The Spoken Karaim CD: Sound, Text, Lexicon and "Active Morphology" for Language Learning Multimedia

David Nathan
Institute for the Study of the Languages and Cultures of Asia and Africa
Tokyo University of Foreign Studies
Tokyo, Japan
http://www3.aa.tufs.ac.jp/~djn/karaim/karaimCD.htm

THE PROJECT

This paper discusses the project "Spoken Karaim: the CD", describing the origin and evolution of the project, and how multimedia materials can support language learning, particularly in the case of endangered languages. The Spoken Karaim CD is one component of a larger enterprise that aims to document and support the Karaim language (see Csat` , this volume). The CD itself aims to support the Karaim community in its efforts to maintain its language, to provide a companion sound resource for Csató's forthcoming book *Spoken Karaim* (Csat` , va gnes, Forthcoming), to preserve valuable spoken texts, and to inform the wider audience about the Karaim's language and community.

The principal participants of the project are va gnes Csat` (Uppsala, Sweden), David Nathan (Tokyo & Canberra), and Karina Firkaviciute (Vilnius, Lithuania). Csat` has worked for several years documenting and analysing the Karaim language, in collaboration with fluent Karaim speakers. The idea of creating an interactive multimedia CD sprang from Csató's determination to create something useful for the community in its attempts to stop the rapid decline in the use of Karaim in the current generation. In Australia, Nathan had been working at the Australian Institute of Aboriginal and Torres Strait Islander Studies, developing multimedia resources for use in Aboriginal communities who wished to maintain languages or attempt to revive languages which had been destroyed. Karina Firkaviciute is a Karaim musicologist who has described traditional Karaim musical techniques and was already collaborating with Csat` (Firkavicuite, 1995).

THE CD'S CONTENT

Today the name "CD" has an interesting ambiguity: on one hand it is simply a large container for digital data; on the other hand it connotes a genre for interactive multimedia as a result of the influence of the successful products published in CD form: games and encyclopedias (Dersot 1998). We have attempted to combine these two aspects, taking advantage of the CD's high capacity in order to present high quality (CD-audio) sound recordings of Karaim spoken texts, while also providing a resource that looks attractive, and can be used in multiple ways. We also wanted to

further explore the capabilities of interactive multimedia for language preservation, presentation, and learning.

The primary content consists of spontaneous monologues spoken by three Karaim speakers (Mykolas Firkovicius, Juzef Firkovich, and Lidija Mashkevic). It amounts to around 200 MB (about 30 mins) of voice recordings. The spoken recordings were selected from original digitally recorded interviews conducted by Csat` , which were redigitised, and then edited to clean up unwanted effects such as coughs and extraneous noises. Csat` transcribed the texts, and, following the design of the data structures that would underlie the CD's functionality, provided detailed linguistic analysis.

The core design connects the spoken text to a Karaim dictionary via the transcriptions. Every word of the transcription texts is linked to a morphological/lexical analysis. In addition, there is access to the dictionary from an English index and a list of grammatical functions. We have provided multiple links between texts, words, linguistic analysis, and thematic essays, so that users can easily navigate to investigate what is of interest to them.

The learning potential of the linguistic content is reinforced by other media: around 150 original photographs and 15 minutes of video are linked thematically to the texts, and Karaim songs and instrumental music specially recorded for the CD by Mykolas Firkovicius and Karina Firkaviciute are accessed through the CD's "songbook".

We are enormously grateful to be able to publish this rich, original material; those making CD projects can find that obtaining rights to reproduce others' materials consumes the bulk of the project's resources (Boaz & Boaz 1996).

FUNDAMENTAL DESIGN CONCEPTS

The progress of the project was aided by the early formulation of a core set of operational rules. The first was: keep "faithful" to the spoken as the primary content. This rule helped keep us on track and make several crucial decisions including the choice of authoring software (Macromedia Director), and prioritising CD space allocation, sound quality (we used 44Kz, 16 bit mono sound, equivalent to audio CD mono). It also provided guidance for building the data model that underlies the whole design (the sound unit resolution is meaning-based, ie its chunks are intonation units or story paragraphs).

The second rule was that a map of the Karaim Street in Trakai, Lithuania, would form the central design metaphor for accessing the various parts of the CD. Each "unit" of the CD (we call them "stations") corresponds to a particular house or location in Trakai. This not only allows users to easily navigate the CD's contents, but it also means that the way that a particular user navigates the CD will to some extent reflect his or her knowledge of and relationship to the Karaim community. This is a useful property for materials that support endangered languages (see also below).

The third rule was that we should employ appropriate interface methodologies for supporting language learners. The use of the Karaim Street in Trakai as a simple, concrete metaphor is one aspect of our attempt to employ known and effective methodologies for multimedia interface design and language learning. The other interface strategies focus on providing rich sets of hypertext links (see next section).

The fourth rule directed the use of database concepts and methodologies. Using databases provides several advantages:

- it helps develop a common terminology and a workspace for the project participants to develop a shared model for the data and its processing.
- it provides a roughly neutral representation of the CD's core data to allow maximum flexibility for interface design, and restructuring and repurposing for other projects. For example, the concordance function in the CD was not planned at the early stages of the project but was easy to implement due to the transparent and robust data structures already in place; and the first web dictionary for Karaim is currently being produced as a collateral product, based on the same data that is used by the CD's internal dictionary database.
- as a strategic methodology, a database approach allows participants to work in parallel, developing the models and applications for data storage and processing while project partners work on refining the primary content such as sound files, translations, dictionary entries, and so on. These can then be imported into the database and further work carried out using the database application.
- it helps to ensure the robustness of the final application. Complex linked data is guaranteed to be syntactically correct if maintained within a well-designed relational database. On the other hand, materials that are basically hand crafted demand that every possible permutation of links be manually checked.
- it is likely that a database developed to support a multimedia project will become an independently useful resource.

Finally, there are special factors to be considered when producing materials designed to support maintenance or revitalisation of endangered languages (see also below).

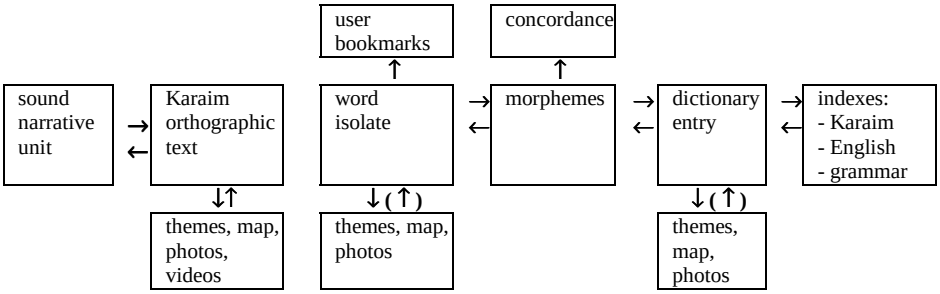
It is difficult at this stage to evaluate the level of success in implementing these rules and methodologies. The third is the most difficult to evaluate because there are few solid rules for effective learning interfaces in this rapidly evolving genre. Interactive multimedia is a new area and there is no theory for the semiotics of interactivity and few generally agreed principles for design (Kress 1997, 73, O'Leary 1998, 1). We will better know about the effectiveness of the CD after it has undergone its user testing (beta) phase and has spent considerable time in the hands of Karaim teachers and students.

INTERFACE: HYPERTEXT

Learners need access to networks of associations between differing types of linguistic and extralinguistic information (Schmitt & Meara 1997). The multimedia

environment is a tailor-made platform for such networks, and the *Spoken Karaim CD* provides a cascading and cross-cutting network of user-navigable and user-configurable links between the CD's structural elements shown in Figure 1.

Figure 1: Links between language elements



Various studies have found that providing glosses and other lookup support improves the effectiveness of vocabulary learning (Lomicka 1998, Hulstijn, J. et al 1996, Chun & Plass 1996, Knight 1994). Hulstijn and colleagues, for example, studied vocabulary acquisition by students using on-line language learning materials. They compared incidental (ie non-directed) vocabulary learning for three groups of students; one group was supplied with marginal glosses, another group could navigate to dictionary entries; the third group had neither marginal glosses nor dictionary. The researchers found that dictionary users acquired more vocabulary than users of the marginal glosses, and that both groups did better than those supplied with no lookup support. However, the authors made an additional finding: due to the extra effort involved in performing dictionary lookup (as opposed to reading marginal glosses placed near the object text), users tended to make less lookups; so the effectiveness of the dictionary is limited by its under-utilisation. The authors concluded that marginal glosses, when frequency of use and effectiveness are both taken into account, were the most efficient means of delivering lookup support; they also called for methods to

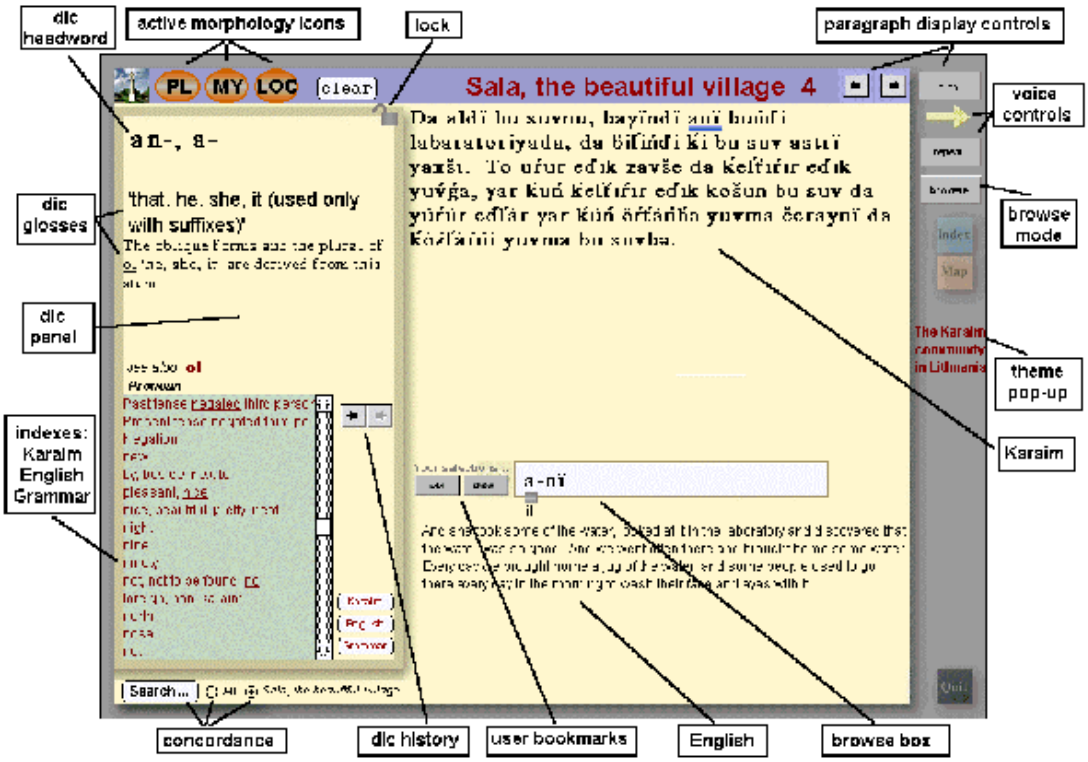
> facilitate the burden of dictionary use because dictionary use interferes with the process of constructing a mental model representation of text meaning

(Hulstijn, J. et al 1996, 337).

On the Karaim CD we have attempted to meet and exceed the optimal lookup support mechanisms suggested by Hulstijn, J. et al. We provide a resource that gives users multiple, easy paths to lookup support. These are summarised in Figure 1; Figure 2 shows parts of the current interface for navigating them. When in the CD's "browse mode", users merely move the mouse over orthographic words in the text to see a morphological segmentation and, below the browse box, the word's *gloss in context*. This is similar to providing marginal glosses, but is both richer and with complete coverage of the content. If the dictionary is not locked, users also see the default morpheme's (typically the first) lexical entry in the dictionary panel. In "normal

mode", users click words to set them in the browse box where they see not only their morphological segmentation and *gloss in context* but can also any morpheme for lexical display. Finally, users can hold down the mouse button over a word to activate a pop up list that allows them to select the word's component morphemes as well as navigate to other associated information that is stored with the word (such as a map for place names, thematic material, or photographs).

Figure 2: Main text screen features



ACTIVE MORPHOLOGY

Karaim is a language with an agglutinative typology and rich morphophonemic alternation. Its inflectional and derivational morphology is typical of Turkic languages. These factors make Karaim morphology suited not only to application of a simple parser but also to visual and interactive manipulation that supports an "active-learner" approach to language presentation. We call this system "active morphology".

An active morphology system does not attempt to *describe* linguistic processes; instead, it *enacts* them via its representation of inflectional morphemes as draggable graphic icons (see Figure 2).

The active morphology system allows users to learn about the morphotactics (constraints on combination of inflections) and the complex morphophonology of Karaim. Users drag graphic representations of inflections onto the dictionary headword to determine the well-formedness of the combination and, if grammatical, the surface form after all the morphophonemic rules have applied. It currently handles three nominal inflections: plural, first person possessive, and locative case.

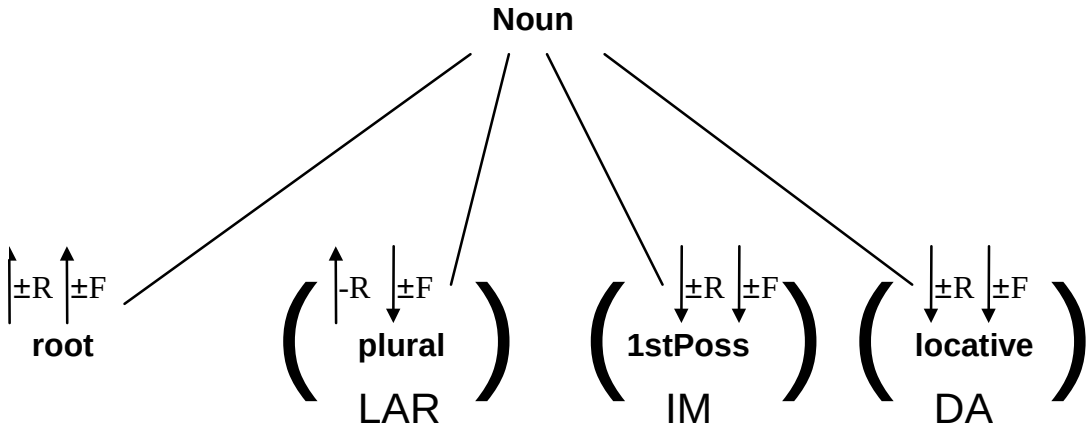
The main steps in the implementation are:

- morphotactic validation
- code object creation
- instantiation of morpheme-specific features

The system's object-oriented string generator receives its input from the user's actions and the currently active dictionary entry. If the inflection can be added, a code object is generated for each component of the combination, including the stem. The code object has the following attributes:

- a form (abstract archiforms for inflections)
- a grammatical category
- feature values for frontness and roundness (can be un- or under-specified)

Figure 3: Feature passing for code objects (R=round, F=front)



An object also inherits from its parent class (which is "Noun" for all cases) these attributes:

- a positional index
- the form of its predecessor
- frontness and roundness features

In addition each inflection code object has two morphophonological processing methods. An inflection-specific method selects its appropriate surface form on the basis of forward chained word-global frontness and roundness features applied with respect to syllable structure. A general method, inherited from the noun class, handles surface liaison processes:

assimilation:	palatalise consonants between front vowels
	root-final back unvoiced consonants are voiced before a vowel
	locative's initial consonant takes the voice-value of its preceding consonant
dissimilation:	a sequence of laterals is dissimilated to velar nasal + lateral

THE RESOURCE APPROACH

Language learning materials such as the *Spoken Karaim CD* should not attempt to usurp the role of the language teacher or specialist. Indeed, when used for language learning, the CD should in most cases be used under the guidance of a suitably knowledgeable Karaim speaker or teacher. There are two main reasons for avoiding directive pedagogy in constructing the *Spoken Karaim CD*. The first comes from the medium itself; the second comes from Karaim's status as an endangered language.

The very notion of interactivity has built into it the view of the user as a collaborative author in a joint construction of an emergent, unique, text (Bolter 1991). Some claim that hypertext gives readers and writers equal participation in creating texts (cf Snyder 1997). This means that the user shares authority and authorship with the original creator of the resources: learning using interactive multimedia enables content to be configured by the user in "deferred negotiation" with the resource creator.

Some, such as Kress, argue that the changes in our ideas about texts and their users are related to changes in fundamental modes of literacy rather than technologies for delivery. Kress finds evidence of a deeper trend through the second half of this century, reflected in many publishing forms including books and newspapers, that is shifting

> an older organisation of *text* to a newer organisation of *resource*;
> from an older concern with knowledge, to a newer concern with gathering information to manage a task

(Kress 1997, 66).

In any case, creators of language learning multimedia need to cater for a multiplicity of users, learning goals, and methods, whilst *minimally* constraining the paths and choices that bring the users' experience into contact with the author's expert knowledge.

A resource-based interactive product also enhances the potential for what I call "hyperinteraction" - socially-integrated interaction between participants *around the computer*, using the target language. This can occur when learners are using a content-based resource collaboratively, neither privately nor competitively.

SUPPORTING ENDANGERED LANGUAGES

The second reason for adopting a resource approach is that in the case of endangered languages we need to channel limited resources toward providing materials for learning rather than materials for teaching (cf Greenhalgh 1996). Many existing CALL materials are exercises that exploit the computer's ability to give immediate feedback but do not supply the content that is so crucial for teaching endangered languages.

While the convergence of media technologies means that learners of major languages increasingly have access to "genuine" cultural teaching content such as radio and TV content, learners of minority and endangered languages enjoy no such support.

The most urgent task is to address teachers needs. Teachers of endangered languages often have incomplete language fluency in the target language, but on the other hand are skilled at making teaching activities if they are supplied with appropriate language content. So we can be wasting precious resources if we build exercises rather than original and rich language content.

It is also important for materials that support endangered languages to "speak at a number of different levels". They need to not only provide a vehicle for learning, but also enhance the *motivation* for learning, by signalling that the language's survival *is* important, indicated by the existence of the learning resource itself.

By focussing on content we can also maximise the delivery of all-important cultural contexts for language meaning and language use (Crozet & Liddicoat 1997). For example, the Spoken Karaim CD contains stories about topics as diverse as religious festivals, Karaim history, and vegetable growing.

Furthermore, consider some of the more abstract messages that young Karaim people using the *Spoken Karaim CD* may receive:

- knowledge of Karaim culture is valuable
- recognition of Karaim speakers as experts about their language and culture
- recognition of Karaim speakers as entitled to make canonical texts
- the "author's" voice is located at the participatory centre of the culture, not in the office of the researcher, programmer, or editor

CONCLUSION

Interactive multimedia is ready to serve as an important delivery platform for language learning. It at last allows the co-ordinated presentation of sound - the fundamental platform for human language - with text, graphics, and video.

The challenge is to create useable, and *well-used* resources that support communities and teachers in their own efforts. The challenge is great because in language learning perhaps more than any other area, content, methodology, and interaction are tightly interwoven. Therefore we need to rapidly produce such resources for a wide variety of language needs in order to learn about their effectiveness.

We are eager not only to assist in the maintenance of the Karaim language, but also to understand what effect the *Spoken Karaim CD* might have on the future course of Karaim language learning and indeed the Karaim language itself.

REFERENCES

Boaz, J & Boaz, M. (1996). TS Eliot on a CD-ROM: A Narrative of the Production of a CD, *Computers and the Humanities* 30: 131-138.

Bolter, J. (1991). *Writing Space: The Computer, Hypertext, and the History of Writing*. Erlbaum: Hillsdale NJ.

Burston, J. (1996). "CALL at the crossroads: myths, realities, promises and challenges", in *Australian Review of Applied Linguistics* 19, 2, 27-36

Chun, D. & Plass, J. (1996). Effects of Multimedia Annotations on Vocabulary Acquisition. *The Modern Language Journal* 80, ii, pp 182-98

Cooper, A. (1995). *About face: the essentials of user interface design*. Foster City, CA; IDG Books

Crozet, C. & Liddicoat, A. (1997). Teaching culture as an integrated part of language teaching; an introduction. *ARAL: Teaching Language Teaching Culture* Series S No 14, pp 1-22

Csat` , va gnes (Forthcoming). *Spoken Karaim*. Wiesbaden: [Turcolgica], Wiesbaden: Harrassowitz.

Dersot, L. (1998). Textuality at the dawn of the digital era - a brief overview of France. http://www2.gol.com/users/dersot/BC1.HTML

Firkavicuite, Karina (1995). The liturgic chant of the Lithunian Karaims. Master thesis. Vilnius: Academy of Music.

Greenhalgh, M. (1996). ArtServe and the Teaching of Art History. *AUUG 1996 Proceedings*, 172-184. (http://rubens.anu.edu.au)

Hulstijn, J. Hollander, M. and Greidanus, T. 1996. Incidental Vocabulary Learning by Advanced Foreign Language Students: The Influence of Marginal Glosses, Dictionary Use, and Reocurrence of Known Words. *The Modern Language Journal*, 80, iii, 327-339.

Knight, S. (1994). Dictionary Use While Reading: The Effects of Comprehension and Vocabulary Acquisition for Students of Different Abilities. *The Modern Language Journal* 78, iii (1994), pp 284-299

Knobel, M. Lankshear, C. Honan, E. & Crawford, J. (1997). The wired world of second language learning. *Page to Screen: taking literacy into the electronic era*, edited by I. Snyder, 20-50. Sydney: Allen & Unwin.

Kress, G. (1997). Visual and verbal modes of representation in electronically communicated communication: the potentials of new forms of text. *Page to Screen: taking literacy into the electronic era*, edited by I. Snyder, 53-79. Sydney: Allen & Unwin.

Lomicka, L. (1998). “To Gloss or Not to Gloss?”: An investigation of reading comprehension on-line. *Language Learning and Technology* 1, 2, January 1998, pp 41-50

Nathan, D. 1998. “Plugging in Indigenous knowledge - connections and innovations”. Paper read at the 1997 Fulbright Symposium: Indigenous people in an interconnected world. To appear in *Australian Aboriginal Studies* 1998, also at http://www.uiowa.edu/~anthro/fulbright/abstracts/nathanful.htm

Neilson, J. (1995). *Multimedia and Hypertext: The Internet and Beyond*. Cambridge Mass: Academic Press

O’Leary, M. (1998). Review of Technology-Enhanced Language Learning. *Language Learning and Technology* Vol 1, No 2 January 1998, pp 20-22

Schmitt, N. & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Langauge Acquisition* 19, pp 17-36

Snyder, I. (1997). Beyond the hype; reassessing hypertext. *Page to Screen: taking literacy into the electronic era*, edited by I. Snyder, 125-143. Sydney: Allen & Unwin.

Warschaur, M. & Healey, D. 1998. “Computers and language learning: an overview”. *Language Teaching*, 31, pp 57-71. Cambridge University Press

ACKNOWLEDGEMENTS

This paper was written during my stay as a Visiting Associate Professor at the Institute for the Study of the Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies, Tokyo, Japan. I would like to acknowledge the assistance of the following people: Eva Csat` , Minegishi Makoto, Mukai Hirokazu, Lionel Dersot, Hayasi Tooru, Toyosima Masoyuki