Report on the actions taken to give effect to recommendations as formulated in the 2003 October Unesco General Conference concerning the promotion and use of multilingualism and universal access to cyberspace

1 Introduction

Focusing on the developments and activities in Hungary in recent years with respect to the advancement of multilingualism and universal access to information in cyberspace the present report addresses the relevant issues in the order as laid out in the recommendation.[1]

2 Development of multilingual content and systems

In line with the proposed objectives of encouraging the creation and processing of educational, cultural and scientific content in digital form as well as facilitating access to it, recent years have seen an outburst in the creation of digital resources, which are now available in Hungarian on the web. To a large extent constructed as results of cooperative projects funded by extensive national programs, the main developments[2] in the field of language resources and contents are the following:

- *National Digital Data Archives.*[3] A general gateway and search interface to Hungarian digital content available on the Internet. At present about 200.000 documents are available through its service from virtual libraries and museums, and digital archives.
- *Digital Literary Academy.*[4] Created as a "virtual national institute", it is a huge collection of contemporary Hungarian literature comprising almost a thousand works of 63 Hungarian authors. Using up-to-date (XML) annotation technology to store relevant information, it is compiled into a database searchable through a web interface allowing for free access to whole works of the authors. The Academy constitutes a unique resource of 35 million words of Hungarian literature.
- *Bibliotheca Hungarica Internetiana.*[5] A recent related project in progress to make the complete oeuvres of the outstanding authors ranging from

[1] Administrative note (AN): these are to the attention of the personnel responsible for final preparation of the report. Should eventually be deleted.
AN: detailed description is only provided where the necessary competence and proficiency could be ensured. Consequently, certain sections are critically underdeveloped or simply skipped. A note (AN) is supplied when this happens. For those, persons qualified for the respective territory should be consulted.
[2] Several smaller databases have also been developed focusing on various domains. To single out each is beyond the scope of the present report.
[3] http://www.nda.hu
[4] http://www.irodalmiakademia.hu
[5] http://www.bhi.hu

old Hungarian literature through the greatest classics of the early 20th century accessible on the Internet.

- *Hungarian Electronic Library.*[6] An online collection of resources concerning Hungary or the Central European region, in the fields of culture, education and academic research. It mainly consists of text-based resources, but does not exclude other formats such as maps, multimedia etc. The Hungarian Electronic Library also collects periodicals and journals, and links to other relevant resources, services and documents concerning libraries. In the past few years it has become one of the most popular and most significant text-archives of the Hungarian webspace.

Linguistically annotated language resources:

- *Hungarian National Corpus.*[7] A synchronic 190m word reference corpus of general written Hungarian containing language variants form Slovakia, Subcarpathia, Transylvania and Vojvodina. It is divided into five subcorpora by regional language variants, and into five subcorpora by text genres also. The corpus is morphosyntactically annotated with every wordform supplied with stem, part of speech and inflecional information. Being the first Hungarian language resource covering language variants from also beyond the border of Hungary it has recently been identified by the European Science Foundation as a *major infrastructure* of European interest in the field of Humanities.

- *Hungarian Web Corpus.*[8] The largest Hungarian language corpus (700m words), available in its entirety under a permissive open content license. It consists of 18 million pages downloaded from the .hu domain, offering a general representation of contemporary written Hungarian. The current version has also been morphosyntactically annotated.

- *Hunglish Corpus.*[9] Free sentence-aligned Hungarian-English parallel corpus of about 54.2m words in 2.07m sentences, accessible through a sentence search web interface.

- *Szeged Corpus.*[10] General, manually annotated 1.5m word reference corpus. The elimination of errors from automatic annotation by manual correction makes this resource unparalleled among Hungarian language resources and a unique base corpus for all developments of natural language processing applications for Hungarian.

- *Szeged Treebank.*[11] Consisting 82.000 sentences this is to date the only syntactically annotated language resource in the Hungarian language.

Established by the Ministry of Education for, among other purposes, a large scale development of teaching materials covering all fields of school disciplines,

[6] http://mek.oszk.hu
[7] http://corpus.nytud.hu/mnsz
[8] http://mokk.bme.hu/resources/webcorpus
[9] http://mokk.bme.hu/resources/hunglishcorpus
[10] http://www.inf.u-szeged.hu/hlt
[11] http://www.inf.u-szeged.hu/hlt

the *Sulinet*[12] program now provides a complex digital knowledge-base of education materials including those for language education, which are freely accessible and usable through a web interface.[13]

In the past years, several national programs have been established to provide shorter or longer term funding for the development of info-communication technologies, in particular, natural language processing applications and language technologies. This form of project funding has been necessary since the relatively small number of Hungarian speakers does not initiate sufficient market demand for the development of such technologies, most of them not immediately remunerative. However, due to these programs launched mainly by the National Office for Research and Technology and its predecessors, essential applications and technologies have been developed including but not limited to:

Localized software applications

- *UHU-Linux.*[14] The first complete distribution built specifically for the need of Hungarian users of the open source Linux operating system and its applications.

  Note, however, that currently all major software distributions (operating systems and applications, open source and proprietary) are fully localized and available in Hungarian version.[15]

Search engines and online dictionaries

- Two major search engines have to this day been equipped with support for the Hungarian language in the form of performing basic morphological operations of lemmatization or generation of inflectional variants.[16] Both are applying specific language technology applications discussed below.

- As a result of several years of developmental work funded partly by national programs, online dictionaries are now available for Hungarian paired with practically all major European languages.[17]

Linguistic systems

- *English–Hungarian machine translation.* Fully automatic general purpose high quality machine translation systems are extremely difficult to build but it is an important application and has immense potential in multilingual communication. Prototype systems for English–Hungarian translation have already been developed and marketed but their use is lucrative only as a supplementary aid in the translation process rather than an independent mediator between two languages.[18] The development of such a system in the reverse direction is in progress, supported by the National R&D Program.

[12] "school-net" — `http://www.sulinet.hu`

[13] AN: To be extended.

[14] `http://www.uhulinux.hu`

[15] AN: This is just the mere state of facts. The author has no competent knowledge of the measures and role of *national policies* furthering these activities

[16] `http://www.polymeta.hu`; `http://keres.sztaki.hu`

[17] To single out one such resource as an example, `http://szotar.sztaki.hu` may be mentioned.

[18] Online translation service versions of these systems are also available either upon subscription or with limited free use. See eg. `http://www.webforditas.hu` or `http://www.dativus.hu`.

- In recent years, a plethora of natural language processing applications have been developed in numerous Hungarian research centers often forming consortia in the framework of several years national R&D projects. Being absolutely necessary for higher order language processing tasks, these include fundamental applications such as morphological analyzers[19], morphosyntactic annotation systems[20], syntactic parsers[21], sentence level aligner[22], speech recognition systems. Seminal projects have also been working in similar setup on the development of language resources in the field of natural language semantics and general knowledge-bases, preparing a *Hungarian Unified Ontology*[23] and also the *Hungarian Wordnet*.

3 Facilitating access to networks and services

Several national and governmental projects have been initiated to facilitate and promote universal access to the Internet. It is to be noted, however, that telecommunications and Internet costs have only very recently been in considerable decrease making these services affordable for the general public. These costs have long constituted a major obstacle for a wider penetration of the use of the Internet into average households as a means of general communication and primary source of information. The situation has been somewhat favorable for the public sector.[24]

4 Development of public domain content

To enable universal online access to public and government-held records at a national level, a general web portal service has been developed[25] offering an easy-to-access centralized gateway to information on government and public services and allowing efficient on-line administration of citizens' affairs. The development of similar sites bringing together the widest range of public service information and services online has also become a general practice at the level of local councils, with more and more of which now offering on-line access to relevant information and services.

To be extended.[26]

[19] See `http://nokk.bme.hu/resources/hunmorph` for the open source *hunmorph* analyzer, and `http://www.metamorpho.hu` for the HUMOR analyzer

[20] `http://corpus.nytud.hu/postag/`; `http://nokk.bme.hu/resources`

[21] `http://www.inf.u-szeged.hu/oasis/hlt/index_en.html`

[22] `http://nokk.bme.hu/resources/humalign`. Being in principle a language independent tool, it is now used in many projects throughout Europe to develop parallel multilingual corpora.

[23] `www.ontologia.hu`

[24] AN: This section is to be extended. The discussion of measurements on lowering of financial barriers to the use of ICT (taxes and customs duties on informatics equipment, software and services) is also beyond the competence of the present author. The same goes for the issue of ISPs' provision of concessionary rates and also for the technical issues in points 10., 11., 12., 13. and 14.

[25] `www.magyarorszag.hu`

[26] AN: The issue of the promotion of repositories of information and knowledge in the public domain and funding for the preservation and digitization of such content is beyond the competence of the present author (IHM?). The same for 17. 18. 19. (this is an issue of educational policy), and 20., 21., 22.

5 Reaffirming the equitable balance between the interests of rights-holders and the public interest

For establishing easily-available shared resources for research communities and also for the general public it is essential that appropriate copyright legislation be developed.[27]

6 Summary

This report has discussed in brief the initiatives taken to give effect to the norms and principles set forth in the recommendation with special emphasis on the development of language resources in digital form as well as those in computer aided language processing. These developments have coincided with recent improvements in the field of language technology in Hungary allowing for the efficient use of these resources and its technology to facilitate a multilingual and multicultural education and to enable lower thresholds to multicultural and multilingual content, by which means this content is becoming more and more accessible for the general society.

[27] W.r.t. 23., 24., 25.: to be written by qualified competent personnel.