- PC Translator, version 9.0 from LangSoft & SOFTEX is a bi-directional system between Czech and English, translating online with manual entering of inflections. Smaller modules are available for French, German and Italian. This system uses simple morphological analysis and also simple NP analysis making it possible to put whole noun phrases in the target language (Czech) into the proper morphological form. The system supports domain-oriented translation. The main dictionary contains some idiomatic expressions (see also below under *Slovak).*
- SiR Translator from SiR Software is similar to PC Translator. It uses more complex lexico-syntactic data, which unfortunately are not complete (eg verbal valency frames have only one slot). The quality of the translation is slightly better than in the previous case, but the syntactic analysis is

only able to handle simple sentences. It cannot deal with more complex noun phrases, such as "a variety of battlefield factors". Both systems have dictionaries containing about 200 000 word pairs for Czech and English and slightly fewer for other language pairs.

- SKIK, version 2.0, from the company SKIK, is a bi-directional system between Czech and English, with a dictionary of 100 000 entries, running in batch mode, with automatic inflection. A German version is undergoing tests. The system is based on the translation of word chains (2 or 3 words).
- Landi Translator (developed by MKCS company) is a bilingual English - Czech dictionary with automatic translation as an added feature. It translates word for word and does not use morphological analysis.

An important Czech software company producing language tools is LINGEA Ltd. It is a small software firm, run by a programmer, Dr Pavel Sevecek (pavel@lingea.cz), with close relations to NLP research at the Faculty of Informatics and the Faculty of Arts at Masaryk University in Brno. According to our information, LINGEA Ltd can be expected to produce the first commercially successful translator in 1-2 years time, ie by the end of 1999.

The following lingware products are currently sold by LINGEA Ltd:

- hyphenator for Czech, Slovak, English and German,
- spelling checker for Czech, Slovak, English and German,
- lemmatiser for Czech, Slovak, English and German - it is used at the Faculty of Informatics, Masaryk University, for tagging Czech corpus texts (in co-operation with the Institute of Czech National Corpus), it comprises 165 000 Czech stems and covers about 97% of the corpus texts,
- morphological analyser for Czech, Slovak, English and German, which is used in NLP research at Masaryk University as a module within the currently built parser for Czech,
- thesaurus for Czech, Slovak, English and German,
- English-Czech, Czech-English translation dictionary,
- German-Czech, Czech-German translation dictionary.

These tools for Czech (and sometimes Slovak) are integrated in many commercial word processors and DTP systems such as Adobe, Corel, Lotus, Microsoft (especially in MS Word), Quark, Software602 and also in fulltext applications (Fulcrum, Verity).

An important commercial translation company, involved in cooperative R&D projects together with Charles University, is MORAVIA

Translations SA in Brno (see http://www.mtranslations.cz/), with offices in Poland and Hungary for Polish and Hungarian.

The periodical *Language International*[9] recently published an overview of Czech translation tools. The overview focuses on two types of tools:

1 *Electronic dictionaries*

Five bilingual dictionaries on CD-ROM and/or diskettes are presented, all with at least 100 000 entries for English (50 000 entries for other languages, in the case of German and French) or with specialised terms. The products presented are:

1.1 Czech Dictionary Database (Leda Ltd Publishing House)
1.2 Oplatek Software
1.3 WinDict Febra Ltd
1.4 WinDict StormWARE
1.5 GED

2 *Computer-Aided Translation Tools*

Two tools are presented:

2.1 SKIK 2.0 (see above)
2.2 PC Translator 9.0 (see above)

The author, MIROSLAV HERALD, is reasonably critical, mentioning that the latter product "obviously [is] not of much use to a professional translator who can type target text directly into an advanced text editor with full formatting capabilities", but still considering both products of interest as they "allow users to scan foreign-language texts for useful information".

Slovak

1 *Computer-Aided Translation Tools*

Several of the systems mentioned above for Czech, also include Slovak: the best Czech MT system, TRANSEN, was originally developed for English-Slovak. We have no precise information about its current availability.

- PC Translator V 9.0 (see also above under Czech): a bi-directional system (MS DOS + WIN) between Slovak and seven foreign languages

(English, German, Russian, French, Italian, Spanish and Czech), all at least with a dictionary of 100 000 entries, for English and German 200 000 entries. It is a Slovak localisation of the Czech product (firm Language Soft, CZ-687 Zahorovice).

- PC Correspondent V 3.0: a system for generation of business letters. The text is generated from 6 000 ready-made Slovak sentences. The Slovak business letter is immediately available in its English, German and Spanish versions.

Both products are marketed by the firm TEOS Trencin, Brigadnicka 11, SK-911 01 TRENCIN. Tel/fax: + 421 831 436104, +421 831 441615.

A commercially available English-Slovak Machine Translation has been developed by MARIAN JANÍK, Zilina, Slovakia, who works for the company ABAKUS Ware Ltd. Work on this MT program was interrupted for some time but has now been taken up again, and the work on the system should be finished in 6-8 months. At present, the program is not being marketed.

2 *Electronic Dictionaries*

Computer Technique English-Slovak Explanatory Dictionary (Anglicko-Slovensky vykladovy slovnik vypoctovej techniky) for MS DOS.
Contact address: LAUDIC Konzorcium, Mierove namestie 44, SK-911 01
Trencin.
Gold Soft Dictionaries:
Three bilingual dictionaries for MS DOS, all with 40 000 entries for the source language:

- *English-Slovak and Slovak-English Dictionary*
- *German-Slovak and Slovak-German Dictionary*
- *Italian-Slovak and Slovak-Italian Dictionary.*

Contact address: CK Arrangement Consulting, Ltd, Kocelova 15, SK-821 08 Bratislava, Tel: +421 75262375, Fax: +421 75262893.

Information for Slovak has been provided by ALEXANDRA JAROSOVA (SASAJ@juls.savba.SK).

Hungarian

Hungary is very active in cooperation projects with partners in EU in the area of Language Engineering and Multilingual issues, and hosted a very successful seminar on "Language Resources for Language Technology"[10] in Tihany in September 1995.

At Lajos Kossuth University in Debrecen, a small group of linguists and programmers has been working over the past 6-7 years on a Hungarian parser with the intention of further developing it into an interactive Hungarian-to-English MT system, with financial support of OTKA, the Hungarian National Science and Research Fund. The information about the project was provided by Laszlo Hunyadi (hunyadi@llab2.arts.klte.hu):

Based on a *full* morphological analysis, the rule-based parser gives the analysis of any arbitrary Hungarian simple sentence (with a modest dictionary of a few hundred words so far). The result of parsing is given in the form of bracketing (similar to predicate calculus notation). It is expected that the system will be interactive. At the moment, it is implemented in Pascal. Here is the output of the parsing of a simple sentence:

akarok	*kerni*	*a*	*lanyoktol*	*egy*	*eheto*	almat
I-want	*to-ask*	*the*	*girls-from*	a	*edible*	*apple-acc*

```
—1 —
akar < > {1,0}        +ok<akj11>
—1 —
ker   < > {2,0}       + ni <inf>
— 1 —
a   < >  {0,0}
—  2 —
a    < >  {0,0}
— 1 —
lany < > {13,0}        + ok <pl>    +tol<abl>
— 1 —
egy   < > {0.0}
— 2 —
egy   < > {2,0}
— 1 —
e    < > {77,0}       + heto <pot>
— 1 —
alma < > {28,0}        + t <acc>
```

Subject NP: en (*en* ["I"] does not appear on the surface, it is deduced from the analysis). Predicate VP: akar (ker (alma (e, egy) lany (a))).
There have been three publications on the project[11].

Perhaps the most best known commercial company in Language Engineering in CEECs is MorphoLogic in Budapest, Director Gabor Proszeky, PhD (proszeky@morphologic.hu, English homepage:

http://www. morphologic.hu/morphenu.htm). Some recent publications are listed in the notes[12]. Morphologic offers the following commercial products (from http://www.morphologic.hu/products.htm):

Translation Support:		
Intelligent bilingual dictionaries	-	MoBiDic
Proofing Tools:		
Spelling checkers and hyphenators	-	Helyes-e?
Spelling checkers for MS-DOS	-	Helyeske
Grammar checker	-	Helyesebb
Hyphenator	-	Helyesel
Thesaurus	-	Helyette
Integrated package of proofing tools	-	Helyesek
Proofing language change utility	-	Babel
Intelligent Search Support:		
Stemming modules	-	HelyesLem
Text Analysis and Extraction:		
Morphological analyser	-	Humor

MoBiDic, which stands for Morphologic Bilingual Dictionaries and is explicitly mentioned as a "translation support", is a series of bi-lingual electronic dictionaries, to be used for Multilingual word processing, computer-aided translation, machine translation: "All the source language modules contain morphological analysers in order to find the adequate lexical entries in the lexicon instead of the letter-wise alphabetic neighbours of the text words. It can also be used as terminology management system that allows the user to create dictionaries. The system is able to use multiple languages/dictionaries at the same time in one environment, so a lookup can have results (translations) in more than one language".

Types of commercially available dictionaries:

English-Hungarian, Hungarian-English (22 dictionaries, 8 more in preparation)
German-Hungarian, Hungarian-German (11 dictionaries, 7 more in preparation)
French-Hungarian, Hungarian-French (2 dictionaries, 1 more in preparation)
Italian-Hungarian, Hungarian-Italian (1 dictionary, 1 more in preparation)
Russian-Hungarian, Hungarian-Russian (1 dictionary)
Latin-Hungarian, Hungarian-Latin (1 dictionary)

Polish-Hungarian, Hungarian-Polish (1 dictionary in preparation)
Hungarian thesaurus (2 thesauri in preparation).

Among the dictionaries are basic dictionaries with general vocabulary, dictionaries of phrases and of idioms, and several dictionaries for specific subject fields such as business, auditing and computing among others.

Slovenian

MIRO ROMIH, general manager of the commercial company Amebis d o o, Kamnik, Slovenia (Miro.Romih@amebis.si), has sent the following basic information about their work in the field of MT systems:

> *Amebis d o o is the main company in Slovenia on the field of NLP. We develop and market different linguistic software. In the field of language modules and tools* we *cooperate with some other companies, like Microsoft, Corel and Lotus.*
> *We have also been member of Copernicus MULTEXT-EAST project.*
> *Our products are:*
> - *speller for Slovenian and Serbian;*
> - *hyphenator for Slovenian and Serbian;*
> - *thesaurus for Slovenian;*
> - *syntax checker with morphologic analyser for Slovenian.*
>
> *For* MT the next *electronic dictionaries (in cooperation with publishing company DZS d d and Slovene Academy for Science and Arts) are important:*
> - *English-Slovenian (100 000 headwords);*
> - *Slovenian-English (70 000 headwords);*
> - *German-Slovenian (120 000 headwords);*
> - *Slovenian-German (100 000 headwords);*
> - *dictionary of Slovenian literary language (110 000 headwords).*
>
> *Italian will be the third language, that we will support in electronic shape.*
> *The "real" MT* system is not *ready yet, but we have prepared all necessary modules.*
> *We have plans to start developing Slovene-English MT system in the middle of*
> *this year (1998-PA).*

At Fran Ramovs Institute of Slovenian Language, Ljubljana, PRIMOZ JAKOPIN (http://www.uni-lj.si/~ffjakopin), has developed a very comprehensive software package EVA, for morphological analysis and tagging, lemmatization, concordance generation and several other functions, including a kind of translation memory/workbench. EVA is part of TRACTOR resource pool, maintained by RAMESH KRISHNAMURTHY, (ramesh@clg.bham.ac.uk) under TELRI (see note 5) and is also available for research purposes at: http://www.zrc-sazu.si/frisj/pj/eva.

An English summary of a relevant paper on POS tagging for Slovenian can be found at: http://www.zrc-sazu.si/frisj/pj/eva/pos_tagging.html.

The Institute is part of the Scientific Research Centre of the Slovenian Academy of Sciences and Arts.

Another important academic institution within Language Engineering is the department for Intelligent Systems at Josef Stefan Institute (IJS) in Ljubljana. They currently do not work on translation tools, but they have produced multilingual language resources, which can help in the production of such tools - see http://nl.ijs.si/, and more specifically the links from http://nl.ijs.si/GNUsl (many of which are in Slovenian).

An on-line Slovenian-English/English-Slovenian computer dictionary from IJS can be consulted at http://www.ijs.si/cgi-bin/rac-slovar.

Contact person at JSI is TOMAZ ERJAVEC (Tomaz.Erjavec@ijs.si).

A very competent contact person in Slovenia in MT questions, also for other languages than Slovenian, is JARO LAJOVIC (jaro.lajovic@mf.uni-lj.si).

Romanian

In the Panel Report (see note 4), the Romanian representatives mention the Awareness Days on Language Technologies at Tusnad (during the Summer School EUROLAN'97, July 1997) in connection with the two projects from the Action Plan "Awareness Seminars on Language Technology and Information Society" and "Multilingual Support for IS".

They further mention that the "Information Society" R-D-I programme is one of the priorities of the new national R-D-I programme (R-D-I = Research & Development and Innovation), and contains Language Technologies as one of ten subprograms. The Romanian R-D-I programme was restructured in 1997 into modules and objectives in order to make it compatible with the contents of the EU R&D programmes in IT and Telematics, and it makes reference to the two language-related projects from the Action Plan mentioned above.

There is thus a clear political interest in Romania for Language Technology, and there was a visible and appreciated support from the Romanian Academy of Science during the Awareness Seminar on Language & Technology, which took place in Bucharest in January 1996[13].

The project DBR-MAT - An Intelligent MAT System for Structurally Different Languages - is an extension of DB-MAT (see below under

Bulgarian) to Romanian, running May 1996 - 1998. Like DB-MAT, it is funded by Volkswagen Foundation, with WALTHER VON HAHN, University of Hamburg (vhahn@nats.informatik.uni-hamburg.de), as project leader.

The declared aim of DBR-MAT is "Investigation and pilot implementation of a MAT system combining a knowledge-based approach with statistical methods for NLP", building upon the ideas and the demo implementation of DB-MAT.

The Romanian partner is the University of Bucharest, Faculty of Mathematics, Computer Science Department, and the contact person is FLORENTINA HRISTEA (flori@math.math.unibuc.ro).

DBR-MAT results are communicated in research papers[14] (most of them available on the web page) and tested in a pilot implementation with an extendable architecture. Further information: http://nats-www.informatik.uni-hamburg.de/~dbrmat/dbr-mat.html.

Other language tools for Romanian that could be used as building blocks for MT, together with (pairwise) grammars yet to be developed and bi- or multilingual dictionaries:

- at least two commercial spelling checkers (integrated in MSOffice) and a free spelling checker for Xemacs under Unix;
- a very accurate hyphenator (commercial);
- several terminological dictionaries (the largest multi-lingual terminological dictionary, with a concept thesaurus for electrotechnical vocabulary, has been developed by Prof Dr Ing ALEXANDRU TIMOTIN, Universitatea Politechnica din Bucuresti, and is currently being further developed in an EU-funded project, LANGELEC, which will add Polish, Russian and German to the existing Romanian, French and English terms. Contact e-mail: timotin@star.tezaur.pub.ro);
- at least two commercial Romanian-English-Romanian desktop dictionaries (word list associations);
- several morphological analysers for Romanian (partly in cooperation with Xerox; see the first part of this article);
- several concordancers;
- several speech systems (recognition, text-to-speech synthesisers, prosody tools);
- several very efficient parsers: chart parsers, generalised LR parsers (Tomita-like for unification grammars);
- at least 4 text generators (one made in cooperation with LIMSI, Paris, and there are plans to integrate it with LIMSI's morphological processor and chart parser);

- very good tagging results - up to 98.32% accuracy for unrestricted texts, with a very large tagset of 680 morpho-syntactic codes;
- a lemmatiser with > 95% accuracy.

The unification-based linguistic platform EGLU (built in cooperation with ISSCO, Geneva) has been intensively used for the development of morphological models. It has also been demonstrated in translation, but without "real" grammars and "real" bilingual dictionaries.

Although only DBR-MAT is focusing on MAT/MT proper, and is an extension to a German-Bulgarian project, the MT infrastructure is available in Romania, and with a real pair of computational grammars and one or more wide coverage bilingual dictionaries (in appropriate format) with Romanian as one of the languages, serious MT applications could be demonstrated in a short time (not necessary with a breath-taking linguistic accuracy or computational performances).

More information can be found in the two Awareness books (see note 13), or by contacting Dan Tufis, Romanian Academy of Sciences: tufis@valhalla.racai.ro.

Bulgarian

Bulgarian institutions participate in several co-operative projects with partners in EU, in fields such as terminology, text corpora, CALL and electronic dictionaries. Most of the participants work within the structure of the Bulgarian Academy of Science, such as Linguistics Modelling Laboratory, Institute of Bulgarian Language and Institute of Mathematics and Informatics. A small commercial company, Capricorn, is involved in a terminology project with partners in most other CEECs and partners from EU.

Of special interest for Machine Translation is the Linguistic Modelling Laboratory (LML). The head of LML, Prof Elena Paskaleva, started her career in 1965 as participant in the first Bulgarian MT experiment, carried out her PhD in the domain of MT, and wrote a book on MT.

LML is engaged in a project, concerning new approaches to the construction of MAT systems and the modelling of terminology semantics, DB-MAT, a German-Bulgarian Machine Aided Translation project funded by the Volkswagen Foundation (Germany). The project investigates a new MAT paradigm[15], differing from the usual MAHT systems[16] in the organisation of lexical knowledge which is encoded not in the lexicon entries but in a single language-independent knowledge base of conceptual graphs from where NL explanations are generated in different languages[17]. The user support is linguistic and subject related. DB-MAT demo works for texts in a technical domain and represents a laboratory version of user-navigated terminological lexicon[18]. The project has now been extended to Romanian, DBR-MAT, see above. As for DBR-MAT, project leader is WALTER VON HAHN (see under *Romanian).* Contact person at LML is GALIA ANGELOVA (galja@lmlserver.acad.bg). Further information: http://nats-www.informatik.uni-hamburg.de/~dbrmat/db-mat.html.

The other main NLP and MT related activities at LML are concentrated on the development of:

1 Basic LR and LT tools for Bulgarian - a task feasible only on a local level. A large lexical database (60 K entries) and a high-speed morphological analyser (100 K words per second) are the current basis of most LML mono- and multilingual achievements. The Bulgarian LDB is ported to INTEX - the multilingual corpora processing system created in LADL - Paris VII University (in JRP COPERNICUS'94 BILEDITA project[19]). The analyser is integrated in various multilingual tools[20] (see below). The EAGLES standards are applied to Bulgarian morpho-syntactic data (see: http://www.lml.acad.bg/ projects/rus-bg - in construction).

2 Multilingual tools for processing large parallel corpora. MARK ALISTeR is a system for marking, aligning and searching for translation equivalents in parallel corpora. Tested and evaluated on large English-Bulgarian, French-Bulgarian and French-English parallel corpora, MARK ALISTeR is an alignment environment for further multilingual achievements[21]. The first prototype of the system was created in COPERNICUS'94 JRP GLOSSER[22] and extended with dictionary part after the project.

3 Tools for automatic extraction of terminological translation equivalents and automatic compiling of terminological dictionaries. Based on MARK ALISTeR, a tool for automatic extraction of translation equivalents of legal terms was created[23]. The tool was designed and tested for 4 MB French-Bulgarian corpora within BILEDITA COPERNICUS'94 project and applied further on 5 MB English-French corpora (both corpora are legal texts - treaties of Council of Europe)[24].

LML is also engaged in development of Basic Russian LR and LT tools (on a common - methodological and implementational basis for both Slavonic languages), but that is outside the scope of the present overview.

For comments, corrections and requests for further information, please contact:

POUL ANDERSEN
Translation Service & MLIS Programme
EUFO 1197
European Commission
*rue Alcide de G*asperi
L-2920 Luxembourg
phone +352-4301-34324,
fax +352-4301-34655
e-mail: poul.andersen@lux.dg13.cec.be

References

[1] MLIS stands for 'Multilingual Information Society' - this European Commission programme runs between 1996 and 1998, see 'http://www2.echo.lu/mlis'.

[2] The *Language Engineering Directory - A resource guide to Language Engineering organisations, products and services.* Compiled by PAUL M HEARN, 1996, published by Language & Technology, Madrid, Spain, for the European Commission, DC XIII/E, 382 pages, see also 'http://www2.echo.lu/mlis/en/leddesc.htm'.

[3] ELSNET Survey of Natural Language and Speech Organizations. It is not foreseen to publish this survey in printed form. The survey can be found at http://www.elsnet.org/publications/survey/ and was updated in the beginning of 1998 - the statistics in this article were before this update. An earlier survey by ELSNET, Survey *of Language Engineering Organisations in Central and Eastern Europe,* was published in October 1994, but is now both out-dated and out of print. Both surveys were funded by the European Commission, DC XIII, the recent one as part of the Concerted Action *ELSNET goes East.*

[4] EU-CEEC Forum on Information Society - Panel on the Implementation of the Action Plan, 10-11 September 1997, Portoroz, Slovenia - Report. Edited by CENE BAVEC / ANTON SCHRAG / GRAZYNA WOJCIESZKO / Ljubljana, September 1997, 179 pages. The report is available on http://www.mzt.si/med/front.html.

[5] TELRI - *Trans-European Language Resources Infrastructure* is a COPERNICUS Concerted Action, funded by the European Commission, originally for 1995-97 (an application for extension has been submitted). It brings together 22 institutions from 17 European countries, incl all 10 candidate CEECs. See TELRI homepage at 'http://www.ids-mannheim.de/telri/telri.html'.

[6] TELRI - Proceedings of the Second European Seminar "Language Applications for a Multilingual Europe", Kaunas, Lithuania, April 17-20, 1997, ed by RUTA MARCINKEVICIENE / NORBERT VOLZ, IDS/VDU Mannheim/Kaunas 1997, 215 pages.

[7] *Jezyk i technologia* (in Polish), ed by ZYGMUNT VETULANI / WITOLD ABRAMOWICZ / GRAZYNA WETULANI, Akademicka Oficyna Wydawnicza PLJ, Warszawa 1996, 217 pages.

[8] *Translation Equivalents: When neither a Dictionary nor a Corpus helps,* by EVA HAJICOVA / ZDENEK KIRSCHNER, in TELRI Newsletter N° 7, October 1997, p 19/22 (TELRI - see note 5 above).
Czech-to-Russian Transducing Dictionary by BÉMOVÁ A / V KUBON, in COLING-90, Papers presented to the 13th Int Conference on Computational Linguistics, Helsinki 1990, p 314/316.
Fail-Soft ("Emergency") Measures in a Production-Oriented MT System, by HAJICOVÁ E / Z KIRSCHNER, in Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics, Copenhagen 1987, 104/108.

[9] "Country Profile: Focus on Czech Translation Tools" by MIROSLAV HEROLD, freelance interpreter-translator/journalist and independent consulting engineer in applied Computer Science, in *Language International* 9.3 (1997) 38/39.

[10] TELRI - Proceedings of the First European Seminar "Language Resources for Language Technology", Tihany, Hungary, September 1 5 and 16, 1995, ed by HEIKE RETTIG, in collaboration with JÚLIA PAJZS /GÁBOR KISS, 196 pages.

[11] DRAGALIN A G/ HUNYADI L / P UZONYI "On some practical issues of an interactive machine translation system with standardized text EVM i perevod. Tbilisi, 1989
UZONYI P "Egy magyar-angol interaktiv forditasi rendszer morfologiai elemzo moduljarol" Elso Magyar Alkalmazott Nyelveszeti Konferencia. [On the syntactic module of an interactive Hungarian-English translation system. First Hungarian Conference on Applied Linguistics.] Nyiregyhaza, 1991. 709/712.
HUNYADI L "Egy magyar-angol interaktiv forditasi rendszer szintaktikai elemzo moduljarol" Elso Magyar Alkalmazott Nyelveszeti Konferencia [On the syntactic module of an interactive Hungarian-

English translation system. First Hungarian Conference on Applied Linguistics.] Nyiregyhaza, 1991. 713/716.

[12] Publications by GABOR PRÓSZÉKY'S in the Field of Language Engineering (my selection - PA, see also note 22):

- "MoBiDic: A New Language Technology Tool for Translators" Proceedings of the 2nd *Transferre* necesse *est Conference,* Budapest (1997, in preparation)
- "Language Technology in the Service of CALL": New *Horizons in CALL* (Proceedings of EUROCALL 96), 53/64, Szombathely (1997)
- "From Research to Application..." *ELSNews* 6(2), (1997)
- "Morphologic - A Language Engineering Company from Hungary" *ERCIM News* No 26, 31/32 (1996)
- "Humor: a Morphological System for Corpus Analysis" *Tihany Proceedings* (see note 10), p 149/158 (1996)
- "How To Reach the LT Market?" D TUFIS (ed) *Limbaj si Technologie* 197/200 (see next note).
- "Morphologic Bridging Gaps between Academia and Industry in Hungary" *ELSNews* 4(5) (1995)
- "On the Increasing Role of Linguistic Subsystems in Office and Business Applications" R TRAUNMÜLLER / GY KOVÁACS (eds) *Computer & Communications (Creating New Applications in Business, Administration and Society)* R Oldenbourg, Wien/Miinchen, 19/26, (1995)

[13] Full proceedings, mainly in Romanian:

Limbaj si Tehnologie, ed by DAN TUFIS, Editura Academiei Romane, Bucuresti 1996, 270 p.

Selected contributions, partly extended, in English translation: *Recent Advances in Romanian Language Technology,* ed by DAN TUFIS / POUL ANDERSEN, Editura Academiei Române, Bucuresti 1997, 277 p.

[14] WALTHER V HAHN "Machine Translation" *Limbaj si Tehnologie* (see preceding note).

MARIUS POPESCU "Maximum Entropy Model for Dependency Parsing" *Annals of the University of Bucharest, Mathematics-informatics year* XLVI, Bucharest, 1997.

FLORENTINA HRISTEA "On WG Syntactic Analysis With Special Reference to Romanian", to appear in *Annals of the University of Bucharest*

FLORENTINA HRISTEA / MARIUS POPESCU "A Word Grammar Approach to Syntactic Analysis With Special Reference to Romanian", to appear in *Annals of the University of Bucharest.*

See also references in following notes (on DB-MAT).

[15] v HAHN W "Innovative Concepts for Machine Aided Translation" *Proceedings VAKKI* Vaasa, Finland, 1992, 13/25.

[16] COLE R / MARIANI J / USZKOREIT H / ZAENEN A / ZUE V *(eds), Survey of the State of the Art in Human Language Technology* Chapter 8.4: "Machine Aided Human Translation".

[17] v HAHN W / G ANGELOVA "Providing Factual Information in MAT" *Proceedings of the Conference "MT- 10 Years on"* Cranfield, UK, November 1994, p 11/1 - 11/16.

ANGELOVA G / K BONTCHEVA *"NL* Domain Explanations in Knowledge Based MAT" *Proceedings of COLING-96,* August 1996, Copenhagen, Denmark, p 1016/1019.

[18] v HAHN W / G ANGELOVA "Combining Terminology, Lexical Semantics and Knowledge Representation in Machine Aided Translation" *TKE'96: Terminology and Knowledge Engineering.* Proceedings of the Conference "Terminology and Knowledge Engineering", August 1996, Vienna, Austria, p 304/314.

[19] SCHULZ K / STOYAN MICHOV *BILEDITA: A Multilingual System for Alignment and Corpus Analysis* Technical Report, 1998.

[20] PASKALEVA E "Bulgarian Language Resources and Tools in Joint European Initiatives" *Kaunas Proceedings* (see note 6).

[21] PASKALEVA E / ST MICHOV "Second Language Acquisition from Aligned Corpora" Proc *of the Int Conf "Language Technology and Language Teaching",* Groningen, The Netherlands, April 1 997.

[22] Nerbonne J / L Karttunen / E Paskaleva / G Proszeky / T Roosmaa "Reading more into Foreign Languages" *Proc Fifth Applied NLP Conference,* April 1997, Washington, ACL

Stoyan Mlchov "MARK ALISTeR: MARKing, ALIgning and Searching TRanslation equivalents" *Kaunas Proceedings* (see note 6).

[23] Michov St "Automatic Extraction of Translation Equivalents from Aligned Corpora of Legal Texts" *Proc 3rd European Seminar "Translation and Equivalence, Theory and Practice",* Monteccatini, Tuscana, Italy, October 1997 (to appear).

[24] Paskaleva E "Automatic Extraction of Translation Equivalents on Data-Driven Platform - Wishes and Reality " *Proc 3rd European Seminar "Translation and Equivalence, Theory and Practice",* Monteccatini, Tuscana, Italy, October 1997 (to appear).