Computer Science Journal of Moldova, vol.**14**, no.1(40), 2006

Linguistic Resources and Technologies for Romanian Language

Dan Cristea Corina Forăscu

Abstract

This paper revises notions related to Language Resources and Technologies (LRT), including a brief overview of some resources developed worldwide and with a special focus on Romanian language. It then describes a joined Romanian, Moldavian, English initiative aimed at developing electronically coded resources for Romanian language, tools for their maintenance and usage, as well as for the creation of applications based on these resources.

1 Introduction

As we begin the 21^{st} century, unhampered access to information technology is one of the foremost requirements for social development. Human Language Technology (HLT), sometimes called Language Engineering, concerned with providing information to the global community in natural, "human", language by supporting unrestricted access to text and speech media documents, is one of the most active research areas nowadays. To ensure that humans and machines have adequate access to resources expressed in natural language, including those on the Internet, technologies such as information retrieval and extraction, speech recognition and text-to-speech capabilities, machine translation, etc. must be developed for the widest possible variety of human languages and language families. Such a technological level is, at present, critical for languages less electronically visible, in order to ensure their appropriate exposure on the Web, especially on the Semantic Web, in a format and with processing capabilities able to give them quick and

complete integration with the new technological developments recently put on scene or which are being prepared worldwide.

The power of Language Engineering, as a solution to the communication requirements of the present times, mainly resides on the existence of sufficient resources in the languages to be treated. It has been estimated[1] that 98% of the HLT researchers who work with specific algorithms and technologies rely mainly on the resources created by the remaining 2% of the HLT community. Resources are used to make explicit, in symbolic form, the phonological, morpho-syntactic, lexical, semantic and structural information, of the language, expressed in both textual and speech form, as it is used in their most natural settings (novels, media, colloquial use, etc.). Apart from raw text or annotated corpora, in different languages, such resources may include statistical data in the form of language models, but also grammar rules, lexicons, dictionaries, terminological thesauri, wordnets, semantic networks, etc. Such resources have been developed on a large-scale for most Western European languages[2].

But the resources alone are not enough to accomplish the challenges expected from the domain of Human Language Technology. Next to them there must stay software tools capable to process the language at all levels and between levels. Section 2 resumes some of the processing machinery used recently in HLT.

Efforts to develop linguistic resources and processing tools for the Romanian language [16] are being pursued in a few centres of Romania (mainly in Bucharest, Iaşi and Cluj-Napoca), in the Republic of Moldova (Chişinău), and, sporadically, also outside this area,

[1] Rada Mihalcea's on-line presentation in the ConsILR meeting, Iaşi, November 2005; see http://consilr.info.uaic.ro/ro/resources/pre/Mihalcea/ConsilR%5B1%5D.Rada.2005.ppt

[2] Many useful linguistic resources and tools can be found at: LDC (Linguistic Data Consortium) - http://www.ldc.upenn.edu/, ELRA (European Language Resources Association) - http://www.elra.com, CLR (Consortium for Lexical Research) - http://crl.nmsu.edu/Tools/CLR/, ECI/MCI (European Corpus Initiative Multilingual Corpus I) – http://www.elsnet.org/resources/eciCorpus.html, MULTEXT (Multilingual Text Tools and Corpora) - http://www.lpl.univ-aix.fr/projects/multext/.

among the most active being those in Germany – Hamburg and Saarbrucken, in England – Sheffield and Wolverhampton, in Italy – Trento, in Canada – Ottawa and in USA – Dallas. In section 3, an overview of the achievements in the acquisition and development of language resources for the Romanian language is presented.

An initiative for the creation of Romanian resources and the development of dedicated tools has been recently agreed upon among the Faculty of Computer Science (FII) of the "Alexandru Ioan Cuza" University of Iaşi (UAIC)[3], the Academy of Sciences of Moldova (ASM)[4] and the University of Sheffield[5]. In a country such as the Republic of Moldova, known for the diversity of languages spoken by the predominant Moldavian population (Romanian) and different minorities (Russian, Ukrainian, Bulgarian, Turkish), the necessity for advanced language technology able to help the citizen to use her/his native language while also interacting smoothly with the official language administration and the multilingual society is remarkable. The main objectives, structure and expected results of the project, recently agreed for co-financing by INTAS[6] and ASM, called RoLTech, and which will be active between May 2006 and April 2008, are summarised in section 4.

2 Language Technology in the Information Society

The Information Society makes the best possible use of new Information and Communication Technologies (ICTs), which is capable to open to the citizen access to enormous amounts of information. Providing easy, natural, access to ICT services to people became a main worldwide objective. In 1998, UNESCO set up an Observatory on the Information Society, to keep the member states informed about the

[3] http://www.info.uaic.ro

[4] http://www.asm.md

[5] http://nlp.shef.ac.uk

[6] INTAS – the non-profit International Association for the promotion of cooperation with scientists from the New Independent States of the former Soviet Union - http://www.intas.be/

new ethical, legal and societal challenges brought by the Information Society. Since then, this observatory has been gradually updated, until it became an Internet-based gateway to online resources on these matters[7]. Although it has the air of being an invention of modern times, the Information Science and Technology (IST) paradigm, one of the seven major research areas in the classification of the EC, has its origins almost 70 years ago, when, in 1937, the American Society for Information Science and Technology (ASIS&T[8]) has been founded with the intention to help professionals in search for new and better theories, techniques, and technologies, and to improve their access to information. In Europe, the long-run declared objective of the EC with respect to science and technology is to overcome the United States and Japan, for the benefit of the society and its citizens. The EC is permanently boosting this challenge through its research policies and financing actions, and is guiding and supporting the revolution in IST through dedicated activities[9].

Language technologies can provide some of the necessary components for the present day Information Society. It can enrich the digital environment with the expressiveness of the human language that is usually lacking in human-machine interaction. Language technology applications and services can radically improve the efficiency and user-friendliness of the acquisition of information and communication tasks, covering business and leisure activities, government and education, services and life at home, through functions providing, minimally, information retrieval, interpretation and translation.

Language technologies are promoted in Europe at various levels, from the EC policies and dedicated programmes to the specific activities in universities and research centres. The 5th Framework Program (FP5[10]) had a Thematic Program on IST, with a specific action on Human Language Technology, to which has been attributed the largest

[7] http://portal.unesco.org/ci/en/ev.php-URL_ID=7277&URL_DO=DO_TOPIC&URL_SECTION=201.html

[8] http://www.asis.org/

[9] http://europa.eu.int/information_society/index_en.htm

[10] http://www.cordis.lu/fp5/

part of the budget (564 millions Euros) of the Key Action III of IST (Multimedia and Content Tools). EC programs like INCO and MLIS also included themes linked to this area. Currently, Human Language Technologies are addressed through the FP6 EC program[11], IST area ,,Knowledge and interface technologies" and FET objectives. The EC "e-Content" program[12] (2001-2004) was intended to stimulate the development, distribution and use of high-quality European digital content on the global networks. Its continuation, the e-Contentplus programme[13] aims to support the development of multi-lingual content for innovative, on-line services across the EU. In Europe there are many associations, universities and research centres with specific LRT activities[14].

Specific national programmes have also been created for the area of Language Technologies. In Romania, the National University Research Council (CNCSIS[15]) and the Managerial Agency for Scientific Research, Innovation and Technological Transfer Politehnica (AMCSIT-POLITEHNICA[16]) stimulate research in Language Technologies at (inter-) institutional and individual level.

The current approach in processing language nowadays is that language peculiarities should be separated from the algorithm, principle which assures both interchangeability of modules and reusability of language data in diverse settings. All modules designed on this principle receive an input file (usually containing a piece of free or annotated text) and a resource (a file able to configure the module according to the specificities of the language) and outputs a transformed (annotated) file or a graphical functionality for the benefit of an interacting user[17]. Designed on this principle, the processing modules used in HLT

[11] http://fp6.cordis.lu/

[12] http://www.cordis.lu/econtent/

[13] http://europa.eu.int/information_society/activities/econtentplus/

[14] For a list, see http://www.lt-world.org/

[15] http://www.cncsis.ro/

[16] http://www.amcsit.ro/

[17] Sometimes, for reasons of computational efficiency, the resources may be incorporated onto the processing module by a compiling process.

are seen to be language independent, their adaptability to a language or another being assured by the plugged-in resource files.

Following, we will make a rough inventory of language processing modules at different levels. The more basic, sub-syntactic, levels should include:

- tokeniser: a module capable to detect word boundaries, including compound words and abbreviations. Some largely used tokenisers are: the Penn Treebank tokenizer[18], GATE (General Architecture for Text Engineering)[19], and QTOKEN[20];

- morphological analyser: a module that returns lists of morphological features of words, each based only on the information communicated by the affixes, therefore taken isolated from the context. A morphological analyser will output all the morphological "interpretations" of an ambiguous word;

- part-of-speech (POS) tagger: taking as input a lexicon of words and a tagset, the POS tagger is a module capable to identify part of speeches of the words by using various methods: rule-based methods, statistical methods or Transformation Based Learning (TBL) methods [6]. Usually POS-taggers are based on previously collected statistics from a gold corpus (a corpus manually annotated by experts to POS data) – called a language model. The most successful POS-taggers have as core a dynamic programming algorithm, for instance – Viterbi [59]. It is proved that the tag set (minimally identifying the part-of-speech, but maximally a complex of morpho-syntactic features) can be optimised [48]. The Brill's tagger[21], the QTAG tagger[22], and the TnT tagger[23]

[18] http://www.cis.upenn.edu/∼treebank/tokenization.html
[19] http://gate.ac.uk/
[20] http://www.english.bham.ac.uk/staff/omason/software/qtoken.html
[21] http://www.cs.jhu.edu/∼brill
[22] http://www.english.bham.ac.uk/staff/omason/software/qtag.html
[23] http://www.coli.uni-saarland.de/∼thorsten/tnt/

are only few examples of POS taggers. The best results on tagging Romanian texts have been reported by Tufiş and Mason, [49] with a tagger based on QTAG;

- lemmatiser: a module that detects the words' lemmas (the canonical form of a lexeme). Lemmatisers, most often, work in combination with POS-taggers, since a lemma of an inflected word may not be unique and may depend on the context. If the context is not considered, the more elementary module is called a stemmer. Largely used lemmatisers are Ellogon[24] and TreeTagger[25]; useful stemmers are Heart Of Gold[26] and Snowball[27];

- chunk parser: a module that detects chunks of text, like noun phrases (NPs), verb phrases (VPs), or prepositional phrases (PPs). Chunks are non-overlapping spans of text, usually consisting of a head word (such as a noun) and the adjacent modifiers and function words (such as adjectives and determiners). The detection of chunks does not necessitate mechanisms more sophisticated than regular expressions [5, 2]. Well known tools including chunk parsers are: fnTBL, a customizable, portable and open-source machine-learning toolkit, primarily oriented towards NL-related tasks, currently trained for English and Swedish[28]; YamCha, a generic, customizable, and open-source text chunker oriented towards a lot of NLP tasks (POS tagging, Named Entity Recognition, base NP chunking, and Text Chunking)[29];

- segmenter: a module that detects sentence or clause boundaries. Most algorithms in this category use lists of key words (segments markers), which are words (expressions) manifesting delimiting

[24] http://www.ellogon.org

[25] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[26] http://heartofgold.dfki.de/

[27] http://www.snowball.tartarus.org/

[28] Freely available at http://nlp.cs.jhu.edu/∼rflorian/fntbl/

[29] Can be accessed at http://www2.chasen.org/∼taku/software/yamcha/

functions. Sometimes, the key words may also indicate the type of relation existing between the two segments it separates.

Above this level the syntactic processing should be considered:

- syntactic parser: a module that produces syntactic trees of the input sentence, either as constituency structures (for instance, Penn Treebank[30]) or as dependency structures. (an example is the Prague Dependency Treebank[31]).

On top of the syntactic level, processes addressing the semantics of natural language and the discourse level should be placed:

- word sense disambiguator (WSD): detects word senses according to a list of senses, as those in a dictionary or a wordnet[32];
- Named Entities Recogniser (NER): identifies expressions that can be classified according to a set of predefined categories, such as entities (organizations, persons, locations), temporal expressions (time, date), quantities (monetary values, percentages, numbers); Ellogon, Yamcha, GATE, Heart of Gold are some NER systems freely available (links above);
- semantic role detector: a module responsible for filling up semantic roles of main verbs (as described in FrameNet[33], for instance);
- discourse parser: a module that assembles discourse trees, usually applying decision criteria depicted from theories of discourse

[30] *Penn Treebank* was built at the University of Philadelphia – Pennsylvania, by Mitchell Marcus (http://www.cis.upenn.edu/∼treebank/home.html).

[31] *Prague Dependency Treebank* is currently being developing at the University of Prague, by Eva Hajičova and her team (http://quest.ms.mff.cuni.cz/pdt/).

[32] Among the projects related to wordnet development there are: the Princeton WordNet (http://wordnet.princeton.edu/), EuroWordNet (http://www.illc.uva.nl/EuroWordNet/), or Balkanet (http://www.ceid.upatras.gr/Balkanet/).

[33] *FrameNet* is a project initiated by prof. Charles Fillmore at the University of Berkley (http://portal.acm.org/citation.cfm?id=980860).

structure, e.g. Rhetorical Structure Theory [34]. Discourse parsing is based on segmentation at elementary discourse unit level and the use of a class of key words known to have discourse significance [36];

- summariser: a module capable to produce a summary from one or more documents. Summaries can be general, displaying most significant facts of a (collection of) document(s), or focussed, giving an insight of the reason for which a certain discourse entity is mentioned in the document(s). Summaries may be classified as extracts (created by reusing portions of the input text(s)) or abstracts (created by re-generating the extracted content) [33];

- anaphora resolver: a module capable to find the discourse entities that anchor referential expressions, such as pronouns, common and proper nouns, etc. Detection of text entities referred by pronouns, or nouns having a referential role (anaphors), is a vital process in many applications of text processing. For instance, in automatic translation, in order to correctly translate pronouns from a source language, in which pronouns have few forms, in a target language, which is richer in pronoun forms, it is of prime importance to know which entities those pronouns refer to. Only few translation systems are nowadays capable to correctly interpret a discourse that is longer than a single sentence, because they do not have means to recognise anaphoric relations. Summarization systems produce better outputs when they incorporate also anaphora resolution mechanisms. Other domains that make heavy use of anaphora resolution are: information retrieval, inter-document summarization and automatic question-answering.

A different type of processing is dealing with multilingual and parallel texts:

- language detection: the identification of the language of a span of text (mainly based on statistics drawn on the occurrence of letters or detection of words as belonging to certain vocabularies);

- sentence and word aligner: receives pieces of parallel texts and outputs their alignments at sentence and word level.

In the field of tools for processing language, of a significant importance is the GATE system [24], developed at the University of Sheffield – a modularised framework for the configuration of pipe-line architectures composed of language processing modules. The processing modules are supposed to be language independent (in the sense described above), therefore easily adaptable to Romanian by integration of adequate resources.

3 The language technology applied to Romanian language

Elaboration of lexical resources makes an important part of researches directed on Romanian language. Although not to the extent of other languages with greater electronic visibility, efforts have been invested by researchers in different places (Romania, Republic of Moldova, United States, United Kingdom, Germany, Italy, etc.) to develop Romanian linguistic resources such as corpora, dictionaries, wordnets and collections of linguistic data in both symbolic and statistical form (n-gram tables to configure language models, sets of grammar rules, name entity lists, etc.).

Research in HLT in Romania is being pursed in several centres, among which: in Bucharest, at the Research Institute for Artificial Intelligence of the Romanian Academy (AR-ICIA)[34] and the University of Bucharest; in Iaşi, at UAIC-FII and the Institute of Computer Science of the Romanian Academy – the Iaşi branch (AR-IIT)[35] ; in Cluj-Napoca, at the "Babeş-Bolyai" University, etc.

The research institutes in linguistics and philology, as those of the Romanian Academy, which, until very recently, employed only classical methods of study in the acquisition of linguistic dictionaries and

[34] http://www.racai.ro

[35] http://www.iit.tuiasi.ro/iit/index.php

thesauri, begun to show an ever-growing interest for digital techniques, mainly in using the computer for lexicographic tasks and the processing of linguistic material. For instance, [29, 30] report preparatory activities performed at the "Al. Philippide" Institute of Romanian Philology in Iaşi towards the computer-aided acquisition, in electronic form, of the Dictionary of Romanian Language (DLR), edited by the Romanian Academy. When this remarkable work, started in 1965 within the three academic institutes (the Institute of Linguistics ,,Iorgu Iordan – Al. Rosetti" – Bucharest, the Institute of Romanian Philology "A. Philippide" – Iaşi and the Institute of Linguistics ,,S. Puşcariu" – Cluj-Napoca), will be finished (expected for 2007), the new series of the DLR will contain 26 volumes, counting almost 12,000 pages, and including the letters D, E and L – Z[36]. Among the advantages of an electronic version of DLR, some obvious ones are: the possibility to add, correct, and modify any entry of the current version, to align word senses with similar entries of other resources (very useful for combined search), to exploit the large collection of examples associated with definitions of senses in order to acquire abilities to disambiguate, by statistical means, semantically ambiguous words in contexts (as required, for instance, in machine translation), to enlarge the database of texts/attestations used in the old series of the dictionary, and, finally, to print and publish easier, including on the Internet, the whole DLR or only instances of it, as imposed by different scientific or commercial needs. The cooperation between linguists and computer scientists at the "A. Philippide" Institute has made possible to develop a dedicated tool, DLRex [29], an instrument able to acquire, process and browse the electronic version of DLR.

In 2001, research groups from Iaşi, Bucureşti and Chişinău have founded the *Consortium for the Romanian Language: Resources & Tools*[37] – an initiative aiming to synergise the efforts of linguists and

[36]The rest of the letters form what is actually called the old series of the Romanian Language Dictionary, or the Dictionary of the Academy (DA) and has been published before 1949.

[37]The Consortium portal is at http://consilr.info.uaic.ro/, with Romanian and English versions

computer scientists who work on Romanian language, mainly by promoting to linguists the software tools for linguistic processing developed by computer scientists, and to computer science people – the resources created by linguists. Two fundamental aspects regarding the processing of a language are mainly taken into consideration in the activities of the Consortium: the elaboration of tools capable to process Romanian in conformity with the established international standards, and the creation and maintenance of proper linguistic knowledge (resources). At the last workshop dedicated to the activities of the Consortium, in November 2005[38], it was extremely encouraging that so many groups and individual researchers have presented their activities related to the development of resources and tools dedicated to the Romanian language. Their achievements will be briefly presented below.

At AR-ICIA, a large collection of corpora, mainly annotated, has been created. Among the manually validated corpora there are:

- *NAACL 2003* and George Orwell's novel 1984 are parallel English-Romanian corpora containing about 1.6 mil., respectively 250,000 tokens; the corpora are segmented, morpho-syntactically annotated and lemmatised. The 1984 corpus is word-aligned and annotated to word senses using the Princeton WordNet;
- Plato's *Republic* (250,000 tokens), *Evenimentul Zilei* (news, about 92,000 tokens), and *ROCO* (news, 7.1 mil. tokens) are parallel English-Romanian corpora, morpho-syntactically annotated.

At the same institute, using specially developed tools, the collection of automatically annotated parallel corpora contains:

- the Romanian FrameNet: 1,094 sentences from the original FrameNet 1.1. corpus (translated in Romanian at UAIC-FII), annotated morpho-syntactically and lemmatised;
- Timex: the Romanian translation (realised at UAIC-FII) of the TimeBank 1.1 corpus [42] – 186 news articles with 72,000 Roma-

[38] http://consilr.info.uaic.ro/en/index.php?showpage=030103

nian tokens; the corpus is lemmatised, morpho-syntactically and temporally annotated in English and partially in Romanian;

- RoSemCor: a parallel English-Italian-Romanian corpus[39], aligned to word-senses, including 12 articles from the SemCor[40] corpus. The alignment methodology will be used for another 80 articles from the same English corpus. The translations have been realized at UAIC-FII and the University of North Texas;

- Acquis Communautaire (about 12,000 Romanian documents; 6,256 parallel English-Romanian documents); the corpus is lematised, POS-tagged, sentence- and word-aligned.

The dictionaries/lexicons developed at ICIA include *WEB-DEX* – the explanatory Romanian dictionary, containing about 65,000 entries, XML encoded according to CONCEDE, *tbl.wordform.ro* – an ASCII file with cca. 546,000 occurrences of Romanian words, *Romanian paradigmatic morphology* – an unification-based description of the morphology with a lexicon of cca. 40,000 lemma, *RoWordNet* – the semantic network of Romanian, aligned to concept level with the Princeton WordNet, and containing about 30,000 synsets, from which about 20,000 were created in cooperation with UAIC-FII during the BalkaNet FP5 project [54, 20], *Romanian-French dictionary* – 16,710 entries XML encoded, and *EUROVOC* – a multilingual thesaurus containing bilingual dictionaries from 21 languages included in the Acquis Communautaire corpus.

The ICIA tools dedicated to HLT, will partially be available soon as integrated web services based on the WSDL, UDDI and SOAP protocols, and will include:

- EGLU – an integrated programming application, implemented in Common Lisp, based on unification of complex NLP systems; it

[39]For references, see http://multiwordnet.itc.it/english/home.php

[40]For the original corpus, see http://multisemcor.itc.it/semcor.php or references on Rada Mihalcea's pages at http://www.cs.unt.edu/∼rada/downloads.html.

includes a compiler of linguistic descriptions and modules for morphological analysis and generation, syntactical analysis – CKY, syntactic generation - Head driven, lexical or structural transfer for machine translation [47];

- DIC – a compiler of electronic dictionaries initially created for the automatic generation of the XML Concede encoding of DEX; with minimal modifications it can be used for compiling dictionaries with structures similar to those of DEX;

- TTAG (Tiered Tagset) – a system of automatic projection of an optimal tagset of morpho-lexical descriptors, implemented in Perl [48, 56];

- TT&CLAM (Tiered Tagging and Combined Language Models) - a system of morpho-lexical disambiguation on two levels, that uses combined language models; it includes a specialized editor [51];

- TTL – a Perl module that allows the text segmentation at sentence/word level, the lemmatization and morpho-syntactic annotation; the module is language independent and uses regular expressions and Markov models;

- WSDTool – a Perl application which annotates at sense level every word of an XCES parallel corpus; any parts from the corpus for which aligned wordnets exist can be annotated [32];

- TREQ – a Perl application that extracts translation equivalents dictionaries from parallel corpora [52];

- YAWA – a Perl lexical aligner, language independent for the modules that do not require specific alignments between the languages involved. It uses a parallel XCES corpus, morpho-syntactically annotated and lemmatized, and translation dictionaries obtained through TREQ [53];

- WN-Builder, WN-Correct – a set of programs, used during the BalkaNet project, dedicated to the development and correction of wordnets aligned with the Princeton WordNet [55];

- Ro-Hyphenator – a Romanian syllable splitter;

- DIAC – a system that automatically recovers the missing diacritics from Romanian texts [50];

- Multilingual thesauri aligner – a C# application used to align the English version of Eurovoc with the incomplete Romanian version;

- MTKit – a C# integrated application for the annotation/lexical alignment of XCES (Extended Corpus Encoding Standard)[41] files. It creates statistical translation models and, through a friendly graphic editor, helps to browse the lexical alignments and the properties of each constituent of the alignment (such as POS tags, lemma, chunks, word sense, and definition of the synset). The incomplete or incorrect alignments can be modified. Using a gold-standard, MTKit automatically calculates the alignment accuracy (precision, recall, F-measure) [57];

- Sentence Aligner – a C# application intended for statistical sentence alignment [7];

- LexPar – a Perl application which determines the structure of a given sentence as a graph of dependencies;

- XCESGen – a group of Perl modules that generates XCES format for parallel corpora;

- Google Screen Scraper – a library of instruments dedicated to automatic web search using Google capabilities.

[41] http://www.cs.vassar.edu/XCES/

Among the HLT modules developed at UAIC-FII, some being available on the ConsILR portal[42], there are:

- AnMorph – an environment for the development and updating of the paradigmatic morphological model of a non-agglutinative language (the basic inflected word model sees a word as composed of a root and an ending). Currently, the database of the application covers only partially the Romanian morphology. During the interaction with the system, a trained user works with a friendly interface to classify new words into already existent inflecting paradigms, or fill in, by examples, new paradigms. The program continuously compares the forms introduced by the user in the automatically drawn tables with those that can be generated from the stored paradigms and, if a matching is confirmed, generates the rest of the forms. When this happens, the user has only to check and validate the automatically filled-in parts of the tables. Apart from the developing/updating interface, the environment offers an editor for the dictionary and the collection of paradigms, a component enabling consistency checks of the data and a lemmatiser [12];

- Occurrence Finder – a tool aiming to identify occurrences of lexical sequences that are subject to restrictions of different kinds, in raw or annotated texts. This application could be of extreme interest to lexicographers, which build their dictionaries entries by looking for adequate examples in text corpora, but can be of help in any linguistic research activity that uses large collections of textual data. The system incorporates a specialized restrictions language – SXPath, which is based on the Xpath standard (the XML Path language). The incorporated search engine evaluates an SXPath expression in one single pass of the XML document, without having to load the whole document into memory. It goes through the document serially and keeps in memory only

[42] Applications available at http://consilr.info.uaic.ro/en/index.php?showpage=060101

On-line tools available at http://consilr.info.uaic.ro/en/index.php?showpage=0604

those parts that are relevant for the current work context. The engine returns the list of elements that observes the restrictions expressed by the SXPath expression [44];

- An environment for the processing of parallel XML annotated documents, including a program allowing the definition of annotation schemas (tag sets with corresponding attributes). A consistent set of schemas are placed in a hierarchy (lattice). Different annotations over the same text can be mixed on the same output document, or vice-versa, different partial annotations can be extracted from a complex annotation [18];

- RARE – a framework allowing the development and testing of anaphoric models [41, 19, 21]. Its architecture offers the user a build-in library of functions that can complement most of the foreseeable models for entity tracking in texts. It also allows a user to define its own strategies, specific to the application settings in which this functionality is a must. The engine recognises the first mentionings of characters/objects in a text and links their subsequent ones to these, outputting co-referential chains;

- Learning-based clause-level text splitter – is a tool that receives a raw text as input and produces an XML annotated version of it in which clause borders are marked. The architecture includes two modules: a learning module and the actual segmenter module. The learning module infers rules for the recognition of clause borders using a corpus of texts annotated adequately. Both modules make use of a list of markers, which is kept as an external resource of the tool. The segmenter can be used independently or integrated in any NLP application [38];

- Rule-based clause-level text splitter – a similar tool for detection of clause borders based on symbolic rules [43];

- Discourse Parser – a tool intended to reveal the (tree) structure of a text [39, 22]. In such a structure the inner nodes represent rhetorical relations and terminal nodes represent elementary text

spans, as discourse units. Once the rhetorical structure of a text is obtained, intelligent browsing over the semantic structure of the document can be performed: general document summaries, summaries focused on specific characters or objects (discourse entities) of interest, data mining in texts, tracking of characters over events and situations distributed in text, determination of temporal positioning of events described by the text as absolute intervals or points on the time axis or relative to one another, detection of relations among events and situations, detection of spatial positioning and/or spatial relations of discourse entities, etc. The discourse parser incorporates a number of natural language processing modules, including a part-of-speech tagger, a segmenter establishing bounds between discourse units, a module able to recognise noun phrases (shallow parser) and RARE (the AR-engine described above). The parser implements rules and heuristics, which combine information contributed by words or expressions, having a role in the identification of structure, with referential chains detected by RARE;

- Summarisation Tool – is a Java based application [40], with a web-interface, that receives an XML file annotated to discourse structure (no relations between the entities involved) and gives focused summaries using Veins Theory [14, 17, 23]. The summaries can be focused on discourse entities mentioned in the text;

- GLOSS – a visual instrument intended for corpus annotation according to its discourse structure and referentiality. The application allows simultaneous annotation of multiple documents [15, 13, 58];

- XML annotator – an interface helping a manual annotation activity over a text, resulting in an XML document. It allows the definition of tags and their attributes, insertion of links between tags and navigation along them [9];

- Tracker of temporal expressions in texts – a program that detects expressions denoting events, their instances and signals specific

links between events. The application is under development, the current version being capable to partially annotate temporal information in Romanian texts, using 3 out of the 7 tags proposed in the TimeML standard [45]. The application is trained on a manual annotated version of a part of the 1984 corpus [28];

- RPC (Romanian Page Categorization) – a system that marks the category of a webpage. The categorisation process is guided by a database containing a set of key words specific to each category, which is automatically generated from webpages and is used in the training phase, based on the words' frequencies [46];

- Language identifier – a program capable to detect the language of a span of written text based on tables of bigrams and trigrams (sequences of two and three letters);

- Word sense disambiguator using WordNet – an application [31] that disambiguates the words' senses in texts using a knowledge base and context information from a training corpus (Orwell's 1984, annotated to POS and syntax structure);

- XML converter – a Java application developed in the frame of the LT4eL FP6 project[43], intended to automatically transpose the html documents representing the learning objects from the 9 languages involved in LT4eL (Bulgarian, Czech, Dutch, English, German, Maltese, Polish, Portuguese, Romanian) to the agreed XML format, that will be further used in the next phases of the project;

[43]The LT4eL project (http://www.let.uu.nl/lt4el/) is an FP6 project that will run between 2005 and 2007 and will provide Language Technology based functionalities to support semi-automatic metadata generation, for the description of the learning objects, on the basis of a linguistic analysis of the content. Semantic knowledge will be integrated to enhance the management, distribution and retrieval of the learning material. Ontologies, key elements in the architecture of the Semantic Web initiative, will be used to structure the learning material within Learning Management Systems (LMSs), by means of descriptive metadata.

- Keyword extractor – a language-independent Java application implementing a tf-idf algorithm [35] built for the benefit of the LT4eL project, which takes as input a collection of UTF-8 encoded text documents and produces a list of key words.

Among the collection of linguistic resources and tools developed by the research group in HLT at AR-IIT, we indicate only few:

- 16 files (summing up 36,357 words) out of the 44 files (with 95,455 words) of the 1984 corpus marked for NPs, and 10 files (21,468 words) marked for referential expressions, using the PALinkA annotator[44]. The whole corpus was segmented using the SCD algorithm [26], into 15,014 clauses and 14,609 implicit VPs;

- A part of the Romanian version of Hemingway's novel "Farewell to arms" (20,148 words out of 61,262 words) morphologically annotated using the TexTag editor;

- A collection of parsers based on Java grammars for DEX and DLR [25].

At AR-IIT, another group is working on Speech Technologies. Recently, they have initiated the creation of a freely-available Electronic Archive of the Romanian Sounds[45]. The site of the archive represents a joint realisation of the Speech Technologies group of AR-IIT, UAIC-FII and the Technical University "Gh. Asachi" of Iaşi (the CERFS Excellence Centre). The Sounds Archive will be used to improve the algorithms for speech recognition and (non-concatenative) synthesis and is planned to host the following speech resources for Romanian:

- a database with recordings belonging to both professional and non-professional Romanian speakers, inhabitants of the Iaşi county, as well as pronunciations specific to other geographical areas.

[44] http://clg.wlv.ac.uk/projects/PALinkA/

[45] http://iit.iit.tuiasi.ro/romanian_spoken_language

Based on it, the Speech Processing Group at AR-IIT will organise statistical studies on the Romanian sound system and, more generally, on various aspects of the spoken Romanian language;

- a Romanian speech database with records taken from persons manifesting different speech pathologies;

- a collection of syllable and word recordings to be used by concatenative synthesizers and speech recognition systems;

- an electronic dictionary with Romanian pronunciations will be correlated with the Atlases of the Romanian language [1], developed as a joint work at AR-IIT and the Institute of Romanian Philology "A. Philippide", Iaşi[46]. A performing editor for dialectal texts, *EditTD*, has been created at AR-IIT and has been used for editing the Linguistic Atlas.

In the Republic of Moldova, the main centre of research in language technology is the NLP group at the Institute of Mathematics and Informatics of the Moldavian Academy of Science (ASM-IMI). This group has significant contributions to the development of resources and tools for processing Romanian language. Among them, the software package "Tools for linguistic applications" [3, 4] has been used in the implementation of the Romanian Spelling Checker RomSP [11]. This package, at the moment of its distribution, was one of the first major processing tools dedicated to the Romanian language, among those developed in the Republic of Moldova and Romania. The resources developed at ASM-IMI include:

- a database with linguistic information for Romanian at the word level. The database accepts queries formulated in SQL;

- an extensive collection of Romanian word forms (approximately 1,000,000);

[46]The volumes of the Atlas are published in the NALR/ALRR series – *The New Romanian Linguistic Atlas / The Atlases of Romanian on Regions.*

- a Romanian lexicon (70,000 lemma entries), with morphological, syntactic and syllable splitting information [10];

- a dictionary of synonyms;

- translation dictionaries for Romanian, Russian and English;

- a Romanian grammar, containing 866 grammatical rules and a set of 320 affixes, which have been used for the development of a morphological vocabulary of cca. 30,000 words.

All resources developed at ASM-IMI are paired with specialized tools for browsing and maintenance (editing, modifying and updating). Another set of tools at ASM-IMI are dedicated to the assisted learning of the Romanian morphology and syllable splitting.

Other research and acquisition of language thesauri are being performed at the Department of Informatics and Foreign Languages, belonging to the Faculty of Informatics and Microelectronics of the Technical University of Moldova, Chişinău. One of the most recent efforts towards the development of linguistic resources in this department is oriented towards the creation of a corpus of 885 decisions of the Justice Court (789,520 words). The texts were scanned, segmented, morphologically annotated, chunked (for NPs and VPs, using regular expressions), and the named entities were marked. A manual correction of the texts will be followed by a semantic annotation, by aligning it with a juridical ontology, specially developed in the same group, which includes 44 concepts with 140 slots and is coded in the RDFS format.

At the Academy of Economical Sciences in Chişinău, in a collaboration with UAIC-FII, pronunciation (wav file), images and video files, illustrating part of the word senses in a Romanian dictionary, have been added, resulting in a Multimedia Dictionary of Romanian [8]. The resulted Microsoft Access database is structured on 7 fields: word, word_with_accent, definition (text fields), as well as audio_word, audio_definition, image and video (OLE fields).

At the University of North Texas, USA (UNT), the following resources dedicated to Romanian language have been developed and are

available on the ConsILR portal and the UNT web-site[47]:

- ROCO – a Romanian news corpus with 400 million words, tokenized and POS-tagged (at AR-ICIA); for the Senseval evaluation the corpus was also sense-tagged;

- Sense-tagged resources – used as Romanian resources in the Senseval 2003 and ACL 2005 competitions, and accessible also as web-collections by means of the *Word Games* interface, developed together by UNT (Rada Mihalcea), USC Information Sciences Institute (Tim Chkloski), University of Ottawa (Vivi Năstase) and AR-ICIA (the group coordinated by Dan Tufiş);

- the Romanian-English parallel news corpus (850,000 words), containing a collection of the dailies *Evenimentul Zilei* and *Monitorul*, and aligned at word level;

- Romanian-English dictionary with 38,000 entries.

Another American group working on Romanian is located at the Department of Computer Science, University of Southern California (USC-DCS). Marcu and Munteanu [37] have presented and demoed, during an invited talk in Eurolan-2005[48], the first Romanian-English statistical translation system. Starting from 15k docs, containing 10M English words, and 170K docs, containing 85M Romanian words, mainly collected by students at UAIC-FII during a term project, they have filtered an accurate collection, contained 6.5M words in parallel texts, aligned at sentence level. This parallel corpus was used in the statistical machine translation system to learn probabilistic word/phrase-based translation rules and to produce from them translations from Romanian into English of a fairly good quality (Bleu score ranging from 20.9 for heterogeneous genre and 49 for the European legislation).

[47] http://www.cs.unt.edu/∼rada/downloads.html#romanian.

[48] Eurolan 2005 – the 7^{th} Biennial International Summer School on *The Multilingual Web: Resources, Technologies, and Prospects*, http://www.cs.ubbcluj.ro/eurolan2005/

The NLP group at the University of Hamburg (Germany) has recently developed G.E.R.L. – a German-English-Romanian lexicon[49], using the Parole/Simple model. Intended mainly for didactic usage, the corpus contains many technical terms and incorporates morphological, syntactic and semantic (synonymy, verbal thematic roles, collocations) annotations.

4 RoLTech – a project dedicated to the Romanian Language

RoLTech – Platform For Romanian Language Technology: Resources, Tools And Interfaces – is a project co-financed by INTAS and ASM, which will proceed for 24 months, between May 2006 and April 2008. Participants in the project are UAIC-FII – as coordinator, the NLP Laboratory of the University of Sheffield and ASM-IMI. The project aims to acquire electronic resources for Romanian language, to develop corresponding tools for their maintenance and use, and to create applications based on these resources.

The project has the following general objectives:

1. to gather and integrate on a dedicated portal the existent resources in electronic form for Romanian language (including dictionaries, thesauri, corpora – raw and annotated, as well as linguistic data) and to develop new ones;

2. to build a platform that will group tools dedicated to process Romanian language at morphological, syntactic and lexico-semantic levels, and that will support integration of these tools into complex applications;

3. to build interfaces able to offer to the citizen (including the native non-Romanian speaker) access to the resources in a friendly and interactive way (including access through the Web).

[49] http://nats-www.informatik.uni-hamburg.de/view/Main/GerLexicon

The intended Web-portal will store and give access to:

- reusable linguistic resources for language technology (including dictionaries, thesauri, corpora – annotated and raw, as well as symbolic and statistical linguistic data),

- language technology tools for Romanian (both open source and authored code),

- documentation related to Romanian language (documents, references to external resources or tools, to significant projects and scientific events dedicated to the domain of HLT, titles of books, collections of papers on Romanian language, significant scientific events, etc.).

A prototype version of the portal will be set-up at the beginning of the project and will be enhanced till the end of the project and, hopefully, maintained permanently after.

All the resources created during the project lifetime will be permanently integrated into the Web-portal. They should conform to formats that will make them re-usable for integration in different NLP applications dedicated to Romanian language. This is the reason why they are called in the project Romanian Reusable Resources for Language Technology (3RLT). Three types of applications will be developed in the project:

- the first type will target the non-native speakers of Romanian. An example is an adaptable e-learning system for Romanian morphology to be used by students, with interfaces and teaching materials in Romanian, English and Russian, with multimedia elements and with possibilities to extend to other languages. The main beneficiary of these applications will be the minority citizens of Moldova, mainly the Russian speaking population;

- the second type of applications will focus to ordinary Romanian speakers. Among them, one application will aim to enhance the

search output on collections of Romanian documents as intermediated by existing search engines, by exploiting the morphological variations of words and synonyms. Another one will offer an interactive Romanian spell-checking service over the Web;

- finally, a third category of applications will be dedicated to the expert users of Romanian language. One example is a support system for dictionary development, including advanced lexicographic operations as, for instance, abilities to use the context for detection of word occurrences in corpora, to support complex browsing among dictionaries of different types and multimedia presentations of linguistic material on Romanian. The package is addressed to experts working on the language technology, but which, until very recently, have used only classical pencil and paper methods to acquire linguistic data, sort and organise them onto dictionaries and other printed lexicographic sources.

A clear Project Management and Dissemination plan will foster the communication strategies among the members of the RoLTech consortium and will increase the transfer of knowledge and the distribution of the project results.

The portal will enable Web access to resources related to the Romanian language and their use in computer linguistics and human language applications, together with dictionaries or pointers to dictionaries, tools for language processing, interactive learning environments for Romanian, links to conferences, books, significant papers and other materials on Romanian computational linguistics and Romanian language technology, accessible through the Internet. The web portal will have a bilingual interface, in English and Romanian.

We are confident that the portal will contribute:

- to enhance the undergraduate, master and doctoral level research as well as the education at these levels in Computational Linguistics and Human Language Technology, in both Romania and Moldova, as is now taught in different universities, including the "Al.I.Cuza" University of Iaşi, the "Babeş-Bolyai" University of

Cluj-Napoca, the University of Bucharest, the Technical University of Chişinău, and others;

- to bring closer the collaboration between linguists and computer scientists that are working and are doing researches on the Romanian language, particularly by emphasising how computational methods can speed up the acquisition of linguistic data and enhance their quality, and by raising the awareness that modern linguistics is bound to using computational methods;
- to disseminate the research on Romanian CL and HLT beyond the geographical area where the language is mainly spoken, this way stimulating world wide collaboration on Romanian language;
- finally, it could become a site used by people wanting to learn Romanian or to enhance their knowledge on the language.

The learning system for Romanian language is intended for use by non-Romanian native students. It will have, for this purpose, a trilingual (English, Romanian, and Russian) interface, and will include teaching materials in all these languages. The environment will be able to adapt to the student's individual needs and perspectives, first, by enabling her/him to choose from among several options for teaching materials. Once the student is acquainted with the teaching material, the system then offers a choice of several activities (tasks, exercises, tests, and games) and lexical material. The student selects a topic, learning and testing activity, and the lexicon to be used that best suits her/his interests and learning style. As the student goes through the material, she/he can also scan the log of previous work and summary reports. Thus the functionally is that of an autonomous learning system, providing a mechanism for self-learning and self-assessment.

One of the particular features of Romanian is the richness of its inflexions. Romanian, belonging to the group of Romance languages, not only resembles its distinguished Romance sisters in the prodigious morphology of verbs, but inherits also from Latin the declinations of nouns and adjectives (feature lost by the other members of this family, but existing, for instance, in the Slav languages, from which is has

got also many influences). This is why it is very important for the user searching Romanian documents to have access to search engines which incorporate the ability of finding information based on morphological derivatives of the word. This feature can be manifested either at the level of the input, by allowing inflected word forms as search criteria, or at the level of the searched documents, by offering the retrieval of documents including variations of the word presented in the input. Optionally, synonyms of the input word could be used in search. The project includes also the creation of morphological interfaces for Web search engines and an adaptation of a Web-service for Romanian spelling checking.

The dictionary development support system is intended to help the interactive creation of dictionaries and lexicons, using the 3RLT. Mainly, it will be addressed to lexicographers supposed to be involved soon in the elaboration of DLRI, starting from the printed editions of DLR (as explained in section 3). Its development will be based on the existing version of DLRex [30], which is presently capable of editing and browsing functions adapted to lexicographic activities. It is supposed that at the end of the project, the activity for the elaboration of DLRI will already be initiated, based on a research plan [27] expected to be elaborated by the Romanian Academy in collaboration with UAIC-FII. But this interactive specialized working environment could equally be promoted among the Moldavian lexicographers working on Romanian.

The multimedia elements needed for a language learning system can include sounds (pronunciation of words), pictures (illustrations), and video clips (animated illustrations). Such elements, adequate for 3RLT, usually require some additional programming. They have a relatively big volume and necessitate binary representation. They can be kept in the same database as the linguistic information itself (so-called BLOBs), or the database can contain only pointers to them (URLs).

The proposed project goes in-line with these researches and elaborations and moves ahead the development of lexical resources on Romanian as well as the elaboration of tools to be integrated into applications or for direct consumption by the end-user. The absolute novelty

of the proposed project is its large-scale integration of resources and tools for Romanian language, with immediate applications intended to the society and the citizen. For the first time, in Romania and in Moldova, neighbouring countries which had, during the history, long periods of common co-habitation, modern technologies of language processing will be used for the benefit of the society – by preparing the technological integration of Romania into the EU, and the citizen – by helping Moldavian minorities to acquire the official language (same as in Romania).

5 Conclusions

Most of the knowledge currently generated in the Information Society is encoded in natural language, more specifically in the form of electronic written texts or speech data. Computational systems able to process and support this huge amount of knowledge have to use large scale and reliable language resources and tools. Creating, maintaining and disseminating the language resources and tools are challenging processes, with impact in many fields of the human activity and civilisation, such as science, culture, economy, and politics. When such resources and tools exist for a given language, they contribute strongly to the promotion of the national identity and the intercultural integration. When, on the contrary, they lack, the visibility of that language on electronic media is very poor, situation which could trigger extremely harmful effects against the people who speak that language and on the language itself. The modern times showed very clearly already that languages aggressively promoted on electronic media, as is the case of English, for instance, could influence important segments of an official language of a country, as is for instance the business jargon or the computers and communication jargon. In the past, when a language or dialect disappeared, the cause had to be looked for on specific social or political conditions (movement of population towards economically more developed zones, wars, etc.). Recently, to the traditional dangers, a new one seems to have appeared: the poor presence of the language on electronic media. One reason for this is the extreme attractiveness of

the Internet and other electronic communication media to the youngsters, therefore that segment of the population which will configure the shape of the language as will be it spoken by the next generation. It is difficult to foresee in detail the impact that a poor representation on the Internet could have on a language, but an alarm signal should be triggered already.

Languages are entities alive. They are transformed in the same rhythm in which the people that speak them are transformed by age and the renewal of generations. Sometimes, they are fragile and apparently minor changes influence them a lot, usually irreversible. If inappropriately cultivated, languages may even die[50] . With the Web technologies developing vividly, but especially with the Semantic Web, which is expected to open incredible computer-mediated content-based intelligent search and retrieval possibilities and which can be already perceived growing around the corner, languages are more and more dependent of their representation on the Web. It becomes a truism that the existence of language technologies for a language becomes nowadays a must in order for that language to "survive" in the Information Society.

The goal of this paper was to show the level of the research attained in the domain of Romanian LT, in order to foster its further development. It has begun by giving a short overview of the state-of-the-art of the language technologies in general, then it continued by presenting the main achievements on both resources and processing tools dedicated to Romanian, and finally it described a project that has just started. RoLTech, by its interdisciplinary nature (combining computer science and linguistics) and its international involvement (grouping researchers, native speakers of Romanian, belonging to Romania, Moldova, but also outside this area) is thought to make a significant step forward in the direction of the coordination of research actions addressing the computational aspects of the Romanian language from the modern perspective opened by the Information Society. Such a process,

[50]For information on endangered languages in Europe, for instance, see the UNESCO Red Book at http://www.helsinki.fi/%7Etasalmin/europe_report.html#Romansch

although having been initiated by several sporadic meetings that took place in the last three years in Romania as well as in Moldova, did not reached yet the maturity from which common actions could have arisen. The creation of a Web-portal intended to host language resources and processing tools, based on which applications will be developed, will help to achieve a much wanted common view on the future activities, to raise the quality of the research at European standards, and finally to offer the results to the citizens, as the final beneficiary.

6 Acknowledgements

In this paper we have used information presented on different occasions (as the EUROLAN-2005 Summer School and the November 2005 ConsILR meeting, organised by UAIC-FII), published or directly offered gently, on our request, by different people or groups. We are grateful to all of the following:

- our graduates, master students in Computational Linguistics and Ph.D. students at UAIC-FII Iaşi, past and present members of the Natural Language Processing Group. Special thanks are addressed to Bogdan Aldea, Cătălina Barbu, Roxana Bejan, Cosmin Bejan, Cristina Butnariu, Costel Coşman, Ovidiu Crăciun, Iustin Dornescu, Daniela Dudău, Gianina Dumitriu, Laur Gheţu, Vlad Horbovanu, Oana Hamza, Alex Hriţcu, Marinela Hriţcu, Maria Husarciuc, Ana Masalagiu, Livia Miron, Alex Moruz, Ciprian Niţă, Oana Postolache, Ionuţ Pistol, Georgiana Puşcaşu, Marius Răschip, Ioana Sandu, Violeta Sereţan, Valentin Tablan, Iulian Tănăsescu, Amalia Todiraşcu, Diana Trandabăţ, Daniel Ţorin, Cristian Ursu, and many others...
- the Natural Language Processing group of the AR-ICIA Bucharest, headed by prof. dr. Dan Tufiş, Correspondent Member of the Romanian Academy;
- the Speech Processing group of the AR-IIT Iaşi, headed by prof. dr. Horia Nicolai Teodorescu, Correspondent Member of the Romanian Academy;

- the Natural Language Processing team at AR-IIT Iaşi, headed by Nicolae Curteanu;
- the Natural Language Processing Group at ASM-IMI Chişinău, headed by dr. Svetlana Cojocaru and dr. Constantin Ciubotaru;
- dr. Cristina Florescu, dr. Gabriela Haja, and the DLRI/DLRex project teams from the "Al. Philippide" Institute of Romanian Philology[51] , Romanian Academy, Iaşi;
- Natalia Burciu and Natalia Elita from the Department of Informatics and Foreign Languages[52], belonging to the Faculty of Informatics and Microelectronics of the Technical University of Moldova, Chişinău;
- the NLP group at the University of North Texas, directed by dr. Rada Mihalcea;
- Valentin Tablan and Cristi Ursu, from the University of Sheffield, members of the GATE team;
- Dr. Daniel Marcu and Dragoş Ştefan Munteanu at the University of Southern California, authors of the Romanian-English statistical translation system;
- Constantin Orăsan, from the Research Group in Computational Linguistics, University of Wolverhampton[53], England, author of PALinKA;
- the Natural Language Systems Division[54], Department of Informatics, University of Hamburg, working on German-Romanian resources, coordinated by dr. Cristina Vertan.

The section 4 of this paper describes the INTAS project RoLTech no: 05-104-7633.

[51] http://www.iit.tuiasi.ro/philippide/index.html
[52] http://www.utm.md/en/3-1-3-1.html
[53] http://www.clg.wlv.ac.uk/
[54] http://nats-www.informatik.uni-hamburg.de/Main/WebHome

References

[1] V. Apopei, F. Rotaru, S. Bejinariu, F. Olariu. 2003. Electronic Linguistic Atlases. In *Proceedings of the International Conference on Information and Knowledge Engineering. IKE'03*, June 23-26, 2003, Las Vegas, Nevada, USA, Volume 2, pp. 628–633, CSREA Press 2003, ISBN 1-932415-08-4.

[2] S. Bird, E. Klein, and E. Loper. 2005. Chunk Parsing. Available at http://nltk.sourceforge.net/lite/doc/en/chunk.html.

[3] E. Boian, S. Cojocaru, L. Malahova. 2000. Instruments pour applications linguistiques. La terminologie en Roumanie et en République de Moldova, Hors série, No. 4.

[4] E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, V. Demidova, L. Malahova. 2003. Lexical resources for Romanian – a project overview. In: *Symposium on Intelligent Systems and Applications*, September 19–20, Iaşi, Romania. Eds.: H.N.Teodorescu, G.Gaindric, E.Sofron. Publisher: Tehnici si Tehnologii, Iaşi.

[5] T. Brants. 1999. Cascaded Markov Models. In *Proceedings of EACL 1999*.

[6] E. Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 1995.

[7] A. Ceauşu, D. Ştefănescu, D. Tufiş. 2006. Acquis Communautaire sentence alignment using Support Vector Machines. (to appear) in *Proceedings of LREC 2006*, Genoa, Italy.

[8] A. Chiorescu. 2005. The Explanatory Graphical Multimedia Dictionary. D.E.I. Multimedia (in Romanian: *Dicţionarul Explicativ Ilustrat Multimedia. D.E.I. Multimedia*). MSc. thesis, Faculty of Computer Science, "Al. I. Cuza" University of Iaşi.

[9] V. Ciubotariu. 2002. An XML annotation environment. Diploma thesis. Faculty of Computer Science, "Al.I.Cuza" University of Iaşi.

[10] S. Cojocaru. 1997. Romanian Lexicon: Tools, Implementation, Usage. In: Dan Tufiş & Poul Andersen (eds.). *Recent Advances in Romanian Language Technology*. Romanian Academy Publishing House, pp. 107–114.

[11] A. Colesnicov. 1995. The Romanian spelling checker ROMSP: the project overview. *Computer Science Journal of Moldova*, v. 3, Nr. 1(7), pp. 40–54.

[12] C. Coşman. 2002. The paradigmatic morphology of Romanian. Development and updating tool. (in Romanian: *Morfologia paradigmatică a limbii române. Mediu de dezvoltare / actualizare*). Dissertation thesis, Faculty of Computer Science, "Al. I. Cuza" University of Iaşi.

[13] O. Crăciun. 1998. GLOSS: Visual Instrument for discourse annotation (in Romanian: *GLOSS: Instrument vizual pentru adnotarea discursului*). Diploma thesis. Faculty of Computer Science. University Al. I. Cuza of Iaşi, Romania.

[14] D. Cristea, N. Ide, L. Romary. 1998. Veins Theory. An Approach to Global Cohesion and Coherence. In *Proceedings of Coling/ACL '98*, Montreal.

[15] D. Cristea, O. Crăciun, C. Ursu. 1998. GLOSS-A Visual Interactive Tools for Discourse Annotation. In *Proceedings of the Workshop on Annotation Tools, ESSLLI'98*, Saarbruecken.

[16] D. Cristea, D. Tufiş, 2002. "Romanian Linguistic Resources And Information Technologies Applied To The Romanian Language" (in Romanian). In Ichim O., F.T. Olariu (eds.) *The Identity Of The Romanian Language In The Globalisation Perspective* (in Romanian), Romanian Academy, the ,,A. Philippide" Institute for Romanian Philology, Trinitas Publishing House, Iaşi, pp. 211–234.

[17] D. Cristea. 2003. The relationship between discourse structure and referentiality in Veins Theory, in W. Menzel and C. Vertan (Eds.): *Natural Language Processing between Linguistic Inquiry and System Engineering*, "Al.I.Cuza" University Publishing House, Iaşi, pp. 9–22.

[18] D. Cristea, and C. Butnariu. 2004. Hierarchical XML representation for heavily annotated corpora, in *Proceedings of the LREC 2004 Workshop on XML-Based Richly Annotated Corpora*, Lisbon, Portugal.

[19] D. Cristea and O. Postolache. 2004. Designing Test-beds for General Anaphora Resolution, in *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium – DAARC*, St. Miguel, Portugal.

[20] D. Cristea, C. Mihăilă, C. Forăscu, D. Trandabăţ, M. Husarciuc, G. Haja, O. Postolache. 2004. Mapping Princeton WordNet Synsets onto Romanian Wordnet Synsets. In *Romanian Journal on Information Science and Technology*, Dan Tufiş (ed.) Special Issue on BalkaNet, Romanian Academy, 7(1-2).

[21] D. Cristea, and O.D. Postolache. 2005. How to deal with wicked anaphora, in António Branco, Tony McEnery and Ruslan Mitkov (editors): *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, Benjamin Publishing Books.

[22] D. Cristea, O. Postolache, I. Pistol. (2005): Summarisation through Discourse Structure, In Alexander Gelbukh (Ed.): *Computational Linguistics and Intelligent Text Processing, 6th International Conference CICLing 2005*, Mexico City, Mexico, February 2005, Proceedings, Springer LNSC, vol. 3406, ISBN 3-540-24523-5, pp. 632–644.

[23] D. Cristea. (2005): Motivations and Implications of Veins Theory, in Bernadette Sharp (Ed.) *Natural Language Understanding and Cognitive Science, Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science,*

NLUCS 2005, in conjunction with ICEIS 2005, Miami, U.S.A., May 2005, INSTICC Press, Portugal, ISBN 972-8865-23-6X, pp. 32–44.

[24] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02). Philadelphia, July.

[25] N. Curteanu, E. Amihăesei. 2004. Grammar-based Java Parsers for DEX and DLR Romanian Dictionaries. In *Proceedings of the Third European Conference on Intelligent Systems and Technologies - ECIT'2004*, Iaşi, Romania.

[26] N. Curteanu, E. Zlăvog, C. Bolea. 2005. Sentence-Level and Discourse Segmentation / Parsing with SCD Linguistic Strategy, H.-N. Teodorescu et al. (Eds.), *Proceedings of the Intelligent Systems Conference*, Performantica Press, Iaşi, p. 153–168.

[27] C. Florescu. 2005. Proposals and suggestions for the Informatised and Unified Dictionary of Romanian (DLRI). (in Romanian: *Propuneri şi sugestii privind Dicţionarul limbii române informatizat şi unificat (DLRI)*). Oral presentation at the Workshop *Language Resources and Tools for Romanian Language Processing*, Iaşi, November 2005. Available at http://consilr.info.uaic.ro/.

[28] C. Forăscu, D. Solomon. 2004. Towards a Time Tagger for Romanian. In *Proceedings of the ESSLLI Student Session*, Nancy, France.

[29] G. Haja, E. Dănilă, C. Forăscu, B.M. Aldea. 2005. The Dictionary of Romanian Language (DLR) in Electronic Format. Acquisition Studies. (in Romanian: *Dicţionarul Limbii Române (DLR) în format electronic. Studii privind achiziţionarea.*) Alfa Publishing House, Iaşi.

[30] G. Haja, C. Forăscu, B.M. Aldea, E. Dănilă. 2006. The dictionary of Romanian Language: steps toward the electronic version. (to appear) in *Proceedings of Euralex 2006*, Torino, Italy.

[31] V. Horbovanu. 2002. Word sense disambiguation using WordNet. (in Romanian: *Dezambiguizarea sensurilor cuvintelor folosind WordNet*). Diploma thesis. Faculty of Computer Science. University Al. I. Cuza of Iaşi, Romania.

[32] R. Ion and D. Tufiş. 2004. Multilingual Word Sense Disambiguation Using Aligned Wordnets. In *Romanian Journal on Information Science and Technology*, Dan Tufiş (ed.) Special Issue on BalkaNet, Romanian Academy, vol. 7, no. 2-3, pp. 198–214, ISSN 1453-8245.

[33] I. Mani. 2001. Automatic Summarization. John Benjamins.

[34] W.C. Mann, and S.A. Thompson. 1987. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. In *Text 8(3)*.

[35] Ch. D. Manning and H. Schütze. 2002. Foundations of Statistical Natural Language Processing. The MIT Press.

[36] D. Marcu. 2000. The Theory and Practice of Discourse Parsing and Summarization. The MIT Press.

[37] D. Marcu and D.Ş. Munteanu (2005). State of the Art in Statistical-Based Machine Translation: A Romanian-English Experiment, invited talk, EUROLAN-2005, Cluj-Napoca, 25 July – 6 August.

[38] I. Pistol. 2003. Automatic discourse segmentation (in Romanian: *Segmentarea automată a discursului*). Diploma thesis. Faculty of Computer Science of the "Al. I. Cuza" University of Iaşi.

[39] I. Pistol. 2005. Automatic parsing of linguistic discourse (in Romanian: *Parsarea automată a discursului lingvistic*). MSc. thesis.

Faculty of Computer Science of the "Al. I. Cuza" University of Iaşi.

[40] O. Postolache. 2001. Automatic Summarisation (in Romanian: *Sumarizare automată*), Diploma thesis, Faculty of Computer Science, "Al.I.Cuza" University of Iaşi.

[41] O. Postolache. 2004. RARE – Robust Anaphora Resolution Engine. MSc thesis, Faculty of Computer Science, "Al.I.Cuza" University of Iaşi.

[42] J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro M. Lazo. 2003. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*.

[43] G. Puşcaşu. 2003. Segmentation in elementary discourse units (in Romanian: *Segmentarea în unităţi de discurs elementare*). MSc. thesis, Faculty of Computer Science, "Al.I.Cuza" University of Iaşi.

[44] M. Răschip. 2003. Occurrence finder (in Romanian: *Ocurenţiator*). Diploma thesis, Faculty of Computer Science, "Al.I.Cuza" University of Iaşi.

[45] R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, J. Pustejovsky. 2005. TimeML Annotation Guidelines, Version 1.2.1, October 2005.

[46] S. Tamba. 2002. Web-page Categorisation. (in Romanian: *Categorizarea paginilor Web*). Diploma thesis, Faculty of Computer Science, "Al.I.Cuza" University of Iaşi.

[47] D. Tufiş. 1997. A Generalised Environment for Unification Based Natural Language Processing. In W. Teubert, R. Markincevicene (eds.) *Proceedings of the European Seminar on Language Resourses,* Kaunas.

[48] D. Tufiş. 1998. Tiered Tagging. In *International Journal on Information Science and Technology*, vol. 1, no. 2, Romanian Academy Publishing House, Bucharest.

[49] D. Tufiş and O. Mason. 1998. Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger. In *Proceedings of LREC 1998.*

[50] D. Tufiş, A. Chiţu. 1999. Automatic Insertion of Diacritics in Romanian Texts. In *Proceedings of the 5th International Workshop on Computational Lexicography COMPLEX*, Pecs, Ungaria.

[51] D. Tufiş. 2000. Using a Large Set of Eagles-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, Athens.

[52] D. Tufiş, A.M. Barbu. 2001. Automatic Construction of Translation Lexicons. In *Proceedings of the WSEAS and IEEE International Conference on Multimedia, Internet, Video Technologies*, ISBN: 960-8052-40-8, Malta.

[53] D. Tufiş, A.M. Barbu, R. Ion. 2003. TREQ-AL: A word-alignment system with limited language resources. In *Proceedings of the NAACL 2003 Workshop on Building and Using Parallel Texts; Romanian-English Shared Task*, Edmonton, Canada.

[54] D. Tufiş, E. Barbu, V. Barbu-Mititelu, R. Ion, L. Bozianu. 2004. The Romanian Wordnet. In *Romanian Journal on Information Science and Technology*, Dan Tufiş (ed.) Special Issue on BalkaNet, Romanian Academy, 7(1-2).

[55] D. Tufiş, E. Barbu. 2004. A Methodology and Associated Tools for Building Interlingual Wordnets. In *Proceedings of the 5th LREC Conference*, Lisabona, pp. 1067-1070.

[56] D. Tufiş, L. Dragomirescu. 2004. Tiered Tagging Revisited. In *Proceedings of the 4th LREC Conference*, Lisabona.

[57] D. Tufiş, A. Ceauşu, R. Ion, D. Ştefănescu. 2005. An integrated platform for high-accuracy word alignment. *JRC Enlargement and Integration Workshop: Exploiting parallel corpora in up to 20 languages*, Arona, Italy.

[58] C. Ursu. 1998. GLOSS: Visual Instrument for discourse annotation: validation and unification (in Romanian: *GLOSS: Instrument vizual pentru adnotarea discursului: validare şi unificare*). Diploma thesis. Faculty of Computer Science. University Al. I. Cuza of Iaşi, Romania.

[59] A.J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13(2):260–267. (The Viterbi decoding algorithm is described in section IV.)

D. Cristea[1,2], C. Forăscu[1,3],

Received May 2, 2006

[1]University "Al. I. Cuza" of Iaşi,
Faculty of Computer Science

[2]Institute for Computer Science,
Romanian Academy, Iaşi – Romania

[3]Institute for Artificial Intelligence,
Romanian Academy, Bucharest – Romania

e-mail: *dcristea@info.uaic.ro*,
corinfor@info.uaic.ro