Maria Gavrilidou / Penny Labropoulou / Stelios Piperidis

National Report on Language Technology in Greece

1. Introduction

The present document reports on Language Technology related activities in Greece. After a general introduction to LT and its benefits, it presents the evolution of the field in Greece and the current state of affairs, with an extensive reporting on Language Resources and Technologies developed for Greek. The report concludes with the presentation of ongoing infrastructural initiatives operating at the European level with the participation of Greek institutes.

2. Language Technology in use

Language Technology (LT, also referred to as Human Language Technology, HLT) covers a wide range of software components, data, tools and technologies, techniques and applications aimed at processing human natural language. Typical examples of such tools are tokenizers and sentence splitters, morphological analyzers, part of speech taggers and lemmatizers, syntactic analyzers, etc. The term Language Resources (LRs) denotes language data in digital format, usually of considerable size, for use by any type of research and development targeting linguistic study and language technology applications, as well as by all fields where language plays a critical role. Typical examples of LRs are spoken, written or multimodal/multimedia corpora, lexica, grammars, terminological thesauri or glossaries, ontologies etc. Nowadays, the term is extended to cover basic language processing tools used for the collection, preparation, annotation, management and deployment of LRs.

The change of perspective from the native speaker's intuition to original data and the analysis of language in actual use has been a landmark in linguistic research. Language data collection had started as a tendency in the '50s, but was led to success by the dramatic improvements in hardware technology and the advent of the web. Besides the constant need for general language and domain specific data, technologies that help quickly and efficiently analyze huge bulks of data are of critical importance.

LT is a valuable aid in many fields where research is based on language material, whether language is the object of research or the means that carries information for the research; even simple procedures such as the compilation of the list of words of a given text and its comparison with the word-list of a general language corpus can lead to insightful observations which would be missed by traditional methodologies. LT reduces the amount of time needed for the initial processing of the research material, leaving, thus, more time to researchers for the qualitative and interpretative processing of the data. Additionally, the use of LT facilitates access to secondary material, such as literature on the research subject (e.g. intelligent full-text search aiming to locate specific sections of interest).

Most importantly, LT can play a crucial role serving the needs of laymen in all aspects of their everyday life, as it enables communication across languages, and increases access to information and knowledge for users of any language. For instance, the use of LT in the access of language resources offers many advantages to the public: natural language queries are friendlier to the lay user than specialized interfaces; machine translation systems integrated in search engines produce a rough translation ("gisting translation") allowing the users to have an idea about the content of a foreign language text, although they are usually unable to convey a complete understanding of it. It is also clear, however, that not all LT tools and applications are mature enough to provide high-level services and in a user-friendly way.

3. Historical overview of LT in Greece

Greece has a thirty-year tradition in LT research and development, starting with the EUROTRA project in the mid-80's. EUROTRA was a very ambitious EU-funded project aiming to create a fully automatic high quality translation system for all of the originally seven and, later, nine European official languages. Although the project did not succeed in fulfilling the set goal, its main legacy (apart from the lexica and grammars produced) lies in the creation and training of groups of LT experts in all the involved countries.

At approximately the same time, EU-funded projects have inaugurated speech processing research in Greece, focusing on speech synthesis at first.

The decade 1990-2000 saw a critical increase in the amount of public funds invested in LT in the country, besides the EU funds. Several national programmes resulted in the creation of resources, tools and infrastructure as well as small and medium-scale applications in the field of language and speech processing. The results included text and speech databases, speech processing tools, Natural Language Processing tools, Machine Translation tools and systems, but also multimedia, LT-aware educational material for the teaching of Greek as mother tongue and foreign language. During the same years, infrastructural programmes catered for the introduction of this educational material in primary and secondary schools. Programmes with dedicated funding for resource creation resulted in the production of lexicographic material, such as computational lexica for HLT, mono-/bi-/multi-lingual multimedia dictionaries for human users, pedagogical dictionaries for Greek, terminological resources for various domains etc. A few medium- and large-scale EU infrastructural projects have also contributed to the development of monolingual resources (corpora and computational lexica) with common specifications for all EU official languages.

Through national funding, the development of the Hellenic National Corpus (HNC, http:// hnc.ilsp.gr) was made possible. HNC, accessible through the web via an interface designed for non-expert users, boosted research on linguistics, lexicology and lexicography as well as education.

These pioneer endeavours of the 90's have inspired the construction of new textual, speech, but also multimodal/multimedia resources, for general language as well as for specialized domains. Regarding general language text corpora, for instance, two en-

deavours, which saw the light in subsequent years (the first by the Centre for the Greek Language and the second by the University of Athens) made available more Greek language resources. Relevant initiatives and results are presented in the following section.

During the next decade, 2000-2010, the national programmes mainly addressed the wider sector of Information and Telecommunications Technologies, although specific activities targeted to LT have also been launched. Their objective was the development and enhancement of the LT infrastructure (e.g. creation of digital corpora and computational lexica) as well as applications in the general framework of human-machine interaction (e.g. voice-enabled dialogue systems for information extraction, intelligent human-machine interfaces, authoring aids, optical character recognition for manuscripts, automatic subtitling of multimedia content, multimedia search etc.). Obviously, the monolingual dimension was prominent in the nationally funded projects; a few bi-/multilingual resources and applications have also been produced, with English as the second language primarily. Additionally, bilateral cooperation programmes for the Balkan region have resulted in the creation of a set of resources and applications incorporating also languages from the area (Bulgarian, Serbian, Romanian, Albanian, etc.).

Recently, Greece has faced new challenges, as the volume of digital (textual and multimedia/multimodal) content has increased rapidly. Many digitization projects nationally funded, aiming at the preservation and the promotion of Greek cultural heritage, have created new requirements on the LT use and, thus, new impetus on related R&D. The results of these projects are (or will be) available over the Internet in the form of digital libraries. Researchers now have access to all types of data through their computers, but the amount of information available is so huge and so dispersed, that, without the appropriate tools, it becomes unmanageable. Furthermore, the use of language resources and tools is not extensive, not because of their quality, but because they are difficult to locate and sometimes even more difficult to use. In order to perform a specific task (e.g. to use a summarization tool, a morphological or syntactic analyzer, a speech synthesis tool etc.) the users have to know the exact tool needed and the organization or the person they need to contact in order to get the appropriate license, to review the terms of use, to download the tool or the data, to convert the format of the data to render it interoperable with the tool, to learn how to use it and so on. This situation can discourage the most dedicated researcher, let alone the ones that are not digitally literate and/or the general public that wishes to have access to digital cultural collections. Cultural informatics is a domain currently attracting LT interest.

As far as EU-funding on LT is concerned, the majority of projects currently on-going in Greece cater for application-oriented research in machine translation, information extraction, data mining, semantic web-based technologies, cognitive systems etc. Greek is not necessarily the focus of these projects, but one of the languages used as test-bed for the applications. As for data resources, the focus is on multilingual lexica and, more recently, on conceptual resources (e.g. ontologies, semantic lexica) as well as corpora; these resources are mainly domain-specific, given that most of the projects target small and medium-scale applications.

4. Current LT activities in Greece

Constant funding through national and EU sources as outlined in the previous section has resulted in a steady, increasing and dynamic evolution of LT research and development in Greece. Thus, human resources employed in the LT area have increased over the last few years, with the advent of new research groups and units in universities and research organizations dedicated to LT research. The private sector has also invested on LT; a small (but important for the dimensions of the country) number of private companies are active in the field, some of which are spin-off companies of research centres that engage in the areas of speech recognition and synthesis, machine translation, media monitoring, ePublishing, eLearning and intelligent content analysis.

A key parameter for the progress of LT has been its introduction in higher education: over the years it has been introduced in the form of modules in the curricula of under- and postgraduate studies in universities, in linguistics and technological departments alike (obviously taught from different perspectives); in addition, a post-graduate two-year interdepartmental course, summer schools and seminars dedicated to LT methodologies and applications constitute an important asset in the dissemination of LT know-how.

LT for a less-widely spoken language like Greek poses additional challenges. Notably, whereas some of the research and development work carried out in Greece is based on English data-sets and/or uses language-independent algorithms, the majority of the research endeavours has focused on Greek, attempting to model linguistic phenomena, to create Greek training data and to develop language-sensitive applications. This is reflected in the high number of research groups who are active in the country as well as abroad, trying to tackle language processing problems from the morphographemic and phonetic level to technological solutions for access to information and content.

In fact, LT research and development concerning the Greek language has spread over the years in a multitude of areas. Taking a closer look at the way it has evolved in Greece, we can discern the main driving forces: the LT domain per se (engaged in Natural Language Processing and speech related research), research in theoretical linguistics (mainly focusing on the analysis of written and spoken language), the use of LRs and LTs in language learning and, more recently, applications for the cultural domain. It is under this prism that we can explain the range and variety of research activities in which Greek LT researchers are involved.

As evidenced from the following summary, the community has moved on from the more "traditional" word/sentence-based research to new challenges (web content, various modalities, emotional language etc.). The following should be seen as points of interest rather than a full synopsis of all research activities of the LT community in Greece.

Important progress has been made in the *LR building domain*. The processes of manual collection, typing and/or OCRing, conversions from typeset material for the construction of corpora, manual selection and encoding for the construction of general language and domain specific lexica etc. are complemented and increasingly replaced by new methods and techniques. The development of (semi-)automatic tools catering for knowledge acquisition from various sources (texts, images, video etc.) are exploited for LR construction where possible: for instance, lemma and term extraction from mono and bi-/

multilingual text corpora, ontology building from textual content, web crawling methods used for spotting candidate texts for the construction of monolingual and bi-/multilingual corpora (both parallel and comparable), new OCRing methods for manuscripts. As a consequence, manual effort is more efficiently spent on the more challenging tasks (e.g. annotation with semantic and pragmatic information). Moreover, most of these techniques and methods are integrated in LT applications and systems serving end-user needs (e.g. keyword extractors for the automatic construction of indexes and thesauri to be used in accessing cultural collections).

As far as *speech* is concerned, both speech recognition and analysis are the objects of extensive research. Current interests of the community include voice interactive systems, speech-only user interfaces, speech synthesis from documents and web content, emotional speech synthesis, implementation of prosodic features etc., going even beyond speech to research on sound and music.

In the wider areas of *text mining*, *information extraction and knowledge acquisition*, the focus is on cross-lingual information retrieval, sentiment analysis, textual entailment and processing, automatic text categorization, text genre detection (including web genres), authorship attribution, spam filtering, multimedia information processing (image/video and/or audio processing for information extraction, automatic metadata extraction and fusion from various modalities), exploitation of cognitive modeling techniques, etc.

Natural language generation activities currently include research in document summarization, image-based summarization, user-adaptive management and presentation of information, monolingual and multilingual subtitling, question answering systems, spoken dialogue interaction etc.

Machine translation research addresses both aids for human translation (e.g. translation memories) and fully automatic machine translation (e.g. corpus-based machine translation approaches exploiting mono- and bi/multilingual corpora).

Developing *assistive technologies for disabled persons* (with visual and/or hearing impairments but also with learning difficulties) is the objective of several research groups in the country.

Finally, research into the use of LT for the benefit of the specialized public but also of the broad public is ongoing: for instance, in educational software and applications, authoring aids (e.g. spelling and style checkers, controlled language applications), eGovernment applications etc.

5. LRTs for the Greek language

As a result of the research efforts described above, there is a significant number of LRs for Modern Greek; most of these are available for educational and research purposes. More specifically:[1]

[1] This section presents a synopsis of results from various surveys on LRT for Greek, the most recent of which has been conducted in the framework of the preparatory stage for the Greek counterpart of the CLARIN project (cf. section 6). The results of this survey can be found at www.clarin.gr/clarinmaps (site in Greek, accessed 29/3/2011).

– as far as *textual data* are concerned:
  - there are three *general language corpora* of considerable size, namely: (a) the Hellenic National Corpus (HNC, http://hnc.ilsp.gr/), which was compiled in the early 90's but continues to be enriched; it currently includes 47 million words solely of written texts from various sources and it can be accessed via a web interface; (b) the Corpus of Greek Texts (CGT, http://sek.edu.gr/index.php?en) comprising around 30 million texts, including transcribed oral texts; the corpus is available for downloading; and (c) the newspaper corpora of the Centre for the Greek Language, of a total of 10 million words, made available through the Portal for the Greek Language (http://www.greek-language.gr/greekLang/modern_greek/tools/corpora/index.html);
  - *domain specific corpora* of small and medium size, an important proportion of which are bi-/multilingual (with English as the most frequent other language), are also available via the internet and/or distributed by the creators, covering a wide range of domains (e.g. biomedicine, health, tourism, press, literature, academic speech etc.);
  - *dialectal* material that has been collected and transcribed in the framework of linguistic research activities;
  - an important number of *cultural text collections* has become available following a digitization programme funded by the Greek state over the last few years. Although most of these texts have been digitized as images and necessitate OCR processing in order to be fully processable by LT tools, the accompanying metadata descriptions can benefit from LT.
– *linguistically annotated resources* include aligned bi-/multilingual text corpora, aligned transcriptions of audio data and text data annotated with various types of linguistic information; the annotated text corpora include morpho-syntactically tagged ones, some of which are manually disambiguated and validated, a treebank and various corpora annotated with semantic information (e.g. ontological class, named entities, event type etc.); obviously, the deeper level annotations are manually performed while morpho-syntactic tagging is usually automatic;
– most recently, a small but increasingly significant number of *multimedia/multimodal resources* has been produced; most of these resources, mainly video with accompanying audio and/or text equivalents, have been annotated with various types of modality-dependent information (e.g. speaker turn, gesture annotation etc.), while their textual counterparts are also linguistically processed (e.g. morpho-syntactically, semantically tagged);
– as far as *lexical/conceptual resources* are concerned, there are a few bi-/trilingual lexica of small and medium size intended both for computer and human use, three large monolingual morphological computational lexica, various small-size computational lexica endowed with syntactic and semantic information, usually developed for specific applications (e.g. ontologies, lists of acronyms and named entities, lexica with event types, semantic classes etc.) and a number of terminological/domain-specific lexical resources (e.g. for biomedicine, science etc.);

– available LTs can be classified in two broad categories:
  - *tools and software components that can be used to manage and process resources* (e.g. grammar/lexicon authoring tools, annotators etc.): here, we include morpho-syntactic taggers, chunkers, dependency parsers, lemmatizers and stemmers, manual annotation aids for text and multimodal/multimedia resources, named entity recognizers, text aligners for bilingual texts etc.; most of these are available for academic research and can be accessed via the internet and/or by permission of the creators; some of these tools address the Greek language, either employing a lexical/corpus resource of Greek or having been developed by the use of statistical techniques on Greek training data; the use of these tools is primarily intended for LT research and applications but it can also be extended to serve needs of end users with appropriate tuning/customization (e.g. lemmatizers deployed to facilitate lemma-based search, named entity recognizers to mark person and place names etc.);
  - *LT applications/technologies/systems* that can be used *for the benefit of the end user*: here we include authoring aids (e.g. spelling and syntactic checkers), speech recognizers, speech synthesizers, statistical information extraction tools, term extractors, speech transcribers, language detectors, summarizers, machine translation tools, etc.

A significant set of LRTs catering for Greek Sign Language (multimedia lexica, corpora, terminological resources etc.) has been compiled during the last decade.

Finally, important digital text resources but also tools and systems (OCRing tools, morphosyntactic taggers etc.) for older variants of Greek (ancient, medieval, early modern Greek etc.) are at the heart of research projects in Greece as well as abroad (cf. Perseus, http://www.perseus.tufts.edu/hopper/ and TITUS, http://titus.fkidg1.uni-frankfurt.de/framee.htm?/search/query.htm#Etabelle, two large repositories including ancient Greek resources).

6. Current initiatives for the promotion of LT

In the previous sections, we have given an overview of the LT field in Greece and the LRs that exist for the Greek language. However, although it is obvious that the field has progressed a great deal in the last years, the impact and the significance of LT for research but also for everyday life has not actually reached crucial audiences, that is, researchers at large, the broad public and, last but not least, the policy makers. The main drawbacks are:

– fragmented scenery as regards the availability of LRs:
  - although most of these are supposedly available for research and/or educational purposes, they are mainly distributed through the creators themselves and quite often they are poorly "advertised" (i.e. dissemination of their existence is at best limited to specialized conferences); interested users have to search the internet in various web sites and/or communicate with all LT institutes to find the resources they need;
  - moreover, access and usage rights are not always clear, so, even when they find them, users are not sure if they can indeed use these LRs;

  - finally, technical issues also need to be tackled before LRs are used: some resources can only be accessed through specific tools that users do not have; in other cases, the operation of the tools is scarcely documented and/or too difficult to be understood by LT illiterate users; or, even in cases where resources and relevant processing tools are both available, they are not compatible and require some customization.

  The infrastructure that puts resources together and sustains them is still largely missing; interoperability of resources, tools and frameworks at the organizational, legal and technical levels has recently come to be recognized as perhaps the most pressing current need for language processing research.

– lack and/or improvements of specific tools and datasets: although most of the basic processing tools and data resources have been developed, there is still need for extensions, enrichment and/or improvements thereof and development of new ones, especially for higher level processing (e.g. semantic annotation, discourse processing, sentiment analysis, etc.); recording of existing tools and resources in surveys like the one presented here is the first step towards the solution of this problem; however, identification of the gaps and prioritization thereof in accordance to user needs must be made in a well organized way, as well as attracting the funds that will support their development.

Bridging the gap between the LRT community and the research community at large is the task of certain initiatives that have been launched lately at the European and at national levels. The European projects META-NET and CLARIN have the aim to prepare the ground and to provide the necessary infrastructure that will offer services based on LT to the research community and to the public. A third initiative, FLaReNet, on the other hand, has a different scope than the other two: it addresses the policy makers, its results being mainly recommendations based on extensive analysis of the field according to several parameters.

More specifically, META-NET (A Network of Excellence forging the Multilingual Europe Technology Alliance, www.meta-net.eu) is a Network of Excellence that brings together researchers, commercial technology providers, private and corporate LT users, language professionals and other information society stakeholders. It constitutes a concerted, substantial, continent-wide effort in LT research and engineering which aims to create an open distributed facility for the sharing and exchange of resources, to build bridges to relevant neighbouring technology fields, as well as to prepare the strategic research agenda of the field for the years to come.

META-NET is supporting these goals by pursuing three lines of action:

– fostering a dynamic and influential community around a shared vision and strategic research agenda (META-VISION),
– creating an open distributed facility for the sharing and exchange of resources (META-SHARE),
– building bridges to relevant neighbouring technology fields (META-RESEARCH).

META-SHARE is a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources. It targets existing but also new and

emerging language data, tools and systems required for building and evaluating new technologies, products and services. In this respect, reuse, combination, repurposing and re-engineering of language data and tools play a crucial role. META-SHARE will eventually be an important component of a LT marketplace for HLT researchers and developers, language professionals (translators, interpreters, content and software localization experts, etc.), as well as for industrial players, especially SMEs, catering for the full development cycle of HLT, from research through to innovative products and services. META-SHARE will start by integrating nodes and centres represented by the partners of the META-NET consortium. It will gradually be extended to encompass additional nodes/centres and provide more functionality with the goal of turning into an as largely distributed infrastructure as possible.

Similar to META-NET but catering for the Social Sciences and Humanities researchers, the European project CLARIN (Common Language Resources and Technology Infrastructure, www.clarin.eu) is structured as a network of organizations that offer LRT for all European languages. It is a research infrastructure that aims to make LRs and LTs available though web services to researchers with little or no technical experience; services include all aspects of resource creation and use (technical, legal, administrative etc.).

When finalized, the infrastructure will constitute a platform on which

- LRT providers will be able to upload their resources and their technologies, to describe them according to a common metadata schema, to get help on legal issues such as licensing or property rights;
- LRT consumers (Social Sciences and Humanities researchers, users, developers, etc.) will profit from unified access to data and tools which physically might exist in different distributed repositories and will be able to: harvest metadata in the process of LRTs identification; browse samples or whole resources; sign usage licenses; save the resources on their computers; run a tool and save the results of the process etc.

The participation of Greece in this network will cater for the integration in the infrastructure of LRs and tools developed for the Greek language. Given that CLARIN will serve as a dynamic, constantly updated atlas of LRTs, it will constitute a valuable tool that will register the gaps that need to be tackled in what concerns the Greek language and that will evaluate the performance of the data and technologies offered for Greek in the domain of Social Sciences and Humanities.

In the framework of the national counterpart of the project, CLARIN-EL, a charting of the field has been initialized, which has recorded user needs and current practices, information on existing resources, tools, LRT organizations and research teams;[2] the national network has also been drafted. The vision of CLARIN-EL is to gather the resources and technologies that have been developed for Greek in one virtual repository and to transform them into web services which are characterized by interoperability, stability, accessibility and extensibility and which will be available to the users.

[2] The results of the CLARIN-EL survey have fed the current report.

The mission of the third initiative that aims at the unification of the LRT scenery, FLaReNet (Fostering Language Resources Network, www.flarenet.eu) is to identify priorities as well as long-term strategic objectives and provide consensual recommendations in the form of a plan of action for EC, national organizations and industry. Its outcomes are essentially of directive nature, aimed at policy makers at all levels. FLaReNet analyses the sector along various dimensions: technical, scientific but also organizational, economic, political and legal. It aims to bring together major experts from different areas, reach consensus, make the community aware of the results and disseminate them in a fine-grained, pervasive way. Work in FLaReNet is inherently collaborative.

These concerting actions have as a goal to help the egression of LRT from the boundaries of the scientific domain and its percolation through other domains, including everyday life. They aspire to introduce the benefits of LRT use to the researcher but also to the lay-man, whose work, whether scientific or not, may be facilitated and accelerated and its quality enhanced. The active participation of Greek LT research institutes in these initiatives is of paramount importance to the progress of the field in the country.