# Crosslingual Language Technology

# in service of an integrated multilingual Europe

## - 20 Years on -

The political process of European integration, which led to a rapidly growing European Union membership since the early 1990ies, had not only a strong political, economic and social impact , but also created new challenges for research and education. Multilinguality became a key issue amplified further by the introduction of Web-based applications into almost all sectors of life.

In the area of language technology a gap between eastern and western European languages became obvious, which had to be filled by appropriate research programmes.

Due to long-term research efforts like EUROTRA, linguistic resources and tools for the initial 12 European-Union languages have been, at least in prototype versions, generally available. In sharp contrast, technological support for most of the Central and East-European languages was still at a very early state. Moreover, the new member states introduced completely new language families (Slavic, Baltic) with their specific linguistic problems. In other cases, languages like Romanian and Hungarian, although belonging to the Romance and Finno-Ugric languages, borrowed so much from their geographic neighbourhood that special solutions had to be developed.

Multilingual language technology evolved over the last decades from an academic field of research into an indispensable technology for powerful applications in economy, social-media, European law, cultural preservation and education.

While at the beginning of the 90ties only some pioneering programmes like DBR-MAT, MULTEXT-EAST were prepared to deal with the new languages, nowadays, there is a concentrated effort

through a unified development of tools and resources all over Europe. Large Research and Infrastructure Networks like META-NET, FLARE-NET, CLARIN, enable researchers from all European countries not only to participate in this scientific endeavour, but provide them with an opportunity to shape the future of the field on a long term.

Despite the enormous progress made in the sector of multilingual language technology, a great number of unsolved key issues remains. Cross-lingual applications, in particular, which try to combine information from documents in different languages, are still quite rare and need to be improved considerably. Not only do they lack an appropriate linguistic foundation, which would require a systematic contrastive study of the different language pairs. They also suffer from the difficulty to port the existing tools and resources across languages and domains.

The aim of the current event is be to bring together researcher from all over Europe, strongly involved in the development of multilingual applications and to explore the steps and actions to be taken in order to ensure the development not only of multilingual platforms but also of real cross-lingual applications.

We are happy to organise this conference with the occasion of the 70$^{th}$ birthday of our colleague, Prof. Dr. Walther v. Hahn, who leaded since the 90ties some of the very first projects in machine translation and multilinguality, which involved languages from central and eastern Europe

Hamburg 04.05.2012 Cristina Vertan & Wolfgang Menzel

## FRIDAY 04.05.2012

## Language technology at the crossroads

Kimmo Rossi
European Commission
Unit E1 of DG INFSO (from 1/7/2012: Unit G3 of DG CONNECT)

In late 2011, the European Commission put forward legislative proposals for new funding programmes for research, innovation and infrastructures - Horizon 2020 (H2020) and Connecting Europe Facility (CEF). The proposals are currently under discussion with stakeholders and decision-makers, most importantly the EU member states and the European Parliament.

These programmes will cover the period 2014-2020, and they represent a significant change in the logic and structure of EU technology funding programmes. To better align the structure of the European Commission to the new tasks and to better respond to emerging policy priorities, a major reorganisation of the Directorate General "Information society and media" has just been announced. As of July 1st, DG INFSO will become DG CONNECT which stands for Directorate General "Communications Networks, Content and Technology".

The new setup offers big challenges and opportunities for language technologies, which will now be placed in the context of the *Media and Data Challenge.* One of the questions we need to ask ourselves in the coming months is this: in which ways can language technology contribute to the access, management, processing, analysis and re-use of online media, content and data - both structured and unstructured. And, in a wider perspective, how does all this contribute to the creation of a vibrant digital single market, providing more jobs, growth and prosperity for Europe. We need a fresh and updated vision, strategy and roadmap for the language technology field. It will be crucial input to the ongoing process of defining the

implementation details and priorities of the new funding programmes.

Marking the last year of the current RTD framework programme (FP7) and preparing the changeover to H2020 and CEF, the FP7 Work Programme 2013 and the related calls for proposals will be announced in summer 2012. This will offer not only a funding opportunity, but a chance for the language technology constituency to contribute to shaping up the future.

http://ec.europa.eu/dgs/information_society/connect_en.htm
http://ec.europa.eu/research/horizon2020/
http://ec.europa.eu/budget/refonn/documents/com2011__0657__en.pdf
http://cordis.europa.eu/fp7/ict/language-technologies/

# Section 1: Multilinguality and Crosslinguality in Europe

## Language Resources and Evaluation for a Multilingual Europe

Joseph Mariani
LIMSI-CNRS
&
Director, Institute for Multilingual and Multimedia Information (IMMI)
*Joseph.Mariani@limsi.fr*

Since the divine punishment of Babel, mankind must live with the wealth of a multitude of languages and cultures. The difficulty and cost of sharing information and communicating despite the language barriers, while preserving these languages, could benefit from the support of automatic language processing systems (or Language Technologies (LT)). At the European level, the needs related to multilingualism are very numerous, and the same is true at the world level. Add to this picture the many needs related to the accessibility of information by the visually or hearing impaired, requiring the translation of information from one medium to another one, and more generally to the access to information for people who do not speak fluently the language in which it has been encoded, including, notably, migrants. The extent of these needs shows very well that existing or even future human resources of the professions dealing with language processing cannot cover them all.

Taking into account multilingualism is not a top priority in any economic sector. However, if we add up the small priorities in each area where it is necessary to take it into account, then we reach a very large sum. This therefore requires, in our opinion, some thoughts and political actions to bring out this awareness and provide appropriate responses Even when multilingualism is seen as a necessity, its cost is still very important. It is this gap that calls for

the development of LT and their utilisation when their performance is up to the needs of the targeted applications. Although numerous applications resulting from those technologies are now in everyday use, it should be noted that LT have not yet reached maturity for all languages, with strong imbalances among languages, and that they cannot replace humans, in activities such as translation, for example. LT can be monolingual or crosslingual. They cover the processing of written, spoken and signed languages. They can be crossmedia, ie translating from one medium to another, with applications to enable accessibility for the disabled.

It is crucial for conducting-research aimed at developing LT to provide a base that includes both language resources (LR) and evaluation methods for the technologies that are developed. LR are both necessary for conducting research investigations in linguistics and for training automatic language processing systems that are based on statistical methods. The larger the data, the better the model and therefore the higher the system performances. The interoperability of language resources needs standards in order to organise, browse, and transmit data. It is also necessary to have a means for evaluating these technologies in order to compare the performance of systems, using a common protocol with common test data, in the context of evaluation campaigns. This allows for comparing different approaches and having an indicator of the advances of a technology and its adequacy to meet the needs of an application. There is currently a two-speed situation and a "digital divide" between languages for which LT and LR exist, and others.

Different initiatives can be identified aiming at producing the LR and LT that are needed to address multilingualism: the activities of big companies, the national programs in some countries and those of the European Commission (EC), the international efforts to network the actors of the field, to better coordinate activities and to promote greater sharing of resources, and the establishment of LR distribution agencies. In the European Union, if one considers the number of languages or language pairs that are to be addressed, and multiplies it by the number of technologies, we see that the size of the effort is probably too large for the EC alone. It would therefore be interesting

to share this effort among Member States (MS), or regions, and the EC, in perfect harmony with the "principle of subsidiarity". LT are well suited for a joint effort. The EC would have the primary responsibility for overseeing and ensuring coordination of the programme and of developing and assessing core LT. Each MS or region would have as a first priority to ensure the coverage of its language(s): to produce the LR essential for the development of systems and to develop or adapt LT to the specificities of its language(s). This model would be easily adaptable to an international framework, combining the efforts of the participating countries and of international organizations.

## How can we shape a Linguistic Schengen Area to take away the language barriers in Europe

Steven Krauwer,
University of Utrecht, the Netherlands,
&
CLARIN
*s.krauwer@uu.nl*

In my talk I will try to explain the big vision that drives CLARIN: the creation of a Schengen area for the European Humanities and Social Sciences scholar, where existing barriers become invisible, and scholars get access to language data and state of the art tools across national, discipline, language and other frontiers right from behind his or her desk. I will show which barriers we feel are hard, which ones are easy, where we feel we can move forward and where we may have to compromise. I will pay special attention to the multi- and cross-lingual dimension.

## Cross-lingual Technology: European Needs and Chances

Hans Uszkoreit,
DFKI, Germany,
&
METANET
*uszkoreit@dfki.de*

Recent advances in machine translation have changed the way people and organizations are dealing with language barriers. Types of information get translated today that had rarely been translated before and large numbers of people use the technology who would never pay for any professional translation. On the other hand, almost none of the text types that have been regularly translated by human translators are now being automatically translated by machines.

Europe is the region in the world with the greatest needs for machine translation and other cross lingual technologies. At the same time, our continent is also the region with the greatest potential for true progress in quality machine translation.

In my talk I will explain the special situation of the European society with respect to language technologies and outline both demands and opportunities. I will summarize the central findings of META-NET concerning the status of existing core technologies for European languages and then present recommendations of the META Technology Council on priority themes for the next few years of research.

Finally, I will argue that the special technology needs of our multilingual society and the massive research efforts proposed for filling these needs also demand a new research paradigm.

This research paradigm has three central components:

1. Involvement of translation professionals in research as providers of data, insights and evaluation judgements but also as the second central component in a high quality translation process responsible for creativity and knowledge

2. A multidimensional quality metrics for quality assessment of both human and machine translation.

3. The systematic and analytical concentration on near misses and on empirically determined quality barriers.

## MULTISAUND -Bridging Multilingual Technology

Alper Kanak,
TUBITAK, Turkey
*alperkanak@uekae. tubitak.gov.tr*

Turkey, one of the G20 countries known as bridging continents, play a critical role in cultural diffusion, socio-economic relations and knowledge transfer with a special focus on language-related developments. There exist more than fifteen active languages spoken in Turkey inspiring from each other, i.e. Turkic languages, Persian, Kurdish, Arabic, Bulgarian and Greek. Within all these, Turkish is widely spoken by more than 100 million natives and there is an increasing interest on learning Turkish. This trend has exploded the demand on advanced language technologies bridging Turkish with other languages. Correspondingly, the framework strategy of European Commission for multilingualism promotes multilingual technologies interoperating within well-studied, widely-spoken and less-studied languages that can increase the scientific and technological impact.

MULTISAUND (MULTilingualism Integrated to Speech and Audio UNDerstanding), an FP7 capacity building project, reflects the European needs to make Turkish a part of the European language technologies family. MULTISAUND aims to increase the research

potential of TUBITAK to produce concrete technologies complementing the Europe's multilingualism framework and related research strategies in language technologies domain. This talk will give a brief overview of the current state of the play in Turkish-specific research activities in Turkey, MULTISAUND activities to increase awareness and further opportunities to make Turkish a real living language in Europe as it has already been spoken more than 6 millions of people in European countries.

## Section *2.3:* Machine Translation

# Historical survey of machine translation in Eastern and Central Europe

John Hutchins
EAMT
*wjh@hutchinsweb,me.uk*

The first half concentrates on machine translation (MT) in Russia between 1954 and 1990. The first Russian research groups were formed in Moscow and Leningrad during the 1950s, at the Institute for Precision Mechanics and Computer Technology (Panov, Bel'skaya), at the Steklov Mathematical Institute (Kulagina, Mel'chuk), at the Institute for Foreign Languages (Rozentsvejg), and at Leningrad University (Andreev, Piotrovskij). These groups were all known outside the Soviet sphere, but there was also a less known group supported by the KGB (Motorin, Marchuk). In the early 1970s most of the Russian groups were brought together as the Ail-Union Translation Centre under Yurij Marchuk and later Ivan Oubine. Also at this time the Speech Statistics Group was founded in Minsk with branches in many republics. Whereas the focus of most Russian groups before the mid 1970s had been on theoretical studies - primarily because of the lack of computer facilities, after 1974 the focus shifted to the practical delivery of translations whatever the quality of the MT systems. Outside Russia, during the dominance of the USSR, there was significant MT activity in Czechoslovakia (Sgall, Kirschner), the German Democratic Republic (Agricola, Kunze), and Bulgaria (Ljudskanov).

After 1990, there emerged two commercial groups from the Leningrad centre: the STYLUS (later PROMT) systems by Svetlana Sokolova, and the PARS systems by Mikhael Blekhman. Research on MT and translation aids grew rapidly in the Czech Republic,

revived in Hungary, Poland, Bulgaria and Romania, and began for the first time in the Baltic states. Much of this research activity since 1990 has centred on statistical machine translation, but the older rule-based 'tradition' has continued (particularly for translation between closely related languages), and there has been considerable attention paid to the creation of language resources.

## Pragmatics as the Ultimate Challenge to High Quality MT

Dr. David Farwell,
Computing Research Laboratory, New Mexico State University,
Las Cruces, New Mexico
&
Catalan Institute for Research and Advanced Studies (ICREA) and
&
Centre for Speech and Language Technologies and Applications (TALP), Polytechnical University of Catalonia [UPC), Barcelona, Spain
*farwell@lsi.upc.edu*

The goal of this presentation is to demonstrate that dealing with pragmatic phenomena is essential for achieving high quality machine translation and to provide an outline of the core computational apparatus required. After briefly introducing what a pragmatics-based approach entails (e.g., modeling beliefs, ascription and reasoning), we motivate the need for such an approach by both briefly reviewing a number of traditional problems requiring a pragmatic solutions (e.g., reference resolution and the interpretation of ellipsis, metonymy and metaphor) but also by way of an in-depth analysis of a particular example in which we postulate three levels upon which a translation is based: locutionary, illocutionary and perlocutionary acts.

Next we present certain concepts related to a pragmatics-based approach to Machine Translation and informally sketch a

processing model. The concepts include notions of discourse context and utterance context which underlie a two-stage translation process that consists of:

-- first, a source language (SL) interpretation stage involving embedded belief spaces for the SL author about the context and about the SL addressees' beliefs about the context

-- and, second, a target language realization stage involving embedded belief spaces for the translator about the context and about the addressees of the translation beliefs about the context.

Given this overview of the process, we introduce, on the one hand, the notion of "user-friendly translation", i.e., customizing translation to the background knowledge and needs of the consumer of the translation, and, on the other, the notion of pragmatics-based translation equivalence.

Finally, we provide two further case studies which are best described by way of a pragmatics-based model. The first demonstrates that variations in what the translator sees as the general intent of the author of the SL text result in different patterns of lexical selection in the resultant translations. The second demonstrates that variations in the translators' higher-level beliefs about the world lead to different patterns of lexical selection in the resultant translations.

## Recent Improvements in Turkish-English Machine Translation

İlknur DURGAR EL-KAHLOUT,
TUBITAK, Turkey,
*idurgar@uekae.tubitak.gov.tr*

Automatic translation of one natural language into another natural language is called as machine translation. Dating back almost sixty years, machine translation is one of the major, oldest, and most

active areas in natural language processing. The last decade and a half have seen the rise of statistical approaches to address the machine translation problem. Statistical approaches automatically learn translation parameters from bilingual and monolingual texts and thus eliminate the labor intensive rule writing done in previous rule-based approaches.

Although statistical machine translation (SMT) has been well-studied for some language pairs, there has been comparatively little research for the Turkish-English language pair. In this talk, we present the results of our investigation and development of a state-of-the-art SMT prototype between Turkish and English. Developing a Turkish-English SMT system is an interesting and challenging problem due to several reasons. First and foremost, English and Turkish are typologically rather distant languages. English has very limited morphology and rather fixed Subject-Verb-Object (SVO) constituent order. However, Turkish is an agglutinative language with very flexible (but Subject-Object-Verb (SOV) dominant) constituent order, and a very rich and productive derivational and inflectional morphology where word structures might correspond to complete phrases of several words in English when translated.

In this talk; we will investigate i) how different morpheme-level representations on the Turkish side impact statistical translation results; ii) which statistical approaches are better suited to Turkish-English statistical machine translation; iii) a novel word alignment approach that outperforms the state-of-the-art EM method; and iv) Turkish local word ordering and question inversion to bring the word order of specific Turkish features in line with the corresponding English words in order to aid word alignment.

Talk will focus on the major challenges introduced by Turkish morphology, the suggested different morpheme-level segmentation methods in order to increase the Turkish-English SMT translation quality, the Turkish-specific preprocessing methods, the performance of hierarchical phrase-based (HPB) SMT on Turkish-English language pair, and our Bayesian-based word alignment

approach as an alternative to the widely-used EM estimation technique.

In our implemented prototype, we improved from 50.76 to 62.00 BLEU points in our HPB model corresponding to a relative improvement of 22.1%. Our state-of-art Turkish-English SMT system (to the best of our knowledge) is the only HPB system for this language pair and outperforms the best Turkish-English system in the IWSLT 2010 BTEC evaluation. Exploring the same techniques in the reverse translation direction (English-Turkish), we improved from 35.86 BLEU baseline to 40.59 BLEU corresponding to about 13.1% relative improvement.

## Machine Translation among related Slavic languages

Vladislav Kuboň,
Charles University Prague, Czech Republic
*vk@ufal.mff.cuni.cz*

The talk will provide an overview of the experiments concerning machine translation among various pairs of Slavic languages carried out at the Charles University in Prague. It will start with the oldest one, a Czech-to-Russian system RUSLAN. The system was rule-based, implemented in Colmerauer's Q-systems. It contained a full-fledged morphological and syntactic analysis of Czech, a transfer and a syntactic and morphological generation of Russian. The original assumption that the close relatedness of both languages would manifest itself especially in the simplification of the transfer phase turned out to be incorrect. The result of this experiment clearly indicated that the architecture of the system did not exploit the similarity of both languages to the desired extent.

The lessons learned from RUSLAN inspired a much simpler architecture of the Czech-to-Slovak system Česílko. It aimed at an exploitation of the similarity of both languages to much greater extent. The system used the method of direct word-for-word

(actually, lemma-for-lemma and tag-for-tag) translation, the use of which was justified by the similarity of syntactic constructions of both languages. The key role was played by a stochastic tagger which resolved the morphological ambiguity of the source language text. No syntactic rules were used for this language pair.

Next part of the talk will describe the development of the system through subsequent experiments which, in the course of the years, concerned the addition of a module of partial syntactic analysis and, finally, a modification of the architecture consisting of the removal of the tagger and the exploitation of the stochastic ranker.

The last part of the talk will address a still open issue of reasons why the quality of machine translation between certain language pairs of related languages provides much lower quality of results than other language pairs. A case-study of the translation problems between Czech and Russian will illustrate at least some aspects of the problem.

## Machine Translation in a content management system - a multilingual case study involving Polish, Bulgarian, Romanian, and Greek

Cristina Vertan
University of Hamburg, Germany
*cristina.vertan@uni-hamburg.de*

The integration of a machine translation service into a content management system is a very challenging task, as the service should be robust versus change of domain and text genre, and able to handle diachronic texts. Current state-of-art in machine translation relies on the other hand on domain-specific training data. Additionally these data are quite sparse when one deal with other language pairs than English-French, English-Arabic or English-Chinese. Within the ATLAS system we integrate a user-guided approach, which means

we make use of metadata extraction and text categorization technology, in order to select already trained models for a certain domain. If such model does not exist the user is informed that the translation may not be reliable. In this talk I will present the general architecture of the system, the main LT-modules, with particular focus on the machine translation engine. I will also introduce the set of experiments we performed in order to set-up the system architecture.

## Multilingual information management for special purposes

Susanne Jekat
Zurich University of Applied Sciences, Institute for Translation and Interpreting, Switzerland
*susanne.jekat@zhaw.ch*

My contribution presents a method of using free tools to draw on the WWW as the largest database (Corpas Pastor 2007). This method assists translators in working with specialised texts. An important part of the method is the compilation and analysis of target language corpora.

## Multilingual Aspects of Information Extraction from Medical Texts in Bulgarian

Galia Angelova and Svetla Boytcheva.
Bulgarian Academy of Sciences, Sofia, Bulgaria
*galia@lml.bas.bg, svetla.boytcheva@gmail.com*

This talk presents recent achievements in automatic processing of free texts in Bulgarian hospital discharge letters with focus on the semi-automatic construction of multilingual dictionaries of medical terminology. Today the international medical vocabulary contains many terms with Greco-Latin origins; in Bulgaria original Latin

terms still occur in hospital discharge letters. Terminology in English is also used, esp. generic drug names. In this way the analysis of medical texts is a challenging task which has to cope with a mixture of Bulgarian, Latin, and English terminology, given in Cyrillic and Latin alphabets. We note that Latin words and abbreviations occur quite often in some sections of the discharge letters, for instance more than 37% of the diseases in the section "Diagnoses" are described using Latin terms.

Some specific techniques have been developed for the semi-automatic extraction of Bulgarian-Latin medical terminology (given that no aligned corpora and electronic resources for Latin and Bulgarian are available). English is used as an intermediate language that enables to establish translation correspondences between Latin and Bulgarian medical terms.

The quality of the constructed bilingual dictionaries is evaluated via the automatic recognition of diagnoses and drugs in 1300 discharge letters. The extractors assign codes to the entities identified in the text: ICD-10 codes to the diagnoses (ICD-10 is the International Classification of Diseases, version 10) and ATC codes to the drug names (ATC is the International Anatomical Therapeutic Chemical Classification System). Diagnoses are recognised with F-measure 84.5% and drug names - with F-measure 98.42%.

## SATURDAY 05.05.2012

# Formal Models and Practice of Annotation

Eva Hajičová and Petr Sgall
Charles University, Prague, Czech Republic
*hajicova@ufal.mff.cuni.cz, sgall@ufal.mff.cuni.cz*

One of the aspects of "cross-linguality" is universality, which is a highly theoretical concept and one should distinguish between universality of functions and universality of means. Speaking about a cross-lingual language technology, two questions are at stake: (1) facing different *language data* do we need different *theories?,* and (2) should different *theories* lead to different *annotation schemes?*

In our contribution, we will concentrate on corpora and their annotation as one of the necessary resources for natural language technology. We will first point out what we mean by optimal qualities of language resources and we will consider the question whether schemes for corpora annotation can be "theory-neutral". Passing over to issues of language universals, we will claim that from the point of view of the function of language, one of the basic universals concerns the communicative function of language, namely to communicate something ABOUT something. In linguistics, this function has been treated in terms of the information structure of the sentence, the study of which is one of the main concerns of the Praguian approach. When evaluating different formal theories of language, it is therefore necessary to ask which formalism offers a suitable and at the same time a flexible framework for the description of this primary communicative function, in spite of the different means present in different languages.

In this connection, the existence of annotated parallel multilingual corpora is a very important resource of information for the build-up of cross-lingual language technology. One of the general

assumptions is that languages are closer to each other on the underlying (deep) level. This assumption is attractive both from theoretical aspects as well as from the point of view of possible applications in NLP. It was already present at the pioneering stages of machine translation research and application, when it was proposed that an underlying level of language description considered to be "common" (at least in some basic features) to several (even if typologically different) languages might serve as a kind of a "pivot" language (Bernard Vauquois). It is then an interesting task to design an annotation scheme by means of which parallel text corpora can be annotated in an identical or at least an easily comparable way. To illustrate this point we will present some experience from annotation of a parallel Czech-English corpus, namely the Prague English Dependency Treebank, which is an English counterpart of the Czech Prague Dependency Treebank. The aim of the project is to test a representation that would make it possible to capture synonymous constructions in a unified way (i.e. to assign them the same underlying representations, both in the same language and across languages) and to appropriately distinguish different meanings by the assignment of different underlying representations.

To sum up, (i) an application of a perspicuous theory is a help rather than an obstacle, (ii) annotated language resources are an irreplaceable resource of information for the build-up of grammars as well as of functioning NLP systems, and (iii) a consistently designed annotation scheme of the underlying structure of sentences is a precondition for the possibility to (carefully) apply it cross-linguistically. All this creates a good background (also) of cross-lingual language technology.

## Section 4 Language for Special Purposes

# Terminological Ontologies in Multi-lingual Cross-domain Communities of Practice

Gerhard Budin
University of Vienna; Austrian Academy of Sciences
*gerhard.budin@univie.ac.at*

This paper is a report on ongoing research projects that focus on building multilingual terminological ontology resources for the "multi-domains" of risk management and of public administration.

Terminological ontologies are derived from various sources, such as from terminologies modeled in structured vocabularies such as thesauri, classification systems, nomenclatures, and the like, from structured terminology databases and from extraction from text corpora and other forms of data mining and data analytics. Various methods and formats have been proposed for these processes of ontology building. Multilinguality has been a special problem for ontologies (in particular in domains where cross-lingual term equivalence is often not given) and various data models have been proposed for this. Communities of practice (CoP), in particular in so-called "multi-domains", face the challenge of cross-disciplinary understanding when different domain knowledge models and domain terminologies meet each other in discourse. The semantic dynamics of social meaning construction is then mirrored in texts that are resulting from collaborative work in such CoP, which in turn is a problem to be solved in corpus-based ontology learning.

## Arts and Sciences ?? New ways of thinking: Where are they born?

Prof. em. Christer Laurén
University of Vaasa, Finland
*chl@uwasa.fi*

1. The relation between Arts and Sciences will be discussed from the point of view of innovation in thinking. Do innovative thinking start in domains of Arts or in domains of Sciences?
2. Contrasting two sociological idiolects: One of them characterized by avoidance of the usage of terms, and the other one by traditional LSP of sociology.
3. Platon and the LSP of Philosophy
4. ?Oriental pearls at random strung? in the Arabic and Persian Literature:
5. Polythematic Texts in LSP and in Literature.
6. Realism and Literature in a long historical perspective
7. Cf. Balzac and Zola; cf. Linné and Darwin
8. Modernism and theories of modern physics

## Integration and harmonisation of multilingual terminological data sets

Chiocchetti Elena
EURAC-Bolzano, Italy
*Elena.Chiocchetti@eurac.edu*

Many terminology databases in the domain of law have been created to support communication in multilingual contexts, favour a consistent use of terminology and/or contribute to a better understanding of different legal systems. However, since every terminology database has its own peculiar history, objective and target users, the data models and entry structures may vary, even between term banks that present similar content. The practical consequence of this lack of homogeneity is that it is difficult to

handle similar entries from different source repositories. Which entries can be merged, which should be kept or eliminated? These decisions can often only be taken by human experts in a time-consuming manual check of similar entries or doublettes. The contribution briefly presents the approach proposed by the LISE project, which is financed by the European Union through the ICT-PSP programme with the aim of supporting collaborative and interinstitutional terminology work. LISE aims at creating a collaborative platform integrating three different tools, one for terminology extraction based on translation memories (Fillup), one for merging and harmonising term collections (OMEO) and one for cleaning databases (Cleanup). They purpose of the tools is to perform some of the tasks usually done by terminologists, such as looking for and eliminating double entries, in a semi-automatic way, thus sparing the staff a lot of time that can be dedicated to other tasks.

## Sections 5.6.7: Language Resources and Tools in a Crosslingual environment

# Language resources and tools for semantically enhanced processing of Slovene

Darja Fišer,
University of Ljubljana, Slovenia
*Darja.Fiser@ff.uni-lj.si*

In this talk I will present language resources and tools that were developed for semantically enhanced processing of Slovene. I will start with a description of the approaches used to automatically develop Slovene wordnet and a semantically annotated corpus. The Slovene wordnet is based on Princeton WordNet and was created by recycling several existing bilingual and multilingual resources, such as bilingual dictionaries, parallel corpora and Wikipedia. Probabilistic methods, by training a maximum-entropy classifier, and distributional semantics information, collected from a reference corpus, were also used to automatically extend the core wordnet and identify the noisy literals in the generated synsets. The corpus, on the other hand, was annotated manually by trained annotators who selected the best sense for each occurrence of 100 most frequent nouns in the josl00k corpus, resulting in slightly more than 5,000 annotated tokens.
In the second part of my talk I will focus on the tools and applications that were derived from the developed resources. sloWTool is a wordnet browser, editor and visualizer that can be used to study the lexico-semantic properties of Slovene in comparison to other languages. I will conclude the talk with the results of a case-study in which wordnet was used to improve the machine translation of polysemous words on the lexical level.

# Multilingual Resources and their Application for the Lithuanian Language

Andrius Utka,
Vytautas Magnus University, Kaunas, Lithuania
*a.utka@hmf.vdu.lt*

The presentation will overview the situation of multilingual language resources in Lithuania, enumerating existing multilingual corpora, machine translation systems, and related multilingual tools.

As less-resourced languages often meet certain restrictions absent in dominant languages, the presentation will focus on some of these restrictions, especially presenting certain design problems of parallel corpora. The limited availability of translated material for "less-resourced-to-dominant" and "less-resourced-to-less-resourced" language pairs makes it difficult to meet all necessary requirements that are valid for "dominant-to-dominant" and "dominant-to-less-resourced" language pairs.

This causes a number of difficulties for corpus creators who have to overcome such problems as the small size of the corpus and the retention of balance. Acknowledging that universal requirements for corpus representativeness do not always work in such projects, we will discuss some different strategies that may be used in this respect.

As most observations and ideas in the presentation come from two recent dissertations "A corpus-based approach towards translation of author-specific neologisms in Lithuanian-English Corpus of Prose" and "Grammatical multi-word lexemes in Lithuanian: evidence for description from the bilingual corpus", they will also be presented.

# Recent Advances in the Development and Sharing of Language Resources and Tools for Latvian

Andrejs Vasiļjevs, Tatiana Gornostay, Inguna Skadiņa, Daiga Deksne, Raivis Skadiņš and Mārcis Pinnis
TILDE, Letonia
*Andrejs@Tilde.lv, tatiana.gornostay@tilde.lv, Inguna.Skadina@tilde.lv, Raivis.Skadins@Tilde.lv, Daiga.Deksne@Tilde.lv, marcis.pinnis@Tilde.lv*

This chapter presents an overview of recent advances in the development and sharing of language resources and tools for the Latvian language - the sole official language in the Republic of Latvia, an official language of the European Union, and one of the oldest European languages with about 1.5 million native speakers worldwide. Although there are only several active research and development institutions in the language technology field in Latvia, strong progress has been achieved in the creation of basic language resources and tools as well as advanced applications, such as machine translation.

The first section briefly describes linguistic and sociolinguistic characteristics of Latvian, the history of language technology for Latvian that can be traced back to the end of the 50's as well as national and EU cooperation activities in language technology for Latvian.

Terminology resources are among the most widely used language resources and the second section introduces the concept of terminology entry compounding that solves the problem of a unified representation of multiple potentially overlapping term entries that are present in the consolidation of a huge number of multilingual terminology sources.

The Latvian language is a synthetically inflected language and exhibits some specific linguistic characteristics such as rich morphology due to inflection, a rich set of derivational means,

relatively free word order with morphological means for marking syntactic relations, and others. The third section overviews the development of Latvian morphological tools and discusses approaches to morphological analysis and maximum entropy-based morphological tagging for Latvian, and the fourth section focuses on the applied grammar checking methods for the Latvian language.

In recent years, several machine translation systems have been built for Latvian and the fifth section reports on recent research in the combination of knowledge-based and data-driven approaches in machine translation, including factored models for statistical machine translation and application of spatial ontologies to improve the translation of toponyms.

The availability, sustainability and interoperability of language resources are important issues that recently have been discussed not only in the framework of the Latvian language, but at the European level and around the world.

The final section provides an overview of activities in Latvia to create an open infrastructure for distribution and sharing of language resources and tools.

Finally, we conclude that current development of language technology for Latvian has reached the level where it can be implemented in practical applications addressing the needs of large user groups in a variety of usage scenarios.

## Construction and exploitation of X-Serbian bitexts

Cvetana Krstev, Duško Vitas
University of Belgrade, Serbia
*cvetana@matf.bg.ac.rs, vitas@matf.bg.ac.rs*

In this presentation we give details about aligned corpora in which one language is Serbian developed at the University of Belgrade,

Faculty of Mathematics - their size, domain, format and number of languages involved and units of alignment.

Next we present several programming environments for exploitation of bitexts. Some of them were locally developed: *Acide,* a tool for production of aligned texts, and *LeXimir* and *Biblisa,* programming tools for search in aligned texts that rely on various lexical resources - first of them works off-line while the second is a web application. A corpus management and querying system IMS-CQP used for the web access to the Corpus of Contemporary Serbian (developed likewise at the University of Belgrade, Faculty of Mathematics) is also used to access a collection of bitexts on the Web.

The focus of our presentation will be the possibilities offered by Unitex, a corpus processing system that relies on rich lexical resources and uses finite-state methodology, to process aligned texts. Namely, instead of preprocessing texts with aim of inserting tags and annotations, Unitex can apply lexical resources - e-dictionaries and local grammars - to both texts in a bitext which enables sophisticated queries to be posed in both languages in order to produce aligned concordances.

## Language Technology Methods Inspired by an Agglutinative, Free-Phrase-Order Language,

Gábor Prószéky
Pázmány Péter Catholic University
&
Morphologic, Budapest Hungary
*proszeky@morphologic.hu*

A crucial requirement for doing machine translation, machine-aided translation, information retrieval or computational information processing of texts of agglutinative languages, in general, is an efficient computational morphology. Having done the morphological processing of the word-forms of sentences, the sentence structure

itself should be identified. This step is also common in all human languages. In case of agglutinative languages, like Hungarian, morphological processing provides a lot of information expressed by syntax in other, less-inflectional languages. Syntactic processing of highly inflectional languages, therefore, can use a lot of syntactic information identified by the morphological subsystem. It is needed, because the agglutinative nature of morphology results in free phrase order on the sentence level. The first part of the presentation deals with Humor, a morphological description formalism and algorithm inspired by an agglutinative language, Hungarian. In the second part of the presentation a syntactic parsing formalism and algorithm will be described. The method has been inspired by Hungarian, a free-phrase-order language, and I show some examples how it is used in the Hungarian-English MetaMorpho machine translation system.

## Maltese: Mixed Language and Multilingual Technology

Mike Rosner and Jan Joachimsen,
University of Malta,
*mike.rosner@um.edu.mt, jan.joachimsen@um.edu.mt*

Maltese is the national language of Malta and an official language of the European Union. It is widely used throughout the Maltese archipelago by some four hundred thousand native speakers. There are also sizeable Maltese-speaking communities in Australia, Canada, the United Kingdom and the United States, bringing the total to number of speakers to around seven hundred thousand. As a national language, it is extensively used, in written and spoken form, both colloquially and in more formal discourse. The broadcast media are predominantly in Maltese. There are several Maltese newspapers and the literary and poetic aspects of the language are assured by a vibrant literary community.

Linguistically, Maltese is derived from Siculo-Arabic (the Arabic dialect that developed in Malta and Sicily between the ninth and the fourteenth centuries). One of the most evident characteristics of Maltese is its pervasively mixed nature. The lexicon includes a large

number of words of Semitic origin and the grammar includes characteristics and features which resemble those found in other Semitic languages. At the same time the lexicon has also borrowed extensively from Romance and, more recently, from English, and indeed, some of the grammatical features of both these language classes enter into Maltese grammar. Maltese is also the only Semitic language whose standard written form uses a modified Latin alphabet, including vowels.

The first part of this talk will concentrate on the language itself and will exemplify the main defining characteristics of the language as well as point out examples of the inherently "multilingual" nature of Maltese. It will also consider the unusual relationship between Maltese and English since it is only with respect to this relationship that an appropriate strategy for multilingualism can be developed.

Maltese has been something of a laggard when it comes to language technology, and the second part of the talk will give an overview of previous research and assess current and future developments with emphasis on multilingual aspects including machine translation and machine-aided translation.

## Multilingual Linguistic Workflows

Dan Cristea, Ionuţ Cristian Pistol
Faculty of Computer Science, University "Al. I. Cuza" of Iasi,
Romania
&
Institute for Computer Science, Romanian Academy, Iaşi,
Romania
*{dcristea,ipistol}@info.uaic.ro*

The field of Natural Language Processing (NLP) has seen important developments over the later years, most significantly in efforts intended to raise the quality and quantity of resources, to enhance the

performance and diversity of tools and to open accessibility to both resources and tools as largely as possible. The demands of multilinguality at the level of language technology impose the necessity of reusing the processing modules performing specific linguistic tasks for different languages. The language a module is able to interpret becomes thus commanded by the resources it is fuelled with. Such a view on interoperability requires a standardization of the processing steps and an efficient building and execution of workflows.

But the issue of language technology addressing the needs of multilinguality has a lot more facets than strict reusability of resources and tools, as for instance, the easiness of adopting resources in resource-poor languages from resource-rich languages, abstracting the ways in which language dependency issues are regarded in approaches involving language processing tasks, or a uniform way of looking at annotation schemas provided by different schools and overcoming language barriers.

Standardization of metadata formats and of the NLP software were, among others, the goals of projects such as CLARIN[1] and FLaReNET[2], as well as several national and international workshops and conferences. Meta-systems, capable of offering diverse users access to libraries of processing modules, as well as interfaces that help building complex processing architectures out of these modules, are two of the most desired behaviours in the NLP field. Systems, such as GATE[3] and UIMA[4] and research efforts such as PANACEA[5] and "Heart of Gold"[6] represent some of the most prominent efforts in this direction.

[1] http://www.clarin.eu/external/
[2] http://www.flarenet.eu/
[3] http://gate.ac.uk/
[4] http://uima.apache.org/
[5] http://www.panacea-lr.eu/en/
[6] http://www.delph-in.net/heartofgold/index.html

Almost all of the most influential NLP frameworks respond very well to requirements specific to different languages. To take just one well known example, UIMA is used as an integration and unifying framework in many multilingual projects. The project ATLAS[7], for instance, builds complex processing chains that perform translation and summarisation of documents in 7 languages: Bulgarian, Croatian, English, German, Greek, Polish and Romanian, and uses UIMA as a compatibility standard. Another project, METANET4U[8], among other things, updates, enhances and disseminates a large spectrum of language resources and tools in at least 6 languages: Catalan, English, Spanish, Maltese, Portuguese and Romanian and UIMA is also a central interest of the consortium. More and more resources in more and more languages are accumulated and/or advertised on big portals[9]. The more numerous these resources will be, the bigger the need for interconnectivity in complex multilingual applications.

ALPE (the Automated Linguistic Processing Environment) has been reported as being a format representation and processing environment which makes use of annotation schemas arranged hierarchically in a hyper-graph in order to automatically compute workflows out of a pool of processing components. The model, called the *Formats and Modules Hierarchy* (FMH) is designed to help users to build complex processing architectures, by involving minimum of expert skills.

Although rather well studied from different perspectives, the FMH model has potential as yet unexplored sufficiently in attaining the multilingual aspects of language processing. In this paper we describe the FMH model with a special emphasis on its capacity to deal with multilingual aspects of linguistic processing.

---

[7] PSP-ICT grant #250467, http://www.atlasproject.eu/atlas/project/en

[8] PSP-ICT grant #27089, http://www.meta-net.eu/projects/METANET4U/

[9] See, for instance, the META-SHARE (http://www.meta-share.eu) initiative of the META consortium.

## Bootstrapping NLP and MT Resources for under-resourced languages

Damir Cávar
Eastern Michigan University, USA
&
Institute of Croatian Language and Linguistics, Zagreb, Croatia
*dcavar@me.com*

## Polish Language Resources and Tools: Towards Multilinguality

Adam Przepiórkowski
Polish Academy of Sciences IPIPAN, Warsaw, Poland
*adamp@ipipan.waw.pl*

In this talk I'll give an overview of Polish linguistic resources and tools (LRTs): corpora, lexica (including two wordnets), morphological analysers, taggers, grammars and parsers, etc. I'll focus on recent work on multilingual LRTs, including parallel corpora, especially, on work carried out at the Polish Academy of Sciences on grammar induction from such parallel corpora.

## Language Technology for Portuguese: progress and prospects

Antonio Branco
University of Lisabon,
*Antonio.Branco@di.fc.ul.pt*

The research community working on the computational processing of the Portuguese language have been steadily developing and maintaining and set of core resources and tools for language

technology applied to this idiom. In this talk, I will start by providing and overview on this research landscape in preparation to focus on one of the latest undertakings, of a novel type and of special relevance for multilingualism. This is a cutting edge initiative by means which a deep linguistic treebank of Portuguese is being developed, containing linguistically principled, fully fledged grammatical representations, which is parallel and aligned to similar data banks for the English and Spanish languages.

## Section 8 Crosslingual APPLICATIONS

# Crosslingual search for assistive products

Gregor Thurmair,
Linguatec, Germany,
*g.thurmair@linguatec.de*

**Assistive technology** to support elderly people or people with disabilities is of growing importance. Public portals offering information on assistive technology are being set up in many European countries. Several national portals have joined into a **common European portal,** www.eastin.eu. providing information on more than 65,000 products, accessed by professionals and end users.

The EASTIN-CL project creates a language technology front-end to the EASTIN portal, to make interaction easier, by offering **Multilingual Support** (Users can forward natural language queries in their native language, and the documents found are automatically re-translated into their native language) as well as **Multimodal Support** (Users can interact both in written and in spoken mode).

The talk will cover the following aspects:

- Challenges in the collection of the **terminology** of the Assistive Technology domain, a structured hierarchy following the ISO9999 classification, and offering more than 12,000 terms in seven European languages.
- Design of the **query processing:** The setup of the AT portal provides specific requirements for multilingual search
- Adaptation and integration **of Machine Translation** components to retranslate the retrieved documents.
- Evaluation and **test** results in the portal tests

# Cross-lingual extraction of concepts and relations in the medical domain

Galia Angelova and Ivelina Nikolova
Bulgarian Academy of Sciences, Sofia, Bulgaria
*galia@lml.bas.bg, iva@lml.bas.bg*

To automatically analyse medical narratives, one needs linguistic and conceptual resources which support capturing of important information from texts and its representation in a structured way. Thus the conceptual structures encoding domain concepts and relations are crucial for the development of reliable and high-performance information extraction systems.

We present research work enabling automatic extraction of relations between medical concepts. The lack of conceptual resources with Bulgarian ontological vocabulary provoked us to reuse already existing resources with English labels, more especially the UMLS (the Unified Medical Language System) Metathesaurus. We form a terminological dictionary of the Bulgarian terms of interest, translate them to English and extract their UMLS definitions which are short English statements in free text. These definitions are processed automatically by a semantic parser; afterwards we apply additional extraction, alternation and validation rules and built a set of new relations to be inserted in our conceptual resource.

The talk presents the input data and available tools, the knowledge chunks extracted from UMLS and their processing, as well as a discussion of the present results. Experiments for automatic extraction of the IS-A and AFFECTS relation have been performed. The extraction of the IS-A relation was done with 81% precision, some 45% of the extracted IS-A relations are newly created. The extraction of the AFFECTS relation performed worse than the IS-A extraction because the variety of the corresponding expressions is much higher and the arguments might be positioned at a longer distance (or in adjacent sentences). These experiments prove the

readiness of the available NLP tools to serve for developing of new conceptual resources.

## Cross-lingual information access to multilingual documents -- an overview of JRC's media monitoring applications

Maud Ehrmann
JRC Lago Maggiore, Italy
*maud.ehrmann@jrc.ec.europa.eu*

At the Joint Research Centre (JRC), we have been using Language Technology since 1998 to fight the information overflow and to overcome the language barrier with the purpose of supporting the European Commission and Member State institutions. To this end, a number of text gathering (retrieval), analysis and visualisation tools have been developed with a focus on high multilinguality, on multilingual and multi-document information aggregation, and on tools to provide cross-lingual information access. These text analysis tools have been integrated with the news gathering engine Europe Media Monitor (EMM) to produce several complex applications. In this talk, I will stress the need for multilinguality in media monitoring and analysis, give an overview of the EMM family of applications and present the design principles and NLP technologies underlining our in-house developments.