Multiliteracy, past and present, in the Karaim communities

Éva Á. Csató and David Nathan

1 Literacy in the Karaim communities of Eastern Europe

1.1 Introduction

For endangered languages, the topic of orthography often arises because the language concerned has no existing or standardised writing system and the development of one is seen as a step towards creating a practice of literacy and creating a corpus of written texts (see, e.g. Ostler and Rudes 2000; Grenoble and Whaley 2006:103). The case of the Karaim communities in Eastern Europe is different; in fact, it runs opposite to these trends. Thus, we hope that this paper may contribute to a broader understanding of the variety of roles and dynamics played by orthography and literacy in endangered language contexts.

Several papers in this volume discuss the formulation of an orthography, but, since in virtually every case, people will want to use computers to prepare texts, there are also technical aspects to consider. These include the identification of existing character sets that contain the characters that are needed, and selecting a suitable font, or creating such resources from scratch for the project at hand. Such technical considerations may even go as far as to cause reconsideration of some choices in the design of the orthography.[1]

This paper traces the history of literacy in the Karaim communities of Eastern Europe, and how it influences today's efforts towards language revitalisation. Literacy in these Karaim communities has developed over many centuries. The 20th century saw massive cultural and political influences including reversing tides of occupation in World War 2 and, less than two generations later, and all within a decade, the deSovietisation and independence of their countries, entry to the European Union with new roles for minorities, a surge of interest in endangered languages, and the arrival of the new communication technologies – Internet and multimedia. This paper cannot do justice to these massive changes. However, by observing how communities respond to such impacts, and how they react to the needs of language revitalisation caused by the rapid attrition of the last generation of speakers with full mastery of the speech, scripts, and scriptures, we can learn about how we linguists can assist them.

In the first part of the paper, we introduce the Karaim communities of Eastern Europe, and summarise their orthographic practices over centuries. We outline some particularities of the Karaim communities, including their liturgical tradition, and

[1] This paper does not discuss several other related technical issues such as input methods, keyboards, spellcheckers, or font design.

show how the tides of history that have flowed across them are reflected in a variety of co-existing orthographies.

The second part of the paper describes the implementation of a 'Turcological notation' for the interactive multimedia CD-ROM *Spoken Karaim*, which has been described elsewhere (Nathan 2000, Nathan and Csató 2006). The section provides a brief tutorial on handling characters, raising many of the issues that will face anyone who wants to create and present electronic texts in non-mainstream writing systems.

In the third part of the paper, we document orthographic aspects within a shorter time span – the five years since the release of *Spoken Karaim* and the commencement of regular annual language summer schools in Trakai, Lithuania. In this period, the participation of a wider array of community members, and the bringing out *into public* of individuals' varied skills with and attitudes to orthographies, provided a new theatre for the examination of orthographic preferences and efficiencies, made all the more complex as contemporary political events influenced attitudes to the national language, Lithuanian. In this environment, we saw that the Turcological notation was possibly not an optimal one for community use, and we undertook new work to address this for *Spoken Karaim*. However, more importantly than that, we found that the key issue was not so much choosing a *better* orthography, or the *right* orthography, but was the provision of a *variety* of orthographies.

1.2 Karaim orthographies through the ages

The Karaims follow Mosaic beliefs that were developed in the Middle East in the 8th and 9th century; for more about the history of this movement, see e.g. Gil (2003). Adherents of this non-Rabbinic branch of Judaism are generally called Karaites. The Karaite communities of Eastern Europe – in Lithuania, Poland, Russia and Ukraine – have been speakers of Kipchak Turkic varieties and are called Karaims in contrast to the non-Turkic speaking Karaites living today mainly in Israel. Their Turkic language is called Karaim; see a presentation of Karaim e.g. in Csató (2001).

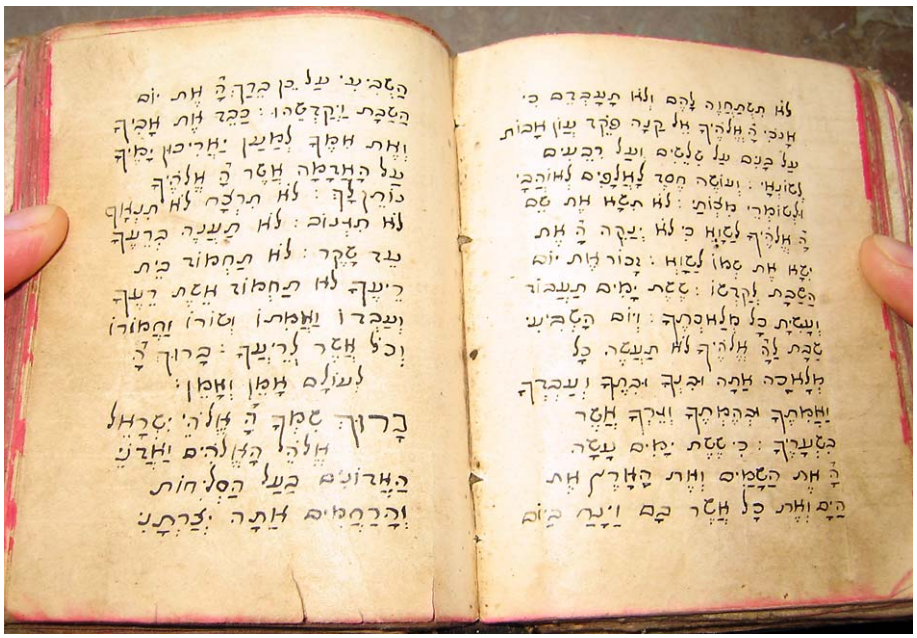
Today, the only living variety of Karaim is spoken in Lithuania. While it is highly endangered, the two other major varieties, those of the Crimea and Halich, are practically extinct. Whereas the total number of Karaims in the world is about three thousand, the number of speakers is not more than forty people, nearly all of whom are elderly. Following the collapse of the Soviet Union, the communities in Eastern Europe, i.e. in Lithuania, Poland, Russia and the Ukraine (Crimea and Halich), have developed an ardent interest in the revitalisation of their cultural heritage including the language. Their motivation is partly to use the language again to emphasise community identity, but even more importantly, to continue to be able to use the Karaim language in their religious practice.

In traditional Karaim religious practice, members of the community read Old Testament texts in both Hebrew and Karaim translation. Turkic speaking Karaims started to translate these biblical texts into their native language long ago; the Karaim

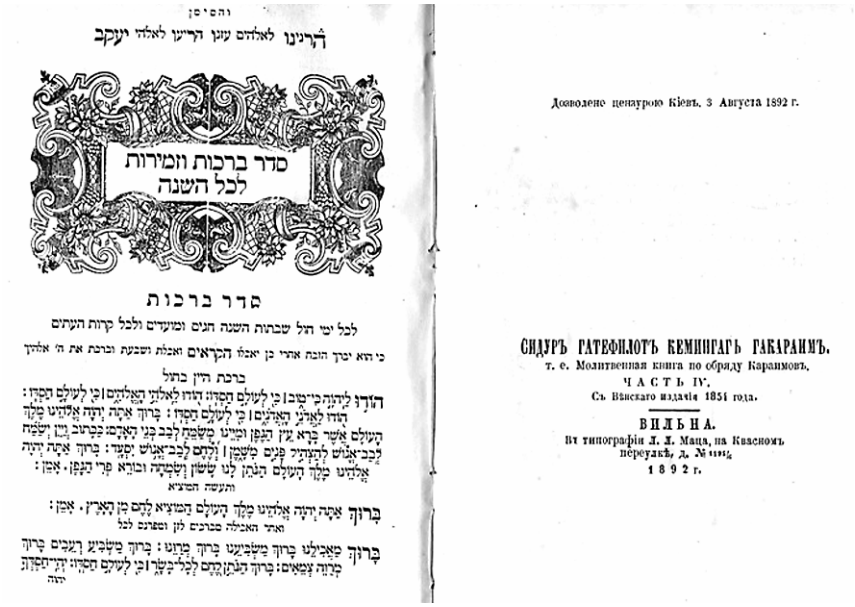
prayer-book printed in Hebrew in 1528-1529 contained a religious hymn in the Karaim variety spoken on the Crimea. Although the first Bible translations into Karaim were printed as late as the 18th and 19th centuries, the language of these books is archaic, indicating a long tradition of translating biblical texts into Karaim. The Hebrew literacy tradition was dominant in the Karaim communities where Hebrew was also the language of scholarship. Karaim scholars wrote important works in Hebrew discussing religious issues with the Rabbanites (Walfish 2003).

As Hebrew was the language of Karaim scholarship, the Karaim used the Hebrew script to write their community language, too. Although Karaim communities spoke different varieties, and the language of Karaim texts differed from community to community, all communities wrote Karaim using the Hebrew script. This common orthographic tradition helped to bridge dialectal differences so that communities could use the same prayer-books; specific features of each Karaim variety did not present any great problem in reading the religious texts. The two illustrations, Examples 1 and 2, show a handwritten Karaim prayer-book containing text in Hebrew and Karaim and a printed *Siddur*, i.e. the Karaim version of the Hebrew prayer-book containing prayers and additional information relevant to the daily liturgy during the whole calendar year, published in Vilnius in 1892.

Example 1. *Hand-written prayer book in Hebrew script from Halich*



Example 2. *Karaim Siddur printed in Vilnius in 1892*



סדר ברכות וזמירות
לכל השנה

סדר ברכות

Дозволено цензурою Кіевъ. 3 Августа 1892 г.

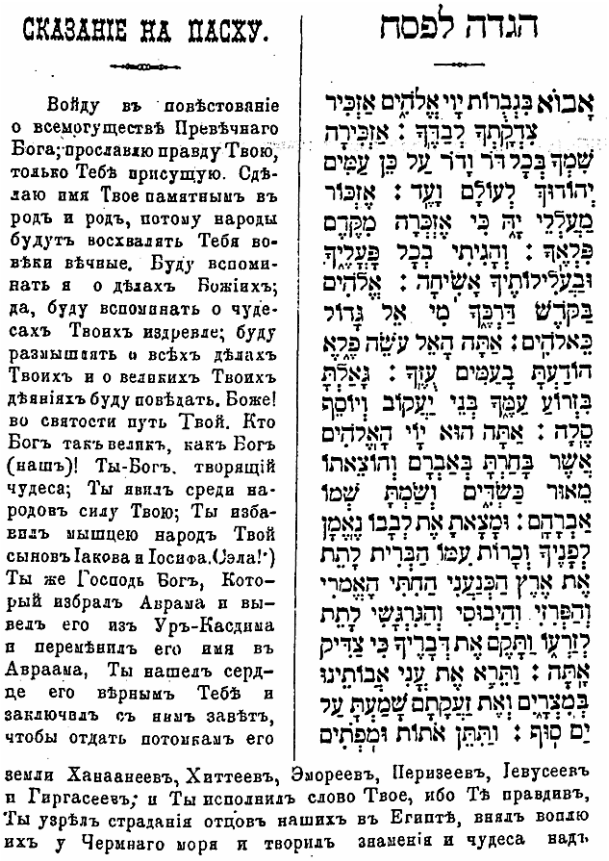
СИДУРЪ ГАТЕФИЛОТЪ КЕМИНГАГЪ ГАКАРАИМЪ.
т. е. Молитвенная книга по обряду Караимовъ.
ЧАСТЬ IV.
Съ Виленскаго изданія 1851 года.

ВИЛЬНА.
Въ типографіи Л. И. Маца, на Квасномъ
переулкѣ, д. № 1894
1892 г.

Before the Soviet era, children learned in the Karaim religious school, the *midrash*, to write and read the Hebrew script. Thus, this script also came to be used for writing Karaim in other contexts, e.g. writing private letters.

The Hebrew literacy tradition was first broken in the Crimean community. In the early 19th century, Crimean Karaims switched from speaking the Karaim language to Crimean Tatar and Russian, lost their competence in reading the Hebrew script, and began to use Russian in their religious practice. Example 3, in which the Hebrew text is translated into Russian, exemplifies this development. A new literacy tradition in Russian developed; many members of the Crimean Karaim community also lived in Moscow where they published on Karaim issues in Russian. This tradition continues today amongst the Karaims of Moscow.

Example 3. *Page 1 of Karaim Haggadah for Passover Eve According to the Custom of the Karaites with a translation in the Russian Language by Shlomoh Prik, Odessa 1901.*



СКАЗАНІЕ НА ПАСХУ.

Войду въ повѣствованіе о всемогуществѣ Превѣчнаго Бога; прославлю правду Твою, только Тебѣ присущую. Сдѣлаю имя Твое памятнымъ въ родъ и родъ, потому народы будутъ восхвалять Тебя вовѣки вѣчные. Буду вспоминать я о дѣлахъ Божіихъ; да, буду вспоминать о чудесахъ Твоихъ издревле; буду размышлять о всѣхъ дѣлахъ Твоихъ и о великихъ Твоихъ дѣяніяхъ буду повѣдать. Боже! во святости путь Твой. Кто Богъ такъ великъ, какъ Богъ (нашъ)! Ты-Богъ, творящій чудеса; Ты явилъ среди народовъ силу Твою; Ты избавилъ мышцею народъ Твой сыновъ Іакова и Іосифа. (Сэла!) Ты же Господь Богъ, Который избралъ Аврама и вывелъ его изъ Уръ-Касдима и перемѣнилъ его имя въ Авраама, Ты нашелъ сердце его вѣрнымъ Тебѣ и заключилъ съ нимъ завѣтъ, чтобы отдать потомкамъ его земли Хананеевъ, Хиттеевъ, Эмореевъ, Перизеевъ, Іевусеевъ и Гиргасеевъ; и Ты исполнилъ слово Твое, ибо Ты правдивъ, Ты узрѣлъ страданія отцовъ нашихъ въ Египтѣ, внялъ воплю ихъ у Чермнаго моря и творилъ знаменія и чудеса надъ

הגדה לפסח

אָבוֹא בִּגְבֻרוֹת יְיָ אֱלֹהִים אַזְכִּיר
צִדְקָתְךָ לְבַדֶּךָ : אַזְכִּירָה
שִׁמְךָ בְּכָל דּוֹר וָדוֹר עַל כֵּן עַמִּים
יְהוֹדוּךָ לְעוֹלָם וָעֶד : אֶזְכּוֹר
מַעַלְלֵי יָהּ כִּי אֶזְכְּרָה מִקֶּדֶם
פִּלְאֶךָ : וְהָגִיתִי בְכָל פָּעֳלֶךָ
וּבַעֲלִילוֹתֶיךָ אָשִׂיחָה : אֱלֹהִים
בַּקֹּדֶשׁ דַּרְכֶּךָ מִי אֵל גָּדוֹל
כֵּאלֹהִים : אַתָּה הָאֵל עֹשֵׂה פֶּלֶא
הוֹדַעְתָּ בָעַמִּים עֻזֶּךָ : גָּאַלְתָּ
בִּזְרוֹעַ עַמֶּךָ בְּנֵי יַעֲקֹב וְיוֹסֵף
סֶלָה : אַתָּה הוּא יְיָ הָאֱלֹהִים
אֲשֶׁר בָּחַרְתָּ בְּאַבְרָם וְהוֹצֵאתוֹ
מֵאוּר כַּשְׂדִּים וְשַׂמְתָּ שְׁמוֹ
אַבְרָהָם : וּמָצָאתָ אֶת לְבָבוֹ נֶאֱמָן
לְפָנֶיךָ וְכָרוֹת עִמּוֹ הַבְּרִית לָתֵת
אֶת אֶרֶץ הַכְּנַעֲנִי הַחִתִּי הָאֱמֹרִי
וְהַפְּרִזִּי וְהַיְבוּסִי וְהַגִּרְגָּשִׁי לָתֵת
לְזַרְעוֹ וַתָּקֶם אֶת דְּבָרֶיךָ כִּי צַדִּיק
אָתָּה : וַתֵּרֶא אֶת עֳנִי אֲבוֹתֵינוּ
בְּמִצְרָיִם וְאֶת זַעֲקָתָם שָׁמַעְתָּ עַל
יַם סוּף : וַתִּתֵּן אֹתוֹת וּמֹפְתִים

The two other Karaim communities, in Halich (Ukraine) and in Lithuania, continued to use the Hebrew script until the Soviet times but in addition they developed an orthography based on the Polish script using Polish Latin characters. They published journals, books and a dictionary in this Polish-based system. As both communities were literate in Polish, which was the language of education at that time, this literature could be used both in Halich and in Trakai (in Polish, Troki). See the example taken from a Karaim journal published in Luck in 1930 (Example 4). The page contains two poems, the first written by Ribbi Itzhak son of Abraham in Troki and the other by Ribbi Shemoel son of Moshe in Halich. This shared Polish-based literacy broke down after World War 2, when new borders placed the Karaim territories within the Soviet Union.

Example 4. *From the Karaim publication "Zemerłer" 1931.*

ZEMERŁER

NE BYŁA UTRUŁAJYM.*)

Tizidi ribbi Icchak uwłu Awrahamnyn, Troktan.

Ne była utrułajym
Seni, bijim Tenrim,
Kurbanym ornuna
Jalbarsyn sa ernim.

Tizet mana yzym,
Tefillegio tursam,
Bolhyn bołusłuhum
Isimdo men talsam.

Bijim jaratuwcum,
Bosat jazyhymny,
Kieter awurłuhum
Tapma azyhymny.

Bakkyn Sen kułuna
Acuwłandyn da ese,
Ał dzanym ez alnyna
Jazykły da ese.

Ułułuhun kiergiz
Bar dusmanlaryna,
Ki umsunctur Sen
Bar isanuwcularyna.

Kipło jarłylarny,
Usajt tellerni,
Galut bolhanlarny
Jomdar, bijim, alarny.

AZIZ KIEKTE MAŁACHŁAR.

Tizidi ribbi Szemoeł uwłu Moszenin, Halicten.

Aziz kiekte małachlar
Birlihin Tenrinin sarnajdłar,
Bir kawanaba ajtadłar:
Ułudur bijimiz!

Da gałgałłar safirli
Hem ki dzanlar hermetli
Machtaw była bijencli:
Ułudur bijimiz!

Aziz Tenri machtalhan
Izlewcihe tabulhan
Ij rachmetlerinni oł sunulhan.
Ułudur bijimiz!

Kier chajifsin israelni,
Sahyn ic ataha sertinni,
Kajir biżgie sawahatynny.
Ułudur bijimiz!

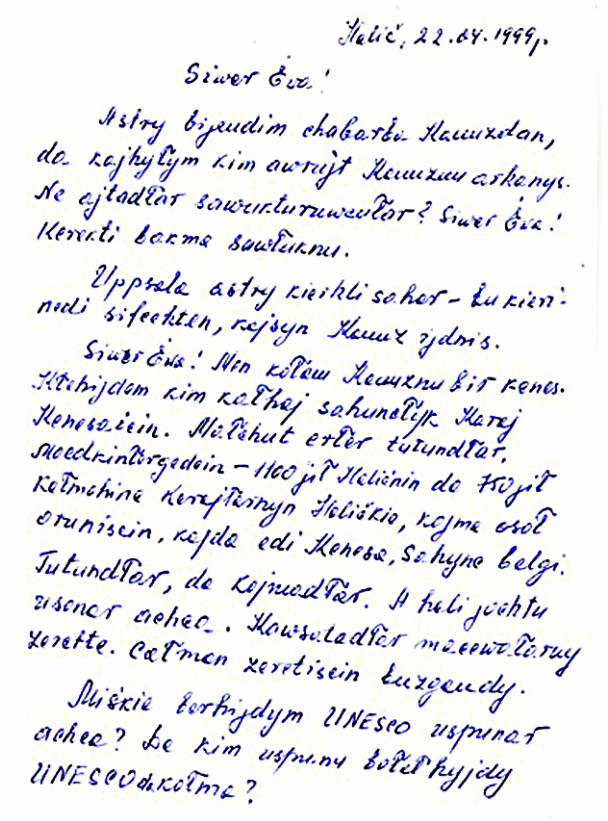
Sira ajtajik Tenrigio
Ceber awaz była birgio:
Kondar mikdaszynny israelgie.
Ułudur bijimiz!

Bunu kiergiej chanłyklar
Hem ki imencli bołhajłar,
Jumusuna kohenler turhajłar.
Ułudur bijimiz!

*) Uchuład szabat—kin.

After World War 2, the Hebrew literacy tradition was completely abandoned. Practice of religion was dangerous and the midrash was closed down by the authorities. The languages of education were Lithuanian and Russian in Lithuania, and Russian in the Ukraine. The communities were dispersed and there was little demand for publications. The majority of Halich Karaims left for Poland and the few who were left did not have the capacity to publish in Karaim. They continued to use the Polish-based Latin orthography in private use. See Example 5.

Example 5. *Private letter in Halich Karaim written in 1999 in the Polish literacy tradition*



Lithuanian Karaims then began using a Cyrillic orthography to publish literary texts in Karaim. The use of Cyrillic orthography, although not new, was restricted to those who were educated in Russian. Karaim communities in Poland – outside the Russian speaking territories – continued to use the Polish orthography. In 1974, the Soviet Academy of Sciences compiled a Karaim-Russian-Polish dictionary in which they also included transcriptions in a Latin-based Turcological notation (Baskakov et al. 1974); see Example 6). Thus, literacy in the Karaim communities had become split between Polish-based Latin and Russian-based Cyrillic literacies.

Example 6. *Karaim-Russian-Polish dictionary 1974*

карайка *th* [karaịka K, karajka M] караимка|Karaimka.

карайкачэх *t* [karaikačex O II, 17 v. 12] *уменьш. от* **карайка** караимочка, молодая караимка|młoda Karaimka, Karaimeczka; **утруладым игит карайкачэхны** я встретил молодую караимку|spotkałem młodą Karaimeczkę.

карайлык *h* [karaịłyx MK I/1, 19 v. 4] караимский народ; караимская вера|lud karaimski; wiara karaimska; *ср.* **карайлых.**

карайлых *t* [karaịłyx O III, 18 v. 39] караимский народ; караимская вера|lud karaimski; wiara karaimska; **ташламады сагъышын карайлых ўчюнь** он не бросил свои мысли о вере караимов|nie wyzbył się myśli o karaimiźmie; *ср.* **карайлык.**

After the fall of the Soviet Union the independent Lithuanian republic was re-established, and new freedoms began to be enjoyed. From 1991 the official language of the new Lithuanian republic became Lithuanian, written in Latin characters; see Example 7. For Karaim people, there was now little motivation for continuing to use Cyrillic, nor for reverting to the old Polish system, and therefore a new Lithuanian-based orthography was introduced for writing Karaim. However, this has had important ramifications for the *other* Karaim communities, as we discuss below.

Example 7. *The Lord's prayer in Karaim, written in the new Lithuanian orthography*

Atamyz ki kiokliardia
machtavlu bolhej birligi adyjnyn
da kip bolhej bijligij
da kliagij kiokliardia johartyn
da jer üśtiunia ašahartyn.
Kiuńdiagi öt'miagimiźni biergiń
biźgia
da bošatchyn bar jazychlarymyzny.
Tiuź jollaryjdan azaštyrmahyn biźni,
ančach kutcharhyn biźni
azhyrtuvčudan,
amień.

Today, only the Lithuanian Karaims still have competence in speaking and writing the Karaim language. Competence in Hebrew is practically nonexistent. The late hazzan, religious and administrative leader of the Lithuanian Karaim community, Mykolas Firkovičius, published the most important religious books, prayers, psalms, and calendar after transliterating them from the Hebrew script into the new Lithuanian orthography (Firkovičius 1993, 1994, 1996, 1998-1999, 2000). His Karaim language-learning textbook is also written in this script (Firkovičius 1996). The result, however, is that Karaims who live outside Lithuania (and are therefore unfamiliar with the Lithuanian orthography) cannot easily use these texts.

The divergence of the Karaim communities' literary traditions has thus resulted in a very complex situation. Lithuanian Karaims write in the Lithuanian system, the Karaims of Poland and Halich in the Polish one, other Karaim communities of Ukraine and Russia write in Russian, and Karaims of the diaspora beyond these countries have difficulty in reading any of them. There is no move to introduce Hebrew literacy in any of the communities, partly due to the strong emphasis that most Karaims place on their Turkic ethnic identity. Another contributing factor is the negative attitude of some Jewish writers towards the Turkic speaking Karaims.[2] In turn, these factors have strengthened the role of the Lithuanian script in which all the religious texts now exist.

All Karaim communities are today interested in revitalising their language. Supporting them in this endeavour is a response to the self-declared needs of the Karaims, and not a case of 'salvation linguistics' (Matras 2005) (see also discussion on the role of outsiders in Grenoble & Whaley 2006: 192-197). Karaims would like to use their own various writing systems, and teaching materials must be developed. However, currently the teaching of Karaim is almost entirely restricted to the annual Karaim Summer School held in Trakai, Lithuania[3] where members of all Karaim communities and diaspora gather in order to revitalise the language and culture, in the only place where the language still lives in an everyday sense. Thus, Karaims from all communities are now learning this variety. However, we have found that although former dialectal differences in Karaim are no longer important, the text materials we are developing in the local orthography cause difficulties for many of the students. In addition, these materials cannot be used for local efforts in Moscow, Warsaw and the

[2] The hostility of some Rabbinic communities towards the small 'sectarian' Karaim communities has a long and sad history. Today this attitude is still prevalent; for example, D. Shapira, in a handbook on Karaite Judaism, states: "The Turkic identity adopted by the East European Karaites *in the course of the twentieth century* [emphasis ÉC & DN] may have saved them from physical destruction by the Nazis, but it was this very identity that caused their ultimate disappearance as a collective ..." (Polliack 2003: 701). The author not only distorts the historical facts but also denies the existence of today's Eastern European Karaim communities.

[3] The main sponsor of the Karaim Summer Schools has been the Swedish Institute, Stockholm.

Crimea. Thus there is a strong need to present Karaim texts not only in Lithuanian, but also in Polish, and Russian and a neutral orthography for Karaims in other countries.

2. Implementing an orthography for *Spoken Karaim*

2.1 Choosing the orthography

Against the backdrop of orthographic history we discuss the development of *Spoken Karaim*, an interactive multimedia CD-ROM that combines sound, text, linguistic information, images, and video, and allows users to navigate among these resources. It has become a flagship resource in the revitalisation of the Karaim language. When we began designing it, in 1997, we decided to use an orthography based on a Polish Turcological tradition (Kowalski 1929), where each grapheme is fully distinguished for phonetic detail. It is a Latin script, but is not associated with any particular language other than Karaim. It had also been used in the Karaim-Russian-Polish dictionary (Baskakov et al 1974) and in the comprehensive Karaim corpus published in Kowalski (1929). We felt it was quite apt for *Spoken Karaim* because the CD is centred on sound recordings; a learner/user can listen to the voices whilst reading the texts at the same time, with quite transparent correspondence between the two. In other words, the orthography is oriented towards the sound of Karaim, not towards the writing system of any national language. We also felt that, being a shallow orthography (i.e. it is governed by representation of sound, not morphological, lexical or etymological factors), it would also better assist learners.

2.2 Strengths and weaknesses of computing for orthographies

Orthography design, promotion, and usage involve pragmatic interweaving of linguistic, political, practical, and technical issues,[4] any of which may require compromise on the part of any of the others; see also Grenoble & Whaley 2006: 137-159; Mosel 2004: 43. The introduction of computers brings new opportunities and new problems to dealing with orthographies. We now typically input, store, process, exchange, display, and disseminate texts mainly using computers, and the main issues can be summed up as follows:

- A system might allow users to always work with an orthography using only the graphemes intended, i.e. WYSWYG ("what you see is what you get"), with details of the implementation kept opaque to the user[5]
- However, the above is not feasible for most writing systems. Until it is achieved, the capacities and constraints of computer representation of characters need to be recognised. We may need to work differently with text

[4] … as well as learning and pedagogical issues, which we do not discuss here.

[5] Unicode may eventually offer this capability.

depending on whether we are inputting (creating) it, processing, displaying, or disseminating it.

Once an orthography has been designed and potential users have been consulted, the technical aspects of implementation can begin. Methods for inputting, storing, exchanging and processing text will need to be identified, and a font that can display all the characters required needs to be found or created. The main point to remember is that if there are a number of non-standard characters to deal with, a lot of future anguish will be avoided if you distinguish between the following:

- inputting text
- representation and storage of text
- processing and exchanging text
- display of text, such as in various printed or electronic publications

For *Spoken Karaim*, we needed a system for text management that allowed us to input text and to robustly store and exchange it across a variety of computer hardware and software systems, since we use equipment ranging from MacOS to Windows of various releases, and in language versions including German, Japanese, English, Turkish, and Lithuanian. The one constant among all of these, and most computers in the world, is the ASCII character set, which consists of the characters a-z, A-Z, numbers, and some punctuation, currency and other basic symbols. This set is available from virtually any computer's keyboard (and therefore has an *input method*) and is consistently encoded and displayed by all computers. Despite its Americo-centric basis, its importance as a globally consistent set cannot be underestimated, not only for its role in the processes listed above, but also for ongoing and long term preservation.

2.3 Text management: a character-sequence system

Drawing on the strengths of ASCII, we designed a system that defines a specific sequence of ASCII characters to correspond to each single grapheme in the Turcological writing system. The system uses similar principles to John Wells' X-SAMPA[6] and makes a clear separation between the entry, storage and processing of text on the one hand, and the display of it on the other.

Example 8: *Character sequence system for writing, storing and processing Karaim Turcological notation*

[6] For X-SAMPA, see http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm. See also http://emeld.org/school/case/ega/x-sampa.html for the application of this method to Ega.

∠	**α#**	α–δοτ
♣	**A#**	A–δοτ
®	**βə**	β–παλ
♦	**Bə**	B–παλ
©	**χə**	χ–παλ
♥	**Xə**	X–παλ
™	**χ#**	χ–η
♠	**X#**	X–η
∏	**χ#ə**	χ–ηπαλ
↔	**X#ə**	X–ηπαλ
√	**δə**	δ–παλ
←	**Δə**	Δ–παλ
·	**ε∴**	κ–σχηωα
¬	**ε∴#**	σχηωα–δοτ
∧	**φə**	φ–παλ
→	**Φə**	Φ–παλ
∨	**γə**	γ–παλ
↓	**Γə**	Γ–παλ
⇐	**γ#**	λχγαμμα
±	**Γ#**	Γαμμα
⇑	**γ#ə**	λχγαμμα–παλ
″	**Γ#ə**	Γαμμα–παλ

◦	**Γ∴**	Γ–βακ
⇒	**ι∴**	ι–δοτσ
≥	**I∴**	I–δοτσ
⇓	**κə**	κ–παλ
×	**Kə**	K–παλ
◊	**λə**	λ–παλ
⊇	**Λə**	Λ–παλ
®	**μə**	μ–παλ
÷	**Mə**	M–παλ
©	**νə**	ν–παλ
≠	**Nə**	N–παλ
™	**ν#**	λχενγ
≡	**N#**	Ενγ
∑	**ν#ə**	λχενγ–παλ
≈	**N#ə**	Ενγ–παλ
⏐	**o#**	o–δοτ
—	**O#**	O–δοτ
⎩	**o∴**	o–δοτσ
↵	**O∴**	O–δοτσ
⎡	**πə**	π–παλ
ℵ	**Πə**	Π–παλ
⏐	**ρə**	ρ–παλ

ℑ	**Pə**	P–παλ
⎣	**σə**	σ–παλ
ℜ	**Σə**	Σ–παλ
⎧	**σ#**	σ–η
℘	**Σ#**	Σ–η
⎨	**σ#ə**	σ–ηπαλ
⊗	**Σ#ə**	Σ–ηπαλ
⎩	**τə**	τ–παλ
⊕	**Tə**	T–παλ
⏐	**υ#**	υ–δοτ
∅	**Y#**	Y–δοτ
□	**υ∴**	υ–δοτσ
∩	**Y∴**	Y–δοτσ
〉	**ϖə**	ϖ–παλ
∪	**ςə**	ς–παλ
⏐	**ζə**	ζ–παλ
⊂	**Zə**	Z–παλ
⎭	**ζ#**	ζ–η
⊆	**Z#**	Z–η
∈	**Z#ə**	Z–ηπαλ
	_∴	φακεσπαχε

In Example 8, the first column shows the grapheme in Turcological notation, the second column shows its corresponding sequence, and the third column shows a mnemonic that assists with disambiguation and preservation. Example 9 shows a short passage as represented in each system.

Example 9: *Sample text in sequential system, and its Turcological display form*

Sequential system	*Turcological notation*
Kayda karaylar t 'ir 'il 'a#d 'l 'a#r bu oram Karay_\orami\ in 'd 'a#l 'a#t '. Kac#an 'es ' da bar vaxtlarni\ de\ ek 'in 'c# 'i dunya yat te\ dunya b 'iz ' d 'e\# in 'd 'a#r_\ed 'ik bu Karay_\orami\n Karaims#c#i\znabe\.	Καψδα καραψλαρ ⎩ι⏐ ιφ∠√φ∠ρ βυ οραμ Καραψ οραμ⇒ ι©√∠φ∠⎩. Κα™α©⎣ δα βαρ ϖαξτλαρν⇒ δ· ε⇓ι©∏ι δυνψα ψατ t· δυνψα ®ι⏐ √¬ ι©√∠ρ ε√ικ βυ Καραψ οραμ⇒ν Καραιμ⎧™⇒ζναβ·.

Such a system provides several advantages. Text is:

- easy to input
- precise

- robust
- portable across different operating systems
- mnemonic and moderately readable
- machine readable, i.e. usable as the basis for various computations[7]

Of course, at some stage, the sequential system text has to be converted into its display form. Here, careful design of the system plays its role, for the conversion can be done simply as a set of ordered search-and-replace operations, for example implemented as a macro in a word processor.[8] However, note that at this stage we have not yet described exactly *what* is produced as the output of the conversion process – we turn to that topic in the next section.

2.3 Creating the display

The second major task in developing the text management for *Spoken Karaim* was to create a font for the Turcological notation. There was no existing mainstream font that provided all the required characters (nor is there today). Some fonts could have provided the right display, by using character combinations – that is, by writing a roman letter followed by a diacritic in the form of a *combining character* that merges graphically with the letter:

Example 10. *Combining characters*:

Base character ("n") plus combining character ("◌́") displays a complex character:

[Type in] **n** [followed by] ◌́ [results in display:] **ń**

Such combining characters are provided in several specialist fonts, as well as in Unicode. This method may have followed quite naturally from the sequential system we used to encode the characters, since in most cases each extra character in the sequential system can be taken to represent a diacritic. However, it is generally *not* a good idea to use combining characters in a computational or interactive environment, because systems do not reliably recognise that the underlying sequence of base character plus combining character represents one character in display.[9] This can be

---

[7] See Section 2.4 for examples.

[8] An excerpt from the MS Word macro we used shows its simplicity, aside from the need for ordering:

```
ChangeChar$("b'", "210")
ChangeChar$("c'", "211")
ChangeChar$("c#'", "213")
…
ChangeChar$("c#", "212")
ChangeChar$("e\#", "216")
ChangeChar$("e\", "215")
```

[9] It can also cause problems with quite basic tasks such as search and replace.

crucial in the case of interactive software, such as Director[10] that was used to create *Spoken Karaim*. This software provides interactivity – for example, where users click a word to do something – by acting upon either the word's content or its serial position. Having one display character correspond to multiple underlying characters creates ambiguity between underlying and display forms, and therefore a potential source of serious error. It was extremely important to avoid this possibility in a product that aimed to be intensely interactive.

In addition, when complex and specialised fonts are required, it is best to spare the user from having to install them. Director offers the convenience of embedding fonts directly in the application, so that the whole matter is hidden from the users. However, Director could not reliably embed fonts using combining characters, and there are copyright restrictions in some cases.

Although we faced these issues nearly 10 years ago, they largely still exist today.[11]

In order, then, to meet the need for simplicity and robustness, and to avoid copyright problems, we created a new font, using Macromedia Fontographer (now owned by Fontlab). The new font, KaraimT, was based on a Times-style font and replaced 74 characters (or, strictly speaking, glyphs)[12] in the slots from 167 to 246 (these are typically slots used for characters other than standard ASCII, such as for various European languages).[13]

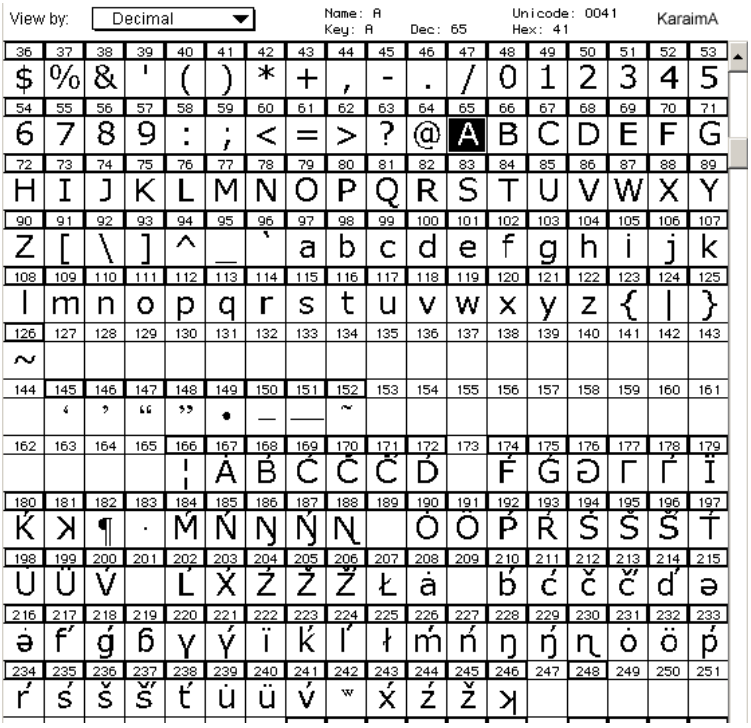
[10] *Spoken Karaim* was developed using the multimedia authoring software Macromedia (now Adobe) Director.

[11] Despite advances in Unicode. Director does not currently support Unicode, although Adobe has stated that the next release in 2007 will do so.

[12] Technically, a character set is a mapping between an ordered number and a character concept, such as "a, the first letter of the Roman alphabet, in its lower case version", rather than a graphic shape; to distinguish the latter, it is known as a *glyph*.

[13] Some slots were not used, since they potentially encode useful characters such as the paragraph mark "¶". For more information about character sets, see Korpela [www].

Example 11. *Using Fontographer to create characters in the "upper ASCII" area*



This method – of "upper ASCII character substitution" has commonly been used to provide specialised character sets, perhaps with even just one or two characters. It has been widely used for linguistic purposes, in cases where just a few ad hoc characters were required, and especially in projects where font creation/modification has been simpler to achieve, such as those using Macintoshes. In some cases, even standard ASCII characters have been replaced by special characters – for example, institutions in the Arnhem Land area of Australia replaced the backslash character ("\") with the glyph for "eng" ("ŋ"). This, of course, was also a strategy to encourage literacy by allowing typing using ordinary keyboards.

Nowadays, such methods are generally frowned upon. There is less need for them because of the development of Unicode and capabilities of most modern software to work with it. The development of the Internet and subsequent mobility of data means that files do not stay within the original domain where particular individuals can ensure that certain fonts are available and are applied to (specific parts of) texts. More importantly, from the perspective of data management and preservation, such substitutions suffer the weakness of being inexplicit and are frequently undocumented. Text files can move from one environment to another, only to be displayed with a sprinkling of meaningless rectangles, or, even worse, meaningful characters that are not those originally intended. Nevertheless, within the constraints of a particular project, where the font is only used within an application and does not expose the user to any of the abovementioned problems, character substitution can present a practical, if not

ideal, solution.[14] In addition, later developments, including our experience of running three annual Karaim Summer Schools, have allowed us to capitalise on its simplicity.

2.4 Other developments

In the previous sections we described methods for dealing with Karaim text that we developed specifically for the initial development of *Spoken Karaim*. However, they also provided opportunities for subsequent developments.

The first development was an extension of *Spoken Karaim's* capabilities that we called "Active morphology". This exploited the close correspondence between the sequential coding system and morphophonological phenomena of Karaim ("vowel harmony"), to build a computational morphophonological model that generates affixes with correct representation of vowel harmony (Nathan 1998, Nathan 2006).

The second development focused on standardisation and preservation. While our text management methods satisfied all the needs of our CD project, it was neither sufficiently documented nor structured to meet current trends towards data portability (Bird and Simons 2003). These include a preference for encoding text data with structural information as XML and character information as Unicode or Unicode-compatible. To achieve this, we documented the KaraimT font/character set fully, and then used *Spoken Karaim* as a computational platform to convert its internal text to representations in XML. Example 12 shows the full morphological encoding of the first three words shown in Example 9 above; notice in particular the encoding of the characters of the third word '⌊ı◊∠√◊∠ρ' as XML-compliant "character entities" that utilise the descriptive mappings given in Example 8 .

Example 12. *XML encoding of Karaim interlinear including character entities*

```
<SEQITEM NUM="1">
      <ORTHO>kay-da</ORTHO>
      <CONTEXTGLOSS>where</CONTEXTGLOSS>
      <MORPHEME>
        <SRCLG>kay-</SRCLG>
        <TGTLG>pronominal stem</TGTLG>
        <LEXSRC DOCID="skcd" ID="41"/>
      </MORPHEME>
      <MORPHEME>
        <SRCLG>-da</SRCLG>
        <TGTLG>Locative case</TGTLG>
        <LEXSRC DOCID="skcd" ID="149"/>
      </MORPHEME>
</SEQITEM>
<SEQITEM NUM="2">
      <ORTHO>karay-lar</ORTHO>
      <CONTEXTGLOSS>Karaims</CONTEXTGLOSS>
      <MORPHEME>
        <SRCLG>karay</SRCLG>
```

[14] In fact the only substantial problem arising from this method applied to *Spoken Karaim* was that there was no support for an Input Method for the Turcological notation – i.e. no way to type the characters in directly. While *Spoken Karaim* does not require the user to type in text, it would have been a useful facility for other associated language activities.

```
      <TGTLG>Karaim</TGTLG>
      <LEXSRC DOCID="skcd" ID="70"/>
    </MORPHEME>
    <MORPHEME>
      <SRCLG>-lar</SRCLG>
      <TGTLG>Plural</TGTLG>
      <LEXSRC DOCID="skcd" ID="160"/>
    </MORPHEME>
</SEQITEM>
<SEQITEM NUM="3">
    <ORTHO>&t-pal;i&r-pal;i&l-pal;-&a-dot;&d-pal;&l-pal;&a-dot;r</ORTHO>
    <CONTEXTGLOSS>they live</CONTEXTGLOSS>
    <MORPHEME>
      <SRCLG>&t-pal;i&r-pal;i&l-pal;-</SRCLG>
      <TGTLG>live</TGTLG>
      <LEXSRC DOCID="skcd" ID="96"/>
    </MORPHEME>
    <MORPHEME>
      <SRCLG>-adlar</SRCLG>
      <TGTLG>Present third person plural</TGTLG>
      <LEXSRC DOCID="skcd" ID="169"/>
    </MORPHEME>
</SEQITEM>
```

3. The Karaim Summer School and *multiliteracies*

3.1 Conversion to Lithuanian-Karaim orthography

The Turcological notation had been well received by linguists and also by Karaims living in and outside Lithuania, and *Spoken Karaim* had become widely known, distributed and used in the communities. However, members of the Lithuanian community raised the question whether it would be possible to use the Lithuanian orthography in *Spoken Karaim*. They had several reasons for coming to prefer the Lithuanian Karaim orthography. By the time the first version of *Spoken Karaim* had been published, Mykolas Firkovičius had already created the Lithuanian orthography, started to teach Karaim children to use it, and had begun publishing a series of important religious and non-religious texts. Community members felt a loyalty to his endeavour. Also, with the changing social and political climate, Lithuanian orthography became more attractive. Consequently, when we ran the first Karaim Summer School in Trakai it became clear that the community's preference was for learners to be trained in the Lithuanian orthography, and we started to produce learning materials in it.

A primary target for the new orthography was *Spoken Karaim*. We decided that we would be able to automatically generate the Lithuanian Karaim text from the existing Turcological text contained in the CD, due to the fact that the Turcological notation is fully specified for relevant phonological features at the grapheme level, and the regular nature of both the Turcological and Lithuanian Karaim orthographies. Developing a system to do this involved several steps. First, we wrote a set of transformation rules for converting from the Turcological to the Lithuanian system. This served to both confirm that one could be generated from the other, as well as to provide the specifications needed by the conversion program. We encoded the rules using a simple syntax, shown in Example 13. This syntax allows the linguist to specify a mapping between characters, and, if required, an environment constraint. For

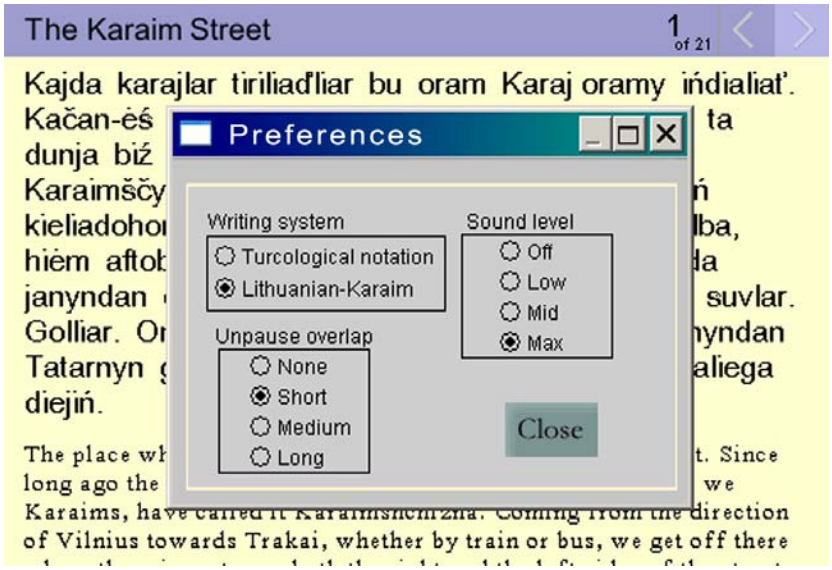
example, `t'>ti:e` means: `change t' to ti where t' precedes e`, and results in, for example, the conversion of ⌊ε®ιρ (“iron”) to tiemir. Then, a program module within *Spoken Karaim* instantiates these rules to convert text in real time from Turcological to Lithuanian notation. The user simply uses *Spoken Karaim*’s preferences panel to select which orthography they prefer to see; they need not be aware that the Lithuanian orthography is actually generated from an underlying Turcological text (see Example 14).

The implications of this conversion go beyond the content of *Spoken Karaim*. The same system can in principle be used as a general-purpose converter to take any input text in Turcological notation and convert it to Lithuanian. While there is a limited body of existing text in Turcological notation, the notation might best serve as an interlanguage, i.e. an intermediate representation within a conversion process. Further investigation is needed to see how fruitful this might be.

Example 13. *Sample subset of rules employed to convert the Turcological notation into the Lithuanian orthography*

```
y>j
x>ch
--  depal before e
t'>ti:e
k'>ki:e
d'>di:e
l'>li:e
n'>ni:e
-- depal before a#
t'>ti:a#
k'>ki:a#
d'>di:a#
l'>li:a#
n'>ni:a#
-- other consonants
n#'>l'
n#>l'
s#'>s#
z#'>z#
-- dissimilation geminate
l'>l:l
n'>n:n
-- catch the rest
b'>b
c'>c
c#'>c#
c#>c#
d'>d'
f'>f
g#'>hi
g#>h
g'>g
k'>k
m'>m
n'>n'
p'>p
r'>r
t'>t'
v'>v
```

Example 14. *The user selects a preferred writing system in the Preferences panel*



3.2 Development of multiliteral publications

The latest version of the Karaim CD thus now allows a choice between Turcological notation and the Lithuanian orthography. This will make the CD more useful for Karaims who are competent in Lithuanian.

However, the urgent task of teaching Karaim to members of *all* Karaim communities requires more work. The divergence of the script traditions in the Karaim communities described in section 1.2 seriously affects today's learners, who come together from various locations to learn the language. Karaims participating at the Summer School come from Moscow, Crimea, Kiev, Warsaw, and elsewhere, with their literacies varying by location. In the present critical situation for the language we believe it is crucial to focus on the maximal effectiveness of the language teaching opportunities. In particular, it is both necessary and urgent to provide textbooks and other language resources, such as *Spoken Karaim*, in the four orthographies that are relevant to the learners:

Example 15. *Four Karaim orthographies*

Spoken Karaim: "Turcological notation"

Τατα| ι‹⇓ι©ι© ∨| ◊| αλν⇒να ψολ™·ξ βαρατ ∨| ◊ αρτ⇒να, βαρ⇒β· ψ⇒ρα ξ ⌊|›|◊

Lithuanian orthography

Tatariškiniń gioliu alnyna jolčech barat giol' artyna, barybe jyrach tiuviul'.

Polish orthography

Tatariszkiniń gioliu ałnyna jołczech barat giol artyna, barybe jyrach tiuwiul.

Russian orthography

Татаришкинин гёлью алнына йолчэх барат гёль артына, барыба йырах тювюль.

Several of the Lithuanian Karaim children do seem to cope with multiple orthographies, perhaps because they speak Lithuanian, Polish and Russian. We have observed them in language learning situations jumping adeptly between the Turcological notation of *Spoken Karaim* and the Lithuanian orthography used in computer language games we have created.

However, we also need to ensure that the older Lithuanian Karaims who are fully-fledged speakers can participate fully in language training activities. That generation's childhood schooling was in Polish or Russian orthographies; they do not easily deal with the new Lithuanian orthography. One older speaker, who we recorded reading Karaim prayers, was competent in the Polish orthography, but could not fluently read the Lithuanian material in our textbook and was embarrassed about his hesitations.[15] Another of the older speakers who taught in the Karaim Summer School had a passive

[15] Older Karaim also experience frustration when, e.g. they are asked to help their grandchildren to read or write in (Lithuanian) Karaim.

knowledge of the new orthography, but could effectively write only in the Polish system, which he used on the blackboard. He also sends email messages in Karaim to his grandchildren, using the Polish orthography (with some interference from the Lithuanian); see Example 16 showing a message sent to one of the authors of this paper.

Example 16. *Email message in the Polish-based orthography*

> Abajly Eva! Iszanam ki kajttyj koduj esianli juvgia. Bizgia astry czebiar bolur kabul etmia kodujdan chabarczechlar. Anyn u"cziun ijam o"z adresymny.[16]
> ("Dear Eva! I hope that you have safely returned home. It will be a great pleasure for us to receive messages from you. Therefore I send my address.")

We now feel that instead of expending time and resources teaching and learning "new" orthographies, it will be more effective to provide each participant with material in the orthography which is best for him/her. Using the experience gained through converting the Turcological notation into the Lithuanian-based orthography, we hope to create further computational tools for converting Karaim texts between Polish, Lithuanian and Russian orthographies. We have begun catering for multiple literacies (and providing a bridge for those with Russian literacy) with the development of a short Karaim to Russian web dictionary (Csató and Nathan 2006).

4 Conclusions

For virtually any resource for endangered languages, including computer-based materials, different writing systems may be needed to meet the needs or skills of particular audiences, including members of the language community, linguists, and others. There is an array of potential resources for many languages; Trosterud (1997) points out that "as a result of the work of philologists and comparativists, huge bodies of fairy tales, mythological texts, legends … etc., are compiled … These texts should be translated from the phonetic transcription they probably are written in, and into the official orthography that hopefully exists for the language today ..." For some language communities, such as described in this paper, this also means transliterating into multiple co-existing orthographies.[17] In addition, as this paper exemplifies, community preferences do change, and it is important to reflect them to effectively support language revitalisation.

This paper has also described the powerful and flexible capabilities of electronic resources to deal with text. Technologies such as Unicode and XML/XSLT are steadily advancing, are generally free to use, and do not require large amounts of programming. Therefore, and especially in the context of products that require large amounts of resources to develop, the additional capacity to handle multiple writing systems creates

[16] Notes: The Lithuanian *v* is used instead of the Polish *w*. Other special characters, as *ń, ź* are missing due to constraints imposed by the email program, or are substituted for using character combinations such as *u"* for *ü* and *o"* for *ö*.

[17] In addition, it can be advantageous to provide materials in a writing system used by the dominant or contact community, so that, for example, local government or planning authorities can understand the value of the resource.

relatively little expense if planned from the outset. We could therefore propose a principle for electronic resource development: 'no monorthographism': language resources should be designed with the potential to host multiple writing systems. The principle is strengthened for multimedia: audio is neutral in regard to orthography and can therefore 'add value' to any writing system that is included.

Although with hindsight we see that the Turcological notation did not provide a complete or ideal solution for Karaim orthography, it did provide the greater advantage of stimulating us to create a general solution to implementing complex orthographies in an interactive computer environment. This in turn bore fruit by allowing us to provide multiple orthographies to support the variety of literacies of today's Karaim people. The developments we have outlined in this paper reinforce the need to be responsive to a community's history, its contemporary social environment, and the linguistic needs and preferences of its members.

References

Baskakov, Nikolaj A., Ananias Zajončkovski and Seraja M. Šapšal 1974. *Karaimsko-russko-pol'skij slovar'*. Moskva: Russkij jazyk.

Bird, Steven and Gary Simons 2003. Seven dimensions of portability for language documentation and description. *Language* 79: 557-582

Csató, Éva Á. 2001. Karaim. In Thomas Stolz (ed.) *Minor languages of Europe.* [Bochum-Essener Beiträge zur Sprachwandelforschung 30], 1-24. Bochum: Brockmeyer.

Csató, Éva Á. [with Meili Fang] 2006[2]. *Karajče ürianiabiź*. Trakai.

Csató. Éva Á. and David Nathan 2004. Multimedia and the documentation of endangered languages. In Peter K. Austin (ed.) *Language documentation and description, Vol. 1*, 73-84. London: School of Oriental and African Studies.

Csató. Éva Á. and David Nathan [with Karina Firkavičiutė] 2004. *Spoken Karaim* (version So1) [Interactive multimedia CD-ROM]

Csató. Éva Á. and David Nathan 2006. *Online Karaim-Russian dictionary*. http://www.dnathan.com/language/karaim/dic/karaim-russian/index.html

Firkovičius, Mikolas (ed.) 1993. *Karaj koltchalary*. [Karaim prayers] Vilnius.

Firkovičius, Mikolas (ed.) 1994. *David' bijniń machtav čozmachlary*. [David's psalms] Vilnius.

Firkovičius, Mikolas 1996. *Mień karajče ürianiam*. [I learn Karaim] Vilnius: Danelius.

Firkovičius, Mikolas (ed.) 1998-1999. *Karaj dińliliarniń jalbarmach jergialliari 1-2*. [The prayer-book of the Karaim believers] Vilnius.

Firkovičius, Mikolas (ed.) 2000. *Šelomonun mašallary*. Ankara: Türk Dil Kurumu.

Gil, Moshe 2003. The origins of the Karaites. In Meira Polliack (ed.) *Karaite Judaism: A guide to its history and literary sources. Handbuch der Orientalistik 73: Handbook of Oriental Studies. Section 1: The Near and Middle East*, 73-118. Leiden: Brill.

Grenoble, Lenore A. and Lindsay J. Whaley 2006. *Saving languages. An introduction to language revitalization.* Cambridge: Cambridge University Press.

Himmelmann, Nikolaus 1998. Documentary and descriptive linguistics. *Linguistics* 36:161-95.

Korpela, Jukka [www] A tutorial on character code issues. http://www.cs.tut.fi/~jkorpela/chars.html

Kowalski, Tadeusz 1929. *Karaimische Texte im Dialekt von Troki*. Krakow: Polish Academy of Sciences.

Matras, Yaron 2005. Language contact, language endangerement, and the role of the 'salvation linguist'. In Peter K. Austin (ed.) *Language documentation and description Volume* 3, 225-238. London: School of Oriental and African Studies.

Mosel, Ulrike 2004. Dictionary making in endangered language communities. In Peter K. Austin (ed.) *Language documentation and description*, *Vollume 2*, 39-54. London: School of Oriental and African Studies.

Nathan, David 2000. The Spoken Karaim CD: Sound, text, lexicon and 'Active Morphology' for language learning multimedia. In Aslı Göksel and Celia Kerslake (eds.) *Studies on Turkish and Turkic languages*, 405-413. Wiesbaden: Harrassowitz.

Nathan, David 2006. Thick interfaces: mobilising language documentation. In Joost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds.), *Essentials of language documentation*, 363-379. Berlin: Mouton de Gruyter.

Nathan, David and Éva Á. Csató 2006. Multimedia: A community oriented information and communication technology. In Anju Saxena & Lars Borin (eds.) *Lesser-known languages of South Asia. Status and policies, case studies and applications of information technology*. [Trends of Linguistics Series], 257-277. Berlin: Mouton de Gruyter.

Ostler, Nicholas and Blair Rudes (eds.) 2000. *Endangered languages and literacy. Proceedings of the Fourth FEL Conference.* University of North California, Charlotte. 21-24 September 2000. Foundation for Endangered Languages.

Polliack, Meira (ed.) 2003. *Karaite Judaism: A guide to its history and literary sources. Handbuch der Orientalistik 73: Handbook of Oriental Studies. Section 1: The Near and Middle East.* Leiden: Brill.

Shapira, Dan 2003. The Turkic languages and literatures of the East European Karaims. In Meira Polliack (ed.) *Karaite Judaism: A guide to its history and literary sources. Handbuch der Orientalistik 73: Handbook of Oriental Studies. Section*

1: The Near and Middle East, 657-707. Leiden: Brill.

Trosterud, Trond 1997. On supporting threatened languages. *Ogmios* 4. http://www.ogmios.org/4.htm

Walfish, Barry D. 2003. Karaite press and printing. In Meira Polliack (ed.) *Karaite Judaism: A guide to its history and literary sources. Handbuch der Orientalistik 73: Handbook of Oriental Studies. Section 1: The Near and Middle East*, 925-959. Leiden: Brill.

Wells, John www. *Computer-coding the IPA: A proposed extension of SAMPA* http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm