Tools for the CEEC languages, an overview

Poul ANDERSEN
DG XIII/E/4 - Linguistic applications
SdT AGL/4 - Translation Service
e-mail: poul.andersen@lux.dg13.cec.be
Fax: +352 4301 34999
Tel: +352 4301 34324
EUROPEAN COMMISSION
Directorate-General XIII
Information Society:
Telecommunications, Markets, Technologies - Innovation and Exploitation of Research

Directorate XIII/E (Director: Frans DE BRUÏNE)
Information Society Technologies: Content, Multimedia Tools and Markets

Unit XIII/E/4 (Head of Unit: Roberto CENCIONI)
Linguistic applications, including the "Multilingualism" Programme
XIII/E/4 - Linguistic Programmes

Research & Development

5th Framework Programme 1999-2004
Thematic Programme:
Information Society Technologies (IST)
- incl. Human Language Technologies (HLT)

Candidate CEECs fully associated to programme

Awareness - Promotion

The Multilingual Information Society (MLIS) 1996-1999
CEEC participation possible, without EU funding
“Translation Tools for the CEEC candidates for EU membership - an Overview”

- *Terminologie et Traduction*, vol. 1. 1998
  (journal published by the European Commission’s Translation Service)
Translation tools for CEEC languages - *Users*

- **CEEC - users** with basic knowledge of English (or French or German) do not need full translation, but *dictionary look up* - ‘click on the word’

- **EU - users** have no knowledge of Slavic or Finno-Ugric (Estonian, Hungarian) languages, and a *primitive word-for-word translation* can tell them at least ‘what a text is about’
Commercial products:

SMEs (Small and Medium-Sized Enterprises):
- Filosoft (Estonia), Lex (Poland), Lingea (Czech Republic), Abakus (Slovakia), MorphoLogic (Hungary), Amebis (Slovenia)
- in cooperation with academic teams:
  - Tartu University (Filosoft), Poznan Univ. (Lex), Brno Univ. (Lingea), Budapest Univ. (Morphologic)
Translation tools for CEEC languages - Developers/2

Research prototypes

Academic teams (Universities and Academies of Science):
- in cooperation projects with EU research teams:
  INCO-COPERNICUS projects (EU funding):
  - GLOSSER (Estonia, Hungary, Bulgaria - with Groningen University and Xerox)
  - STEEL (Poland, Czech Republic - with Xerox and Tübingen University)

Commercial funding:
- DBR-MAT, funded by Volkswagen foundation (Bulgaria, Romania - with University of Hamburg)
Translation tools for CEEC languages - Information sources/1

General survey
The Language Engineering Directory - A resource guide to Language Engineering organisations, products and services.


http://www2.echo.lu/mlis/en/direct/home.html
**CEEC survey**

**ELSNET goes East** (1995-97) - a European Commission funded project in the field of Natural Language and Speech.

**Aim**: to build a research infrastructure in CEEC and the Newly Independent States of the former Soviet Union (C&EE/NIS), by extending a Western European network ELSNET.

**Coordination**: Institute for Logic, Language and Computation, University of Amsterdam.

**ELSNET Survey of Natural Language and Speech Organizations.**

http://www.elsnet.org/publications/survey
Bulgaria
• Dept. of Artificial Intelligence, Institute of Information Technologies - Bulg. Acad. of Sciences, Sofia, Bulgaria.

Czech Republic
• Dept.of Information Technologies, Faculty of Informatics, Masaryk University, BRNO.
• Department of Electronics and Signal Processing, SpeechLab, Technical University of Liberec, Liberec.
• Czech Language, University of Ostrava.
• Department of English and American Studies, University of Ostrava.
• Department of Computer Science, University of West Bohemia, Pilsen.
• Institute of Phonetics, Faculty of Arts, Charles University, Prague.
• The Institute of the Czech National Corpus, Faculty of Philosophy, Charles University, Prague.
Czech Republic (continued)

- Department of Circuit Theory, Speech Processing Group, Czech Technical University in Prague, Faculty of Electrical Engineering.
- Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Praha.

Estonia

- Laboratory of Phonetics and Speech Technology, Institute of Cybernetics, Tallinn.
- Filosoft Ltd., Tartu.

Hungary

- Department of Corpus Linguistics, Research Institute for Linguistics Hungarian Academy of Sciences, Budapest.
Latvia
• Department of Methodology of Latvian Language and Literature, University of Latvia, Riga.
• SIA Tilde, Riga.

Lithuania
• Center of Computational Linguistics, Vytautas Magnus University, Kaunas.
• Speech Research Laboratory, Kaunas University of Technology.
• Department of Dictionaries, The Institute of the Lithuanian Language, Vilnius.
• Department of General Linguistics, Vilnius University.
• Department of Lithuanian Linguistics, Vilnius Pedagogical University.
• Department of Phonoscopic Examination, Lithuanian Institute of Forensic Examination, Vilnius.
• Recognition Processes Department, Institute of Mathematics and Informatics, Vilnius.
Poland

- Department of Applied Informatics, Technical University of Gdansk.
- School of English, Adam Mickiewicz University, Poznan.
- Department of Transmissions Systems, Institute of Telecommunications, Warsaw University of Technology.
- Man-Machine Communication Group, Institute of Computer Science, Warsaw.
- Speech Acoustics Laboratory, Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw.
- Department of Processing and Analysis of Acoustic Signals, Wroclaw University of Technology.
Romania

- Computer Science, Faculty of Mathematics, Bucharest University.
- Electronics and Telecommunications Department, Applied Electronics Group, Bucharest.
- Faculty of Electrical Engineering, Technical University of Bucharest.
- Communications Department, Technical University of Cluj-Napoca.
- Department of Computer Science, Technical University of Cluj-Napoca.
- Natural Language and Multimedia Department, INTERDATA, Iasi.
- Natural Language Processing Department, Computer Science Institute, Romanian Academy, Iasi.
- Department of Computer Science, "Politehnica" University, Timisoara.
Slovak Republic

- Department of Speech Analysis and Synthesis, Institute of Control Theory and Robotics, Slovak Academy of Sciences, Bratislava.

Slovenia

- Department for Intelligent Systems, Jozef Stefan Institute, Ljubljana.
- Laboratory for Digital Signal Processing, University of Maribor, Faculty of Electrical Engineering and Computer Science.
Institute of Formal and Applied Linguistics

The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Arts and later at the Faculty of Mathematics and Physics. We ...

Faculty of Mathematics and Physics,
Charles University
Institute of Formal and Applied Linguistics
Malostranske nam. 25
Praha 1 118 35
Czech Republic
Phone: +420 2 2191 4288
Fax: +420 2 2191 4309
WWW: http:\ufal.ms.mff.cuni.cz
Email: korbay@ufal.ms.mff.cuni.cz
Contact Person: ing. Ivana Kruijff-Korbayova
Translation tools for CEEC languages -

**Information sources/3**

- **TELRI II** - Trans-European Language Resources Infrastructure - *Concerted Action 1998-2001*

  - partners from 9 EU and 15 CEEC/NIS
  - **TRACTOR** (TELRI Research Archive of Computational Tools and Resources:
    - Monolingual, bilingual, and multilingual corpora and lexica in i.a. Bulgarian, Czech, Dutch, English, Estonian, French, German, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Norwegian, Rumanian, Russian, Serbian, Slovenian, Spanish, Swedish, Turkish and Uzbek.
    - Corpus- and lexicon-related software (parsers, taggers, morphological analyzers, corpus retrieval tools, concordancers etc).
    - Individuals and academic, public or industrial organizations can join the **TRACTOR User Community (TUC)** for a nominal annual fee of 50 EURO (Western Europe) or 20 EURO (rest of Europe).

**Coordinator: Wolfgang TEUBERT / Ann LAWSON**

_Institut für deutsche Sprache, Mannheim - [http://www.telri.de](http://www.telri.de)_

---

DG XIII Information Society: Telecommunications, Markets, Technologies - Innovation and Exploitation of Research

Poul ANDERSEN

E-Mail: poul.andersen@lux.dg13.cec.be
Translation Experts Ltd. - products

- **Word Translator for Windows**
  
  *WordTran is an bi-directional electronic dictionary, spell-checker and word-for-word translator.*

- **InterTran™ - Internet/Intranet/Extranet Translator**

- **Neural Translator for Windows (not released yet)**
  
  *NeuroTran is a high technology software intended for natural language processing*
Translation Experts Ltd. - contacts

Creative Technology (Micro Design) Ltd.
Park House, Park Street, Uttoxeter
Staffordshire ST14 7AG                United Kingdom                .
Tel: +44-(0)1889-567-160               Fax: +44-(0)1889-563-548
Internet: UK@tranexp.com
Worldwide Web: http://www.translation-experts.co.uk

Translation Experts Croatia d.o.o.
Rusanova 14                10 000 Zagreb, Croatia                .
Tel/Fax: +385-1-221-980               Internet: Croatia@tranexp.com
Worldwide Web: http://tranexp.cro.net

+ distributors in Poland, Hungary a.o.
WordTran English-Russian dictionary:
Russian translations of one English verb, with Russian keyboard displayed, which makes it easy to use even for those who are not used to it

Source: http://main.amu.edu.pl/~sipkadan/wt.htm
InterTran™ /1

InterTran provides a browser-based interface to Translation Experts' translation dictionaries.

The system offers word, phrase and document translation from within email programs and web browsers.

The system allows you to add words and phrases to the translation dictionary to build a highly customised translation environment for your business.

Source: http://www.translation-experts.co.uk/in_info.htm
InterTran™ /2

English to
Russian (CP 1251)
Croatian, Hungarian, Polish, Czech, Slovenian (CP1250)
Serbian (Latin)

German to
Polish, Hungarian Croatian

Russian to
English, Hungarian

Croatian to
English, German, French, Italian

Hungarian to
English, German, Russian, Spanish, French

Polish to
English, German

Czech to
English

Serbian to
English

Slovenian to
English

French to
Croatian, Hungarian

Italian to
Croatian

Spanish to
Hungarian

Source for language list + demo:
http://www.tranexp.com/InterTran.cgi
**English original:**

Word Translator for Windows is an interactive translation and learning tool. It can help you to:

- Translate web pages written in East European languages such as Russian, Ukrainian, Polish, Czech, Slovak, Bulgarian, Romanian, Hungarian, Croatian, Serbian, Bosnian, Macedonian, Slovenian, Albanian, etc.
- Translate e-mail messages to/from a foreign language.
- Read/Write letters, facsimile, reports and memos in a foreign language.
Polish translation:

Word Translator dla Windows jest [an] wzajemnie oddziaływujący tłumaczenie
<ALT="Przekładać, Konwertować, Przemieszczać, Odkodować, {Translate}">
i wiedza narzędzie
<ALT="mechanizm, środek, instrument, program narzędziowy, {tool}">
• Ono puszka metalowa
<ALT="móc, umieć, potrafić, puszka blaszana, blaszanka, konserwować, możesz, mogą, może, puszka, puszkować, konewka, bańka, {can}">
  pomagać ty wobec:
• Tłumaczyć *kanina* urządzenia wzywające do telefonu pisemny w Wschód Europejski języki taki jak Rosjanin, Ukraiński, **Polski**, Czech, Słowak, Bułgar, Rzymski, Węgierski, Chorwat, Serbski, Bośniacki, Macedończyk, Słoweniec, Albański, [etc].
• Tłumaczyć przesyłać wiadomość pocztą elektroniczną wiadomości wobec/z pewien.
Czech translation:
Slovo Překladatel do Okna >> be neurč. člen dialogový dešifrování >> AND operation Účenost jet
<ALT="jet, náčiní, nářadí, nástroj, obdělávat, otesat dlátem, razidlo, užívat nástroje, výstroj, zdobit ražením, {tool}".}
• Ono hajzl
<ALT="konev, konzerva, konzervovat, mohu, nalít do konve, plechovka, {can}", pomoci tebe až k:
  Chápat tkanivo blok napsáný do Orient Evropan jazyk jako takový Rus, [Ukrainian], Hladit,
<ALT="Lesknout se, Naleštit, Polský, Polština, Stát se hladkým, Uhledit, Zdobit, Zjemnit, Zkrášlit, Zušlechtit, {Polish}".>
Čech, Slovák, Bulhar, Říma
Hungarian translation:

Word Translator Windows alá van egy egymásra kölcsönösen ható fordítás és tanulás szerszám. Tud segít ön -hoz:

- Lefordít pókháló oldalak írott -ban Kelet Európai nyelvek mint Orosz, Ukrán, Fényesít, Cseh, Szlovák, Bolgár, Románia, Magyar, Horvát, Szerb, Boszniai, Macedón, Szlovén, Albán, stb.
- Lefordít elektronikus posta üzenet -hoz/-ból egy idegen nyelv.
- Olvas/Ír irodalom, hasonmás, beszámol és memorandum -ban egy idegen nyelv.
Russian translation:
Слово Переводчик для Окно быть сильная форма, грамматически неопределенный член находиться во взаимодействии перевод и ученость рабочий. Он мочь помогать ты к:
• Переводить с одного языка на другой ткань паж
<ALT="мальчик, служитель в законодательном собрании, сопровождать в качестве пажа, вызывать кого-либо, громко выкликая фамилию, страница, нумеровать страницы, {pages}>"
• писать в Восток Европейский язык такой как Русский, Украинский, Полировать
NeuroTran/1

- is a piece of software intended to "do things with words" (…) :

- lookup a word and its L2 equivalent(s), with their respective grammar and usage labels,
- reproduction of the sound of a word,
- generation of all inflection forms of a word,
- lookup of all words with common part of speech, or some other grammatical feature
- lookup of all words with common subject-matter field or common usage features,
- lookup of synonyms and antonyms,
- sentence-to-sentence translation,
NeuroTran/2

- spell-checking,
- determining the type of text and choosing appropriate word/translation for it,
- parsing,
- qualitative (content) text analysis.

Languages available under NeuroTran:
English (in both L1 and L2 position) with Croatian, Czech, French, German, Polish, Russian, and Serbian (in both L1 and L2 position).
Languages to be included:
Danish, Swedish, Norwegian, Bulgarian, Old Church Slavonic, French, Portuguese, Italian, Hungarian.

Source: http://main.amu.edu.pl/~sipkadan/nt.htm
but try and find a child who does not wish to see the inner workings of his toys

Little Mark is running very fast towards his big house. Because he is a good boy, his mother has made him a delicious lunch. Like any child, Mark likes to play with his wonderful toys which are in his room. It is true that most of his toys do not last longer than a few months, but try and find a child who does not wish to see the inner workings of his toys. His grandfather and grandmother spoil him, all the time, by letting him get away with most things.

It is time for Mark to go to bed so that he can get up early tomorrow morning and go to

Mali Marko trči vrlo brzo k njegovoj velikoj kući. Zato jer on je dobar dječak, njegova majka je napravila njemu ukusan ručak. Kao svako dijete, Marko voli igrati se sa svojim divnim igračka koje su u njegovoj sobi. To je istina da većina od njegovog igračka ne traje duže nego nekoliko mjeseci,
NeuroTran shows the parse tree of the sentence that is currently being translated:

Little Mark is running very fast towards his big house

mali Marko trči vrlo brzo k njegovoj velikoj kući

Source: http://main.amu.edu.pl/~sipkadan/ntp.htm

NeuroTran contact:
Danko Sipka, Ph.D
Slavic Department
Adam Mickiewicz University, Poznan, Poland
sipkadan@hum.amu.edu.pl
NeXTMail: sipkadan@plpuam11.amu.edu.pl
CEEC country case 1: Latvia/1

Latvian situation: small country, one active centre, close cooperation with institutional users, strong language policy, orientation towards Scandinavia (Sweden, Denmark)

Contact:
Dr. Inguna Greitane
Artificial Intelligence Laboratory (Head: Andrejs Spektors)
Institute of Mathematics and Computer Science
University of Latvia
Raina bulv. 29
Riga LV1459 LATVIA

e-mail: inguna@ailab.mii.lu.lv
CEEC country case 1: Latvia/2

LATRA - development of automated translation tools for Latvian:

1. Work on bilingual (English, Latvian) text corpora
   Since there are several types of ambiguities which LATRA is unable to handle, we work on elaboration of LATRA and incorporation of statistical methods.

2. Development of terminology database - appr. 115,000 terms
   Project carried out for Translation and Terminology Center (TTC) and funded by TAIEX. Implementation with Trados Multiterm. The results are now available at www.ttc.lv.

3. UNL (Universal Networking Language) project of United Nations University: developed grammar rules for generation of Latvian sentences from UNL as well as lexicon for domain of soccer.
The marathon of the increase of statuary capital which the most of the commericals bank must begin characterised the beginning of the year also.
CEEC country case 2: Czech Republic/1

Czech situation: (relatively) large country, several centres, international orientation - contacts with many countries

Academic centres:

1 Charles University, Prague:

- Institute of Applied and Formal Linguistics, Faculty of Mathematics and Physics
  - English-Czech MT (since 1974), Contact person: Alexandr Rosen (rosen@ff.cuni.cz).
  - Czech-Russian MT (since 1985), Contact persons: Eva Hajičová (hajicova@ufal.mff.cuni.cz), Vladislav Kubon (vk@ufal.mff.cuni.cz).
CEEC country case 2: Czech Republic/2

1 Charles University, Prague (continued):

- Institute of Theoretical and Computational Linguistics, Faculty of Philosophy.
  Contact person: Vladimir Petkevicz

- Czech National Corpus, headed by
  Contact person: Frantisek Cermak

2 Masaryk University, Brno:

- Department of Information Technologies, Faculty of Informatics, Masaryk University, Brno
  Contact person: Karel Pala

In EU Telematics project EuroWordNet 2, producing a lexical database, Czech WordNet, to be linked to WordNets in English, German, French, Dutch, Italian, Spanish and Estonian.


CEEC country case 2: Czech Republic/3

Commercial products / 1

1 TRANSEN, from company POSY s r o.
- contains a knowledge base describing syntactic relations among words.
- can identify the subject and object of a sentence, and contains a rich system of morphemic classes, for the generation of morphologically correct text.
- can learn from sentences which it was not able to analyse.
- originally built for English-Slovak translations and later adapted for Czech.

2 PC Translator, vers. 9.0 from LangSoft & SOFTEX
- a bi-directional system Czech-English, translating online with manual entering of inflections. Smaller modules for French, German and Italian. This system uses simple morphological analysis and also simple NP analysis making it possible to put whole noun phrases in the target language (Czech) into the proper morphological form. The system supports domain-oriented translation. The main dictionary contains some idiomatic expressions.
CEEC country case 2: Czech Republic/4

Commercial products / 2

3 SiR Translator from SiR Software

- similar to PC Translator. It uses more complex lexico-syntactic data, which unfortunately are not complete (e.g. verb valency frames have only one slot). The quality of the translation is slightly better than in the previous case, but the syntactic analysis is only able to handle simple sentences.

Both systems have dictionaries containing about 200,000 word pairs for Czech and English and slightly fewer for other language pairs.

4 SKIK, version 2.0, from the company SKIK,

- a bi-directional system Czech-English, with a dictionary of 100,000 entries, running in batch mode, with automatic inflection. A German version is undergoing tests. The system is based on the translation of word chains (2 or 3 words).
CEECE country case 2: Czech Republic/5

Commercial products / 3

5 Landi Translator (developed by M K C S company)
- a bilingual English–Czech dictionary with automatic translation as an added feature. It translates word for word and does not use morphological analysis.
Company profile

**Lingea, Ltd.**

- a small software firm, run by a programmer, Dr Pavel Sevecek (pavel@lingea.cz);
- cooperation with Masaryk University in Brno.

- *first commercially successful translator expected by the end of 1999* - ???

- The following lingware products are currently sold by LINGEA Ltd:

  - *hyphenator* and *spelling checker* for Czech, Slovak, English and German
  - *lemmatiser* for Czech, Slovak, English and German - used at Masaryk University for tagging Czech corpus texts (in co-operation with the Czech National Corpus), - 165 000 Czech stems, covers about 97% of the corpus texts
  - *morphological analyser* for Czech, Slovak, English and German, - used in NLP research at Masaryk University within the currently built parser for Czech
  - *thesaurus* for Czech, Slovak, English and German
  - *English⇔Czech, German ⇔ Czech* translation *dictionaries*
CEEC country case 3: Hungary/1

Academic centre:
Lajos Kossuth University, Debrecen

- work over the past 6-7 years on a Hungarian parser with the intention of further developing it into an interactive Hungarian-to-English MT system, with financial support of OTKA, the Hungarian National Science and Research Fund.

- Contact person: Laszlo Hunyadi (hunyadi@llab2.arts.klte.hu)

- Based on a full morphological analysis, the rule-based parser gives the analysis of any arbitrary Hungarian simple sentence (with a modest dictionary of a few hundred words so far).

- Implementation in Pascal.
CEEC country case 3: Hungary/2

Lajos Kossuth University, Debrecen (continued)

- The output of the parsing of a simple sentence:
  - akarok kerni a lanyoktol egy eheto almat
  - I-want to-ask the girls-from a edible apple-acc.

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

Subject NP: en (*en* [“I”]) is deduced from the analysis.
Predicate VP: akar (ker (alma (e, egy) lany (a))).
Company profile

MorphoLogic, Budapest  
Director: Gabor Proszeky, PhD (proszeky@morphologic.hu),  

Cf. separate presentations by Proszeky (22.04.99, 11.20-11.50 & 13.00-13.40)

- Intelligent bilingual dictionaries with morphological analysers - MoBiDic
  - English↔Hungarian, German↔Hungarian, French↔Hungarian, Italian↔Hungarian, Russian↔Hungarian, Latin↔Hungarian, Polish↔Hungarian, Hungarian thesaurus

- Basic dictionaries with general vocabulary, dictionaries of phrases and of idioms, and dictionaries for specific subject fields such as business, auditing and computing.
CEEC country case 4: Romania/1

1 DBR-MAT - An Intelligent MAT System for Structurally Different Languages


- funded by Volkswagen Foundation

- aim: “Investigation and pilot implementation of a MAT system combining a knowledge-based approach with statistical methods for NLP”

- project leader: Walther von Hahn, University of Hamburg (vhahn@nats.informatik.uni-hamburg.de)

- Romanian partner: University of Bucharest, Faculty of Mathematics, Computer Science Department, - contact person: Florentina Hristea (flori@math.math.unibuc.ro or fhristea@mailbox.ro)
CEEC country case 4: Romania/2

1 DBR-MAT (continued)

- final results were presented at a seminar in Sofia mid January 1999;
- two parts

1. Romanian morphology - a continuation of what had been already achieved for German and Bulgarian within the DB-MAT project.

2. Romanian syntax - completely new, implemented only for Romanian (but can be easily adapted to other languages), based on STATISTICAL PARSING (Eisner - Coling 1996);

- testing was performed using texts belonging to a restricted technical field but results were quite good, considering the small corpus which was available (38865 words). Parsing was successfully performed in 69.60% cases and we expect this result to improve when using a larger corpus (hopefully future developments).