Multimedia and the documentation of endangered languages

Éva Á. Csató (Uppsala University, Sweden)
David Nathan (University of Tsukuba, Japan)

The decline of linguistic competence

Today, both language-speaking communities and the linguistic community face the sad fact that the transmission of competence and expertise in human languages has stalled. The increasing degree of language endangerment resulting in the potential loss of the majority of human languages is well known. Moreover, the transfer of expertise in lesser-used languages has also become an endangered enterprise as so-called small-language courses are closed down at academic institutions, specialisation in such languages becomes professionally unrewarding, and linguists' expertise in endangered languages is untapped.

The resultant loss of knowledge about lesser-used languages, communities' linguistic competence in them, and opportunities to learn them, has become an enormous challenge. The Hans Rausing initiative to establish an Endangered Languages Project at the School of African and Oriental Studies at the University of London may provide the impetus for new efforts in the documentation and maintenance of the linguistic competence that still survives today.

In the present article we wish to present some aspects of our experience in designing, creating and using a multimedia product for the documentation and dissemination of linguistic knowledge and behaviour for a highly endangered language, the Turkic language of the Karaim of Lithuania. Our aim was to use a community-focused approach to produce a multimedia CD for the Karaim community, as well as provide the linguistic community with access to Karaim data that is transparent and accountable to further linguistic analysis. A version of this CD, *Spoken Karaim*, accompanies this publication.

Linguistics and documentation

In recent years there has been a convergence of interest between linguists working with endangered languages and those who use computers as a primary tool. A ten-year period has seen language endangerment placed on the linguistic agenda as a problem for language diversity, as an issue of human rights and resource distribution, and, most recently, increasingly framed by two complementary areas: the need for researchers to make richer documentation of a variety of language phenomena, and computer-based approaches to data—its archiving, portability, encoding, and standardisation. Amongst all of this, the voices of members of endangered language communities have not become amplified. In addition, issues in the cognitive and educational aspects of electronic language resource delivery-concrete matters like software development and interface design have received scant interest.

The idea of a new documentary linguistics advocated among others by Himmelmann (1998) had a considerable influence on subsequent archiving and data encoding projects, and ultimately the establishment of a dedicated academic programme. However, our own work with multimedia resources for Karaim and for other endangered languages has led to the identification of two crucial aspects of a successful documentary linguistics that were not part of Himmelmann's proposal and thereby perhaps limited its impact so far.

Firstly, a documentation methodology should build in the participation and interests of the language community (as well as the linguistic community) as a core component of practice and outcomes. This can assist in providing interaction between the two constituencies, as well as highlight those interests that are shared, and those that are distinct. More importantly, it is part of recognising that languages are social entities rather than mere data, and suggesting that linguists'

work be somehow measurably accountable to the state of endangered languages in their communities.

Documentary linguistics is expected to evolve into a specialised pursuit whose success will be measured in part by the vitality of the languages described and by the successful impetus to new research and publication on the language. It should be differentiated from a linguistics that works with derived data in pursuit of theoretical, technical, or even archival concerns.

Secondly, the results of documentation should be realised, disseminated, mobilised, and scrutinised through the creation of new genres of documentation products. It seems remarkable that while the idea of documentary linguistics has been widely embraced as a way to deal with global language endangerment in an era characterised by global networking and near-seamless multimedia and communication technologies, there remains a tacit assumption that the traditional genres of inscription and dissemination—text-based, linear, and page-oriented—remain the only adequate tools. Documentary linguistics—and members of endangered language communities, teachers and others—need more than more data and better ways to encode, transmit and process it. It needs an evolution of interfaces and software to deliver the richness and diversity of collected materials, as well support a diversity of users.

Building in the community

Others before Himmelmann have urged the collection of comprehensive records that could potentially support a variety of research disciplines and theoretical approaches. Goldman-Segall, for example, describes her use of video as a tool in providing a "thick description" for ethnographic research (Goldman-Segall 1994: 258), noting Margaret Mead's use of film "as data" over 70 years ago, and Mead's later predictions about the merit of emerging technologies:

> The emerging technologies of film, tape, video, and, we hope, the 360 degree camera, will make it possible to preserve materials ... long after the last isolated valley in the world is receiving images by satellite.[i] (Mead 1975:9, quoted in Goldman-Segall 1994)

Goldman-Segall further suggests that the camera can change hands, so that the "researched" play a role in documentation (1994: 257, 269) to create a better, "thicker" record, by putting the recorders themselves at the centre of a range of authentic communicative activities.

Actually, it now seems that multimedia provides a sufficiently transformative platform that it matters less who is holding the camera than whether the recorded events are appropriately presented in a documentation product. A sound or video recording of a particular speaker, suitably presented and acknowledged as a personal performance, provides a relatively rich, authentic, holistic and multifaceted linguistic resource. For linguists, such material makes linguistic data and analysis derived from it openly accountable to the original linguistic events, while also making the documentation product accountable to the scrutiny of the language community.

It is notable that Himmelmann, while recognising that recording spontaneous communicative events provide the most authentic materials, gave the "pessimistic assessment" that participants will usually not consent to recording (1998: 176ff, 187). We object to his assessment, as a result of successful experience in recording materials, including our recordings for Karaim and in Aboriginal communities in Australia. A truly worked-out discipline of documentary linguistics would grapple with the methodological and socio-political questions of how its authentic type of data can be collected, rather than burdening community members with responsibility for its scarcity.

Community members' involvement in the process of documentation need not be limited to the two polarised possibilities of holding the camera or telling the researcher to turn it off. They can be eager users of the products of documentation. The problem of how

> communities can be actively involved in the design of a concrete documentation project ... in such a way that the community not only accept it but also shape it in essential aspects (Himmelmann 1998:188)

can be solved not only by consultation and collaboration within the project planning and execution phase but also by making sure that the community understands, appreciates, and looks forward to concrete *outcomes* of a documentation project.

Many endangered language communities have long seen their languages as the objects of linguistic research but have not found linguists to be active partners in supporting the ongoing health of their languages. In creating electronic materials, many linguists might for the first time be adapting or producing materials for the community as a primary audience/readership. In such cases, a linguist's contribution of resources that are accessible and attractive to community members will provide a valuable demonstration of the linguist's commitment to the language and a symbol of wider recognition of the language's relevance.

In our project, the CD *Spoken Karaim*, we were primarily motivated by a desire to channel resources and competence that have been gained in collaboration with Karaims back into the community. Our aim has been to help revitalisation efforts by producing electronic materials and also demonstrating to the community that we intend to interact with them in new ways. Whereas linguists working in the community earlier contributed through traditional academic publishing we have chosen to use multimedia development to highlight our relationships with the community.

In order to address the second "missing plank" of documentary linguists, we describe below our example of a multimedia, community-based approach to linguistic documentation that provides linguists with rich information on the Karaim language, while significantly contributing to the Karaim communities where the documentation has its origins. We begin with a short description of the Karaim community of Lithuania.

The Karaim community

The Karaim community in Lithuania is relatively clearly defined by their unique confession Karaism (also called the Karaite confession), which, according to their traditions, diverged from Rabbinic Judaism about thousand years ago and developed into a distinct belief. They are speakers of a Kipchak Turkic language also called Karaim. The language plays an important role in their religious practice, as believers are required to read their scriptures in it. Other Karaims, with similar cultural heritage, live in Poland and in a diaspora scattered widely across the world.

The Karaim in Lithuania were traditionally communal farmers sharing communal lands; the most important such settlement has been in Trakai, near the Lithuanian capital Vilnius. These people migrated to Trakai at the end of the 14th century, and lived in their communities until their lands were confiscated by the Soviet regime and families were forced to seek their living in towns. Their religion was also forbidden, resulting in the breakdown of intergenerational language transfer. Following the collapse of the Soviet Union, much community property has been returned and the religious life of the community has been revived. The new Lithuanian government has acknowledged the Karaim minority and supports the maintenance of its cultural heritage.

Today's Karaim community in Lithuania, consisting of about 300 people, needs to reacquire its language competence. About forty people, most of them over seventy years old, still have some knowledge of the language.

Before recent documentation efforts, the linguistic community had little information on the Karaim language and it was widely believed that the language was already dead. Karaim is a significant language for Turcologists and linguists in general. It has retained some archaic features (important clues in the reconstruction of historical stages of Turkic), and has partly undergone a typological metamorphosis due to changes induced by intensive contact with Slavic and Baltic languages. For more about the Karaim community and language, see Csató (1999a), (1999b), (2000a), (2000b), (2001a), (2001b), (2002a), and (2002b).

The Spoken Karaim CD

The *Spoken Karaim* CD is based on linguistic and cultural materials recorded and collected in the Karaim community of Lithuania by community members and by Éva Á. Csató.[ii] The CD itself has been produced as a result of international cooperation, originating at the Institute of the Languages and Cultures of Asia and Africa at the Tokyo University of Foreign Studies, who funded its initial development through cooperation between Nathan and Csató (Nathan 2000a). It was further developed by Nathan while at the Australian Institute of Aboriginal and Torres Strait Islander Studies, and by Csató supported by the University of Cologne and the University of Uppsala.

The *Spoken Karaim* architecture comprises at its core of a corpus of high quality DAT recordings of Karaim monologues from a small number of remaining speakers (as well as about 40 edited video clips) made by Csató. Starting from this core, the project team created various layers of explication and support, including transcription and interlinear glossing, dictionaries, concordance, and grammar. Further user support includes searching, bookmarking, and interactive exploration of morphology. Finally, there is an outer layer of cultural, historical and geographical information that not only provides social context to the inner layers of "data" but also forms part of an architecture providing multiple paths of access between information of various types. The CD architecture, then, resembles the layers of an onion:

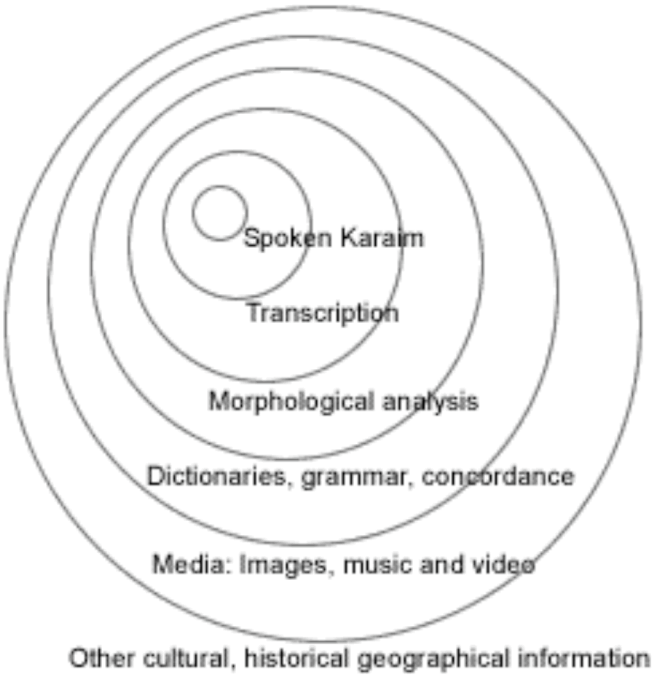

Figure 1. The architecture of the Spoken Karaim CD is centred on its voice recordings. Surrounding the recordings are various layers of linguistic support as well as culturally-based media—music, photo, video—and other support information.

When designing the Karaim CD we aimed to present a variety of resources with several ways to access them, so that different types of users are encouraged to explore it. Additional features such as bookmarking allow users to store their own selections of words and encourage them to interactive creatively with the materials.

Multimedia as a collaborative project

The creation of multimedia products encourages new patterns and styles of cooperation. Project participants include not only linguists and programmers but also community members, graphic designers, artists, archivists, sound recordists, funding and educational bodies etc.

New kinds of collaboration occur between linguists and computing practitioners. For the computing practitioner, this often means learning about linguistic concepts that are implemented on the CD. This learning can range from a minimal level (reaching a mutual, often partial, understanding of a shared vocabulary for discussing data) to an advanced level where the computing practitioner can alert the linguist to ways of presenting, representing, or processing data. Clearly, it is helpful if the computer practitioner has linguistic knowledge so that less time is wasted on learning concepts from the linguist and repairing inappropriate implementations. Anecdotal reports from several other projects indicates that computer practitioners with little linguistic knowledge can either expend a lot of project effort on attempting to learn relevant linguistic concepts or else push linguists into developing along lines that reflect the computer practitioner's own skills, preferences, or imperfect understanding of the task.

For the linguist, it can be interesting to learn how to write hypertext materials, to experience how writing for the screen varies from traditional writing, to see how linguistic distinctions and processes are represented computationally, or to find new linguistic insights as a result of the computationalist's reinterpretation of their data.

In addition, the multimedia project provides opportunities for linguists to have new interactions with speakers, materials, and learners. One of the key characteristics of multimedia projects is that they reconfigure the relationship between data providers, product developers, and the product's users. For example, linguists who are eliciting or selecting sound or video materials for multimedia publication are likely to pay attention to factors like sound quality, demographics (for instance, choosing a balance between male and female speakers), and eliciting a pedagogically balanced corpus of materials—factors that are less important when eliciting and recording linguistic data as evidence for purely linguistic descriptions or arguments (in this respect, the multimedia project may provide a better approach to a broader documentation and archive of the language than traditional linguistic elicitation). There may be opportunities to work with sound digitisation and editing, where a basic knowledge ought to be necessary for virtually any modern linguist.

When the multimedia product (or any of its evolving prototypes) is presented to community members, linguists can learn much from the various types of feedback, learning, and other usages of the materials—some of which, invariably in our experience, are unpredicted. In many cases of language revitalisation, the most important actors are children who are learning the language, and the older generation whose members retain language competence. In 2001, at a summer school for young Karaims in Trakai organised by the Karaim community and financed partly with the help of the Swedish Institute in Stockholm, it became clear that the multimedia CD could provide an effective catalyst for the young and the elderly to interact and to complement each other's competence. The young were adept in exploring the language materials while the older speakers

could help them in understanding and using the materials. Nathan has observed similar dynamics promoted by the presence of language multimedia in Aboriginal communities in Australia (Nathan 2000b).

The *Spoken Karaim* CD will also be employed in preparing students in field linguistics for a summer course in the Karaim community. By bringing linguists to the community we hope to motivate young Karaims to learn the language and to contribute to linguists' interest in exploring Karaim and maintain expertise in it. We hope to create a fruitful collaboration between linguists who can help the Karaims to learn the language and the Karaims who may supply further valuable information about their community.

Multimedia and archiving

We would like to comment on current developments in the linguistic community in relation to encoding, standardisation and archiving of data for endangered languages. Linguists have been urged to encode their data using standard schemas (which are being researched and published by projects such as OLAC nd), and to make proper provision for its safe archiving. We have been similarly concerned about preservation of the materials in the *Spoken Karaim* CD; for example, the linguistic data has been kept separated from the computer code within the application, and a project has been initiated to place the primary data with the Oxford Text Archive.

However, preservation of complex linguistic materials has several dimensions, not all encompassed by the imperative to encode data within particular schemas or in dedicated archives. Firstly, we believe that the primary means of preserving a language lies in enabling its community to continue to use it, to enjoy its ability to express their identity, and to continue its living evolution within their changing culture. Secondly, it is possible that wide distribution of a published CD is a viable way to ensure its preservation, in the same way that widely published canonical paper documents have survived hundreds or even thousands of years due to the widespread awareness of them, and demand and regard for them.[iii] Thirdly, a highly complex, interactive and interconnected application has a holistic integrity that is not adequately preserved by extracting and archiving some of its data—if the application has been well produced, it ought to be already the best way to digitally preserve the materials!

Fourthly, if the data representations within the project and its application are well-designed both computationally and linguistically, then an effective life of the materials is assured. For example, while the *Spoken Karaim* CD uses a new orthography, with many complex diacritic characters not previously available or included in the Unicode character set, we created and archived the text data using a custom-designed system of ASCII sequences similar to X-SAMPA used for Ega (Gibbon nd).[iv] Rigorous design and implementation of the data structures underlying the glossed interlinear text and lexicon enabled rapid and seamless conversion of that data to XML.

Multimedia: integrating cultural and linguistic materials

Integrating linguistic, cultural and community materials is a crucial enterprise for documentary linguistics. The Karaim CD includes many different types of information: recordings of language events and linguistic description and analysis; descriptions of the community, its history, religion, food, literature, and language background; secular and religious music; and videos and photos of people and local features.

The CD opens with a photo of community members gathered around the Karaim temple, the *kenesa*, in Trakai. The late *hazzan*, Mykolas Firkovicius, stands in the middle. This opening has

been chosen in order invoke the personal interest of the community members, who know each other well, and thus recognise people on this and other photos. This is a strong motivation to browse for other photos on the CD, which therefore functions as a new kind of family/community album.

The metaphor used to integrate different clusters of information is a map of the Karaim settlement, showing the peninsula in Trakai in the middle of which the Karaim Street stretches down to the castle. The Karaim have lived here for the last six hundred years; this map reflects the Karaim's unique setting for community life and cultural and social activities. The user moves around it visiting different sites: the entrance of the Karaim street, the Karaim temple, *kenesa*, the house of a Karaim speaker, the Karaim restaurant, the Karaim cemetery, the house of a famous writer and religious and administrative leader, and the castle of Vytautas, the Grand Duke of Lithuania who invited the Karaims to Trakai at the end of the 14th century in order to defend his newly built castle. Knowledge of this landscape is an indispensable part of Karaim identity; Karaim community members will use this knowledge to help navigate the CD, and the visitor will learn about the locality while encountering the cultural richness of Karaim life, including the Karaim language.

In this architecture all elements of the Karaim heritage are integrated and the access from one type of information to another is based on associations that may be linguistic, historical, religious, or based on location of family ties.

There are few full speakers of Karaim today. Moreover, competence in the religion and liturgical music is also very limited. The CD includes video and sound recordings of the late leader of the community, Mykolas Firkovicius, who unfortunately recently passed away. He was the last *hazzan* who had both the musical ability and the traditional training to perform the religious songs. Access to these recordings is very important for many people. For the many Karaims who live isolated from the community, for small Karaim communities such as in Halich, Ukraine, that do not have a *hazzan*, for children who never have participated at Karaim prayers, or for those who cannot read the Karaim prayer books, these recordings provide the only description of and access to Karaim religious practice. The recorded Psalms are played at Karaim burials. One elderly Karaim from Halich described the situation: before receiving her CD, she could pray to God only by opening her prayer book and placing it silently on the table.

Karaim songs also play an important role in maintaining community identity. When community members gather they sing in Karaim. For young members, who are often ashamed because they do not know the text, it is important to provide recordings to help learn these songs.

Linguists working with speakers of endangered languages are familiar with the emotional throes of 'rusty' speakers who have only partial language competence. Although they regard themselves as native speakers, they are often ashamed to acknowledge—or find it revealed—that they no longer have full competence. Multimedia, cultural-based materials make it easier and more comfortable for such speakers to refresh their competence through self-learning.

The *Spoken Karaim CD* recordings are of natural colloquial speech in which copies of Slavic and Baltic lexical items are used. In contrast, older materials, such as written textbooks, typically present purified forms of the language. This has several drawbacks. While speakers are used to creatively copying non-Karaim words and expressions as their communicative needs require, the purism of the written language creates the impression that 'good Karaim' does not employ 'foreign' elements; this inhibits speaking and discourages older speakers from teaching their grandchildren. After completing the first lesson on the CD we asked a Karaim child what he had learnt in Karaim: his answer was *aftobus**nun** stanciya**sï*** "bus stop". The lexical elements were already known to him, what he had learned was the Turkic morphology. This was a good start to learning spoken Karaim.

The Karaims are in many ways untypical of the world's endangered language communities but rather typical of European communities. Many community members have higher education and access to computers. Historically, they have enjoyed long, stable and prosperous periods within a dominant community, and have long possessed dictionaries and grammars. However, due to historical circumstances, the Karaim's orthography has been changed several times. First, the traditional Hebrew script was replaced by a Cyrillic one, then by Polish, and later by a Lithuanian-based Latin script. The available written resources for language learning reflect these changes and have discouraged students wishing to learn the language. However, new multimedia genres are encouraging Karaims to reconnect with the resources for their language. Due to the rich multimedia content of the CD, learning processes are not hindered by an unfamiliar orthography, because users can simultaneously hear and read the texts, as well as use the video and graphical materials to reinforce the linguistic information.

The CD also provides the linguistic community with access to linguistic data for Karaim, and to some generalised linguistic phenomena. Linguists can use this data in creative ways, or apply new modes of analysis. Or the data, together with the specialised linguistic functions on the CD, can be used for linguistic training; for example, students can be assigned to find morphophonological rules for describing the distribution of suffix variants.

The results of the collaboration and computational work on the *Spoken Karaim* CD are not only that we are starting to understand the scope of a workable documentary linguistics through an active contribution to the Karaim and the linguistic communities, but also that we now have a 'template' that can be adapted for creating similar CDs for other endangered languages.

References

Bird, S. and Simons, G. 2003 (to appear). "Seven dimensions of portability for language documentation and description", *Language 79*.

Csató, É. Á. (1999a) "Should Karaim be 'purer' than other European languages?", *Studia Turcologica Cracoviensia,* 5:81-89.

— (1999b) "Analyzing contact-induced phenomena in Karaim", in S.S. Chang, L. Liaw and J. Ruppenhofer (eds.) *Twenty-Fifth Annual Meeting of the Berkeley Linguistic Society, Special Session: Caucasian, Dravidian, and Turkic Linguistics*. BLS 25S, 54-62.

— (2000a) "Some typological features of the viewpoint aspect and tense system in spoken North-Western Karaim", in Ö. Dahl (ed.) *Tense and Aspect in the Languages of Europe.* Berlin: Mouton de Gruyter, 671-699.

—(2000b) "Syntactic code-copying in Karaim", in Ö. Dahl and M. Koptjevskaja-Tamm (eds.) *The Circum-Baltic Languages: Their Typology and Contacts.* Amsterdam: John Benjamins, 265-277.

—(2001a) "Karaim dictionary on CD-ROM", in N. Demir and E. Yılmaz (eds.) *Uluslar Arası Sözlükbilim Sempozyumu Bildirileri.* Gazimagusa: Dogu Akdeniz Üniversitesi, 35-40.

—(2201b) "Karaim", in Th. Stolz (ed.) *Minor Languages of Europe*. [Bochum-Essener Beiträge zur Sprachwandelforschung 30] Bochum: Brockmeyer, 1-24.

—(2002a) "Karaim: A high-copying language", in M.C. Jones and E. Esch (eds.) *Language Change. The Interplay of Internal, External and Extra-Linguistic Factors* [= Contributions to the Sociology of Language 86]. New York & Berlin: Mouton de Gruyter, 315-327.

—(2002b) "The Karaim community in Lithuania", in W. Maciejewski (ed.) *The Baltic Sea Region. Cultures, Politics, Societies*. Uppsala: Baltic University Press, 272-275.

Gibbon, D. nd. EGA WEB ARCHIVE http://coral.lili.uni-bielefeld.de/LangDoc/EGA/

Goldman-Segall, R. (1994) "Collaborative virtual communities: Using learning constellations, A multimedia ethnographic research tool", in E. Barrett (ed.), *Sociomedia: Multimedia, Hypermedia, and the Social Construction of Knowledge*, Cambridge, MA: MIT Press, 257-296.

Good, J. and Sprouse R. 2001. "Creating a database and query-tools for the TELL multi-speaker linguistic corpus". In Bird, S. Buneman, P and Liberman, M. (eds) 2001. *Proceedings of the IRCS Workshop on Linguistic Databases*, IRCS University of Pennsylvania, December 2001, pp 82-91

Himmelmann, N. (1998) "Documentary and Descriptive Linguistics", *Linguistics* 36:161-95.
Mead, M. (1975) "Visual anthropology in a discipline of words", in P. Hockings (ed.) *Principles of Visual Anthropology*, The Hague: Mouton, 3-10.
Nathan, D. (2000a) "The Spoken Karaim CD: Sound, text, lexicon and 'Active Morphology' for language learning multimedia", in A. Göksel and C. Kerslake (eds.) *Studies on Turkish and Turkic Languages,* Wiesbaden: Harrassowitz, 405-413.
— (2000b) "Plugging in Indigenous knowledge - connections and innovations", *Australian Aboriginal Studies* 2000, 2:39-47.
OLAC nd. http://www.language-archives.org/OLAC/

[i] Today's Virtual Reality software (e.g. Quicktime VR) functions like a 360 degree camera.
[ii] Csató's research was carried out at the Linguistics Department of the University of Cologne under the supervision of Hans-Jürgen Sasse, an initiator of intensive research on endangered languages. The project was financed by the Deutsche Forschungsgemeinschaft.
[iii] While it is often argued that digital materials are rendered obsolete systems if they are not "refreshed" to keep up with changes in computer software—clearly an important factor to be taken into account—it is also true that traditional documents have been "refreshed", for example through the invention of paper, printing, and through continued translations into other languages and contemporary versions of changing languages.
[iv] see http://coral.lili.uni-bielefeld.de/LangDoc/EGA/Formats/egasampa.txt. The SAMPA alphabet was developed in the late 1980s by John Wells. Our system is similar to, but more phonetically detailed than the orthographic representations used for Turkish in the TELL project (Good and Sprouse 2001).