Computer Science Journal of Moldova, vol.**19**, no.2(56), 2011

Romanian Linguistic Resources On Very Large Scale

Dan Cristea

1 Introduction

This paper suggests a methodology for building a technological environment for linguistic processing, intended to conserve, update and exploit, for research, for public and for commercial purposes, strategic linguistic resources of the Romanian language, rooted in textual data contributed daily and in the long run by important editorial houses and mass-media institutions. In essence, it describes a technology able to receive, store and continuously process large amounts of textual data, received from voluntary contributors, on a daily basis. Apart from storing linguistic data *à la longue* for the benefit of preserving the language, the results of the processing will be returned to three categories of users: the researchers working on Romanian language and computational linguistics, the contributors of the resources, and the public at large.

Such an initiative is motivated by the growing needs for linguistic resources, including textual data and processing tools, which are manifested in social sciences and humanities, and which should bring the Romanian language[1], now still less-resourced, to the level of technologically-rich languages of Europe. Raising the quantity of resources dedicated to different languages was a constant preoccupation in Europe over the past 15 years[2], triggered by the necessity to boost

[1]Spoken by 28.000.000 people (from which 23.501.683 native), according to Latin Union report http://www.ethnologue.com/show_language.asp?code=ron

[2]Here are some of the most important projects and coalitions aimed at building linguistic resources and processing capacities, financed by the EC, since 1995 till

the language industry to the level that makes text and speech fully machine interpretable media, in more and more complex applications.

The solution proposed in this paper will not only satisfy today, but should build the roots for a continuous observation of the language, in its evolution. Indeed, the permanent change of the language (Romanian of today is no more the same as that spoken or written in the middle years of the previous century, and this happens, to a greater of a less extend, to all living languages) makes language resources extremely volatile, as they become obsolete very quickly. As the language evolves, and sometimes our vision with respect to the linguistic phenomena changes, the resources, themselves, get old. Language resources should be kept aligned with the language evolution and the continuous update of the theoretical and computational views on language.

The proposal fulfils also important targets in the direction of language monitoring. Keeping a language under surveillance can be compared with monitoring a volcano which manifest some activity from time to time. Same as a volcano must be kept under strict observation, and different physical and chemical parameters must be continuously recorded and interpreted in order to signal possible eruptions, as such preventing damages on the population before being too late, a multitude of features of a language can be recorded and the direction of its evolution can be identified. Significant events in the evolution of a language have to be signalled, as are the acquisition of new words, new expressions, or the emergence of new senses. Tendencies must be perceived, such as a possible invasive influence from a foreign language, caused by its exaggerated exposure on public channels (TV, social web, etc.). If these factors are notified and signalled in time by a specialised service, then the adequacy and moment when an act should be engaged remains the attribute of appropriate decision factors (Academia, mass-media organisations or the university education system), in order to preserve the language and to keep its original spirit alive.

The scientific knowledge on language processing has reached to-

today: TELRI, MULTEXT, MULTEXT goes EAST, EUROWORDNET, BALKANET, LT4eL, CLARIN, FlareNet, and the ongoing METANET with its 4 satellite projects: T4ME, METANET4U, CESAR and META-NORD.

day an advanced level of competence internationally, also doubled by notable technological performances. Languages that are today most vividly supported theoretically and technologically do make use of rich collections of linguistic resources, continuously updated. These resources include corpora, i.e. collections of texts in original form, but also texts supporting annotation by experts, reflecting human competence over linguistic phenomena, which can be incorporated, through learning mechanisms, into automated systems. As resources are needed dramatically and many of them are very expensive, the issue of acquiring them should cease being episodic and must be driven by a national determination. Our initiative reflects the point of view that the linguistic resources of the languages spoken in a country should be considered of national interest.

2 Previous work

The proposal advanced in this paper is scientifically supported by a number of research achievements in the field of Romanian language technology and resources.

ALPE (Automated Linguistic Processing Environment) (Cristea and Butnariu, 2004; Cristea and Pistol, 2008; Pistol, 2011) is a theoretical framework for organising and exploiting the annotation added to texts. A hierarchical organization of a universe of annotation schemas and processing links makes possible the design of complex linguistic processing workflows. The model is intended to increase human skills to design linguistic processing tasks, to manipulate and re-use resources and tools, covering expertise levels that range from the expert down to the novice. The ALPE philosophy can stay at the base of the processing capabilities of the Portal, the storing and processing entity described below, featuring intelligent human-computer interaction capabilities.

The elaboration of eDTLR, the electronic form of the Thesaurus Dictionary of Romanian Language, where approximately half of its tremendous number of citations (about 1.3 million) have been linked onto the original scanned books (Cristea and Răschip, 2008; Cristea et al., 2009; Haja and Cristea, 2010), has opened a huge field of investi-

gation in Romanian lexicography and computational linguistics. The inventory of almost all words that have been used in the language in written form since the first known documents in Romanian, the extremely fine collection of word senses and their associated definitions, the high number of citations selected for each sense of each word, ordered chronologically, the indication of the etymological sources, and, as remarked, the associated sources, in scanned and OCR form, makes eDTLR one of the richest sources of linguistic data for Romanian language. Finally, there are also other outstanding resources for Romanian language: DEX (the Explanatory Dictionary of Romanian Language) with its online form[3], Romanian FrameNet (Trandabăţ, 2010; for English FrameNet, see Fillmore et al., 2002), Romanian VerbNet (Moruz, 2010; for English VerbNet, see Kipper-Schuler, 2005), Romanian WordNet (Tufiş et al., 2004; for the English WordNet, see Fellbaum, 1998), etc. These resources are more or less complete, but even when they will reach a satisfactory coverage of the language there will still remain the need to keep them updated with the evolution of language. The dynamics in language have to be mirrored in the strategic resources of the language as well.

In (Cristea, 2010), the idea of promoting a legislative initiative was investigated, that would impose to the producers of written texts (called *resourcers* in the paper: editing houses, recording houses, studios, etc.) the obligation to donate their linguistic resources in electronic form for the benefit of language research, without, this way, inducing any harm to them (producers or authors) as, for instance, induced by weakening their property control over the resources, or by commercial losses. The proposal advanced in this paper prepares the field for such a large scale implementation of the daring initiative to acquire language resources and to process them continuously, by considering only voluntary donations of publications in electronic form (mainly books, journals, magazines, newspapers and web publications).

[3]www.dexonline.ro

3 The Portal and its Repository

Technologically, the enterprise of sustaining a continuous flow of linguistic data can be fulfilled by a platform (let's call it *Portal*) capable to receive, store, process and make accessible to researchers, the public and the language industry large amounts of linguistic data on Romanian language.

The storage section of the *Portal* (let's call it *Repository*), basically, includes the following three important types of resources:

A). original documents – linguistic data in electronic form contributed by the voluntary donors (let's call them *resourcers*). The originals of these data are usually distributed on paper and/or electronic form on the culture and mass-media channels by the *resourcers*. In order to be easily retrieved, the source files need to suffer a series of transformations, including classification, indexing, statistical processing, archiving, assignment of persistent addresses, etc.

B). a set of representative, specialised and diachronic corpora of Romanian language, continuously updated. These are documents selected from the *Repository* by applying rigorous rules of corpus consistency and balance (Sinclair, 1996), and covering a long period in time. Part of these corpora will be automatically annotated (minimally, for token, lemma and part-of-speech) and metadata will be generated according to generally accepted standards, such as TEI (Burnard and Sperberg-McQueen, 1995).

C). a collection of synchronised Romanian linguistic thesauri, intended to be offered to researchers in the fields of social sciences, humanities and computational linguistics, kept updated out of the components A and B of the *Repository*. Minimally, there are: eDTLR, RoWordNet, RoVerbNet, and RoFrameNet. The synchronisation should be assured by automatic procedures capable to signal the occurrence in the *Repository* of pieces of language data that would fit as adequate entries in these resources.

Sections B+C make up the collection of *strategic resources*.

A significant library of web services, procedures that could be called online to perform elementary language processing tasks, facilitate the access to the Repository. A bunch of processing modules need to be integrated in an ALPE-like hierarchy, based on their input-output XML signatures. Although the idea from which ALPE emerged has been published 7 years ago (Cristea and Butnariu, 2004), a thorough theoretical model has only recently been finalised (Pistol, 2011) and a decisive implementation still waits to be realised.

4 Designing the functionality of the Portal

The realisation of the Portal should be based on a solid technological infrastructure. Moreover, the enterprise should be sustained by a coalition of researchers, from the Romanian speaking areal as well as from outside it, able to communicate and work together. It should not be neglected the formation of alliances with similar initiatives in Europe to envisage exchange of data, synchronisation of the data formats of web-services, creation of protocols for complex processing flows involving more languages, annotation standards and adoption of a unified processing framework (such as ALPE, for instance).

As for the Portal itself, the design of its functionalities should address *communication*, *storing* and *processing*.

The communication section addresses the exchange of information between the *Portal* and the community of users and third parties. The design of this section should see the *Portal* as a factory that processes words. On one side, in input, the raw textual material is received from the *resourcers*, voluntary contributors of the resources, and on the other side, in output, the consumers should be served. These are social sciences and humanity (SSH) researchers (in need for corpora, for aligned dictionaries and for sophisticated access onto the *strategic resources*), the contributors themselves (in need for services that would allow them to raise their profit, a kind of reward for their offered data), and the general public (usually browsing the primary textual data and dictionaries for contexts of occurrence, linguistic and cultural knowl-

edge, statistical data on the language, etc.).

The rapid accumulation of resources on the *Portal* imposes a perfect organisation of the storage section – the *Repository*. A farm of storing devices should be kept running continuously, on which efficient indexing algorithms should be used. To prevent data losses due to unexpected events (disasters, technical failures) a redundant storing architecture and mirroring techniques must be used.

A preliminary estimation of the hardware support needed for storing, based on data acquired by Serediuc (2010), shows that a capacity of 1 PB would be sufficient to host the electronic format of all books printed in Romania for a period of one century, including safety and backup.

It is important to note that the textual data should be recorded not only in their original format. Since the main intend is to allow targeted searches in the collection of text material of the *Repository* (including metadata, words and compounds, expressions, part-of-speeches, name entities, time anchored events, semantic relations, collocations, frequent terms and n-grams, syntactic structures, first or last known occurrences, absolute and relative frequencies, etc.), techniques to represent textual data in complementary form to its original string format, including indexes and XML-based formats shall be used. In this respect, the technology of Word Sketch (Kilgarriff et al., 2004; Macoveiciuc and Kilgarriff, 2010), which uses additional annotation attached to the word form, could be inspiring. In contrast with other massive initiatives for storing and processing textual data, which obtains the character strings of the primary data after processing the paper format (as are the Google Books initiative[4] or the Gutenberg project[5]), my proposal takes into consideration only accurate textual data, therefore clean texts. This is because texts will be contributed by the editing houses that own the original data, therefore avoiding the scanning process followed by transforming the images onto character strings by OCR.

[4] "History of Google Books", http://books.google.com/intl/en/googlebooks/history.html

[5] http://www.gutenberg.org/

The processing section represents the back spine of the Portal. An ALPE-like framework can implement the interoperability of basic components. Each document, once placed on the portal, should be submitted to a processing chain that includes, minimally: tokenization, part-of-speech tagging, lemmatization and indexing. This makes necessary that each raw text document be paired by a standoff XML annotation referring to it.

The ALPE framework will combine the basic functionalities mentioned above with other superior level functions that could be triggered by more advanced applications. These can include: syntactic parsing, segmentation at sentence and clause boundaries, identification of noun and verb phases, anaphora resolution, discourse parsing, summarisation, etc.

5 Keeping the data aligned and updated

A number of resources, which are considered of a strategic importance in keeping a language technologically fresh, have to be continuously interconnected on the *Repository*. Some of the most significant of these resources are: the electronic version of the Thesaurus Dictionary of Romanian Language (eDTLR), the Romanian FrameNet (RoFrameNet) and the Romanian VerbNet (RoVerbNet). None of them are finalised, but even when this will happen, they should be maintained updated with the evolution of language. In this section we propose a methodology that allow them mirror the significant changes that the language could stand. It is obvious that changes usually manifest slowly, and often there are controversies whether a certain linguistic or syntactic tendency should be recorded as being accommodated by the language. The same academic institutions will continue to take the final decisions, the way they do it now, but the technology can help them to better monitor the language, to track frequencies of occurrences, to detect first uses or depreciations.

We believe that each lexical item of the language must be represented by a *word file* on the *Repository*, and this record should include references in all the strategic resources. As such, to take an example,

the verb v's record is linked to its corresponding entry in eDTLR, where the inventory of senses is recorded, and these senses are aligned to the corresponding entries in RoVerbNet and RoFrameNet.

The Dictionary in its paper format, concentrates a significant and exquisite effort of the Romanian Academy for over a century (the last volume appeared in 2010). Originally, it has been printed in 36 volumes (19 – on the anastatic edition (Academy, 2010)), it contains more than 15,000 pages and about 175,000 entries, with citations collected from about 4,000 volumes of the written Romanian literature. The electronic version (eDTLR) was created during the years 2007-2010[6]. The entries are XML codified conforming to TEI P5[7].

Some of the most significant functions of the Portal are the dynamic discovery of contexts in the input documents (with or without the recognition of word senses), signalling of new words, signalling of new senses, signalling of obsolete words and senses, identifying the lexical entry in citations, etc. The process which should be placed at the base of recognizing new words and senses, as well as obsolete words and senses, presupposes placing a bag of words under constant surveillance. These are words/senses plausible of becoming recently popular or, on the contrary, becoming under-used. If we take the example of words manifesting a constantly degrading frequency, let's note that the criterion of absolute or even relative frequency, over a certain time interval, could prove not be relevant, because there are words which are very rarely used, although not being in danger of extinction (some science neologisms, for instance). The best way to do this is to associate to each word an individual *word file*, recording a set of dynamic features, among which the frequency of occurrence over time in specific registers (plots, out of which relative frequencies and gradients of deterioration over a constant interval of time, considered always back from the current day, could be computed).

It is evident that eDTLR contains already many features that makes it the perfect host of *word files*. Nevertheless, its entries should cer-

[6]Project no. 91_013/18.09.2007 under National Programmes II, section D, of the Romanian Ministry of Education, Research and Youth.

[7]http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

tainly be revised in order to transform eDTLR onto the repository of *word files*. Due to its peculiar importance on the panoply of Romanian resources, a special section of the *Repository* should be dedicated to eDTLR.

On the other hand, FrameNet and VerbNet are resources initially built for English but which have got a lot of attention world wide because of their ability to capture the semantic structure and syntactic constrains of the verbal constructions. Linking these resources onto eDTLR involves finding specific semantic structures fitted for each word in the dictionary (usually, words triggering events, such as verbs or predicative nouns).

At UAIC-FII, a number of tools intended to align different pieces of the strategic resources of the *Repository* have already been designed[8]. What needs to be done is to see the updating process synchronised with the flow of data continuously uploaded onto the *Portal*.

6 Addressing the users' needs

The technological apparatus described in this paper is useless if considered in isolation of the needs of *researchers*, *resourcers* and *ordinary citizens*. In this section I will discuss the needs of all these categories of users, in turn.

We have in mind different categories of *researcher* users, from experts in computational linguistics and NLP to SSH researchers, considered usually rather unskilful in the use of IT technologies. This last category may include social scientists, archaeologists, historians, geographers, linguists, lexicographers, etc., all in need of working on textual data with sophisticated NLP technology. In a recently ended project[9], Cristea et al. (2011) designed the lines of behaviour of an intelligent help desk addressing the needs of different categories of researchers working with textual material, from experts in NLP and IT

[8]For instance, links connecting eDTLR onto RoWN (Ivnescu, 2011), RoFrameNet (Trandab?, 2010), and RoVerbNet (Moruz, 2011).

[9]CLARIN – EC FP7 project no. 212230.

down to novices. A configurable interface should allow easy interaction of all these categories.

In a basic scenario, a researcher may access the *Portal* to ask for some type of processing on a file stored on the *Repository*, by following a dialogue with the Help Desk interface. In a more sophisticated scenario, the user could be interested to upload her/his own file, and initiate a processing chain using the language technology available on the *Portal*. The dialogue component of the interactive interface drives her/him onto a design process that should end up in putting together disparate processing modules (operating on the Portal or of which the Portal is aware of, although they operate distantly), thus configuring a complete or only a partial solution to the problem at hand.

Another important user type is the provider of linguistic material, the *resourcer* (Cristea, 2010). The services addressing this user type should be, as much as possible, free, as an award for their offer to donate linguistic data to the Portal. Examples of services addressing the resourcers are: advertising on books printed by the editing houses, search facilities to browse electronic versions of books, annotation and retrieval of citations, automatic summarisation, authors' indexing, statistics and plots regarding number of accesses, search criteria, time, location, etc. More elaborate demands, addressing cultivated readers, may include: search for exemplification of certain syntactic structures, occurrence of different types of semantic relations, collocations, frequent terms and n-grams in books, search by name entities of public interest, including VIPs, in different sources, automatic generation of genealogical trees of characters featuring in fiction or chronicles, tracing geographical journeys on travelling and historical books, etc. Many more suggestions of services made possible by sophisticated NLP technology can be collected from the volunteer *resourcers* and, eventually, be implemented on the Portal.

Finally, the third type of user is the *public* at large. We include here categories as school children and university students, people learning Romanian for the first time or trying to improve their foreign language skills on the Romanian language, Romanian natives in search for definitions, orthography rules, contexts of occurrence, old and contemporary

usage of words, but also IT companies interested to access the NLP technology offered by the Portal programmatically, for instance to include in their authored software APIs or NLP library codes. A significant library of Web-services should allow users to access the resources on the Repository, after they have been processed by the *Portal*, for goals other then research. A business model describing the exploitation of the resources on the Repository should be defined for this purpose.

7 Business model and IPR issues

The formation of a coalition of *resourcers*, potential partners supposed to accept to contribute on a continuous basis their resources to the Portal, is of an extreme importance. Without such contributions, the whole skeleton depicted in this visionary paper would prove weak and illusory. Bringing the owners of resources, mainly editorial houses, on-board means to engage a process of persuasion with them. A Memorandum of Understanding should be signed with each possible partner, with the intention to harmonise her/his interests with those of the project and stating clearly that neither the printing houses nor the author's IPRs over the texts will be affected. The main message delivered by the MoU is to discourage the partners' presumption that the alliance with the project would be potentially harmful for them. On the contrary, they should feel that, by entering into this coalition, incredible possibilities to exploit their data will be offered by the new technologies of intelligent text processing.

The contributor of the resource is in the position to decide whether the access of the end user could be free or bound to certain commercial restrictions. The general politics to be adopted is that if a fee should be imposed on the public area use then it should mainly return, as a benefit, to the proprietary of the corresponding resources. In any business model, the *Portal* should keep for itself only that part of the fee that would allow it to support the expenses. After an initial installation, the *Portal* should work on a self-sustained basis. The business model becomes thus essential for the long live of the enterprise.

From a business model point of view, two main types of uses can be

imagined depending on whether the user has or not a commercial interest. The general public, being triggered by a non-commercial interest in exploiting the data on the Portal, should, in principle, be excerpted from fees. The commercial access, in the benefit of organisations and companies needing NLP technologies for all kinds of applications, on the contrary, should be paid, and the benefits should go to the contributors of the resources, as mentioned above.

The other very important issue in collecting resources stays in not. They should become aware that none of their IPR are affected.

8 Conclusions

It is the author's believe that implementing the proposals advanced in this paper will boost the language processing capacities for Romanian language to a level compatible with technologically most advanced languages in Europe.

I am aware that the sophisticated technology described in this paper is not possible to be realised without a wealthy basis. Financial and human resources able to build the Portal and make it work should be looked for and deployed. A proper awareness campaign that would sensibilise responsible political decision makers and stakeholders towards accomplishing this goal, needs also be launched.

The project, if financed, will boost the interest of the Romanian speaking people towards applications based on advanced natural language technologies. It will open to researchers in SSH and computational linguistics and to the public an incredible amount of linguistic data, to be exploited for research on language and on social aspects that can be reflected in language. It will mean also a tremendous step forward for that part of the industry which is oriented on natural language applications and semantic web, more and more keen to bring natural language on mobiles and desk interfaces. The technological background described in this paper will facilitate the acquisition of the much wanted strategic resources for Romanian language and will open the gate towards keeping them aligned with the language evolution.

A recent initiative of the META-NET consortium finalised the

drafting of a series of Language White Papers for 30 European languages. The META-NET Language White Paper series "Languages in the European Information Society"[10] reports on the state of each European language with respect to Language Technology and explains the most urgent risks and chances. Summarising tables[11] indicate cluster-based rankings of all languages in four sample areas: text analysis, speech, machine translation, and resources. In both areas of text analysis and resources Romanian is shown as belonging to Cluster 4 (out of 5), meaning medium support: "research prototypes/resources exist, but quality and coverage varies"[12].

Finally, not least, such a wide enterprise of collecting resources cannot be realised without the consent of the owners of texts to donate their data. A recent investigation spotting few of the most important producers of printed information in Romania revealed that many editing houses are keen to donate their resources for research purposes, if they would gain the confidence of not being exposed to any commercial or property damage. Gaining the owners' trust that, on the contrary, donating linguistic data means thriving, not loosing, is still a task ahead us.

Acknowledgments

I am grateful to the ICT-PSP projects of the European Commission ATLAS (http://www.atlasproject.eu/) and METANET4U (http://metanet4u.eu/) for supporting part of the work described in this paper, and to my colleagues Lucian Gâdioi, Adrian Iftene and Diana Trandabăţ for their contributions to the ellaboration of a project proposal in lines with the objectives described in this paper.

[10] downloadable from http://www.meta-net.eu/intranet/language-whitepapers/files-for-publication/whitepapers

[11] These statistics, very fresh, at the moment of drafting this paper are not yet made public.

[12] Where Cluster 1 "excellent LT support: "technologies/resources exist that are in widespread use and cover practically all linguistic phenomena – vocabulary, compounds, grammar, metaphors etc. – of a language" shows no entries, and Cluster 5 means "low to almost no support: from the drawing board to rudimentary prototypes – very limited quality and coverage, toy systems".

References

[1] *** (2010) *Dictionary of the Romanian Language (in Romanian)*, anastatic edition, following Dictionary of the Romanian Language (DA) and Dictionary of the Romanian Language (DLR), Romanian Academy Printing House, Bucharest.

[2] Burnard, L., Sperberg-McQueen, C.M. (1995). *The Design of the TEI Encoding Scheme*, Computers and the Humanities, 29 (1).

[3] Cristea, D. (2010). *Very large language resources? At our finger!* In Proceedings of the Workshop Language Resources: From Storyboard to Sustainability and LR Lifecycle Management, LREC-2010, Valleta.

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

[5] Cristea, D., Pistol, I. (2008). *Managing Language Resources and Tools Using a Hierarchy of Annotation Schemas.* Proceedings of the Workshop on Sustainability of Language Resources, LREC-2008, Marrakech.

[6] Fillmore, C. J., Baker, C. F., and Sato, H. (2002). *The framenet database and software tool.* In Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas.

[7] Ivănescu, M.-L. (2011). *Automatic Techniques for filling in the Romanian WordNet out of eDTLR (in Romanian),* graduation paper, "Alexandru Ioan Cuza" University, Faculty of Computer Science, Iaşi.

[8] Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004). *The Sketch Engine*, Proc. Euralex, Lorient.

[9] Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon.* Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA.

[10] Macoveiciuc, M. and Kilgarriff, A. (2010). *The RoWaC Corpus and Romanian Word Sketches* In: Multilinguality and Interoperability in Language Processing with Emphasis on Romanian. Edited by Dan Tufiş and Corina Forăscu, Romanian Academy Publishing House, Bucharest.

[11] Moruz, M.A. (2011). *Predication Driven Textual Entailment*, Ph.D. thesis, "Alexandru Ioan Cuza" University, Faculty of Computer Science, Iaşi.

[12] Pistol, I. (2011). *The Automated Processing of Natural Language*, Ph.D. thesis, "Alexandru Ioan Cuza" University, Faculty of Computer Science, Iaşi.

[13] Serediuc, F. (2010). *High Volume Textual Processing (in Romanian)*, graduation thesis, "Alexandru Ioan Cuza" University, Faculty of Computer Science, Iaşi.

[14] Simionescu, R., Cristea, D. (2011). *Help-desk and registry*, CLARIN report M6C-3.3.

[15] Trandabăţ, D. (2010). *Natural Language Processing Using Semantic Frames*, Ph.D. thesis, "Alexandru Ioan Cuza" University, Faculty of Computer Science, Iaşi.

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

Dan Cristea Received October 13, 2011

Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi
Institute of Computer Science, Romanian Academy, the Iaşi branch
E–mail: *dcristea@info.uaic.ro*