Goran Nenadić
(National Centre for Text Mining, Manchester,
and Faculty of Mathematics, Belgrade)

CREATING DIGITAL LANGUAGE RESOURCES

Abstract: We discuss building digital language resources (such as annotated corpora, lexicons, ontologies, terminologies, tools), which are the main prerequisite for successful communication and information management in the e-society of the 21st century. We give an overview of the main requirements and best practices, and point to necessary steps for creation and maintenance of standards-based and reusable language resources for written language. The notion of basic and extended language resource kits are discussed, along with other international initiatives, including the *Declaration on open access to language resources*. We also analyse challenges and responsibilities in creating digital language resources, and identify the need for wider national and international coordination and cooperation.

Keywords: language resources, digitization, human language technologies, lexica, corpora, terminologies, open access

1. Introduction

In the past few decades our society has moved towards the *e-society* in many aspects. A large amount of information is widely available in different electronic forms, ranging from structured databases (e.g. market and trade data, experimental results, image databases, etc.) to free text archives and multimedia (e.g. speech, music and video) libraries. The volume of electronic or digital texts (warehoused in numerous digital archives, libraries, corporate information systems or on the Web) has grown in the size and coverage: it is estimated that even 80% of governmental, scientific and business information is contained in digital textual form [22]. Apart from extensive newswire text archives, business and consumer textual databases and legal document collections, rapid changes in specialised areas (such as biomedicine, telecommunications, computer science, etc.) resulted in huge and constantly increasing repositories of documents, which have already shifted research from the traditional "library study" to computer-based mining of digital literature. The size of digital textual archives is increasing so rapidly that it is impossible for users to locate and assimilate information without automated help. For example, the biomedical literature currently contains over 12 million bibliographic units (predominantly in English, but in other languages as well), growing by more than 2000 abstracts each working day. Therefore, it is doubtful that anybody could process such huge amount of information without automated help, in particular if knowledge spans across domains and across languages. Furthermore, for the foreseeable future, textual communication will still be one of the prevailing methods for representing and communicating knowledge.

Therefore, sophisticated and effective methods are needed to help users to mine and extract useful knowledge from large bodies of text. The availability of electronically available documents has spurred huge interest in human language technologies (HLT) and natural language processing (NLP) applications, such as information retrieval, information extraction and text mining. These technologies aim at helping humans coping with an overwhelming amount of textual information, i.e. with a phenomenon known as the *information overload*. Users of digital archives need systems that can go beyond traditional retrieval of documents relevant to user queries (like in Web search engines), as user needs are more often oriented towards effective extraction of facts, question answering, text filtering and summarisation, and information/knowledge discovery. More precisely, instead of retrieving whole documents that might be relevant, users require systems that can extract either relevant passage(s) that address their information needs, or can analyse large amounts of text and present "digested" information. Such applications are indispensable in many domains, e.g. in legal and corporate information management systems, engineering, e-commerce, e-publishing, translations, media analyses, software industry, etc.

Sophisticated HLT and NLP applications critically depend on availability of *digital language resources* (DLRs) as the most crucial assets for processing information represented in text: the lack of reliable and large-scale DLRs is recognised as the main bottleneck in accessing and processing textual information. While there are a number of available DLRs for widely spoken languages (such as English, French, German, Spanish, Chinese), large-scale resources for other languages are still scarce and not widely available. In this article we discuss the main issues in the creation of DLRs, in particular for minority and less-widely spoken languages. After an overview of the notion and roles of DLRs in Section 2, we examine what constitutes the basic and extended language resource kits (Section 3). In Section 4 we discuss the main issues related to creating DLRs, and briefly overview major world-wide initiatives as well as resources developed so far for Serbian. Finally, the article is concluded with further challenges (Section 5) and recommendations.

2. Digital Language Resources

The term **digital language resources** refers to "a set of speech or language data and descriptions in machine readable form, used e.g. for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users" [8]. DLRs include various computational lexicons, frequency lists and machine-understandable dictionaries, collections of written and spoken language usage (also known as *corpora*), terminological databases for different subject areas, translation equivalents, audios and videos of people conversing, static gesture images, etc. Also, DLRs involve different tools (e.g. stemmers, taggers, spelling checkers, chunkers, parsers, named-entity recognisers, voice recognisers and generators, etc.) that are used for the acquisition, preparation, collection, management, and customisation of language resources. Although DLRs include resources for both spoken and written language, in this article we will concentrate on DLRs for processing written language only.

As opposed to traditional language resources (such as paper or electronic editions of dictionaries and word thesauri, grammar books, etc.), which are intended to support human users in processing and creating text, digital language resources mainly address computational systems. The main aim of a digital language resource is to support development of technologies that will enable automated processing, retrieval and extraction of information from textual collections. In that sense, developing DLRs is a major component of the language engineering process. However, DLRs can be also seen as a supplement to traditional resources, as they are important for supporting linguistic and heritage studies, language education, learning and acquisition.

In general, written language DLRs can be clustered in the following four groups:
  – *lexica*, representing basic lexical knowledge, including various types of machine-understandable dictionaries, thesauri, word networks, etc.,
  – *corpora*, representing examples of language usage, including corpora of general language and specialised sub-languages, e.g. sub-languages of weather forecasts, medical reports, technical manuals, legal texts, etc.,
  – *terminologies*, representing specialised vocabularies, including standardised terminological databases, nomenclatures, ontologies, etc.
  – *tools*, representing software modules that are used in conjunction with other resources for their management, acquisition, integration and employment.

These resources can be *monolingual* (covering one language) or *multilingual* (addressing several languages). They may be used in various applications, where DLRs are considered as a basic "linguistic infrastructure" necessary for software development. For example, monolingual lexicons and word-nets can be used for spelling checks, indexing and information retrieval, spam filtering and e-mail classification, while corpora can be used for lexicon acquisition and consolidation. Terminological resources are indispensable for processing technical and scientific literature, in particular for text mining, question answering, information extraction, computer-aided translation, etc., as well as for software and application localisation. Apart from the research and HLT communities, reliable and large-scale DLRs are crucial for supporting information and knowledge management in educational and governmental institutions (within the e-society and e-government initiatives). Also, they are of the utmost importance for promoting national languages as a functional means of communication in the digitalised environment. Availability of such resources also encourages the development of various industrial sectors (e.g. HLT, e-commerce, e-publishing, etc.).

Several initiatives and organisations have been established to deal with and to promote language resources for language engineering and to evaluate HLT technologies (e.g. the European Language Resources Association, ELRA [8]). They have suggested the main features and key priorities that have to be considered for DLRs. Here we highlight the following issues.

DLRs need to fit into an *open and standards-based framework*, i.e. their development, description, access and distribution should be based on a notion similar to the idea of open source software distribution, and with strict adherence to international language engineering (LE) standards. The international standards and best practices (such as ISLE, TEI, CES, OLIF, Dublin Core, etc.) have to be followed for encoding the resources, maximising their reusability, interoperability and availability. This approach is strongly promoted by all professional LE associations. The framework and methodology used for building DLRs should be generic but should facilitate

representation of language-specific information. This means that an abstract model of the linguistic architecture should be designed, which will allow linking and mapping different resources for various HLT scenarios, providing the environment for easier harmonisation between different resources and languages. Further, national activities have to be co-ordinated with wider (e.g. European) initiatives. Also, the resources need to be thoroughly described and documented in a standard way (*metadata*), so that they can be used and accessed by a wider research/industrial community.

DLRs need to be *reusable*, *of large-scale*, *flexible* and *multi-layered*, so that different users and applications can configure them for their needs. This would also facilitate reduction of development time for DLRs as well as content interoperability. They need to have a large-scale coverage and to facilitate efficient updates. The resources should be based on knowledge-rich linguistic representations, such that available pragmatic and semantic information (e.g. selectional preferences, syntactic/semantic roles, valency, contextual frames, discourse constraints, etc.) is encoded, providing a layered approach to their exploitation.

Digital language resources need to further be *dynamic* and *sustainable*: creating a DLR is a continuous *process* as the resource should be maintained and updated. Therefore, DLRs should not be seen as static repositories, but rather as dynamic entities which are constantly being refreshed and adjusted using methods for (semi)automatic acquisition of linguistic information from various sources.

Finally, apart from being open for *multilingual* integration, digital language resources need to allow linking to/with *multimodal* (e.g. spoken dialogue and/or gestures) and *multimedia* (e.g. graphics, images, video) digital resources, that will facilitate creation of more general digital communication resources.

3. Basic and extended standards-based language resource kits

Recently, a Europe-wide initiative (within the ENABLER[1] project) has been launched to provide recommendations, promote development and harmonise the minimal set of DLRs to be available for each and every language, in particular in Europe [9]. This set is referred to as a *Basic Language Resource Kit* (BLARK), and its main aim is to fulfil the needs for potential HLT applications and build language infrastructure equally for all European languages in order to promote functional multilinguality and cultural diversity. The initiative also includes identification of existing resources and their mapping to the BLARK requirements, and further co-ordinated actions to produce compatible and standards-based resources within Europe.

The initial idea of a minimal set of DLRs was first discussed for Dutch [2], but it is still under discussion what exactly BLARK should include. In general, BLARK is concerned with establishing recommendations both in terms of quantity (e.g. dictionary and corpus size and coverage) and quality (e.g. the level of linguistic annotation) of the basic DLRs. More precisely, BLARK aims at defining a specification of minimal general (both written and spoken) corpus and basic lexical resources (e.g. a morphological dictionary, a dictionary of idioms, a monolingual Wordnet, etc.), but also a specification for a set of basic *tools* and *skills* required for pre-competitive research and applications. The basic tools, for example, include a tokeniser, lemmatiser,

---

[1] ENABLER (European National Activities for Basic Language Resources) is a thematic network funded by the European Commission [9].

morphological analyser, chunker, parser, etc. In addition, an *Extended Language Resource Kit* (ELARK) is also being discussed, which further includes more sophisticated resources and tools, such as multilingual dictionaries and Wordnets, spelling and grammar checkers, named-entity recognisers, discourse analysers, anaphora resolution, document indexing and retrieval modules, machine-translation and computer-aided translation software, translation memories, etc. ELARK is envisaged as a layered system of resources, which may include different "sophistication" levels of DLRs for supporting various types of HLT applications.

The BLARK and ELARK initiatives are also concerned with promoting standardised representations of DLRs, so that interoperability and exchange can be ensured for efficient development of HLT applications. For developing lexical resources suitable for HLT, the European Advisory Group for Language Engineering Standards (EAGLES)[2] has developed a number of widely-adopted standards and guidelines. The EAGLES guidelines have become the de facto standard for computational lexicons in Europe, and have been used for building lexica for many languages (see, for example, the MULTEXT project, Section 4). The guidelines are mainly concentrated on representing morphology, syntax and semantics. More recently, EAGLES suggested a specification for the Multilingual ISLE Lexical Entry (MILE) as part of the International Standards for Language Engineering (ISLE) project[3]. However, EAGLES has not developed guidelines for representations of terminological data for the purposes of HLT. Currently, the ISO TC37/SC1-SC4 committee[4] is being tasked with suggesting such recommendations.

There are various possibilities to represent lexical semantic information. For example, basic relationships among words are typically represented in a framework of monolingual and multilingual Wordnets. A Wordnet is an electronic lexical thesaurus based on word meanings rather than word forms [10]. The most important relation for a Wordnet is similarity in meaning, and that is why it is organised into synonym sets (called *synsets*) classified hierarchically according to specific ontologies (which implement the basic *is_a* and *part_of* relationships). Each synset represents a concept that has become lexicalised in a language, and concepts are characterised by lists of lexical entries that can be used to express the concept in question, with entries associated with each sense of a concept. Wordnets exist for many languages (e.g. the Princeton WordNet for English; EuroWordNet comprising Wordnets for several European languages such as German, French, Spanish, Italian, Dutch, Estonian, etc.; BalkaNet for Bulgarian, Greek, Romanian, Serbian, Turkish, Czech). Although the WordNet framework is not a standard itself, the majority of monolingual Wordnets are implemented using a standard representation, which also includes cross-references to semantically equivalent concepts in other languages using an inter-lingual index (ILI).

For encoding corpora, several recommendations have been suggested. These recommendations specify a minimal encoding level that corpora must achieve to be considered standardised and useful for HLT applications (e.g. for applying machine learning techniques) and/or linguistic research. These include encoding specifications for linguistic annotation, descriptive representation of corpora (i.e. marking of structural and typographic information), and a general corpus structure (so that it can be stored in a database system for efficient access). The widely-adopted standardisation efforts

---

[2] http://www.ilc.cnr.it/EAGLES96/home.html
[3] http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm
[4] http://linux.infoterm.org/iso-e/i-iso.htm, http://www.tc37/sc4.org/

include CES (*Corpus Encoding Standard*)[5], a recommendation that is based on SGML and has been designed for use in language engineering research and applications. Further, the *Text Encoding Initiative* (TEI) Guidelines have been produced "for the encoding of all kinds of textual material in all languages from all times".[6] The TEI P3 Guidelines (1994) have become the de facto standard for encoding of literary and linguistics texts, and other corpora. An XML version of the Guidelines, TEI P4, was produced in 2002, and a new revised version (known as TEI P5) is expected soon. Several recommendations have been further suggested for adopting metadata standards for resource description (e.g. the Dublin Core Metadata Initiative, DCMI)[7].

4. Creating digital language resources

Building a high-quality, large-scale DLR is a laborious, slow and typically expensive effort, which has to include several research and specialist teams. It combines the strengths and knowledge of many traditional (lexicology, lexicography, terminology, terminography) and emerging disciplines (language engineering, computational linguistics, natural language processing, digitisation, computer science) for design, collection, development, acquisition, preservation, referencing and cataloguing services. Also, additional efforts are needed for distribution and development of user support services. The work is also influenced by many other challenges, such as cultural, economic, social, copyright and political issues.

In respond to the needs and priorities of research and industrial communities, numerous DLRs have been developed and distributed so far. Many groups and companies have created from scratch their own DLRs for specific projects they have been involved in, and many of them are reluctant to share the resources with wider community [9]. The vast majority of existing DLRs are monolingual, with general language resources prevailing [9]. Domain-specific DLRs have been created typically in multilingual frameworks, and are frequently funded by the involved industrial sector. The most demanding domain areas include law, finance, biomedicine, e-commerce, information technologies, and media.

4.1 National and international initiatives. Creating basic monolingual DLRs (i.e. digital resources for a specific language) is mainly considered as a task carried out by the respective national institutions. The vast majority of DLRs are produced as part of initiatives and projects funded by national governmental bodies for research, education, culture, justice, telecommunications, industry and technology development. Such projects exist in almost all European countries, providing basic support for European languages. For example, 28 HLT projects with almost 100 participating institutions are being funded in France (from 2002), with nine projects dedicated to the creation of DLRs, and additional five for the development of tools. In the USA, even 14 centres for national language resources have been established. Similarly, large-scale HLT initiatives are in place in many other countries (e.g. *Smartkom* in Germany, and similarly Italian, Dutch, Chinese national programs, etc.), including smaller language communities (e.g. Welsh, Basque, etc.).

---

[5] http://www.cs.vassar.edu/CES/
[6] http://www.tei-c.org/
[7] http://dublincore.org/

Building of DLRs needs to be organised as a concentrated and widely supported action, coordinated by an authoritative national institution. The best practices show that establishing a *national centre for digital language resources* ensures a productive and cost-effective long-term solution, which provides a sustainable *service* for building and updating DLRs, and avoids the repetition of development efforts. Also, it is widely recognised that an alliance among governmental, research and industrial communities is needed, although the focal points for less-spoken languages have to be governmental support. Still, many commercial areas (such as the publishing sector, media, translators, e-commerce, export and import sectors) have huge interest in developing such resources, and should consequently contribute in their development. National centres and initiatives are typically tasked with the following short and medium term priorities:

- Identification of existing resources, and selection of a priority list of DLRs that have to be developed (typically with respect to the BLARK requirements); the list may include basic or extended general language (sub-)corpora and lexical resources (morphological dictionaries and word-nets), and some essential tools such as taggers, lemmatisers and chunkers; also, domain-specific sub-corpora for emerging areas can be considered;
- Identification of criteria that will be used while creating the basic DLRs; the criteria include the selection of the encoding format(s), domains, coverage, supporting sources, methodologies and standards that will be followed, as well as verification and validation procedures and distribution strategies;
- Identification of teams that will be involved in the creation of resources; building DLRs is an collaborative effort, and typically needs substantial man-power, so it needs careful organisation, coordination and detailed planning, definition of funding sources, management, maintenance, and technical support.

Creation of some basic DLRs can be often bootstrapped by transformation and integration from existing traditional resources (e.g. paper/electronic dictionaries, thesauri, etc.), when those are available, although many problems have been reported [24]. Further, as indicated above, creation of DLRs is a process, and it is important to develop methods for continuous development and enrichment of the resources. Also, in particular for less-widely spoken languages, international and regional co-operation, transfer of competence and know-how can substantially accelerate the creation of both monolingual and multilingual resources. In some cases, richer DLRs can be obtained by transferring knowledge from existing compatible DLRs, which have been developed for other languages [21]. This also further highlights the necessity of following international standards.

Problems of standards-based and multilingual approaches to developing and distributing DLRs are in the focus of many international projects and initiatives. For example, both the Fifth (FP5) and Sixth (FP6) Framework Programmes of the European Community address these issues. The FP5 IST (*Information Society Technology*[8]) programme implemented a specific Key Action on HLT, which included cross-lingual information management and knowledge discovery, multilingual communication services and appliances, and development of the multilingual Web. Within FP5, the ENABLER project [9] aimed at establishing a co-operative network of various national

---

[8] http://www.cordis.lu/ist/

programmes, promoting harmonisation efforts and focussing on infrastructural and co-ordination initiatives. HLT within the ongoing FP6 (2002–2006) is covered mainly within the *Knowledge and interface technologies* IST area, focusing on the integration of existing DLRs into interoperable and real-time applications in various domains. An additional EU programme (*eContent*[9], 2001–2005) aims at stimulating and promoting the development and distribution of European digital content, in particular for use in the public sector and in multilingual and multicultural environments.

Several projects have been launched to develop multilingual language resources and standards. For example, MULTEXT (Multilingual Text Tools and Corpora [11]) and MULTEXT-EAST (MULTEXT for the CEE languages [4–6]) projects have developed tools, corpora, and linguistic resources for a wide range of European languages (including Bulgarian, Croatian, Czech, Dutch, English, Estonian, French, German, Hungarian, Italian, Romanian, Russian, Serbian, Slovenian, Spanish, Swedish, etc.). The main results of these projects are freely available DLRs, in particular morphosyntactic lexicons for several languages (using the same EAGLES-based representation format), and a multilingual annotated parallel corpus (Orwell's novel *1984*, which is thoroughly annotated with structural and linguistic information, and encoded using TEI P3). Also, several tool suits are provided for accessing and using these resources.

4.2 DLRs for Serbian. In case of Serbian, a collection of DLRs is being developed, mainly at the University of Belgrade. The resources developed within the HLT group at the Faculty of Mathematics (see [27] for an overview) include a set of sub-corpora[10] of various sizes and annotated at different levels (e.g. newspaper and literature corpora are not annotated, Plato's *Republic* is marked on the sentence level, while Orwell's *1984* is fully linguistically annotated using the TEI P3 guidelines and compatible with the MULTEXT-EAST recommendations [5]). Also, smaller bilingual aligned parallel corpora (e.g. Serbian–French, Serbian–English) are produced. There is also an extensive, fully manually lemmatised diachronic corpus[11] of Serbian (developed jointly by the Institute for Experimental Phonetics and Speech Pathology and the Laboratory for Experimental Psychology, University of Belgrade). For lexical resources, the HLT group has developed a morphological dictionary of simple words and their forms [23-25], as well as dictionaries for specific named entities (namely, toponyms, oronyms, and hydronyms [20]). Within the BalkaNet project, the Serbian WordNet (compatible with the EuroWordNet and encoded using Unicode) is being developed [12, 21]. Further, initial research has been started in the area of building and acquiring of terminological resources in Serbian [18].

Prototypes of several HLT tools have also been developed. Although a stand-alone part-of-speech tagger is not yet available, the Belgrade HLT group use morphological dictionaries [23] and local grammars to tag and disambiguate text [14, 24–27]. Further, available are syntactic chunkers (for some classes of noun [13, 15] and verb phrases [17, 26]), and generic recognisers for some types of named-entities (e.g. institutional and governmental names [16]).

For the time being, these resources have been mainly used separately to support linguistic research within the respective groups, while their further standardisation, combination and integration in larger-scale HLT applications is still to be explored.

5. Applications and further challenges

The availability of DLRs is beneficial for many users. For example, governmental institutions (e.g. legislation and document management departments) and the research and educational communities can benefit from efficient access to textual collections (for information retrieval and extraction). For such users, domain-specific DLRs (mainly terminologies) are needed along with general language resources. DLRs are not only important for accessing information from contemporary information sources (such as newswire, scientific literature, legislation, etc.), but also for enabling access to national heritage and for promotion of the language. In particular, DLRs are important as support not only in studying the language by specialists, but also for foreign speakers in learning the given language. Furthermore, language resources are valuable for support in crossing the language barriers (e.g. ("rough") translation of a Web page into a given language using a machine-translation system).

The Web, in particular, is an interesting and challenging textual collection that can be used both as a source of information and as a potential language resource ("the Web as a corpus"[12]). As its size and coverage grow, managing the information available on the Web becomes even more critical. Current methodologies typically rely on an emerging approach known as the Semantic Web[13], i.e. "an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [1]. It is obvious that the Semantic Web needs extensive language information to be integrated into the Web either manually or automatically using reliable DLRs such as dictionaries, thesauri, ontologies and terminologies. Also, the issue of multilinguality becomes extremely important: it has been reported that by 2005 even ¾ of Web users will not speak English as their native language[14]. Thus, there is an apparent need to support national languages, as users prefer to mine information in their own languages. Therefore, multilingual Web-scale language technology systems that are "the-Semantic-Web-ready" will be needed to enable reliable knowledge retrieval and enhanced concept-based language processing techniques. To achieve this goal, high-quality and large-scale multilingual resources with rich semantic knowledge will be needed. This knowledge needs to involve topological and semantic relations between the lexical elements (words, terms, idioms), with cross-references to other languages. Building such advanced resources would obviously benefit from existing basic DLRs.

Apart from creating a DLR, one of the additional tasks and challenges is to provide sustainable update and effective distribution of the resource. Distribution includes a thorough description of the resources (providing standardised metadata), and typically resolving legal, ethical and copyright issues. It has been already reported that the information about existing resources is typically very scarce [9], which limits their availability and usability as "only a small fraction of them is visible for interested users" [3]. These issues are being recently addressed in a proposal for a new generation of

---

[12] http://www.webcorp.org.uk/
[13] http://www.w3.org/2001/sw/
[14] http://www.hltcentral.org./

DLRs, which are referred to as open-access resources [9]. A statement (known as the *Paris Declaration on Open Access to Language Resources*) suggests that "all projects funded by governmental money that create language resources have the duty to describe them with high-quality metadata according to one of the internationally available standards for language resources and to integrate this metadata into the existing open frameworks" [3]. It is further expected that resources funded by public money are openly available unless there are specific reasons (e.g. legal or ethical). Also, it is important to suggest adequate solutions for promoting existing resources into open-access DLRs in order to avoid the repetition of huge efforts involved in their creation.

6. Conclusion

A huge portion of knowledge that is currently being generated within the e-society is encoded using natural languages, in particular in the form of digital written texts. In order for computational systems to be able to process and support this knowledge, large-scale and reliable digital language resources are needed. It is very important to understand that basic technological solutions and standards (such as the availability of standardised language-specific characters[15] on keyboards and within word-processing systems) are only an elementary support for generating texts, but that DLRs are indispensable for further processing. Similarly, developing the physical infrastructure (e.g. highly throughput networks) is not sufficient for enabling effective communication. Creating quality DLRs (both for written and spoken language) is still the main prerequisite for accessing information and communicating knowledge. National languages that do not have large-scale DLRs can be hardly functional for communication in the e-society, even if a highly advanced technical infrastructure is available. These two aspects do not exclude each other – on the contrary, they need to be viewed and treated rather as elements of the same infrastructure.

Creation of DLRs is an important and strategic task of the utmost national priority. Naturally, one of the main assumptions is that each country wishes to use its language(s) for communication, and that it will support research and development activities that will support the evolution of HLT systems [9]. Although it is a strong EU commitment to support generic aspects that are common for all languages (such as the general infrastructure, standardisation and interoperability), it is the individual countries that have to take into account all "language-specific" issues, including creation of respective monolingual and multilingual DLRs. Still, in many cases it is necessary to make national bodies and policy makers aware of the problems that the lack of DLRs can cause with regard to the overall development, access to information and national and cultural identity. Many research studies have stated that it is extremely important to consider "preparing" national language infrastructures for use in the multilingual Europe as part of the overall harmonisation process for the accession countries [9]. It is expected that a future European Linguistic Infrastructure will be defined soon, covering technical, cross-linguistic, communicational, political and commercial aspects. It is, therefore, important – in particular for less-widely spoken languages – to develop and integrate language resources urgently.

---

[15] See, for example, the Unicode standard (http://www.unicode.org/) and ISO 10646 (http://www.iso.org/).

Creating digital language resources is an emerging, challenging and sensitive process in many aspects, such as scientific, technical, public, economic, cultural, political, etc. The existence of such resources demonstrates and promotes the national identity and responsibility for the future, as cultural and language diversities are preserved and seen under a new perspective, enabling their functionality and potential in the emerging digital environments. The lack of such resources, on the other hand, alerts for actions for "survival" in the e-era.

References

1. T. Berners-Lee, J. Hendler, O. Lassila, 2001: *The Semantic Web*, Scientific American (May 2001)
2. C. Cucchiarini, E. d'Halleweyn, 2002: *How to HLT-Enable a Language: The Dutch-Flemish Experience*, http://www.hltcentral.org/page-996.0.shtml
3. ENABLER Declaration Committee, 2003: *Declaration on Open Access to Language Resources*, Paris, August 2003
4. T. Erjavec, A. Lawson, L. Romary, (Eds.), 1998: *East Meet West: A Compendium of Multilingual Resources*. TELRI-MULTEXT EAST CD-ROM, 1998
5. T. Erjavec, C. Krstev, V. Petkević, K. Simov, M. Tadic, D. Vitas, 2003: *The MULTEXT-East Morphosyntactic Specifications for Slavic Languages*, in Proc. of the EACL 2003 Workshop on Morphological Processing of Slavic Languages, Budapest
6. T. Erjavec, 2004: *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*, in Proc. of LREC 2004, pp. 1535–1538
7. European Advisory Group for Language Engineering Standards (EAGLES), information available at http://www.ilc.cnr.it/EAGLES96/home.html
8. European Language Resources Association (ELRA), information available at: http://www.elra.info/
9. European National Activities for Basic Language Resources (ENABLER), information available at: http://www.enabler-network.org/
10. C. Fellbaum, (Ed.), 1998: *WordNet – an Electronic Lexical Database*, MIT Press, 1998
11. N. Ide, J. Veronis, 1994: *MULTEXT (multilingual tools and corpora)*, in Proceedings of the COLING 1994, Kyoto.
12. C. Krstev, G. Pavlović-Lažetić, I. Obradović, D. Vitas, 2003: *Corpora Issues in Validation of Serbian WordNet*, in: V. Matousek, et al. (Eds): Text, Speech and Dialogue, TSD 2003, LNAI 2807, pp. 132–137.
13. G. Nenadić, D. Vitas, 1998: *Formal Model of Noun Phrases in Serbo-Croatian*, in BULAG 23, Figement et T.A.L., 1998, pp. 297–311
14. G. Nenadić, D. Vitas, 1998: *Using Local Grammars for Agreement Modelling in Highly Inflective Languages*, in Proc. of TSD'98, Masaryk University, pp. 97–102
15. G. Nenadić, 2000: *Local Grammars and Parsing Coordination of Nouns in Serbo-Croatian*, in P. Sojka et al. (Eds.): Text, Speech and Dialogue (TSD 2000), Lecture Notes in Artificial Intelligence, Vol. 1902, Springer Verlag, pp. 57–62
16. G. Nenadić, I. Spasic, 2000: *Recognition and Acquisition of Compound Names from Corpora*, in: D. Christodoulakis, (Ed.): Natural Language Processing (NLP 2000), LNAI 1835, Springer-Verlag, 2000, pp. 38–48
17. G. Nenadić, D. Vitas, C. Krstev, 2001: *Local grammars and Compound Verb Lemmatization in Serbo-Croatian*, in: G. Zybatow, et al (Eds.): Current Issues in Formal Slavic Linguistics, Frankfurt/Main: Peter Lang, pp. 469–477
18. G. Nenadić, I. Spasic, S. Ananiadou, 2003: *Morpho-syntactic Clues for Terminological Processing of Serbian*, in: Proc. of the EACL 2003 Workshop on Morphological Processing of Slavic Languages, Budapest
19. G. Nenadić, I. Spasic, S. Ananiadou, 2003: *Reducing Lexical Ambiguity in Serbo-Croatian by Using Genetic Algorithms*, in: P. Kosta, et al. (Eds.): Investigations into Formal Slavic Linguistics, Linguistik International, Peter Lang, Frankfurt, 2003
20. G. Pavlović-Lažetić, D. Vitas, C. Krstev, 2003: *Dictionary of toponyms in Serbian*, in Proceedings of Sixth INTEX Workshop, Sofia, Bulgaria

21. S. Stamou, G. Nenadić, D. Christodoulakis, 2004: *Exploring Balkanet Shared Ontology for Multilingual Conceptual Indexing*, in Proc. of LREC 2004, 781–784

22. D. Sullivan, 2001: *Document Warehousing and Text Mining, Techniques for Improving Business Operations, Marketing and Sales*, Wiley

23. D. Vitas, 993. *Mathematical Model of Serbo-Croatian Morphology (Nominal Inflection)*, PhD thesis, Faculty of Mathematics, Belgrade

24. D. Vitas, C. Krstev, G. Pavlović-Lažetić, G. Nenadić, 1998: *Recent Results in Serbian Computational Lexicography*, in Bokan, N. (Ed.): Contemporary Mathematics, the Monograph on the 125th anniversary of the Faculty of Mathematics, University of Belgrade, pp. 111–128

25. D. Vitas, C. Krstev, G. Pavlović-Lažetić, 2001: *The Flexible Entry*, in: G. Zybatow, et al. (Eds): Current Issues in Formal Slavic Linguistics, Frankfurt/Main, pp. 461–468

26. D. Vitas, C. Krstev, 2003: *Composite Tense Recognition and Tagging in Serbian*, in Proc. of the EACL 2003 Workshop on Morphological Processing of Slavic Languages, Budapest

27. D. Vitas, C. Krstev, I. Obradović, Lj. Popović, G. Pavlović-Lažetić, G., 2003: *An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts*, in Proc. of Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics, Greece