

**METANET4U** 

**D2.3.ro.en**  
**Language Report for**  
**Romanian**  
**(English version)**

Version 1.2

2011-07-29



# METANET4U

[www.metanet4u.eu](http://www.metanet4u.eu)

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.

This central objective is articulated in terms of the following main goals:

**Assessment:** to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.

**Collection:** to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.

**Distribution:** to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.

**Dissemination:** to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

METANET4U is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META. META is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society.



Deliverable D2.3.ro.en: Language Report for Romanian (English version)

METANET4U is co-funded by the participating institutions and the ICT Policy Support Programme of the European Commission



and by the participating institutions:



Faculty of Sciences, University of Lisbon



Instituto Superior Técnico



University of Manchester



University *Alexandru Ioan Cuza*



Research Institute for Artificial Intelligence,  
Romanian Academy



University of Malta



Technical University of Catalonia



Universitat Pompeu Fabra

Revision History

Version	Date	Author	Organisation	Description
1.0	5-07-2011	Diana Trandabăț, Elena Irimia, Verginica Mititelu- Barbu, Dan Tufiș, Dan Cristea	UAIC, RACAI	Draft version
1.1	17-07-2011	Diana Trandabăț, Elena Irimia, Verginica Mititelu- Barbu, Dan Tufiș, Dan Cristea	UAIC, RACAI	Pre-final version
1.2	29-07-2011	Diana Trandabăț, Elena Irimia, Verginica Mititelu- Barbu, Dan Tufiș, Dan Cristea	UAIC, RACAI	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.



**METANET4U**

**D2.3.ro.en**  
**Language Report for**  
**Romanian**  
**(English version)**

Document METANET4U-2011-D2.3.ro.en  
EC CIP project #270893

Deliverable  
Number: D2.3.ro.en  
Completion: Final  
Status: Submitted  
Dissemination level: Public

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: UAIC, RACAI

Authors: Diana Trandabăț, Elena Irimia, Verginica Mititelu-Barbu,  
Dan Tufiș, Dan Cristea

Reviewer: Amália Mendes, Adrian Iftene, Ionuț Pistol

© all rights reserved by FCUL on behalf of METANTE4U



## Table of Contents

<b>Executive Summary .....</b>	<b>6</b>
<b>A Risk for Our Languages and a Challenge for Language Technology .....</b>	<b>8</b>
Language Borders Hinder the European Information Society.....	8
Our Languages at Risk.....	9
Language Technology is a Key Enabling Technology .....	9
Opportunities for Language Technology .....	10
Challenges Facing Language Technology.....	11
Language Acquisition .....	11
<b>Romanian in the European Information Society .....</b>	<b>13</b>
General Facts .....	13
Particularities of the Romanian Language .....	13
Recent developments.....	16
Language cultivation in Romania .....	16
Language in Education.....	17
International aspects .....	18
Romanian on the Internet.....	18
Selected Further Reading.....	19
<b>Language Technology Support for Romanian .....</b>	<b>20</b>
Language Technologies .....	20
Language Technology Application Architectures .....	20
Core application areas.....	21
<i>Language checking</i> .....	21
<i>Web search</i> .....	23
<i>Speech interaction</i> .....	25
<i>Machine translation</i> .....	26
Language Technology.....	28
Language Technology in Education .....	31
LT Industry and Programs .....	31
LT Research and Education .....	32
Status of Tools and Resources for Romanian.....	34
Table of Tools and Resources for Romanian.....	35
Conclusions .....	37
<b>About META-NET.....</b>	<b>39</b>
Lines of Action .....	39
Member Organisations .....	41
<b>References .....</b>	<b>44</b>

## Executive Summary

Many European languages run the risk of becoming victims of the digital age because they are underrepresented and under-resourced online. Huge regional market opportunities remain untapped today because of language barriers. If we do not take action now, many European citizens will become socially and economically disadvantaged because they speak their native language.

Innovative, language technology (LT) is an intermediary that will enable European citizens to participate in an egalitarian, inclusive and economically successful knowledge and information society. Multilingual language technology will be a gateway for instantaneous, cheap and effortless communication and interaction across language boundaries.

Today, language services are primarily offered by commercial providers from the US. Google Translate, a free service, is just one example. The recent success of Watson, an IBM computer system that won an episode of the Jeopardy game show against human candidates, illustrates the immense potential of language technology. As Europeans, we have to ask ourselves several urgent questions:

- ❑ Should our communications and knowledge infrastructure be dependent upon monopolistic companies?
- ❑ Can we truly rely on language-related services that can be immediately switched off by others?
- ❑ Are we actively competing in the global market for research and development in language technology?
- ❑ Are third parties from other continents willing to address our translation problems and other issues that relate to European multilingualism?
- ❑ Can our European cultural background help shape the knowledge society by offering better, more secure, more precise, more innovative and more robust high-quality technology?

This whitepaper for the Romanian language demonstrates that research in universities and academia was successful in designing particular high quality software, as well as models and theories widely applicable. However, for the further development of the LT domain in Romania, a more vivid implication of the Government through adequate financing should be obtained, as well as promoting attractive collaborations with the industries that use or provide LT.

META-NET contributes to building a strong, multilingual European digital information space. By realising this goal, a multicultural union of nations can prosper and become a role model for peaceful and egalitarian international cooperation. If this goal cannot be achieved, Europe will have to choose between sacrificing its cultural identities or suffering economic defeat.





## A Risk for Our Languages and a Challenge for Language Technology

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in digitised and network communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

*We are currently witnessing a digital revolution that is comparable to Gutenberg's invention of the printing press.*

After Gutenberg's invention, real breakthroughs in communication and knowledge exchange were accomplished by efforts like Luther's translation of the Bible into common language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled an exchange across languages;
- the creation of journalistic and bibliographic guidelines assured the quality and availability of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology helped to automate and facilitate many of the processes:

- desktop publishing software replaces typewriting and typesetting;
- Microsoft PowerPoint replaces overhead projector transparencies;
- e-mail sends and receives documents faster than a fax machine;
- Skype makes Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- search engines provide keyword-based access to web pages;
- online services like Google Translate produce quick and approximate translations;
- social media platforms facilitate collaboration and information sharing.

Although such tools and applications are helpful, they currently cannot sufficiently implement a sustainable, multilingual European information society, a modern and inclusive society where information and goods can flow freely.

### Language Borders Hinder the European Information Society

We cannot precisely know what the future information society will look like. When it comes to discussing a common European energy strategy or foreign policy, we might want to listen to European foreign ministers speak in their native language. We might want a

platform where people, who speak many different languages and who have varying language proficiency, can discuss a particular subject while technology automatically gathers their opinions and generates brief summaries. We also might want to speak with a health insurance help desk that is located in a foreign country.

It is clear that communication needs have a different quality as compared to a few years ago. In a global economy and information space, more languages, speakers and content confront us and require us to quickly interact with new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter and YouTube) is only the tip of the iceberg.

*A global economy and information space confronts us with more languages, speakers and content.*

Today, we can transmit gigabytes of text around the world in a few seconds before we recognize that it is in a language we do not understand. According to a recent report requested by the European Commission, 57% of Internet users in Europe purchase goods and services in languages that are not their native language. (English is the most common foreign language followed by French, German and Spanish.) 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the web<sup>i</sup>. A few years ago, English might have been the lingua franca of the web—the vast majority of content on the web was in English—but the situation has now drastically changed. The amount of online content in other languages (particularly Asian and Arabic languages) has exploded.

An ubiquitous digital divide that is caused by language borders has surprisingly not gained much attention in the public discourse; yet, it raises a very pressing question, “Which European languages will thrive and persist in the networked information and knowledge society?”.

*Which European languages will thrive and persist in the networked information and knowledge society?*

## Our Languages at Risk

The printing press contributed to an invaluable exchange of information in Europe, but it also led to the extinction of many European languages. Regional and minority languages were rarely printed. As a result, many languages like Cornish or Dalmatian were often limited to oral forms of transmission, which limited their continued adoption, spread and use.

The approximately 60 languages of Europe are one of its richest and most important cultural assets. Europe’s multitude of languages is also a vital part of its social success<sup>ii</sup>. While popular languages like English or Spanish will certainly maintain their presence in the emerging digital society and market, many European languages could be cut off from digital communications and become irrelevant for the Internet society. Such developments would certainly be unwelcome. On the one hand, a strategic opportunity would be lost that would weaken Europe’s global standing. On the other hand, such developments would conflict with the goal of equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society<sup>iii</sup>.

*The wide variety of languages in Europe is one of its most important cultural assets and an essential part of Europe’s success.*

## Language Technology is a Key Enabling Technology

In the past, investment efforts have focused on language education and translation. For example, according to some estimates, the

European market for translation, interpretation, software localisation and website globalisation was € 8.4 billion in 2008 and was expected to grow by 10% per annum.<sup>iv</sup> Yet, this existing capacity is not enough to satisfy current and future needs.

Language technology is a key enabling technology that can protect and foster European languages. Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates regardless of language barriers or computer skills. Language technology already assists everyday tasks, such as writing e-mails, conducting an online search or booking a flight. We benefit from language technology when we:

- ❑ find information with an Internet search engine;
- ❑ check spelling and grammar in a word processor;
- ❑ view product recommendations at an online shop;
- ❑ hear the verbal instructions of a navigation system;
- ❑ translate web pages with an online service.

The language technologies detailed in this paper are an essential part of innovative future applications. Language technology is typically an enabling technology within a larger application framework like a navigation system or a search engine. These white papers focus on the readiness of core technologies for each language.

In the near future, we need language technology for all European languages that is available, affordable and tightly integrated within larger software environments. An interactive, multimedia and multilingual user experience is not possible without language technology.

## Opportunities for Language Technology

Language technology can make automatic translation, content production, information processing and knowledge management possible for all European languages. Language technology can also further the development of intuitive language-based interfaces for household electronics, machinery, vehicles, computers and robots. Although many prototypes already exist, commercial and industrial applications are still in the early stages of development. Recent achievements in research and development have created a genuine window of opportunity. For example, machine translation (MT) already delivers a reasonable amount of accuracy within specific domains, and experimental applications provide multilingual information and knowledge management as well as content production in many European languages.

Language applications, voice-based user interfaces and dialogue systems are traditionally found in highly specialised domains, and they often exhibit limited performance. One active field of research is the use of language technology for rescue operations in disaster areas. In such high-risk environments, translation accuracy can be a matter of life or death. The same reasoning applies to the use of language technology in the health care industry. Intelligent robots with cross-lingual language capabilities have the potential to save lives.

There are huge market opportunities in the education and entertainment industries for the integration of language technologies in games, edutainment offerings, simulation environments or training programmes. Mobile information services, computer-assisted language learning software, e-learning environments, self-assessment

*Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates across different languages.*

tools and plagiarism detection software are just a few more examples where language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggests a further need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology represents a tremendous opportunity for the European Union that makes both economic and cultural sense. Multilingualism in Europe has become the rule. European businesses, organisations and schools are also multinational and diverse. Citizens want to communicate across the language borders that still exist in the European Common Market. Language technology can help overcome such remaining barriers while supporting the free and open use of language. Furthermore, innovative, multilingual language technology for Europe can also help us communicate with our global partners and their multilingual communities. Language technologies support a wealth of international economic opportunities.

*Multilingualism is the rule, not an exception.*

## Challenges Facing Language Technology

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. We cannot wait ten or twenty years for significant improvements to be made that can further communication and productivity in our multilingual environment.

*The current pace of technological progress is too slow to arrive at substantial software products within the next ten to twenty years.*

Language technologies with broad use, such as the spelling and grammar features in word processors, are typically monolingual, and they are only available for a handful of languages. Applications for multilingual communication require a certain level of sophistication. Machine translation and online services like Google Translate or Bing Translator are excellent at creating a good approximation of a document's contents. But such online services and professional MT applications are fraught with various difficulties when highly accurate and complete translations are required. There are many well-known examples of funny sounding mistranslations, for example, literal translations of the names *Bush* or *Kohl*, which illustrates the challenges language technology must still face.

## Language Acquisition

To illustrate how computers handle language and why language acquisition is a very difficult task, we take a brief look at the way humans acquire first and second languages, and then we sketch how machine translation systems work—there's a reason why the field of language technology is closely linked to the field of artificial intelligence.

Humans acquire language skills in two different ways. First, a baby learns a language by listening to the interaction between speakers of the language. Exposure to concrete, linguistic examples by language users, such as parents, siblings and other family members, helps babies from the age of about two or so produce their first words and short phrases. This is only possible because of a special genetic disposition humans have for learning languages.

*Humans acquire language skills in two different ways: learning examples and learning the underlying language rules.*

Learning a second language usually requires much more effort when a child is not immersed in a language community of native speakers. At school age, foreign languages are usually acquired by learning their grammatical structure, vocabulary and orthography

from books and educational materials that describe linguistic knowledge in terms of abstract rules, tables and example texts. Learning a foreign language takes a lot of time and effort, and it gets more difficult with age.

The two main types of language technology systems acquire language capabilities in a similar manner as humans. Statistical approaches obtain linguistic knowledge from vast collections of concrete example texts in a single language or in so-called parallel texts that are available in two or more languages. Machine learning algorithms model some kind of language faculty that can derive patterns of how words, short phrases and complete sentences are correctly used in a single language or translated from one language to another. The sheer number of sentences that statistical approaches require is huge. Performance quality increases as the number of analyzed texts increases. It is not uncommon to train such systems on texts that comprise millions of sentences. This is one of the reasons why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, available online information, and translation services such as Google Search and Google Translate rely on a statistical (data-driven) approach.

Rule-based systems are the second major type of language technology. Experts from linguistics, computational linguistics and computer science encode grammatical analysis (translation rules) and compile vocabulary lists (lexicons). The establishment of a rule-based system is very time consuming and labour intensive. Rule-based systems also require highly specialised experts. Some of the leading rule-based machine translation systems have been under constant development for more than twenty years. The advantage of rule-based systems is that the experts can have more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. Due to financial constraints, rule-based language technology is only feasible for major languages.

*The two main types of language technology systems acquire language in a similar manner as humans.*

# Romanian in the European Information Society

## General Facts

Spoken by over **29,000,000 speakers**, Romanian is mother tongue for approx. 25,000,000 speakers: around 21,500,000 speakers in Romania<sup>v</sup> plus approx. 3,500,000 speakers in the Republic of Moldavia<sup>vi</sup> (where the language is officially called Moldavian). The countries around Romania (Albania, Bulgaria, Croatia, Greece, Hungary, The Former Yugoslav Republic of Macedonia, Serbia, Ukraine) and communities of immigrants in Australia, Canada, Israel, Latin America, Turkey, U.S.A. and other European and Asian countries totals around 4,000,000 Romanian native speakers<sup>vii</sup>.

Romanian is an official language also in the Autonomous Province of Vojvodina in Serbia, in the autonomous Mount Athos in Greece, in the European Union and in the Latin Union; it is a recognized minority language in Ukraine.

Romanian has **4 dialects**<sup>viii</sup>: Daco-Romanian/ Romanian, Aromanian (spoken by approximately 600.000 speakers in Albania, Bulgaria, Greece and Macedonia), Istro-Romanian (15,000 speakers in 2 small areas in the Istrian Peninsula, Croatia) and Megleno-Romanian (about 5,000 speakers in Greece and Macedonia). Because of their small number of speakers, these dialects are included in the *UNESCO Red Book of Endangered Languages*.

In Romania there are 18 officially recognized national (ethnic) minorities; in the last Census (2002), the most numerous were Hungarians (1,431,807) and Romas (535,140), followed by Germans, Ukrainians, Lippovan Russians, Turks, Serbs, Croats, Slovenes, Tartars, Slovaks, Bulgarians, Jewish, Czechs, Poles, Greeks, Armenians, etc. For all these minorities, official language policies in Romania guarantee their rights to be protected as language communities and to use their own languages in private and public, culturally and socially, in economy and in communication media. However, article 13 of the Constitution states that “In Romania, the official language is Romanian”. Moreover, Law number 500 from 12<sup>th</sup> November, 2004 stipulates the obligation of any text (either oral or written) that serves public interest to be translated or adapted into Romanian<sup>ix</sup>.

## Particularities of the Romanian Language

Developed at distance from the other languages in the Romance family, Romanian is an eastern Romance language. Elements of the Vulgar Latin from which it descends are more faithfully preserved in this isolated language: it has inherited the Latin morpho-syntactic structure, preserved features that other Romance languages have lost (such as declinations), and incorporated some non-Romance features in its structure (-o vocatives, the neuter gender).

The great part of the Romanian vocabulary has a Latin origin, either inherited from Vulgar Latin or borrowed in modern times from Latin. 60% of the fundamental vocabulary (i.e. the words that are known and used by all speakers of a language) is inherited from Latin.

During Roman colonization of Dacia (106-271 A.D.), the colonizers imposed Latin as the official language. However, comparative stud-



ies of Romanian and Albanian vocabularies reveal a set of around 100 words that have been preserved from the Thraco-Dacian substratum. These words designated fundamental concepts, like body parts, natural elements or food. They are still used today, are very frequent and with rich polysemy and lexical families.

During the migration of Slavic tribes over the territory of nowadays Romania, the language underwent a process of transformation in all its compartments: phonetics, lexis, morphology and syntax. However, morphology, the backbone of a language, remained Latin in most of its aspects. The Cyrillic alphabet was adopted in this period, especially due to the church influence. The old Slavonic was the liturgical language of the Romanian Orthodox Church until the late 18<sup>th</sup> century, when Romanian started a process of re-latinization, modernization and westernization. It is now when many words of other origin are replaced by Latin words, borrowed directly or indirectly, via other Romance languages (French and Italian). French as a language of culture in the last 2 centuries and France as a place where the Romanian aristocracy sent their children to school justify the existence of extremely numerous words of this origin in Romanian.

Lately, English took the place of French and Romanian has many Anglicisms, entirely, partially or at all adapted to its phonetic and morphologic systems. Political, economic and social aspects in the history of Romania explain the words of various other origins in this language: Turkish, Greek, German, Hungarian, Bulgarian, Russian etc. New words have been created in Romanian mostly through suffix derivation. However, recent studies reveal the importance prefix derivation has got lately.

Romanian has 5 letters using diacritics: ă, î, â, ș, ț. For the last 2, two variants have circulated: one with a comma under the letter, and another one with a cedilla. However, only the former is recommended nowadays by the Romanian National Standardization Body (ASRO).

*Romanian has five letters using diacritics: ă, î, â, ș, ț. For the last two, a couple of variants have circulated: one with a comma under the letter, and another one with a cedilla. However, only the former is recommended.*

Many electronic texts are not written with diacritics. In order to automatically introduce diacritics, programs have been created to recover them in such texts.

Romanian inflection is quite rich: for nouns, pronouns and adjectives there are five cases and two numbers, for verbs there are two numbers, each with three persons, and five synthetic tenses, plus infinitive, gerund and participle forms. In average, a noun can have 5 forms, a personal pronoun about 6 forms, an adjective around 6 forms, a verb has more than 30 forms. Besides morphologic suffixes and endings, phonetic alternations inside the root can be also used to inflect words.

Romanian is a subject pro-drop language, like most of its Romance sisters, that is, it allows the deletion of the subject:

*Știu.*

*Know-I*

*"I know."*

The explanation resides in the rich inflectional systems of verbs that have distinctive endings for different persons and numbers.

Nevertheless, subject doubling is also possible in Romanian when a personal pronoun doubles a lexical noun phrase:



*Vine el tata imediat!*

*Comes he father-the immediately!*

*“Father will come immediately!”*

The structure is characteristic of the familiar use of language, marking a certain illocutionary attitude of the speaker: threat, promise, and reassurance.

Romanian has in common with some Spanish dialects and several Balkan languages a structure currently known as ‘clitic doubling’. Pronominal clitic doubling in Romanian may be realized with accusative clitics, with dative ones or with both. For example, in the sentence:

*I<sub>i</sub> l<sub>j</sub>-am dat mamei<sub>i</sub> pe Ion<sub>j</sub> la telefon.*

*Dat.cl. Acc.masc.cl.-have-I given to-mother PE John on phone.*

*“I gave John to my mother on the phone.”*

The noun *mamei* and the Dative clitic *i* refer to the same person, and the Accusative clitic *l-* and the Accusative noun *Ion* are also coreferential. The presence of clitics in such constructions is mandatory, although they do not saturate any verbal valences. However, when the nouns are not present, it is the task of these pronouns to saturate the verbal valences:

*I l-am dat la telefon.*

*To-her him-have-I given on phone.*

*“I gave him to her on the phone.”*

The clitic doubling phenomenon is obligatory with proper names and definite nouns.

Romanian displays both Negative Concord and Double Negation. The presence of the negative marker *nu* “not” in the verbal phrase negates the sentence and licenses negative words in the respective sentence (negative concord):

*Nu am văzut pe nimeni niciodată aici.*

*Not have-I seen PE nobody never here.*

*“I have never seen anybody here.”*

However, certain configurations in which the negative markers and words occur trigger the double negation (that is, the sentence acquires a positive meaning). For instance, a negative main clause followed by a negative subjunctive clause is such a configuration with overall positive meaning:

*Maria nu a vrut să nu spună nimic.*

*Maria not has wanted SĂ not say nothing.*

*“Maria did not want to say nothing.” = “Maria wanted to say something.”*

Case is inflectional in Romanian. However, there are also three case marking prepositions: *pe* for Accusative (conditioned by the animacy, definiteness and specificity features of the nominal phrase), *la* for Dative and *a* for Genitive (both of them conditioned by the presence of numerals in the nominal phrase):

*Romanian is a highly inflected language, with various linguistic particularities: it is a subject pro-drop language (that is, it allows the deletion of the subject), allows for subject and clitic doubling, displays both Negative Concord and Double Negation.*

*L-am văzut pe colegul meu.*

*Acc.masc.cl.-have-I seen PE colleague-the my.*

*“I have seen my colleague.”*

*Am dat cărțile la trei dintre ei.*

*Have-I given books-the LA three of them.*

*“I gave the books to three of them.”*

*Cărțile a trei copii erau noi.*

*Books-the A three children were new.*

*“The books of three children were new.”*

## Recent developments

Analogue to the re-latinization phase in the 19<sup>th</sup> century after the liberation from the Greek and Turkish domination, Romanian language was passing in the last 20 years through a process of transformation from the totalitarian usage (“langue de bois”, unidirectional discourse, etc.) to an open usage in which new linguistic patterns must adapt to the social and cultural transition. Therefore, similar to many other languages, Romanian is going through a continuous process of internationalization under the influence of the Anglo-Saxon vocabulary.

In essential domains like political, administrative and economic sciences, media, advertising, computers, etc. substantial loans and semantic extensions from English occurred; terminologies in new fields are based on English loans, the active vocabulary of educated people contains more and more anglicisms, new intonation patterns can be observed (especially in media).

In some areas, anglicisms have started to replace existing Romanian vocabulary. One example is the use of English titles in job advertisements, in particular for executive positions, e.g. ‘Human Resource Manager’ instead of *Director de Resurse Umane*. A strong tendency to overuse anglicisms can also be detected in product advertisements. Banks in Romania use for promotion slogans such as: *Cu cine faci banking?* or *Prima modalitate de plată contactless*, although banking or contactless are anglicisms that most Romanians aren’t used to.

The example demonstrates the importance of raising awareness for a development that runs the risk of excluding large parts of the population from taking part in information society, namely those who are not familiar with English.

## Language cultivation in Romania

The Romanian Academy, Romania’s highest cultural forum, has as one of its main objectives the cultivation of national language. The major goal of its linguistic institutes was building and publishing *Dicționarul Tezaur al Limbii Române* (the Thesaurus Dictionary of the Romanian Language), a process which took almost a century. The old series, known as *Dicționarul Academiei* (Dictionary of the Academy - DA) includes 5 volumes with 3,146 pages and 44.890 entries, and has been developed between 1913 and 1947. After an

interruption, the work was restarted in the middle of the 7th decade of the last century with the new series, known as *Dicționarul Limbii Române* (the Dictionary of the Romanian Language - DLR). The last volume was finally published by the Editing House of the Romanian Academy at the beginning of 2009. In total, DA and DLR have 33 volumes, more than 15,000 pages and about 175,000 entries. The dictionary was created in the traditional pencil-and-paper way, with citations collected from more than 2,500 volumes of the written Romanian literature.

The Institute of Linguistics “Iorgu Iordan – Al. Rosetti” has a research program focusing on language cultivation. They elaborate normative dictionaries (*Dicționarul împrumuturilor neadaptate* “Dictionary of non-adapted words”, *Dicționarul termenilor oficiali* “Dictionary of official terms”, *Dicționar ortografic, ortoepic și morfologic al limbii române* “Orthographic, orthoepic and morphologic dictionary of Romanian”) and grammars (*Gramatica limbii române* ‘Romanian Language Grammar’, *Dinamica limbii române actuale* “The Dynamics of Contemporary Romanian”).

Law 500 of 12<sup>th</sup> November 2004 states that all written or spoken texts in Romanian that serve the public interest must conform to the norms established by the Romanian Academy.

Institutul Limbii Române (The Institute of the Romanian Language) was created with the aims of promoting Romanian language learning abroad, supporting learners of Romanian and attesting their knowledge of Romanian<sup>x</sup>. There are over 70 international centres where Romanian is taught as a foreign language by Romanian university teachers.

*There are over 70 international centres where Romanian is taught as a foreign language by Romanian university teachers.*

In Romania there is also an increasing interest for studying Romanian among foreigners, not only at diplomatic level (by representatives of various diplomatic missions of different countries), but also by business people. Besides universities that offer Romanian as foreign language classes (usually for foreign students in Romania), there are numerous private firms with classes offered in general to foreigners involved in the economic sector. Romanian summer courses for all levels are organized annually by the Romanian Cultural Foundation in various places of the country and by several high education institutions (such as University “Al. I. Cuza” of Iași).

Language cultivation in the context of accelerated innovation is a priority also for media. National radio and television channels have programmes in which tricky aspects of language are discussed with specialists and explained to the audience.

## Language in Education

According to the New National Curriculum (2000) Romanian is taught for 4-5 compulsory classes per week in secondary school and for 3-4 compulsory classes in high school. Prescriptive aspects of language preservation are combined with communication as skilled behaviour and the language-culture relation is emphasized. Romanian language and literature are compulsory subjects for national exams (graduation exam from secondary school and graduation exam from high school; the latter involves two kinds of examination: oral and written).

Romanian language and literature are studied as major and minor subjects in more than 30 state and private universities throughout Romania.

## International aspects

Romania is internationally known for its literature, the major works of Eminescu (the great national poet of Romania) being translated to over 60 languages. Other known names of the Romanian literature are: Mircea Eliade, the first to write a History of Religions; Eugen Ionesco, one of the foremost playwrights of the Theatre of the Absurd; or Emil Cioran's philosophy.

Nowadays, the large majority of the scientific publications in the LT field are written in English, although a Consortium for the Digitalization of Romanian Language – ConsILR – organizes annually a scientific workshop dedicated to research in LT regarding the Romanian language, with the proceedings written in Romanian. The same situation goes for other domains also, possibly being less prominent for disciplines such as law, philosophy, linguistics or theology.

Similarly, this is true of the business world. In many large and internationally active companies, English has become the *lingua franca*, both in written (emails and documents) and oral communication (e.g. talks), especially in multinational firms with foreign management.

Language technology can address this challenge from a different perspective by offering services like Machine Translation or cross-lingual information retrieval to foreign language text and thus help diminish personal and economic disadvantages naturally faced by non-native speakers of English.

Romanian minorities live in neighbouring countries and in Diaspora communities all over the world. Romania promotes policies for language and cultural identity preservation of the Romanian communities. The “Euxodius Hurmuzachi” Centre offers hundreds of scholarships a year in Romania for Romanian minorities from neighbouring countries. There are many school and academic exchanges, especially with the Republic of Moldavia. The first Romanian school and university extensions through franchising appeared in the Republic of Moldavia in 2000.

In different communities from the Diaspora, there are various initiatives through which those interested can study Romanian language and culture. For instance, Romanian Language School in Kitchener, Canada, teaches Romanian language and culture classes to children and teenagers.

Romanian Cultural Institutes are established in 19 cities all over the world (including Bucharest, New York, Paris, London, Roma, Istanbul, etc.) and they all have as an important concern the promotion of the Romanian through language classes and cultural events of all types.

## Romanian on the Internet

The Internet market in Romania is in continuous growth. In 2010, 44.2% of the Romanians had access to a computer at home, and 35.5% (i.e. 7,786,700 Romanians) were Internet users<sup>xi</sup> (with almost 60% of them using the Internet daily), which places Romania on the 8<sup>th</sup> place in a top 10 of Internet users from European countries<sup>xii</sup>. Over 500,000 websites are registered in the .ro domain.

When compared to the data from 2000, when only 3.6% of the population (800,000) used the Internet, we notice an increase of almost 10 times.

A study of the Latin Union in 2007 states that, similar to most of the neo-latin languages, Romanian had in the 1998 - 2007 period an increase of the language evolution over the Internet. Dividing the web pages percent for every language with the percent of the language's relative presence of speakers in the real world, they computed the vigour of each language (or the weighted presence of the studied languages in cyberspace). Although this coefficient is considered reduced for Romanian (0.62 in 2007, in comparison with English 4.44, French 2.24, Italian 2.93), this is the only language which increased its vigour in the 2005-2007 period (previous to the European Union integration).

### Selected Further Reading

Grigore Brâncuș, *Autochthon vocabulary of the Romanian Language* (in Romanian), Bucharest, Scientific and Encyclopaedic Ed., 1983.

Alf Lombard, *La langue roumaine: Une presentation*, Paris, Klincksieck, 1974.

Ioana Vintilă-Rădulescu, *Romanian Language from the European integration perspective* (in Romanian), [http://www.unibuc.ro/ro/limba\\_romn\\_din\\_perspectiva\\_integrri\\_europene](http://www.unibuc.ro/ro/limba_romn_din_perspectiva_integrri_europene)

Latin Union in collaboration with Funredes, *Languages and cultures over the Internet 2007* [http://dtil.unilat.org/LI/2007/index\\_ro.htm](http://dtil.unilat.org/LI/2007/index_ro.htm)

# Language Technology Support for Romanian

## Language Technologies

Language technologies are information technologies that are specialized for dealing with human language. Therefore these technologies are also often subsumed under the term Human Language Technology. Human language occurs in spoken and written form. Whereas speech is the oldest and most natural mode of language communication, complex information and most of human knowledge is maintained and transmitted in written texts. Speech and text technologies process or produce language in these two modes of realization. But language also has aspects that are shared between speech and text such as dictionaries, most of the grammar and the meaning of sentences. Thus large parts of language technology cannot be subsumed under either speech or text technologies. Among those are technologies that link language to knowledge. The figure on the right illustrates the Language Technology landscape. In our communication we mix language with other modes of communication and other information media. We combine speech with gesture and facial expressions. Digital texts are combined with pictures and sounds. Movies may contain language in spoken and written form. Thus speech and text technologies overlap and interact with many other technologies that facilitate processing of multimodal communication and multimedia documents.

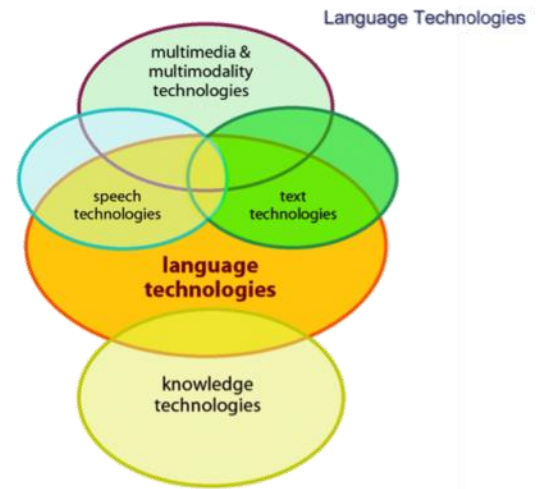


Figure 1: Language Technologies overlap

## Language Technology Application Architectures

Typical software applications for language processing consist of several components that mirror different aspects of language and of the task they implement. The figure on the right displays a highly simplified architecture that can be found in a text processing system. The first three modules deal with the structure and meaning of the text input:

- Pre-processing: cleaning up the data, removing formatting, detecting the input language, replacing the wrong diacritics with the recommended ones (changing ș in ş for Romanian, for example).
- Grammatical analysis: finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
- Semantic analysis: disambiguation (which meaning of “apple” is the right one in the context?), resolving anaphora and referring expressions like “she”, “the car”, etc.; representing the meaning of the sentence in a machine-readable way.

Task-specific modules then perform many different operations such as automatic summarization of an input text, database look-ups and many others. Below, we will illustrate core application areas and highlight their core modules. Again, the architectures of the applications are highly simplified and idealised, to illustrate the complexity of language technology (LT) applications in a generally understandable way.

After introducing the core application areas, we will give a short overview of the situation in LT research and education, concluding with an overview of past and ongoing research programs. At the end of this section, we will present an expert estimation on the

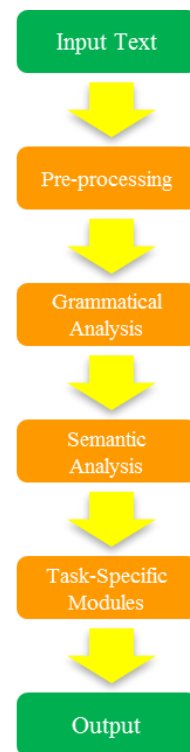


Figure 2: A Typical Text Processing Application Architecture



situation regarding core LT tools and resources on a number of dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of LT for Romanian.

The most important tools and resources involved are underlined in the text and can also be found in the table at the end of the chapter. The sections discussing the core application areas also contain an overview of the industries active in the respective field in Romania.

## Core application areas

### Language checking

Anyone using a word processing tool such as Microsoft Word has come across a spell checking component that indicates spelling mistakes and proposes corrections. 40 years after the first spelling correction program by Ralph Gorin, language checkers nowadays do not simply compare the list of extracted words against a dictionary of correctly spelled words, but have become increasingly sophisticated. In addition to language-dependent algorithms for handling morphology (e.g. plural formation), some are now capable of recognizing syntax-related errors, such as a missing verb or a verb that does not agree with its subject in person and number, e.g. in 'She \**write* a letter.' However, most available spell checkers (including Microsoft Word) will find no errors in the following first verse of a poem by Jerrold H. Zar (1992):

*Eye have a spelling chequer,*

*It came with my Pea Sea.*

*It plane lee marks four my revue*

*Miss Steaks I can knot sea.*

For handling this type of errors, analysis of the context is needed in many cases, e.g., for deciding if a word needs to be written with or without a hyphen in Romanian, as in:

*Plouă întruna de ieri.*

*[It keeps raining since yesterday.]*

*Într-una din zile am să merg la Paris.*

*[One of these days I will go Paris.]*

This either requires the formulation of language-specific grammar rules, i.e. a high degree of expertise and manual labour, or the use of a so-called statistical language model. Such models calculate the probability of a particular word occurring in a specific environment (i.e., the preceding and following words). For example, *într-una din zile* is a much more probable word sequence than *într-una de ieri*, and *plouă întruna* is more frequent than *plouă într-una*, therefore in the second case, the writing without hyphen is recommended. A statistical language model can be automatically derived using a large amount of (correct) language data (i.e. a corpus). However, there are cases when not even this could be of any help:

*Plouă întruna din primele zile ale lui martie.*

*[It keeps raining since the first days of March.]*

*Plouă într-una din primele zile ale lui martie.*

*[It rained in one of the first days of March.]*

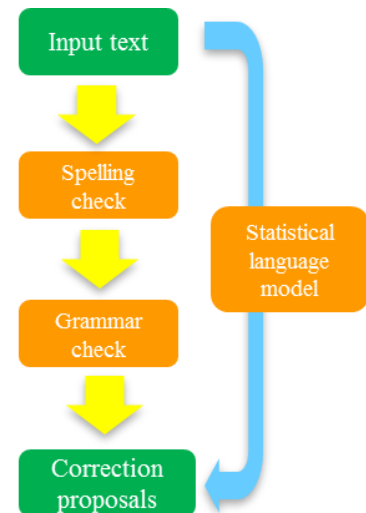


Figure 3: Language Checking (left: rule-based; right: statistical)

The only discriminating element here is the verb. In the first sentence it is in the present tense, with a durative meaning. In the latter, it is in the past tense. The two forms are homographs, although not homophones in Romanian. Only the part-of-speech tag has discriminative value in such examples.

Up to now, these approaches have mostly been developed and evaluated on English language data. However, they do not necessarily transfer straightforwardly to Romanian with its richer inflection and particular constructions.

The use of Language Checking is not limited to word processing tools, but it is also applied in authoring support systems. Accompanying the rising number of technical products, the amount of technical documentation has rapidly increased over the last decades. Fearing customer complaints about wrong usage and damage claims resulting from bad or badly understood instructions, companies have begun to focus increasingly on the quality of technical documentation, at the same time targeting the international market. Advances in natural language processing lead to the development of authoring support software, which assists the writer of technical documentation to use vocabulary and sentence structures consistent with certain rules and (corporate) terminology restrictions.

There are nowadays no Romanian companies or Language Service Providers offering products in this area, although researchers in different natural language processing groups have developed language models tailored for the Romanian language particularities. At the Research Institute for Artificial Intelligence within the Romanian Academy (RACAI), language models for Romanian are created from large corpora. Due to the fact that most of the Romanian texts on the Web are written with no diacritics, RACAI has also developed a diacritics recovery facility<sup>xiii</sup>, intended to indicate the right diacritics form of a word initially written with no diacritics, using a large Romanian lexicon developed by their team and character based 5-gram model to find the most probable interpretation in terms of diacritic occurrences for an unknown word. The approach is taking into account the context surrounding the word in a preliminary process of part-of-speech tagging, which is critical for choosing the right word form in the lexicon. For instance, the word “fata” is transformed in “față” (face, front) in the example below:

*“Teatrul este chiar in față”*

*[You have the theatre right in front of you.]*

but it kept as “fata” (daughter) in:

*“Fata voastră este foarte talentată”.*

*[Your daughter is very talented.]*

This decision is based on the previous step of part-of-speech tagging in which “fata” in the first example is annotated with an adverb tag and the same word in the second example is annotated with a noun tag.

In Romanian, at least 30% of the words in a sentence use diacritics signs, with an average of 1.16 diacritic signs per word. Only approx. 12% of these words can be immediately transformed to their diacritic version (since their non-diacritic form is not a valid word in



the Romanian language dictionary). For the rest of the words, the diacritic discovery program is useful.

Besides spell checkers and authoring support, Language Checking is also important in the field of computer-assisted language learning and is applied to automatically correct queries sent to Web Search engines, e.g. Google's 'Did you mean...' suggestions.

### Web search

Search on the web, in intranets, or in digital libraries is probably the most widely used and yet underdeveloped Language Technology today. The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries world-wide<sup>xiv</sup>. Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, incorporated basic semantic search capabilities into their algorithmic mix<sup>xv</sup>, which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a lot of data at hand and efficient techniques for indexing these data, a mainly statistically-based approach can lead to satisfactory results.

However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential. In the research labs, experiments using machine-readable thesauri and ontological language resources like WordNet (or the equivalent Romanian WordNet<sup>xvi</sup>), have shown improvements by allowing to find a page on the basis of synonyms of the search terms, e.g. *energie atomică* or *energie nucleară* (atomic power or nuclear energy) or even more loosely related terms.

The next generation of search engines will have to include much more sophisticated Language Technology. If a search query consists of a question or another type of sentence rather than a list of keywords, retrieving relevant answers to this query requires an analysis of this sentence on a syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of the relevant documents. For example, imagine a user inputs the query 'Give me a list of all companies that were taken over by other companies in the last five years'. For a satisfactory answer, syntactic parsing needs to be applied to analyse the grammatical structure of the sentence and determine that the user is looking for companies that have been taken over and not companies that took over others. Also, the expression *last five years* needs to be processed in order to find out which years it refers to.

Finally, the processed query needs to be matched against a huge amount of unstructured data in order to find the piece or pieces of information the user is looking for. This is commonly referred to as information retrieval and involves the search for and ranking of relevant documents. In addition, generating a list of companies, we also need to extract the information that a particular string of words in a document refers to a company name. This kind of information is made available by so-called named-entity recognizers.

Even more demanding is the attempt to match a query to documents written in a different language. For cross-lingual information retrieval, we have to automatically translate the query to all possible source languages and transfer the retrieved information back to the target language. The increasing percentage of data available in non-textual formats drives the demand for services

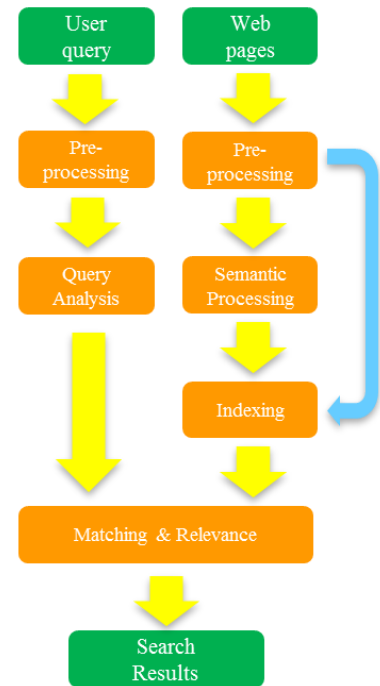


Figure 4: Web Search Architecture

enabling multimedia information retrieval, i.e., information search on images, audio, and video data. For audio and video files, this involves a speech recognition module to convert speech content into text or a phonetic representation, to which user queries can be matched.

In Romania, natural language-based search technologies are not considered for industrial applications yet. Instead, open source based technologies like Lucene are often used by search-focused companies to provide the basic search infrastructure. However, research groups from University Al. I. Cuza Iasi (UAIC) and RACAI have developed different modules that constitute the backbones of a semantic search tool, such as part-of-speech tagger, syntactic parsers, semantic parsers, named-entity recognizers, indexing tools, multimedia information retrieval, etc. Their coverage and outreach, however, is fairly limited so far.

At RACAI, a part-of-speech tagger able to identify the lemma (dictionary form) and the part of speech of words in texts is available as web service<sup>xvii</sup>. For instance, if the user's query for a web search contains "*evenimente*" (events), the root (or lemmatized form) of the word can be used instead for search, i.e. "*eveniment*" (event)<sup>xviii</sup>.

Another module developed by researchers both at UAIC and RACAI is a named-entity recognizer, which, given a text containing persons, companies, organizations, events, etc. (all referred as named-entities), identifies these entities in the text. For the example:

*Maria și-a luat bilet la concertul trupei din vară de la Paris.*

*[Mary bought a ticket for the band's concert this summer in Paris.]*

this system recognizes "Maria" as a female person, "this summer" as a temporal reference, and "Paris" as a place.

A semantic parser developed at UAIC<sup>xix</sup> is also available for the Romanian language, being able to identify, in a given sentence, the different roles entities play. For instance, for the sentence above, the system identifies "Maria" as the doer of the action and "a ticket for the band's concert" as the good being purchased. Similarly, in the example below:

*Maria și-a luat fără ezitare bilet pentru a-și vedea trupa preferată.*

*[Mary bought a ticket without hesitation to see her favourite band.]*

"without hesitation" represents the *manner* in which Mary bought the ticket, and "to see her favourite band" represents the *reason* for the acquisition of her ticket.

Recently, a group of researchers at UAIC have started to look at automatic image detection and annotation, in order to develop a web search image tool<sup>xx</sup>. However, this system is still in an incipient stage.

## Speech interaction

Speech Interaction technology is the basis for the creation of interfaces that allow a user to interact with machines using spoken language rather than, e.g., a graphical display, a keyboard, and a mouse. Today, such voice user interfaces (VUIs) are usually employed for partially or fully automating service offerings provided by companies to their customers, employees, or partners via the telephone. Business domains that rely heavily on VUIs are banking, logistics, public transportation, and telecommunications. Other usages of Speech Interaction technology are interfaces to particular devices, e.g. in-car navigation systems, and the employment of spoken language as an alternative to the input/output modalities of graphical user interfaces, e.g. in smart phones.

At its core, Speech Interaction comprises the following four different technologies:

- ❑ Automatic speech recognition (ASR) is responsible for determining which words were actually spoken given a sequence of sounds uttered by a user.
- ❑ Syntactic analysis and semantic interpretation deal with analysing the syntactic structure of a user's utterance and interpreting the latter according to the purpose of the respective system.
- ❑ Dialogue management is required for determining, on the part of the system the user interacts with, which action shall be taken given the user's input and the functionality of the system.
- ❑ Speech synthesis (Text-to-Speech, TTS) technology is employed for transforming the wording of that utterance into sounds that will be output to the user.

One of the major challenges is to have an ASR system recognise the words uttered by a user as precisely as possible. This requires either a restriction of the range of possible user utterances to a limited set of keywords, or the manual creation of language models that cover a large range of natural language user utterances. Whereas the former results in a rather rigid and inflexible usage of a VUI and possibly causes a poor user acceptance, the creation, tuning and maintenance of language models may increase the costs significantly. However, VUIs that employ language models and initially allow a user to flexibly express their intent – evoked, e.g., by a ‘How may I help you’ greeting – show both a higher automation rate and a higher user acceptance and may therefore be considered as advantageous over a less flexible directed dialogue approach.

For the output part of a VUI, companies tend to use pre-recorded utterances of professional – ideally corporate – speakers a lot. For static utterances, in which the wording does not depend on the particular contexts of use or the personal data of the given user, this will result in a rich user experience. However, the more dynamic content an utterance needs to consider, the more the user experience may suffer from a poor prosody resulting from concatenating single audio files containing different syllables/words. In contrast, today's TTS systems prove superior, though optimisable, regarding the prosodic naturalness of dynamic utterances.

Regarding the market for Speech Interaction technology, the last decade underwent a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artefacts for a given application. There also has been strong market consolidation within the last ten

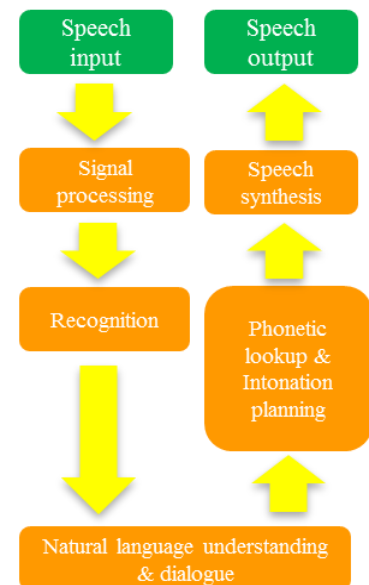


Figure 5: Simple Speech-based Dialogue Architecture

years, particularly in the field of ASR and TTS. Here, the national markets in the G20 countries – i.e. economically strong countries with a considerable population - are dominated by less than 5 players worldwide, with Nuance and Loquendo being the most prominent ones in Europe.

The speech recognition and analysis field is one of the less represented in Romania. On the Romanian TTS market, there are solutions commercialized by international companies (like MBROLA or IVONA), but with reduced accuracy and fluency. Car equipments and telecommunications companies, such as Continental and Orange, have recently started to allocate resources for specialized departments for speech processing, adapting existing solutions to their specific needs. On the other side, research in this direction is performed at University Bucharest and at the Institute for Computer Science within the Romanian Academy, Iasi Branch. Most researchers focus on text to speech synthesis, while the speech interpretation area is not so well developed yet.

Looking beyond today’s state of technology, there will be significant changes due to the spread of smart phones as a new platform for managing customer relationships – in addition to the telephone, Internet, and e-mail channels. This tendency will also affect the employment of technology for Speech Interaction. On the one hand, demand for telephony-based VUIs will decrease, on the long run. On the other hand, the usage of spoken language as a user-friendly input modality for smart phones will gain significant importance. This tendency is supported by the observable improvement of speaker-independent speech recognition accuracy for speech dictation services that are already offered as centralised services to smart phone users. Given this ‘outsourcing’ of the recognition task to the infrastructure of applications, the application-specific employment of linguistic core technologies will supposedly gain importance compared to the present situation.

### Machine translation

The idea of using digital computers for translation of natural languages came up in 1946 by A. D. Booth and was followed by substantial funding for research in this area in the 1950s and beginning again in the 1980s. Nevertheless, Machine Translation (MT) still fails to fulfil the high expectations it gave rise to in its early years.

At its basic level, MT simply substitutes words in one natural language by words in another. This can be useful in subject domains with a very restricted, formulaic language, e.g., weather reports. However, for a good translation of less standardized texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty here lies in the fact that human language is ambiguous, which yields challenges on multiple levels, e.g., word sense disambiguation on the lexical level (‘Jaguar’ can mean a car or an animal) or the attachment of prepositional phrases on the syntactic level as in:

*Politiştul a văzut omul cu telescopul.*

*[The policeman saw the man with the telescope.]*

*Politiştul a văzut omul cu arma.*

*[The policeman saw the man with the gun.]*

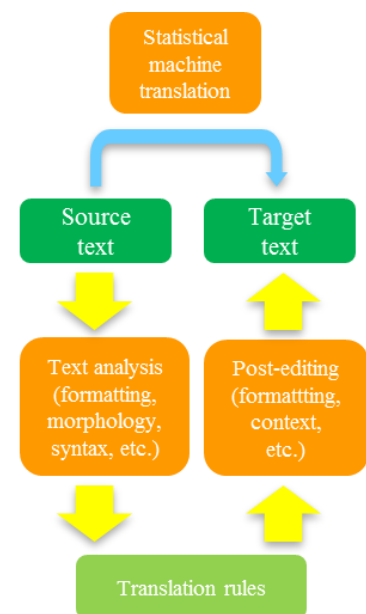


Figure 6: Machine translation (top: statistical; bottom: rule-based)

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be feasible in cases like the example above. But often rule-based (or knowledge-driven) systems analyse the input text and create an intermediary, symbolic representation, from which the text in the target language is generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by a skilled linguist.

Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for MT. The parameters of these statistical models are derived from the analysis of bilingual text corpora, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 11 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text. However, unlike knowledge-driven systems, statistical (or data-driven) MT often generates ungrammatical output. On the other hand, besides the advantage that less human effort is required for grammar writing, data-driven MT can also cover particularities of the language that go missing in knowledge-driven systems, for example idiomatic expressions.

As the strengths and weaknesses of knowledge- and data-driven MT are complementary, researchers nowadays unanimously target hybrid approaches combining methodologies of both. This can be done in several ways. One is to use both knowledge- and data-driven systems and have a selection module decide on the best output for each sentence. However, for longer sentences, no result will be perfect. A better solution is to combine the best parts of each sentence from multiple outputs, which can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

Provided good adaptation in terms of user-specific terminology and workflow integration, the use of MT can increase productivity significantly. Special systems for interactive translation support were developed, while language portals provide access to dictionaries and company-specific terminology, translation memory and MT support.

Evaluation campaigns allow for comparing the quality of MT systems, the various approaches and the status of MT systems for the different languages. The best results (shown in green and blue) were achieved by languages that benefit from considerable research efforts, within coordinated programs, and from the existence of many parallel corpora (e.g. English, French, Dutch, Spanish, German), the worst (in red) by languages that did not benefit from similar efforts, or that are very different from other languages (e.g. Hungarian, Maltese, Finnish).

The quality of MT systems is still considered to have improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories.

Table 1, presented within the EC Euromatrix+ project, shows the pairwise performances obtained for 22 official European Union languages (Irish Gaelic is missing) in terms of BLEU score<sup>xxi</sup>.



	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.0	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Table 1: Pairwise performances obtained for 22 official EU languages in Machine Translation (source: Euromatrix+)

The best results (shown in green and blue) were achieved by languages that benefit from considerable research efforts, within coordinated programs, and from the existence of many parallel corpora (e.g. English, French, Dutch, Spanish, German), the worst (in red) by languages that did not benefit from similar efforts, or that are very different from other languages (e.g. Hungarian, Maltese, Finnish).

The quality of MT systems is still considered to have improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories.

The machine translation field is among the most attractive fields in language technologies for the eyes of industrials. Thus, companies such as Language Weaver work on translating from/to Romanian using various linguistic techniques. The major online translation systems include Romanian as both source and target language, and a multitude of online dictionaries are available for Romanian.

Important research efforts were and continue to be dedicated to Machine Translation with Romanian as a source or target language by Romanian researchers from different centres. Good results are reported for an experiment of Statistical Machine Translation for English-Romanian pair in terms of comparison with contemporary performance of Google Translate for the same pair<sup>xxii</sup>.

Moreover, at RACAI there are already 5 years of experimenting in MT with different approaches like Example-Based Machine Translation, Statistical Machine Translation, extracting Machine Translation data from comparable corpora, etc. Two PhD Theses, accompanied by various papers and supported by different national or international projects like STAR and ACCURAT, are dedicated to this field<sup>xxiii, xxiv</sup>.

## Language Technology

Building Language Technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but provide significant service functionalities ‘under the

hood' of the system. Therefore, they constitute important research issues that have become individual sub-disciplines of Computational Linguistics in academia.

Question answering has become an active area of research, for which annotated corpora have been built and scientific competitions have been started. The idea is to move from keyword-based search (to which the engine responds with a whole collection of potentially relevant documents) to the scenario of the user asking a concrete question and the system providing a single answer: 'At what age did Neil Armstrong step on the moon?' - '38'. While this is obviously related to the aforementioned core area Web Search, question answering nowadays is primarily an umbrella term for research questions such as what *types* of questions should be distinguished and how should they be handled, how can a set of documents that potentially contain the answer be analysed and compared (do they give conflicting answers?), and how can specific information (the answer) be reliably extracted from a document, without unduly ignoring the context. Question answering systems can be also successfully used to identify answers of type: Location, Person, Organization, Date, Measure, Count.

This is in turn related to the information extraction (IE) task, an area that was extremely popular and influential at the time of the 'statistical turn' in Computational Linguistics, in the early 1990s. IE aims at identifying specific pieces of information in specific classes of documents; this could be, for instance, the detection of the key players in company takeovers as reported in newspaper stories. Another scenario that has been worked on is reports on terrorist incidents, where the problem is to map the text to a template specifying the perpetrator, the target, time and location of the incident, and the results of the incident. Domain-specific template-filling is the central characteristic of IE, which for this reason is another example of a 'behind the scenes' technology that constitutes a well-demarcated research area but for practical purposes then needs to be embedded into a suitable application environment.

Two 'borderline' areas, which sometimes play the role of standalone application and sometimes that of supportive, 'under the hood' component are text summarization and text generation. Summarization, obviously, refers to the task of making a long text short, and is offered for instance as a functionality within MS Word. One approach is statistically based, identifying 'important' words in a text (that is, for example, words that are highly frequent in this text but markedly less frequent in general language use) and then determining those sentences that contain many important words. These sentences are then marked in the document, or extracted from it, and are taken to constitute the summary. In this scenario, summarization equals sentence extraction: the text is reduced to a subset of its sentences. Most commercial summarizers make use of this idea.

A drawback of this approach is that it ignores the referential expressions that could occur in the initial text and be kept in the summary. Thus, due to sentence elimination, their antecedents may not be present anymore, resulting in incomprehensive readings. For example, consider the following text to be summarized:

*Hercules, of all of Zeus's illegitimate children seemed to be the focus of Hera's anger. She sent a two-headed serpent to attack him when he was just an infant.*

The summary of this very short fragment, using the sentence elimination method, could be:

*She sent a two-headed serpent to attack him.*

which is really incomprehensible if no explanation is provided of who is “she” or “him”.

One way to increase the coherence of such summaries is to derive first the discourse structure of the text and to guide the selection of the sentences to be included into the summary by a score that considers both the relevance of the sentence in a discourse tree and the coherence of the text<sup>xxv</sup>, as given by solving anaphoric references. For the summary example above, solving anaphoric references means identifying “she” as Hera and “him” as Hercules. Thus, the provided summary becomes readable:

*Hera sent a two-headed serpent to attack Hercules.*

The UAIC summarizer adopted this method, yielding good summaries for relatively short initial texts<sup>xxvi</sup>.

An alternative approach, to which some research is devoted, is to actually synthesize *new* sentences, i.e., to build a summary of sentences that need not show up in that form in the source text. This requires a certain amount of deeper understanding of the text. This method can also be applied in the case of very large texts, such as a whole novel, where neither the determination of most significant sentences based on occurrences of frequent words, nor building discourse structures could be of help. In these cases, other methods, mainly expanding a collection of predefined flexible summary patterns (based for instance on the genre of the novel, or on some data on the main characters of the novel, a time and place positioning, and a rather shallow sketch of the initiation of the action) could be applied in these cases.

All in all, a text generator is in most cases not a stand-alone application but embedded into a larger software environment, such as into the clinical information system where patient data is collected, stored and processed, and report generation is just one of many functionalities.

Romanian, as the focus language in all these research areas is somehow less attractive than English, where question answering, information extraction, and summarization have since the 1990s been the subject of numerous open competitions, primarily those organized by DARPA/NIST in the United States or by CLEF campaigns in Europe. However, Romanian teams from UAIC and RA-CAI have participated after 2006 at question answering competitions with good results<sup>xxvii</sup>. The main remaining drawback is the small size of annotated corpora or other resources for these tasks. Summarization systems, when using purely statistical methods, are often to a good extent language-independent, and thus prototypes are available also for Romanian. At UAIC, a summarization tool based on discourse structure and anaphora resolution, developed for Romanian texts, is available.

Adjacent domains recently attacked by Romanian research teams include computational lexicology, e-learning, and sentiment/opinion analysis.

A consortium of five research institutes and one university (UAIC) has recently been involved in transforming the Thesaurus Diction-



ary of the Romanian Language (about 35 volumes, from 1913 onwards) in electronic form. The main objective was to transform in structured electronic form the approx. 13.000 pages of the Dictionary, allowing complex searches, but also a much more facile editing and continuous updating activity<sup>xxviii</sup>.

More useful access to the lexicographic material of a language is facilitated by semantic networks in the form of wordnets. The Romanian WordNet has been undergoing development for eight years and has more than 52,000 synsets in which almost 60,000 literals occur. They are distributed in four parts of speech: nouns, verbs, adjectives and adverbs. Each synset contains a set of words (with associated sense numbers) that are synonyms. The synsets are the nodes of the network, while its arcs are the semantic relations between synsets: hyponymy (the *is-a* relation), meronymy, entailment, cause, and others. The Romanian WordNet is aligned to the Princeton WordNet<sup>xxix</sup>, the oldest and largest wordnet. The synsets have DOMAINS labels: each synset is labelled with the name of the domain in which it is used. Moreover, Romanian WordNet is aligned to the largest freely available ontology, SUMO&MILO<sup>xxx</sup>. It is also used in various applications developed for Romanian: Question Answering, Word Sense Disambiguation, Machine Translation.

A different domain in which UAIC researchers have been involved is the e-learning domain, by incorporating multilingual language technology tools and semantic web techniques for improving the retrieval of learning material. The developed technology facilitates personalized access to knowledge within learning management systems and support co-operation in content management.

The newest domain of interest in the natural language processing field is sentiment/opinion analysis. Thus, having a text, the software identifies if the text has a positive or negative emotional load. Research in this direction started at RACAI with the development of SentiWordNet, a sentiment annotation of the Romanian WordNet<sup>xxxi</sup>. At UAIC, research in this direction involved collaboration with a private organization, Sentimatrix, in order to develop a system able to monitor the web and extract user's opinion (forum, blogs, social networks, etc.) about different products<sup>xxxii</sup>.

## Language Technology in Education

Language Technology is a highly interdisciplinary field, involving the expertise of linguists, computer scientists, statisticians, psycholinguists, and neuroscientists. As such, it has not yet acquired a fixed place in the Romanian faculty system. Many universities in Romania and in the Republic of Moldova recently introduced natural language processing and computational linguistics courses at bachelor, master and PhD level. Since 2001, a master in computational linguistics was initiated as part of the Faculty of Computer Science at the University "Al. I. Cuza" of Iași. Still, a consolidated higher education system in natural language processing and computational linguistics is yet to be configured.

## LT Industry and Programs

The user and provider industries of LT in Romania are certainly important and vital (BitDefender, Continental, Nokia, etc.), but yet more cooperation with them can be achieved. An important issue to solve is the "secrecy" of LT, which could be solved through a good marketing strategy. Language industry is not a significant employer in Romanian, rather few companies working in the In-

formation Communication Technology (ICT) domain having developed already LT departments.

Previous national programs have led to an initial impulse, but subsequent financial aid missing or not attractive enough lead to a loss of interest from major ICT players and young researchers, formed by universities and the Academy. One of the programs of collaboration between industry and education that has a good impact and results in Romania is the MSDN Academic Alliance, offering free access to students to different Microsoft technologies.



The main research laboratories conducting activities in LT in Romania are RACAI in the Romanian Academy, Bucharest; the Department of Computer Science of the Alexandru Ioan Cuza University in Iasi, and the Institute of Computer Science of the Romanian Academy, also in Iasi, which hosts the Voiced Sounds of Romanian Language – an online repository of recorded Romanian voices. As for research programs, UAIC and RACAI have been involved in several national or international research programs, intended to develop existing or new language technologies. Among these, worth to be mentioned are some European funded projects: ACCURATRO (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation), STAR (A System for Machine Translation for Romanian), the FP7 project CLARIN (Interoperable Linguistic Resources Infrastructure for Romanian), BALKANET (which built a network of aligned wordnets for Balkan languages), the FP6 project LT4eL (Language Technology for e-Learning), the INTAS project RoLTech (Platform For Romanian Language Technology: Resources, Tools and Interfaces), etc. Some of nationally funded projects also existed, such as: SIR-RESDEC (Open Domain Question Answering System for Romanian and English), ROTEL (intelligent systems for the Semantic Web, based on the logic of ontologies and NLP), eDTLR (The Romanian Thesaurus Dictionary in electronic form), among others.

The market for language technologies can only be estimated, and will most probably get a boost by mobile appliances, the Apple iPad and similar products, (educational) games, etc.

## LT Research and Education

The most representative centres in computational linguistics dealing with Romanian language are in Bucharest, Iași, Cluj, Timișoara and Craiova, in Romania, and Chișinev – in the Republic of Moldova. There is a multitude of universities and research centres which include teams working in the domain, such as the Romanian Academy Centre for Artificial Intelligence in Bucharest, the Romanian Academy Institute for Computer Science in Iași, the Department of Computer Science at the “Alexandru Ioan Cuza” University of Iași, the Faculty of Mathematics-Informatics of the Babeș-Boyai University of Cluj-Napoca, etc. Some of these centres work in common national and international projects in the LT domain.

The common meeting points of most researchers in the LT domain are, besides international conferences abroad, a series of international events that intend to bring together young students and mature professionals, linguists and computer scientists, which are being hold periodically in Romania: the ConsILR events – Consortium for the Digitalization of Romanian Language<sup>xxxiii</sup>, the EURO-LAN series of international summer schools, the SPED conferences – Speech Technology and Human-Computer Dialogue, the KEPT conferences - Knowledge Engineering: Principles and Techniques,



ECIT – the European Conferences on Intelligent Systems and Technologies, etc.

Computational linguistics is an exotic topic and is either located in the computer science faculties or in the humanities, being therefore focussed either on the linguistic aspects, or on the engineering ones, the research topics only partially overlapping. Another major drawback of this landscape is the minor involvement of ICT companies in LT research (although they have recently begun to be more present in the educational life).

## Status of Tools and Resources for Romanian

The following table provides an overview of the current situation of language technology support for Romanian. The rating of existing technologies and resources is based on educated estimations by several leading experts using the following criteria (each ranging from 0 to 6).

- 1 **Quantity:** Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.
  - 0: no tools/resources whatsoever;
  - 6: many tools/resources, large variety.
- 2 **Availability:** Are tools/resources accessible, i.e., are they Open Source, freely usable on any platform or only available for a high price or under very restricted conditions?
  - 0: practically all tools/resources are only available for a high price;
  - 6: a large amount of tools/resources is freely, openly available under sensible Open Source or Creative Commons licenses that allow re-use and re-purposing.
- 3 **Quality:** How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Are these tools/resources current and also actively maintained?
  - 0: toy resource/tool;
  - 6: high-quality tool, human-quality annotations in a resource.
- 4 **Coverage:** To which degree do the best tools meet the respective coverage criteria (styles, genres, text sorts, linguistic phenomena, types of input/output, number of languages supported by an MT system etc.)? To which degree are resources representative of the targeted language or sublanguages?
  - 0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases;
  - 6: very broad coverage resource, very robust tool, widely applicable, many languages supported.
- 5 **Maturity:** Can the tool/resource be considered mature, stable, ready for the market? Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Is the performance of such a technology adequate and ready for production use or is it only a prototype that cannot be used for production systems? An indicator may be whether resources/tools are accepted by the community and successfully used in LT systems.
  - 0: preliminary prototype, toy system, proof-of-concept, example resource exercise;
  - 6: immediately integratable /applicable component.
- 6 **Sustainability:** How well can the tool/resource be maintained/integrated into current IT systems? Does the tool/resource fulfil a certain level of sustainability concerning documentation/manuals, explanation of use cases, front-ends, GUIs etc.? Does it use/employ standard/best-practice programming environments (such as Java EE)? Do indus-

try/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?

- 0: completely proprietary, ad hoc data formats and APIs;
- 6: full standard-compliance, fully documented.

7 **Adaptability:** How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases etc.?

- 0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand;
- 6: very high level of adaptability; adaptation also very easy and efficiently possible.

### Table of Tools and Resources for Romanian

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
<b>Language Technology (Tools, Technologies, Applications)</b>							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	5	4	5	5	5	4	4
Parsing (shallow or deep syntactic analysis)	3	3	4	4	4	3	4
Sentence Semantics (WSD, argument structure, semantic roles)	4	3	4	4	3	4	4
Text Semantics (coreference resolution, context, pragmatics, inference)	3	3	5	4	5	5	5
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	3	3	4	3	3	3	3
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	3	4	5	5	5	5	5
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	3	4	4	5	4	4	5
Language Generation (sentence generation, report generation, text generation)	0	0	0	0	0	0	0
Summarization, Question Answering, advanced Information Access Technologies	5	4	5	4	5	4	4
Machine Translation	3	4	4	3	4	4	4
Speech Recognition	2	1	3	2	2	2	2
Speech Synthesis	1	1	2	2	2	2	1

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Dialogue Management (dialogue capabilities and user modelling)	0	0	0	0	0	0	0
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	1	1	1	2	2	1	2
Syntax-Corpora (treebanks, dependency banks)	3	3	5	4	4	4	4
Semantics-Corpora	2	3	3	2	3	3	3
Discourse-Corpora	2	2	3	2	2	3	2
Parallel Corpora, Translation Memories	5	6	5	4	6	6	5
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	3	2	4	2	3	3	3
Multimedia and multimodal data (text data combined with audio/video)	0	0	0	0	0	0	0
Language Models	4	1	4	4	4	3	3
Lexicons, Terminologies	4	3	5	4	5	4	4
Grammars	2	2	3	2	2	3	3
Thesauri, WordNets	4	3	4	4	5	5	4
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	1	2	2	2	2	2	2

The most important goal of the table is not to provide an exhaustive and scientific chart of the field. The table is meant to support abstract, high-level messages, which are further explained in this section. Among these messages are:

- ❑ Even if, in general, all NLP fields are covered, there are three fields that are not yet considered for the Romanian language by researchers: language generation, dialogue management systems, and multimodal corpora building.
- ❑ Although different parsing technologies are available for the Romanian language, a reference Treebank corpus, to be used as benchmark when testing automated parses, is yet unsatisfactory.
- ❑ Speech processing is currently much less mature than NLP for written text, both in terms of corpora and instruments.
- ❑ If relatively significant work can be seen in NLP fields such as tokenization, sentence semantics or question answering systems, NLP fields dealing with more complex phenomena, such

as deep syntactic analysis or advanced discourse processing still need more attention.

- Resources for the Romanian language are less represented than instruments, although they are essential for testing the designed tools.
- With some exception, as the Web services for basic language processing, morphological analysis, question answering tools and machine translation systems, the existing tools for the Romanian language are not completely freely available, nor out of the box systems.
- The NLP tools for Romanian cover wide domains for the sentence semantics and information retrieval fields, while being relatively domain-restricted for the other tasks.
- Among the existing NLP tools for Romanian, the mature ones are provided as being freely available.
- If the different tools are not necessarily further maintained, the few resources for Romanian have good quality and are mostly sustainable.
- Since most tools are based on language models or machine learning techniques, their adaptability is generally good, which is not the case for language resources.
- The scores different experts gave to the same NLP field were usually relatively similar, mostly on availability, which suggest that the existing instruments and resources for Romanian are widely disseminated. Sometimes however, concerning sustainability and coverage, the expert gave scores that differ by more than half the total score. The main areas of disagreement were: reference corpora, semantics corpora, grammars, and ontological resources.
- The raw containing information about language models may be slightly discussable, since some experts gave scores considering the written language models, while other considered models for Romanian spoken language and gave low scores.

## Conclusions

This document describes the state of the art on Language Technology in general and on Romanian language in particular, and the support that exists for the language through Language Technology. What is the situation concerning cross- and multilingual technologies? Where does the language and its LT stand in the European context?

Research in universities and academia was successful in designing particular high quality software, as well as models and theories widely applicable. However, it is nearly impossible to come up with sustainable and standardized solutions given the current relatively low level of linguistic resources. There is a tremendous need for linguistic resources, from raw texts on Romanian till heavily annotated data, where particular linguistic phenomena are highlighted by markings contributed by experts. Since the best known source of raw texts are electronic copies of printed publications, an awareness campaign addressing the publishing houses, in order to persuade them to donate part of their textual productions for research purposes, is very much necessary<sup>xxxiv</sup>.

Many of the resources lack standardization, i.e., even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardize data and interchange formats.

Language generation and dialogue management systems are LT fields where much research is still to be done for the Romanian language. Speech technologies and corpora need also to be closely considered in order to align Romanian to the standards of other European languages.

For the development of the LT domain in Romania, a more vivid implication of the Government through adequate financing should be obtained. This document does not contain an evaluation of the funds allocated to the LT domain by the Romanian Government, but the weak support contributed by the state has been repeatedly remarked by the researchers active in the field, and is reflected in the few number of projects funded from national funds, by the shy implications in the new range of European infrastructures in the area (as CLARIN-ERIC) and the lack of interest for co-funding ICT-PSP projects addressing the field (see the example of the project ATLAS and that of METANET4U, part of the network developed by this very project – META).



## About META-NET

META-NET is a Network of Excellence funded by the European Commission. The network currently consists of 47 members from 31 European countries. META-NET fosters the Multilingual Europe Technology Alliance (META), a growing community of language technology professionals and organisations in Europe.

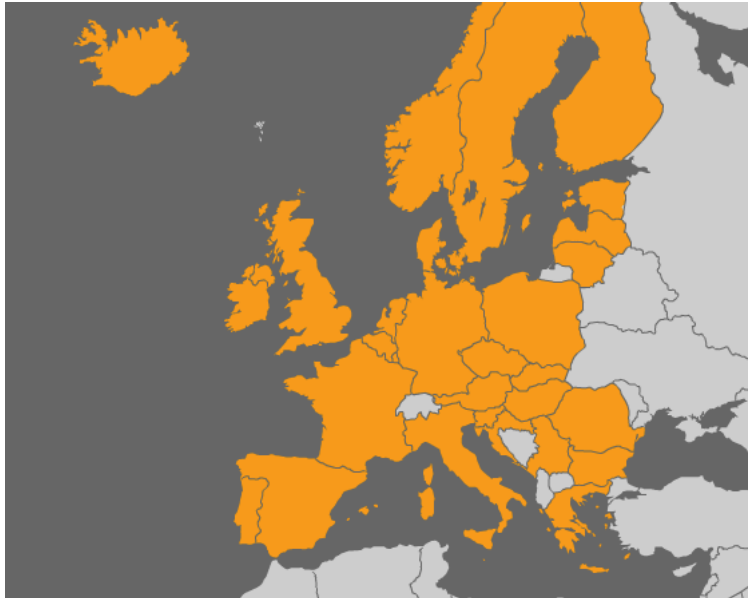


Figure 7: Countries Represented in META-NET

META-NET cooperates with other initiatives like the Common Language Resources and Technology Infrastructure (CLARIN), which is helping establish digital humanities research in Europe. META-NET fosters the technological foundations for the establishment and maintenance of a truly multilingual European information society that:

- ❑ makes communication and cooperation possible across languages;
- ❑ provides equal access to information and knowledge in any language;
- ❑ offers advanced and affordable networked information technology to European citizens.

META-NET stimulates and promotes multilingual technologies for all European languages. The technologies enable automatic translation, content production, information processing and knowledge management for a wide variety of applications and subject domains. The network wants to improve current approaches, so better communication and cooperation across languages can take place. Europeans have an equal right to information and knowledge regardless of language.

### Lines of Action

META-NET launched on 1 February 2010 with the goal of advancing research in language technology. The network supports a Europe that unites as a single, digital market and information space. META-NET has conducted several activities that further its



*The Multilingual Europe Technology Alliance (META)*

goals. META-VISION, META-SHARE and META-RESEARCH are the network's three lines of action.



Figure 8: Three Lines of Action in META-NET

**META-VISION** fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. In the first year of META-NET, presentations at the FLReNet Forum (Spain), Language Technology Days (Luxembourg), JIAMCATT 2010 (Luxembourg), LREC 2010 (Malta), EAMT 2010 (France) and ICT 2010 (Belgium) centred on public outreach. According to initial estimates, META-NET has already contacted more than 2,500 LT professionals to develop its goals and visions with them. At the META-FORUM 2010 event in Brussels, META-NET communicated the initial results of its vision building process to more than 250 participants. In a series of interactive sessions, the participants provided feedback on the visions presented by the network.

**META-SHARE** creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items. META-SHARE targets existing language data, tools and systems as well as new and emerging products that are required for building and evaluating new technologies, products and services. The reuse, combination, repurposing and re-engineering of language data and tools plays a crucial role. META-SHARE will eventually become a critical part of the LT marketplace for developers, localisation experts, researchers, translators and language professionals from small, mid-sized and large enterprises. META-SHARE addresses the full development cycle of LT — from research to innovative products and services. A key aspect of this activity is establishing META-SHARE as an important and valuable part of a European and global infrastructure for the LT community.

**META-RESEARCH** builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, this activity wants to bring more semantics into machine translation (MT), optimise the division of labour in hybrid MT, exploit context when computing automatic translations and prepare an empirical base for MT. META-RESEARCH is working with other fields and disciplines, such as machine learning and the Semantic Web community. META-RESEARCH focuses on collect-

ing data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community. This activity has already clearly identified aspects of MT where semantics can impact current best practices. In addition, the activity has created recommendations on how to approach the problem of integrating semantic information in MT. META-RESEARCH is also finalising a new language resource for MT, the Annotated Hybrid Sample MT Corpus, which provides data for English-German, English-Spanish and English-Czech language pairs. META-RESEARCH has also developed software that collects multilingual corpora that are hidden on the web.

## Member Organisations

The following table lists the organisations and their representatives that participate in META-NET.

Country	Organisation	Participant(s)
Austria	University of Vienna	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernelle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	University of Zagreb	Marko Tadić
Cyprus	University of Cyprus	Jack Burston
Czech Republic	Charles University in Prague	Jan Hajic
Denmark	University of Copenhagen	Bolette Sandford Pedersen and Bente Maegaard
Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi and Krister Linden
France	CNRS/LIMSI	Joseph Mariani
	Evaluations and Language Resources Distribution Agency	Khalid Choukri
Germany	DFKI	Hans Uszkoreit and Georg Rehm
	RWTH Aachen University	Hermann Ney
	Saarland University	Manfred Pinkal
Greece	Institute for Language and Speech Processing, "Athena" R.C.	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi

Country	Organisation	Participant(s)
	Budapest University of Technology and Economics	Géza Németh and Gábor Olaszy
Iceland	University of Iceland	Eiríkur Rögnvaldsson
Ireland	Dublin City University	Josef van Genabith
Italy	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli"	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiljevs
	Institute of Mathematics and Computer Science, University of Latvia	Inguna Skadina
Lithuania	Institute of the Lithuanian Language	Jolanta Zabarskaitė
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Malta	University of Malta	Mike Rosner
Netherlands	Utrecht University	Jan Odijk
	University of Groningen	Gertjan van Noord
Norway	University of Bergen	Koenraad De Smedt
Poland	Polish Academy of Sciences	Adam Przepiórkowski and Maciej Ogrodniczuk
	University of Lodz	Barbara Lewandowska-Tomaszczyk and Piotr Pezik
Portugal	University of Lisbon	Antonio Branco
	Institute for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufis
	Alexandru Ioan Cuza University	Dan Cristea
Serbia	University of Belgrade	Dusko Vitas, Cvetana Krstev and Ivan Obradovic
	Institute Mihailo Pupin	Sanja Vranes
Slovakia	Slovak Academy of Sciences	Radovan Garabik
Slovenia	Jozef Stefan Institute	Marko Grobelnik
Spain	Barcelona Media	Toni Badia
	Technical University of Catalonia	Asunción Moreno
	Pompeu Fabra University	Núria Bel

Country	Organisation	Participant(s)
Sweden	University of Gothenburg	Lars Borin
UK	University of Manchester	Sophia Ananiadou
	University of Edinburgh	Steve Renals

## References

---

- <sup>i</sup> European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 ([http://ec.europa.eu/public\\_opinion/flash/fl\\_313\\_en.pdf](http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf)).
- <sup>ii</sup> European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 ([http://ec.europa.eu/education/languages/pdf/com/2008\\_0566\\_en.pdf](http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf)).
- <sup>iii</sup> UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (<http://unesdoc.unesco.org/images/0015/001503/150335e.pdf>).
- <sup>iv</sup> European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (<http://ec.europa.eu/dgs/translation/publications/studies>).
- <sup>v</sup> “Statistical Yearbook 2009” of the National Institute of Statistics at [http://www.insse.ro/cms/files/Anuar\\_statistic/02/02\\_Populatie\\_en.pdf](http://www.insse.ro/cms/files/Anuar_statistic/02/02_Populatie_en.pdf)
- <sup>vi</sup> Statistical databank of the National Bureau of Statistics of the Republic of Moldova at <http://statbank.statistica.md>
- <sup>vii</sup> [http://en.wikipedia.org/wiki/Romanian\\_diaspora](http://en.wikipedia.org/wiki/Romanian_diaspora)
- <sup>viii</sup> Marius Sala (ed.), *Encyclopaedia of the Romanian Language* (in Romanian), 2nd Edition, Bucharest, Univers Enciclopedic Publishing House, 2006.
- <sup>ix</sup> <http://www.efnil.org/documents/language-legislation-version-2007/romania>
- <sup>x</sup> <http://www.ilr.ro/plr.php?lmb=1>
- <sup>xi</sup> <http://www.internetworldstats.com/eu/ro.htm>
- <sup>xii</sup> <http://www.internetworldstats.com/stats9.htm>
- <sup>xiii</sup> Tufiş Dan and Ceauşu Alexandru (2008). *DIAC+: A Professional Diacritics Recovering System*, In Proceedings of Language Resources and Evaluation Conference, LREC 2008, Marakkech, Morocco. ELRA - European Language Resources Association. ISBN: 2-9517408-4-0
- <sup>xiv</sup> <http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html>
- <sup>xv</sup> [http://www.pcworld.com/businesscenter/article/161869/google\\_rolls\\_out\\_semantic\\_search\\_capabilities.html](http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html)
- <sup>xvi</sup> Tufiş, D., Radu I., Bozianu, L., Ceauşu, A., Ştefănescu, D. (2008). *Romanian Wordnet: Current State, New Applications and Prospects*. In Attila Tanacs, Dora Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen (eds.), Proceedings of 4th Global WordNet Conference, GWC-2008, pp. 441-452, ISBN 978-963-482-854-9.
- <sup>xvii</sup> [www.racai.ro/WebServices](http://www.racai.ro/WebServices).
- <sup>xviii</sup> Tufiş, D., Radu I., Ceauşu, A., Ştefănescu, D. (2008). *RACAI's Linguistic Web Services*, In Proceedings of Language Resources and Evaluation Conference, LREC 2008, Marakkech, Morocco. ELRA - European Language Resources Association. ISBN: 2-9517408-4-0
- <sup>xix</sup> Trandabăţ D. (2011) *Towards automatic cross-lingual transfer of semantic annotation*, in 6e Rencontres Jeunes Chercheurs en Recherche d'Information RJCRI-CORIA 2011, 16-18 March, Avignon, France.
- <sup>xx</sup> Iftene, A., Vamanu, L., Croitoru, C. (2010). *UAIC at ImageCLEF 2009 Photo Annotation Task*. In C. Peters et al. (Eds.): CLEF 2009, LNCS 6242, Part II (Multilingual Information Access Evaluation Vol. II Multimedia Experiments). Pp. 283-286. ISBN: 978-3-642-15750-9. Springer, Heidelberg.

- xxi The higher the score, the better the translation, a human translator would get around 80. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA.
- xxii Munteanu D.S. and D. Marcu. (2005). *Statistical Machine Translation: An English-Romanian Experiment*. Invited tutorial at EUROLAN 2005 Summer School on NLP & HLT: The Multilingual Web: Resources, Technologies, and Prospects, Cluj Napoca, Romania.
- xxiii Tufiş, Dan, Ceauşu Alexandru (2009). *Factored Phrase-Based Statistical Machine Translation*, In Corneliu Burileanu, Horia Nicolai Teodorescu (eds.) Proceedings of the 5th Conference "Speech Technology and Human-Computer Dialogue" SpeD 2009, IEEE Catalogue number:CFP095H-CDR
- xxiv Irimia. Elena (2009). *EBMT experiments for the English-Romanian Language Pair*. International Joint Conference Intelligent Information Systems (IIS 2009). Kraków, Poland, June 15-18, 2009.
- xxv A detailed analysis of the coherence of different texts is presented in Cristea, D., Iftene, A. (2011) *If you want your talk be fluent, think lazy! Grounding coherence properties of discourse*. Invited talk at the University of Sussex, March.
- xxvi Cristea, D., Postolache, O., Pistol, I. (2005): *Summarisation through Discourse Structure*. In Alexander Gelbukh (Ed.): Computational Linguistics and Intelligent Text Processing, Proceedings of CICLing 2005, LNCS, vol. 3406, ISBN 3-540-24523-5, pp. 632-644.
- xxvii See for example the system presented in Iftene, A., Trandabăţ, D., Moruz, A., Pistol, I., Husarciuc, M., Cristea, D. (2010). *Question Answering on English and Romanian Languages*. In C. Peters et al. (Eds.): CLEF 2009, LNCS 6241, Part I. Pp. 229-236. ISBN 978-3-642-15753-0. Springer, Heidelberg.
- xxviii See more details about the digitalization of the Romanian Thesaurus Dictionary in Cristea, D. (2009): *Steps towards an electronic version of the Thesaurus Dictionary of the Romanian language* (in Romanian), ASTRA 2009, Iaşi.
- xxix <http://wordnet.princeton.edu/>
- xxx The content of this lexical resource can be freely browsed at <http://www.racai.ro/wnbrowser/>
- xxxi Baccianella Stefano, Andrea Esuli, Fabrizio Sebastian (2008). *SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, In Proceedings of LREC 2008, pp. 2200-2204.
- xxxii (Gînscă, A. L., Boroş, E., Iftene, A., Trandabăţ, D., Toader, M., Corîci, M., Perez, C. A., Cristea, D. (2011). *Sentimatrix - Multilingual Sentiment Analysis Service*. In Proceedings of the 2<sup>nd</sup> Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-WASSA2011) Portland, Oregon, USA, June 19-24, 2011.
- xxxiii See the proceedings of these events in Linguistic resources and instruments for Romanian language processing from 2006 to 2010, *Linguistic resources and instruments for Romanian language processing* (in Romanian). University "A.I. Cuza" Iasi Ed., ISBN 978-973-703-208-9.
- xxxiv See the detailed proposed solution in Cristea, Dan (2010). *Linguistic resources in a continuous flux* (in Romanian). In Iftene, A. et al. (ed). Linguistic resources and instruments for Romanian language processing, Bucharest, University "Al. I. Cuza" Iaşi Ed. ISSN 1843-911X.