EU and the new languages

Translation – possibilities, policies and practicalities

Prague, April 22-23, 1999

Programme with abstracts

April 22

Introduction to the Workshop

John Hutchins

The European Association for Machine Translation was founded in 1992 to serve the community of people interested in machine translation and computer-based translation tools. It is one of the three regional associations of the International Association for Machine Translation (IAMT). Its goal is to bring together users, developers and researchers in this field of language technology into the fruitful interchange of information and opinion both through its publications and through its annual workshops.

Previous workshops have taken place in Vienna, Copenhagen and Geneva. A major focus of these workshops has been the integration of MT systems and translation tools into the documentation processes of organisations. At each however, there have also been sessions devoted to the economic and managerial issues faced by individuals and companies deciding whether these technologies are suitable or appropriate in their own situations. In the coming years, when countries of Central and Eastern Europe join the European Union translators and translation companies of all kinds are going to encounter translation problems of a nature and on a scale which they have not met before. It is our hope that this workshop will introduce some of the issues involved and suggest some possible solutions through the greater use of computer aids. It will not be the last opportunity - the EAMT committee expect that the questions raised at this workshop will remain with us for many years to come, and will therefore continue to appear on the agenda of our workshops and conferences. These two days are just the beginning.

John Hutchins (president, EAMT)

Introduction to Session 1: EU and Government Language Policy

John Hutchins

The accession of new countries to the European Union has particular mplications for translators, both within the new member-countries themselves and within the Translation Service of the European Commission. The challenges are not just the sheer volume of documentation that has to be translated (initially and on a regular basis), the setting up of the necessary translation organization, the training of competent translators in the languages required, the development of new translation and interpretation procedures, and so forth. There is also the cculturation to the European Union: familiarization with new organizations and regulations, learning new vocabulary, and encountering new meanings for old vocabulary. Some of these changes impinge on everyday life: regulations of the food industry are frequent subjects of amusement or despair, new Euro-words are the butt of journalists' jokes, and familiar words used with new meanings have led sometimes to serious misunderstandings. At the same time, member countries naturally seek to assert their own individualities, not just in their institutions, laws, taxation, traditions, etc. but also in their languages. The result is a formidable set of challenges for the translation professions of new member-countries. Their comfort has to be that their predecessors have successfully risen to the challenges - which does not mean all difficulties have always been overcome - and that the Translation Service of the Commission has the means and the expertise to help them to make sure that the gains of joining outweigh any transitional pains.

This morning we shall hear from Pavlina Obrava about how translators in the Czech Republic are tackling the problems of translating the fundamental legal documents of the Union, and what steps are being taken to cope with terminological issues. In the second talk, Dimitri Theologitis of the Commission's Translation Service will describe the wide range of integrated computer tools and facilities which are now at the disposal of European Commission translators, ranging from lexical resources and multilingual databanks of official and working documents to automatic translation systems, computer-based translator workstations and terminology management systems.

John Hutchins (president, EAMT)

Translation of EC Legislation into the Czech Language

Pavlína Obrová

As an associated country to the European Union, the Czech Republic has the task to provide translations of EC legislation which will become (after authentication by the Council of the European Union and the accession of the Czech Republic to the EU) valid EC legislation applicable in the Czech Republic. It means that it is necessary to provide translations of documents, which will have the same legal effect as Czech Law adopted now by the Czech Parliament, government, ministries, etc.

This presentation deals mainly with development of this activity, describes the state of play, institutional background and methodology for ensuring good quality translations which will fulfil their task.

The main purpose of the Coordination and Revision Centre is to prepare a parallel Czech version to the existing eleven versions of EC legislation. Simultaneous task coming with translations is to create equivalents of EC terms in the Czech language (the Coordination and Revision Centre is responsible for legal terms, while the individual ministries are responsible for technical terms). For terminology management, the products of the company TRADOS are used (MultiTerm, and partly also the Translator’s Workbench). These products were chosen especially because the Translation Service of the European Commission also uses them.

Department for Compatibility with EC Law – Coordination and Revision Centre

Department for Compatibility with EC Law was established in the year 1994 as a part of the Office for Legislation and Public Administration to ensure compatibility of Czech Law with EC Law and to transpose EC legislation into Czech law. For this purpose it became necessary to translate EC legislation (especially directives) into Czech. In the beginning the form of "working" translations was sufficient, because legal experts (mostly in legislative bodies of ministries) used them.

Continuously it became evident that more precise translations (so called "revised" translations) would be necessary so as private companies, citizens, etc. could use them. The European Commission promised to help in setting up translation and revision centres for EC legislation in all associated countries. (This obligation of the European Communities appears already in the Association Agreement - 7/1995 Coll.)

It was natural that the Department for Compatibility played crucial role in establishing formally the Coordination and Revision Centre, because up to 1998 the Department itself performed the tasks of the Centre. The Coordination and Revision Centre was formally established by the Government Resolution in August 1998 and the Project on Providing the Translation of EC Legislation and the Activities of Coordination and Revision Centre was adopted in September 1998 (Government Resolution 645/98).

The Department for Compatibility (including the Coordination and Revision Centre) was till February 1999 part of the Ministry of Justice, now it forms part of the Office of the Government.

Pavlína Obrová (Department for Compatibility with EC Law)

Preparing for new languages at the European Union

Dimitri Theologitis

{Disclaimer: What follows are probable scenarios based on experience from previous accession exercises and the latest thinking. Official decisions have not yet been finalised.}

Regulation No 1/1958 states that all official languages of the Member States are official and working Community languages. While no change is foreseen on the official language front, the Community Institutions will no doubt restrict the number of working languages following a new set of internal procedural rules.

Setting aside interpretation which poses supplementary real-time constraints, the extension of translation activity from 11 to potentially 22 languages creates new issues linked to personnel, recruitment, administration, training, infrastructure, workload, workflow, working methods, outsourcing, translation tools, terminology and phraseology.

Means to cushion the impact of this dramatic increase in language combinations from currently 110 to potentially 462, might include:

The use of relay or pivot languages

Two-way ("aller et retour") translation (i.e. also between the translator's mother tongue and first foreign language with revision of the latter by a native speaker)

Increased use of freelance translation

Temporary translation and revision "field offices" in the candidate countries prior to accession

Increased use of translation technology in the areas of terminology, translation memories and machine translation.

An interesting pre-accession project is the translation of the so-called "acquis communautaire", i.e. all the community legislation, into the language of each candidate country. Special means are set aside for this by the European Commission Translation Service and the Technical Assistance and Information Exchange office (TAIEX).

Budgetary implications are not negligible. However, at around 1 euro per citizen per year currently, this is a small price to pay for clear communication in Europe.

The Translation Service of the European Commission

Situated in Brussels and Luxembourg, the SdT (for *Service de Traduction*) today houses some 1 300 translators, 100 language support staff, approximately 100 management staff and close to 500 secretaries and assistants. It produces well over one million pages per year, in a combination of in-house, freelance and machine production.

A major technological modernisation effort is producing good results. Main translation aids in use, being installed or being overhauled include: the Eurodicautom termbase; the Euramis Linguistic Resources Database and search engines combined with translator's workbenches; Systran machine translation; and a document server with full text search and retrieval possibilities.

Dimitri Theologitis
Head of Unit "Computer Translation Aids"
European Commission Translation Service
CCE JMO B4/74
L-2920 Luxembourg

Tel: +352 4301 33632, fax: +352 4301 34069
E-mail: dimitrios.theologitis@sdt.cec.be

==============

Session 2: Experience from translation of EU documents

Chair: *Anthony Clarke*

Experience from translation of EU documents

Gábor Prószéky

There are three main actors in the ideal translation workflow: the translator, the terminologist and the reviser. All the three roles we have described in terms of their input and output information in the translation workflow and developed a language technology toolkit to help them to solve their tasks as effective as possible. In the project "*Translation EU Legislation Texts into Hungarian"* the translators and revisers belong to various agencies, and the agencies communicate to each other via a central program office. This organisational issue helps to ignore technical difficulties in the workflow caused by the different technical education of the translators.

In the basic system called *MoBiDic* both the dictionary modules based on already published paper dictionaries of various publishers and the translators' own (common) glossaries are represented using existing standards (i.e. SGML, XML, HTML) enabling simultaneous access to multiple lexical resources performing queries through linguistic pre-processing (stemming, etc.). At the end of the project, the common glossary of the translators involved in the project will be published as the largest and most up-to-date "*English–Hungarian Dictionary of Legal Terms*

Used in EU Documents.

In the project of translation of EU legal documents into Hungarian the above-introduced LT toolkit is used supporting the translation of 40,000 pages by around 100 translators. The system has been written in standard C and C++, and is totally portable. Actually the MoBiDic server runs on Windows NT and Unix (Linux, Solaris), and the MoBiDic clients are Windows 98/95 and Windows NT applications. Communication between the clients and servers is based on TCP/IP, consequently web-browsers (like Netscape or Explorer) can also be used as special clients of the *MoBiWeb* system. It also allows parallel lookup in a set of an unlimited number of MoBiDic-dictionaries and dictionaries on various web sites with the help of the embedded *MoBiGloss* subsystem.

Gábor Prószéky

MorphoLogic

Késmárki u. 8, 1118 Budapest, Hungary

proszeky@morphologic.hu

http://www.morphologic.hu

Aligning and extracting translation equivalents from EU documents - a possible look on EU integration

Elena Paskaleva

Within the framework of two EU Copernicus'94 projects, the Linguistic Modeling Laboratory in the Bulgarian Academy of Sciences was charged with the following research tasks: the compiling of a large bilingual aligned corpora base and the creation of a data-driven tool for extracting translation equivalents from these corpora (see the demo presented at EAMT). In the accomplishment of these tasks, besides the general CL challenges in the design and the implementation of the tool, some specific problems arose. They are connected to the choice and the compilation of the document base in the initial phase of the project and to the usage and extension of the tool after the project. Following the general cycle of the involvement of the linguist/translator in these activities (from the design to the application) some observations spanning a 4 year period were made.

The tool. The basic function of the tool (MARK ALISTER) is the alignment in Gale-Church style and the extracting module of the system extracts the translation equivalents of terms using a data-driven algorithm (Ted Dunnin's coincidence method applied to aligned texts). The main features of the tool from the user's point of view are: a large scope of text formats for the input texts, language independence and a linguist friendly interface with editing facilities aiming to compensate the insufficiency of the data-driven method.

The text base.

Structure. Following the general recommendations for the structure of the text base (in Multext) the fiction is only 20%. Three corpora collections were compiled – English-Bulgarian, French-Bulgarian, and English-French. The last two (each of volume 1,2 M words) consist of legal documents, issued by the Council of Europe.

Compilation. In the process of construction of the corpora base the specific problems of acquiring the texts and their translations in electronic form mirrored the degree of our integration in the system of the European legal, social and political values. At the beginning of the project all Bulgarian translations were keyboarded. Five year later there are considerable changes in the availability of Bulgarian translations in electronic form. That is not the simple result of technology advance but a natural issue of the way Bulgaria took to the decisive incorporation into the European political and legal system in 1997.

The feed-back in the triade*:* *creating* the tool and the text base, *extending* the base, and *applying* the tool. The suppliers of the text corpora became the first users of the tool. The real life of the tool became possible in the new joint initiatives of the different institutions: governmental – Ministry of Justice and European Legal Integration, Ministry of Foreign Affairs, etc.; international - Delegation of the European Commission in Bulgaria, different Phare Program Management Units; non-governmental – the newborn European Law Society and even in the commercial software – in a DB systems in law the translation of European documents is included in the base.

The European Law Society being a research and educational institution develops the Center for Legal Translators and Interpreters in close collaboration with the Scool for Training and Education of Translators in Sofia University. In this vigorous activity the tools developed in LML will be placed at translators' disposal in the centers where investigation work and creation of translators' aid facilities is planned – at first place in Sofia University on the ecquipment supplied by TAIEX – EC.

The authors of the tool are satisfied to see their production to leave the domain of research and prototype software construction (generously granted from the sources of European collaboration) and to contribute to the same collaboration, in this way restoring not the financial costs but the noble idea of integration.

Elena Paskaleva

Bulgarian Academy of Sciences

hellen@lml.bas.bg

http://www.lml.bas.bg/

===============

Session 3: Tools providers

Chair: *Bente Maegaard*

Semi-automatic acquisition of lexical resources for new languages or new domains

Jean Senellart

Building lexical resources is a costly problem in terms of manpower, this is particularly true of the description of a totally new language or of the description of a new technical domain in an already described language. This acquisition is necessary to increase translation coverage and accuracy.

We show how finite state automaton can be used for that purpose. Compared with a word-list description, finite-state grammar representations provide a very accurate description of the different syntactic structures and of the lexical similarities of a family of utterances. We present different tools designed to build and maintain databases of local grammar using corpora to which bootstrap algorithms are applied. We illustrate the whole process on examples, such as the description of human occupations in French.

We show that this description is well-fitted to automatic translation and that the quality of the description ensures the quality of the translation process. Conversely, the use of finite-state automaton database in the translation process guarantees the correctness of the description.

To conclude, we show that the description of finite state automata is not limited to the lexical level of technical sub-langages but can also be used for the description of complex sentences. In that case, we discuss the possibility of automatic acquisition of new syntactic structures.

Jean Senellart

SYSTRAN & LADL

Université Paris 7

MT Open to Standards

Alex Murzaku

Machine Translation (MT) is as good as the flexibility it offers to the end user. There have been many approaches to this issue from creating domain specific dictionaries to allowing the creation of user dictionaries.

The former is a task almost impossible to be considered accomplished:

* How many entries are enough entries?

* What's the perfect size for a dictionary?

* What makes an entry valid for being included in the dictionaries shipped with the products?

The latter issue seem to have its own problems as well:

* How many words can an end user enter?

* How does the end user code these words?

* What is a manageable dictionary size from the software point of view?

At Logos we decided to attack all these questions at once - open up to the standards:

We deliver a minimum dictionary that covers the most frequent dictionary entries; we offer to the end user tools that allows for maximum encoding with minimum effort; all our databases reside in a commercial RDBMS which allows for scaleability and for all the benefits that come with it. We put our biggest effort in the grammars that use the information stored in the dictionary and in the tools that allow for quick build-up of the dictionary.

All the intelligence resides in the grammar and in the entry/import/export rules.

This flexibility allowed us to implement a new target language (Portuguese) in a very short time frame. Even the code base in development has become smaller and easier to maintain and upgrade. The linguists can try now "wild" ideas with very little effort.

The L&H approach to development of tools for new languages

Gregor Thurmair / Johannes Ritzke

Extending Language coverage is one of the basic requirements for players in the language technology sector. L&H runs an internet service called iTranslator which covers more than twenty different language pairs.

There are two possibilities to extend the language coverage:

1. The iTranslator provides a multi-vendor platform where a scheduler accepts translation requests, routes them to the respective translation engine, and hands the result back to the user. This architecture has APIs which are supported by all different L&H translation systems (including ones for Arabic and Japanese), and could host additional languages as well.

This technique allows for fast integration of existing translation systems if they can be made compliant with the API descriptions.

2. We also are building up new languages. In doing so, we follow the principle that we need to have tangible results as soon as possible.

The first step would be to collect **corpora** of data for the respective language. These corpora are used for training and testing language technology components. They are a value in themselves in case of speech understanding applications, as a means to create language models for speech systems.

In parallel, we start with **lexicon** work, and lexicon lookup components. We take conventional dictionaries or other resources which are available (e.g. via ELRA).

We then develop **morphological components** for the languages involved. This task includes the definition of tagsets, of inflection paradigms etc. Based on this, we develop lemmatisation and inflection components. The technology which we use is based on learning algorithms, inferring linguistic rules from example material. The examples partly come from lexicons, partly from corpus material. These components need to cover words which are not lexicalised yet, and need to have error rates of < 5%.

Based on this, we can implement tools which do terminological extraction from corpora (both single words and multiwords), so we can validate our lexicon coverage. We can also support small applications like word to word translation, indexing for retrieval, and the like.

The next step is to gather information on the **syntactic** behaviour of the new language; there are two levels of sophistication, the first one being statistical tagging. In order to obtain the statistical material we developed a training tool which allows linguists to quickly decide on the ambiguous word forms with simple mouse clicks. The results are then used for the tagger. With this technique, many applications requiring shallow parsing can be supported. (Statistical tagging is somewhat language dependent, however; languages with rich morphology do not gain as much as languages like English). The second level would be a complete high-level syntactic analysis; such an analysis will use the T1 formalism, and a sophisticated rule editing and testing environment. Benchmark suites and comparison tools are needed to evaluate improvements and deteriorations.

Finally, **transfer** lexicons and links to the target languages need to be developed for a full MT system.

The whole development is supported by a framework of multilingual developments based on the experience with several western and eastern European languages. This experience lead to well-defined component interfaces and clearly exchangeable software and lingware parts.

Current developments cover several Central European languages, as well as new combinations of existing languages.

Gregor Thurmair / Johannes Ritzke

L&H Language Technology

The PROMT Translation Technology for Russian and other languages

Svetlana Sokolova

ProMT is a new name of the next version of our well-known MT Software -STYLUS - the most popular machine translation software for the Russian language. It can translate from Russian and into Russian for many European languages: English, French, German, and Italian. This MT system STYLUS was launched into the Russian market in 1991 as an English-Russian translator for software documentation. At the moment our Russian systems have more than 50 000 users worldwide.

ProMT is a family of several applications with the same Machine Translation kernel inside. The interface solutions are intended for wide range of end users. They are ProMT Internet Kit, ProMT Home, Pocket ProMT, and ProMT Professional.

These applications are language-independent, Machine Translation kernel works as an OLE automation server providing a possibility for text translation and managing system dictionaries both for ProMT applications and for external applications like MS WORD or EXCEL.

Machine Translation kernel combines translation engines for the above mentioned language pairs. The translation engines for different language pairs have an identical structure and include Translator (to provide translation process), Dictionary Editor (to adjust the system to the user's domain) and Linguistic Database dedicated to the language pair. All components are designed to be as much language-undependable as possible and to employ the software tools that have already been developed for some language pairs in the ProMT systems.

Linguistic Database is a heart of the engine. There are three logical components of the linguistic database for each language pair: Morphology, Network Grammar and Dictionaries which are closely related to each other.

Our approach is universal for all the languages in the system and is based on the most complicated Russian morphology. For the Russian language the morphology base combines about 800 morphological types, for German – more than 300 , about 200 for English and more than 350 for French, Spanish and Italian.

The general-purpose dictionary for one language pair contains, as a rule, about 100000 entries. The volume of a domain-specific dictionary varies from 10000 to 60000 stems. The English-Russian, Russian-English domain-specific dictionaries include more than one million entries for more than 30 domains. The morphology model allows to provide a semi-automatic procedure for definition of morphological type of word and allows to speed up the process of customization of the system.

Our experience with Russian-based language pairs and with non-Russian language pairs shows that this technology allows to produce translation engines for new language pairs for the very short period of time as we made it for German-French and English-French engines which have been published under the trade name Reverso by Softissimo. They have been very well accepted by the computer press and the users.

Svetlana Sokolova , graduated from the Leningrad State University as mathematician , obtained a Ph.D. in computer science, had been involved in several machine translation projects as a chief of software design group. In 1991 established the company " PROject MT", Ltd , of which she is currently the president.

PROject MT, Ltd, established 1991 as a private company to develop and to market MT software. The PROject MT team consists of mathematicians, programmers and linguists. Now the company staff includes 42 full-day employees and about 12 freelancers. The leading specialists of the company have extensive experience both in MT research work and development of commercial software.

PROject MT LTD,

199053, p.o.box 632, St.-Petersburg,

Russia

svetlana@promt.spb.su

www.promt.ru

Reverso : a new generation of machine translation software for English-French-English, German-French-German, etc.

Théo Hoffenberg

Developed with ProjectMT, and published worldwide by Softissimo, Reverso is setting a new standard for machine translation. The combined features allow to reach a very large target of users : it’s so simple and powerful, that all PC users whether ignorant of foreign languages or professional translators, can take advantage of it.

Reverso provides:

Very good linguistic output : it allows to understand totally most of the texts : letters, technical documents, fact sheets... and to give a very good basis if you need to publish the translation.

Very high speed (up to 1 page per second): this allows real-time process like web-browsing, or it helps doing iterative process (the translator can modify some words in the dictionary, and get instantaneous result).

Extensive enhancement capabilities : with Reverso, the users or specialists can create specialized lexicons with very powerful and simple coding system : automatic suggestion of conjugations of verbs, complex expressions with choice of the variable parts, coding of word construction (preposition use, ...). There are different levels of access to the dictionaries... Lexicographers can take advantage of the interface to create lexicons (which can be printed, merged, compared...) and to see directly the impact of their work on real texts.

Browser-translator : Reverso allows to translate Web pages on-the-fly, preserving layout, links...

Professional interface : with its „ translation environment ", Reverso allows you to have simultaneous scrolling between source and target text (alignment), to select translation alternatives with a mouse-click, print source and target text together...

Integration into Word and Excel : you can translate documents totally or partially, add words to the dictionaries, select the dictionaries used... without quitting your favorite application.

Reverso was first released as the first French-German-French machine translation software in June 98 ; a new version of this language pair is available since April 99.

Reverso was officially released in April 99 in English-French-English, and it has already been selected by numerous organizations (even with preliminary versions): Compaq, France Telecom, Axa, French Army, French Department of Culture... and has already received rave reviews from the computer press : PC Expert : „ its translation engine is the best ", PC Professionnel 4 stars... „ Translating while browsing, integration into Word and Excel, exceptional speed, translation quality, ... Reverso has a lot of trump cards "

Reverso is based on Promt technology, which has a long experience and leadership in Russian machine translation systems. It can be adapted to new language pairs, in particular those involving languages already handled : German, English and French and also more generally to Roman and Slavonic languages. There are also opportunities to create specialized or customized dictionaries. Reverso can be a very good educational tool for students in linguistics, computer linguistics, lexicography, translation...

Softissimo, established in 1986, has been a pioneer, developer, publisher and distributor in the field of linguistic software for the corporate the mass markets. After electronic dictionaries (Collins On-Line), grammar checkers (Hugo), multilingual writing tools (Unitype, Uniwrite), language learning products (Quick English, Infolangue...), Softissimo is now focusing on machine translation with Reverso, developed within a close cooperation with ProjectMT, and has already established this product as the new reference in its field.

Théo Hoffenberg, CEO Softissimo

theo@softissimo.com

Softissimo

7 rue Auber

75009 Paris

www.softissimo.com

============

April 23

Session 4: Research in MT and translation tools

Chair: *Eva Hajičová*

Introduction

Machine translation may serve as a prototypical example of a domain where the applicational aspects overshadow its possible theoretical implications (or, perhaps, basis?). Even worse: I bet that one can find many specialists to think (and we can find signs of such opinions also in the Czech software community) that if a company hires a linguist for a machine translation project, the work on the project slows down, if not collapses. At the same time, there is a well-founded opinion in the computational linguistic community that such broadly conceived multi-lingual projects in machine translation as e.g. the EUROTRA project in the past decade have offered very exciting possibilities for theoretical investigations, even if the expectations of applicational results not really high.

It has been the intention of the programme committee of the present workshop to make both sides of the issue of machine (or machine-aided) translation visible. The two papers in this session are oriented towards the theoretical side, though with a good applicational background of the speakers: Petr Sgall is the initiator of the machine translation projects in former Czechoslovakia and in spite of being the founder of the group with the official title "Section of the theory of machine translation and algebraic linguistics" (founded at the Faculty of Philosophy as early as in 1959) he has never lost from his sight possible applicational projects (be it English-to-Czech or, much later, Czech-to-Russian). The second speaker, Igor Boguslavskij from Moscow, has long been a major figure in MT efforts in Russia and has always had the reputation of a guardian of the theoretical soundness of the systems his Laboratory has been building, based on the Meaning-Text linguistic theory proposed by Igor Melchuk.

Petr Sgall in his paper revisits the notion of intermediate structure and gives a brief characterization of one possible way how to build such structures; Igor Boguslavskij gives an overview of the system ETAP which was built as a more general natural language processing system with a particular view to machine translation. Both proposals are based on rather closely related starting considerations: they adhere to a stratificationally conceived system of language description, both are based on dependency syntax and both assume that much of grammatical information is to be included in the lexicon. And, in my opinion, both presentations include theoretically well-founded observations from which any applicational machine translation system may profit.

Eva Hajičová

ÚFAL MFF UK
Malostranské náměstí 25
CZ-11800 Praha
Czech Republic

hajicova@ufal.mff.cuni.cz

Translation to and from Russian: the ETAP System.

Igor Boguslavsky.

ETAP-3 is a multipurpose linguistic processor designed to handle natural language texts. The functions carried out by the processor include machine translation, natural language interface with databases, synonymous paraphrasing of sentences and some other things. The most advanced option is that of machine translation. The working languages of the machine translation option are Russian and English for which full-scale morphological and syntactic parsers and generators, as well as 60,000-strong high level syntactic and semantic lexicons, have been developed. There are operational prototypes translating from French into Russian and from Russian into German and Korean.

The system is based on the Meaning <--> Text linguistic theory proposed by Igor Melchuk and further developed and adapted for computer implementation by our laboratory. The salient features of the ETAP-3 system are:

Stratificational approach: every sentence is represented at several linguistically motivated levels which include morphological level, surface syntactic level, and normalized (deep) syntactic level.

Transfer approach: correspondence between the source and the target languages is established at the level of the Normalized Syntactic Structure.

Dependency approach: both surface and normalized syntactic structures are represented by means of dependency trees.

Lexicalistic approach: linguistic information is distributed among the grammar and the lexicon, the latter containing much more information than is usual in NLP, including lexical functions and rules of different types.

In addition to the main mode of operation which yields one (usually the most plausible) translation, a multiple translation mode is developed which offers alternative translations for syntactically and/or lexically ambiguous sentences. E.g., for the ambiguous sentence *They made a general remark that...* two different syntactic structures and respectively, two different translations are obtained: (a) *Oni sdelali obshchee zamechanie, chto...* (≈ They made some common remark that ...) and (b) *Oni vynudili generala otmetit', chto...* (≈ They forced some general to remark that ...). [The example is taken from real ETAP output].

The system allows the tuning of the text processing to different subject domains and different types of texts.

A 15 to 20 words long sentence of an average complexity is translated in 10-15 seconds. The system is implemented on a MicroVAX 3100 computer and is being ported under Windows NT/95. All programs have been written in the C language.

Igor Boguslavsky,

Computational Linguistics Laboratory

Institute for Information Transmission Problems

Russian Academy of Sciences

On Intermediate Structures and Tectogrammatics

Petr Sgall, Prague

Multilingual machine (assisted) translation, the need of which will soon be underlined by the rise of the number of EU languages, requires intermediate structures of a new level. It may be assumed that procedures for analysis and synthesis of the individual languages can be formulated relatively easily on the basis of the methods that have been already tested, mainly on languages from western Europe. However, the transfer procedures should find a new, more economical and perspicuous shape; moreover, most of the central and east European languages differ typologically from languages already handled to a higher degree. Therefore it appears useful to prepare an alternative structure that could serve as a basis for the treatment of grammatical structures of these new languages, or, eventually, of all languages involved.

Such an economical basis for intermediate structures can be found in dependency grammar with complex symbols, describing the (underlying, tectogrammatical) sentence structure in the form of bracketted strings of indexed symbols (with indices corresponding to the values of morphological categories, to syntactic functions and to the topic-focus articulation). An approach underlying such a grammar is the Functional Generative Description, which has been elaborated at Charles University, Prague.

The paper brings a characterization of this approach with an outline of a specification of the syntactic representations both in the form of a generative procedure and of a declarative specification. Illustrations of problems of analysis and synthesis and of their solutions for Slavonic languages will be presented.

Petr Sgall

ÚFAL MFF UK
Malostranské náměstí 25
CZ-11800 Praha
Czech Republic

sgall@ufal.mff.cuni.cz

Machine translation of very closely related languages

Jan Hajič, Jan Hric, Vladislav Kuboň

Although the field of machine translation has a very long history, the number of really successful systems is not very impressive. Most of the funds invested into the development of various MT systems were wasted and did not stimulate the development of techniques which would allow to translate at least technical texts from a certain limited domain. There were of course exceptions, which demonstrated that in certain conditions it is possible to develop a system, which will save money and efforts invested into a human translation.

The main reason why the field of MT did not filled not only the expectations of sci-fi writers, but also the expectations of scientific community, is the complexity of the task itself. A successful automatic translation system requires the application of techniques from several areas of computational linguistics (morphology, syntax, semantics, discourse analysis etc.) as a necessary, but not sufficient condition.

The general opinion is that it is easier to create an MT system for a pair of related languages. In the Czech-to-Russian MT system RUSLAN, which has been developed at the Charles University in the second half of the last decade, we had a chance to test this hypothesis. It turned out that even these relatively closely related Slavic languages require full-fledged syntactic analysis of Czech and a complex transfer phase. On the other hand, the fact that both languages allow high degree of word-order freedom accounted for certain simplification of the translation process.

In the group of Slavic languages there are more closely related languages than Czech and

Russian. Apart from the pair of Serbian and Croatian languages, which are almost identical and were considered one language few years ago, the most closely related languages in this group are Czech and Slovak language.

The recently developed system of an automatic translation between Czech and Slovak ČESÍLKO aims at the exploitation of the similarity of both languages to the full extent and to create a system as simple as possible. The system uses the method of word-for-word translation, which allows omitting a syntactic analysis of input sentences. The hypothesis justifying the use of this method says that the involvement of a syntactic analysis (for example the syntactic analysis of Czech used in the system RUSLAN) will have a negative influence on the quality of results due to the complexity of the task of parsing free-word-order language. On the other hand, the similarity of syntactic constructions of both languages will allow a direct transfer of individual words.

The only substantial problem of the word-for-word translation approach is the difference in ambiguity of individual word forms. Without the analysis of nominal groups it is often very difficult to solve this problem, because for example the actual morphemic categories of adjectives are in Czech distinguishable only on the basis of gender, number and case agreement between an adjective and its governing noun. We have solved this problem by means of a stochastically based morphological disambiguator whose success ration is close to 94%. Our system therefore uses the following modules:

Morphological analysis of Czech

Morphological disambiguation

Domain-related bilingual glossaries

General bilingual dictionary

Morphological synthesis of Slovak

The success ratio of the translation (almost 90% of words are translated correctly) justifies the hypothesis that word-for-word translation might be a solution for MT of really closely related languages. The remaining problems to be solved are problems with the one-to many or many-to-many translation, where the lack of information in glossaries and dictionaries sometimes causes an unnecessary translation error.

The success of the system ČESÍLKO encouraged the investigation of the possibility to use the same method for other pairs of Slavic languages, namely for Czech-to-Polish translation. Although these languages are not as similar as Czech and Slovak, we hope that the addition of simple partial noun phrase parsing might provide results with the quality comparable to the full-fledged syntactic analysis based system RUSLAN.

Jan Hajič, Jan Hric, Vladislav Kuboň

ÚFAL MFF UK
Malostranské náměstí 25
CZ-11800 Praha
Czech Republic

hajic@ufal.mff.cuni.cz

============

Session 5: Tools for the translator

Chair: *Colin Brace*

Tools for the CEEC languages, an overview

Poul Andersen

When the European Union hopefully is enlarged with ten CEEC countries in a few years, 10 new languages (Estonian, Latvian, Lithuanian, Polish, Czech, Slovak, Hungarian, Slovenian, Romanian and Bulgarian) are expected to be added to the current 11 official EU languages (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish).

As each of the 21 languages can be translated into the 20 other languages, this results in 21 x 20 = 420 language combinations. On this background, technological means to facilitate the translation process become even more important, particularly as most of the new languages are little known or understood outside their respective countries.

The present overview is based on the author's experience and contacts from coordinating cooperative research projects in the area of Language Engineering with partners from EU and CEEC during the last five years. It may be of a certain interest to the professional translator who is looking for tools to facilitate his work, but the main 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 CEEC, it might be useful to distinguish between two typical situations :

1.1 Persons in CEEC 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, and they may be helped e.g. with access to an on-line dictionary, which allows them 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 confronted with documents written in a CEEC language. These persons will often not have any knowledge at all of the CEEC languages, most of which belong to other language families than the more familiar EU languages, so that it is not even possible to „guess" the approximate contents through similar words.

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 might be interested in setting up projects for development of Machine Translation or more restricted Translation Tools between CEEC and EU languages.

Examples of EU activities in this area are recently started or planned projects within the MLIS programme (Multilingual Information Society) on development of MT between less widely spoken EU languages, and 'major' EU languages.

A concrete opportunity for joint EU-CEEC teams to submit such proposals are the HLT Action Lines (Human Language Technologies) in the IST Call for Proposals under the 5th Framework Programme (cf. separate presentation at the EAMT Workshop).

Poul Andersen

European Commission

Automatic Translation Lexicon Extraction from English-Czech Parallel Texts

Martin Čmejrek, Jan Cuřín

The primary motivation for our research was to create translation lexicon of terminology of a particular disciplin. Many disciplines lack relevant dictionaries or the dictionaries are obsolete because of the quick development of the discipline. The terminology in target language, in our case Czech, is generated spontaneously, unmethodically. The idea was that the fundamental part of the translation lexicon would be generated from the parallel corpora of up to now translated texts automatically and afterwards it could be manually edited.

We decided to use statistical models of translating between English and Czech. The secondary aim was to use these models for machine translation. This however will not be the topic of my present talk.

Our paper has five parts. In the first part we will describe the material from which the translation lexicon was extracted, i.e. the English-Czech corpus. We used two different bilingual corpora, a) Computer oriented corpus (119,886 sentence pairs) and b) Journalistic corpus (58,137 sentence pairs).

The second part is devoted to the statistical alignment of corresponding paragraphs and sentences in these corpora. We have used the probabilistic model of sentence alignment by Gale and Church (1991).

In the third part we will focus on the noun phrases. The noun phrases in the aligned pairs of sentences were marked using a regular grammar based tools.

This output was used for the training of the statistical model of translation by Brown et al. (1993). We will deal with this model in the fourth part.

The resulting translation lexicon was afterwards processed by several automatic filters. The automatic filters will be shown in the part 5.

Resulting dictionaries' size varies around 6,000 entries. After significance filtering, weighted precision is 86.4% for Computer oriented English-Czech dictionary and 70.7 % for Journalistic English-Czech dictionary.

Martin Čmejrek, Jan Cuřín

ÚFAL MFF UK
Malostranské náměstí 25
CZ-11800 Praha
Czech Republic¨

=================

Session 6: EU support for research and development

Chair: *Anthony Clarke*

Human Language Technologies - possibilities in the EU 5th Framework Programme for Research and Technological Development

Bente Maegaard

The European Commission has launched its new Framework Programme, the Fifth Framework Programme for Research and Technological Development. One of the thematic programmes is the IST programme, 'Creating the user-friendly Information Society'. This programme has a key action called Multimedia contents and tools, under which we find *Human Language Technologies* (HLT). HLT has two active action lines for the call for proposals which opened in March: *Multilinguality in digital content and services*, and *Natural interactivity*. The closing date for the current call is June 16. The presentation will describe the possibilities in this first call, with respect to content, as well as project types and consortia. More information can be found at http://www.linglink.lu/hlt.

Bente Maegaard holds a M.Sc. in Mathematics and French from the University of Copenhagen, 1970. She was employed at the University of Copenhagen, Department of Applied and Mathematical Linguistics, 1971-90, being a research professor 1984-89, and visiting professor at the University of Geneva 1981. She is currently director of the Center for Sprogteknologi (Centre for Language Technology) since its creation 1991. Her main areas of expertise are machine translation, evaluation methodology, dictionaries, corpora. She has held and holds positions as officer in scientific and other associations, member of editorial boards, reviewer for journals and conferences.

She is the coordinator of the EU project EUROMAP the goal of which is to promote language technology and HLT in the 5th FP.

Center for Sprogteknologi

The Center for Sprogteknologi, CST, is a Government Research Institute under the Danish Ministry of Research and Information Technology. The Centre was established in 1991 with the purpose of promoting research and development in computational linguistics and language technology. CST has some 20 employees with expertise in machine translation, general and computational linguistics, computational lexicography, computer science and Danish and a number of other languages. The Centre participates in European and national research programmes, and performs commercial development and consultancy under contracts with Danish as well as foreign companies.

Bente Maegaard

Director, Center for Sprogteknologi

Njalsgade 80

DK-2300 Copenhagen S, Denmark

Tel: +45 35 32 90 74, Fax: +45 35 32 90 89

E-mail: bente@cst.ku.dk

Summary and conclusions

Dimitri Theologitis

==

Post-Workshop Commentary by Vladislav Kubon

New languages are not virgin languages: EAMT’99 Workshop from the "eastern" point of view

The fourth annual workshop of the European Association of Machine Translation was held in Prague in the second half of April. The location of the workshop corresponded with its main topic – the accession of Central and East European Countries into the European Union from the point of view of translation problems. The prospect of the increase of the current 11 official languages to potentially 22 after the accession of all candidate countries will not only mean that the number of language pairs will grow from currently 110 to potentially 462, but it will probably also increase the demand for the whole range of translation support tools, from databases of terms to full-fledged machine translation systems.

The program of the workshop was designed with the intention to map the "possibilities, policies and practicalities" of translation. It ranged from talks of officials both from the European Union and from the Czech government through presentations of scientific results, demonstrations of commercial systems and presentations of industrial companies active in the field of MT.

The workshop had some very positive aspects. The composition of participants (not only presenters and demonstrators) – a mixture of people from the industry, from governmental institutions and from the research area - indicated that MT is still one of the fields of computational linguistics which has the potential to bridge the industry-research gap. There was an even more important aspect to be noticed – the industrial participants did not represent only the developers of tools and systems, but also potential users, who came with the intention to find out what the developers and researchers can provide. The fact that at least some users feel the need to participate in such a workshop and that they actively seek help with an enormous task of translation and localization is quite important. On the one hand it means that there is a good market potential for MT tools and systems, and, on the other hand, it also means that commercially available systems do not provide the help customers expect to get. That could have a positive influence on application oriented research of MT systems in the near future.

The positive aspects of the workshop prevailed from the general point of view. However, from the point of view of researchers investigating the so-called "new" languages, not all impressions from the workshop were entirely positive. During the past decade we had time enough to get accustomed to the European version of market economy. We understand that there is a long journey from scientific research towards a commercial application, that the development of an M(A)T system is very costly and no commercial company is going to invest into such a venture if the market for the product is not big enough. That is quite clear. It is quite clear to us, but it seems that some companies work according to different rules. They do, otherwise Lernout&Hauspie would not invest into Albanian prior to investing for example into Czech, Slovak or Bulgarian. They really do, otherwise they would not try to carry out all the development of systems for "new" languages (for those they are above the threshold of minimal size of the market) from scratch and would try to adapt already existing linguistic resources. As prof. Elena Paskaleva from Bulgarian Academy of Sciences has pointed out: "New languges are not virgin languages." In several candidate countries there are numerous linguistic resources which might be useful for any kind of translation software. They represent many man/years of work and would provide a good basis for a successful commercial application. The example of the two successful companies specialized in NLP (Morphologic, Hungary) and MT (ProMT, Russia) shows that both the developers and resources available in these countries are comparable to those from "old" countries and that the arrogance of some companies is really not wise. The results presented during the workshop had shown that ProMT is a serious competitor to any company well-established on the M(A)T market, while Morphologic represents an example of a company founded by former researchers who are commercially exploiting the results of their previous linguistic research.

There is one more point worth mentioning. The languages of "new" countries belong to language types different from the current official languages of EU and they constitute a slightly different type of challenge for the issue of automatic translation. The attitude of western companies "Our experts can handle everything and our methods are good enough" (presented not only during the panel discussion) may at the end initiate the foundation of more Morphologics and ProMT's in the "new" counties. That would finally be a very positive outcome of the "Let's cooperate, but ..."attitude.

Vladislav Kubon

UFAL MFF UK

Malostranske nam.25

CZ-11800 Praha 1

Czech Republic

vk@ufal.ms.mff.cuni.cz