The/My Philosophy of Romanization

Joan Biella
Association of Jewish Libraries
June 21, 2004

When I decided to call the first part of this paper "Why We Romanize," I found myself thinking about a coffee-table book that came out ten years ago-- Heather Busch's *Why Cats Paint*.

[**Figure 1. Why cats paint**]

If you don't remember this book, I'll tell you that it was published in association with the Museum of Non-Primate Art and has chapters called things like "Theories of Feline Marking Behavior" and "Twelve Major Cat Artists." Some people choose to regard the book as a spoof, and respond to the title *Why Cats Paint* simply by saying "They don't."

I'm afraid there are also people who would respond to the title "Why We Romanize" by saying "We shouldn't."

These are all very learned people who know lots of languages that use scripts other than the roman one, and they feel that if you want to find a book in Hebrew, that means you know how to read Hebrew, and you can look for it under its Hebrew title and read it in Hebrew, so why should obstacles be placed in your way, like first turning the title into roman letters and making you look for it like that?

I can think of at least one reason, though it has ramifications and may even be several reasons. It's the very basic principle of library organization called collocation.

Many years ago, I attended a university with a famous library. The enormous card catalog was on the main floor, and if you looked under "M" you could find a certain number of books by, say, Maimonides, in English and French and such languages. But if you were not satisfied with that, you ventured into the depths of the library building, a long journey across mountains and rivers (as in a Russian fairy tale), and there in the sub-sub-basement in a dark dank corridor you found the Hebrew language card catalog. Many of the cards in it were handwritten. And if you looked under "Mem" you found books, a lot *more* books, by Maimonides, only these were in Hebrew.

I believe there were some musty old scholars who spent all their time in the sub-sub-basement and never consulted the main card catalog at all, but I couldn't do likewise. My coursework required using both. In those days I was not any slimmer or fitter than I am now, and my time was limited because of all

the classes I was taking other than those that required Hebrew. I sometimes resented having to tramp up hill and down to gather information about all the books in the library by Maimonides. I wondered more than once why all the records couldn't be in the same card catalog--preferably the one that was easier to get to. And there were probably patrons who never even realized that the main card catalog didn't show them all the Maimonides books in the library-- knowledge which might have benefited them.

Many libraries attack these problems by putting all the Maimonides books, whatever language they're written in, under the roman letter "M." In other words, they turn the bibliographic information about these books into a roman-script form and thus make it possible to insert records for them among the records for books with already-romanized information, such as the ones in English. This solution doesn't please everybody. If the Maimonides-seeking patron doesn't understand the romanization system for Hebrew, he may even have to consult a librarian. But most people get used to it.

Keeping all the records for books by Maimonides in one place where they can all be surveyed together exemplifies the principle of collocation. The beauty of this principle is partly abstract, but also partly practical because it cuts out so much running up and down stairs.

Furthermore, In those ancient days there were no machine-readable catalogs, even in famous universities. Even years later, when someone had discovered how to make computerized catalogs, the computers couldn't read Hebrew. In order for the Hebrew books in the collection to appear in the machine-readable catalog, someone had to romanize their records if they hadn't been romanized already.

Nowadays, much later, many library computers *can* read Hebrew, and we hear again the idea that we should forget the tedious, outmoded practice of romanization and put raw Hebrew-script bibliographic information into our catalogs. Why not? To this I can only say: As long as there are library catalogs that *can't* display scripts other than roman (and there are still many many of these in the world and even in our own country), we do not serve all the patrons if we don't provide romanized information. Small, specialized libraries may be able to get away with Hebrew-only catalogs; less specialized ones can't.

So for the time being, romanization is still needed, which is good for people like me and many in this room who have invested a great number of braincells in learning how to do it.

Now on to my second topic, "*How* We Romanize." I want to explore two major concepts, "reversibility" and "conversion."

Reversibility concerns the match-up of the source script in the item (the

non-roman, or, to use the technical term, "squiggly" script) and the target script (for us, the roman script, the one we want to express the first in terms of).

If there's an exact one-to-one match between the characters of the source script and those of the target script, we or a computer can usually flip back and forth between them without any effort at all--that is, they're completely "reversible." Look at this chart, which shows various fonts which match one-to-one:

[**Figure 2. Various fonts**]

The top one is a normal sort of script such as you see everywhere. The second line is in an italic script. The third line is bolded. I made the fourth line by selecting the Wingdings font and hitting the keys that make "The egg and I" in the other three fonts. It doesn't look like English, but you can see there's still a one-to-one match-up--the Wingding at the end of "The" is the same as the Wingding at the beginning of "egg," and the two g's of "egg" are the same Wingding as each other.

Now for something a little more complex. Here's a text in Greek and romanized Greek:

[**Figure 3. Greek**]

The correspondences are all straightforward till we come to the theta, which is represented in ALA-LC romanization as "th"–two roman characters representing the single Greek one. But this need not be a problem if those two roman characters never occur together except when they represent theta. The match-up is still essentially one-to-one. Reversibility is not compromised.

How about a Semitic language? Let's take Geez, otherwise known as Classical Ethiopic:

[**Figure 4. Geez**]

In the romanization of this language, *most* vernacular characters have to be represented by two roman ones--as with the Greek theta--because the Ethiopic script is a syllabary, not an alphabet. Each character represents a syllable, consonant plus vowel–one for "ba," one for "bi," one for "bu," and so on. A computer can easily handle these correspondences. But in this script, ambiguity is also possible: the character for any-particular-consonant-plus-short-e can also represent that-consonant-plus-no-vowel. Only a fairly sophisticated knowledge of the language can determine that "kebra" is not "kbera" and "nagasht" is not "nagashete." Though a computer could perhaps be trained to recognize the contexts determining the choice between consonant-plus-short-e and consonant-plus-no-vowel, the programing would have to be elaborate and

complex.

At this point, a traditional lecturer would probably bring up the romanization of Chinese to illustrate still-knottier problems. For this audience, I'll take another Semitic language, Akkadian, written in cuneiform. Take a very simple cuneiform sign:

[Figure 5. Aš]

That's about as simple as they get. But look at what Labat's cuneiform dictionary says about romanizing this sign:

[Figure 6. Labat on aš]

Its most common values are aš, dil, til-four (til-one through -three are values of other signs), t-with-a-dot-til, rum, ru-with-grave-accent, and so on and so on. Furthermore, not all of the values given are Akkadian--some are Sumerian, the language for which the signs were invented. This sign can be read as Sumerian "didli" or as Akkadian "ma'dûtu," both meaning "crowd." And sometimes it represents yet other whole words, including Akkadian "aplu" = "son" and "nadanu," the verb "to give."

If you're reading aloud about "crowds" in Akkadian and come to this sign, you naturally don't say "didli," you say "ma'dûtu." It's not a "foreign" word, such as might be printed in italics in English text. It's just a Sumerian sign used as a shortcut in Akkadian for the multiple signs that could also be used to spell out the Akkadian word "ma'dûtu."

But if, in addition to sometimes being pronounced "madûtu," this sign might also be "aš" or "dil" or maybe "aplu," how can cuneiform text be interpreted at all? Clearly, much depends on the context, and it would take a computer just about as complicated as an Akkadian's brain to sort out the right value for such a multivalent sign from its context. This script, practically speaking, is not reversible. People do romanize it, but it sounds very odd when read aloud--all that "til-four" and "ru-with-grave-accent" stuff interrupts the flow.

Now let's turn to Hebrew. When I first studied it, my textbook was Lambdin's *Introduction to Biblical Hebrew.* This book employs a monumentally reversible romanization system. It provides sufficient diacritical marks to distinguish nineteen Biblical Hebrew vowels—not to mention all the consonants. Here's the table of vowels in that system:

[Figure 7. Hebrew vowels according to Lambdin]

And here's a sample of romanized text:

[Figure 8. Text romanized à la Lambdin]

Here's a comparison of a single word romanized à la Lambdin and à la the ALA-LC romanization table:

[Figure 9. Comparison of Lambdin and ALA/LC]

The Lambdin system is *so* reversible that we (or a computer) could flip flawlessly both from romanized text to fully vocalized Hebrew text and also the other way around. ALA-LC romanization is reversible enough that we can apply it to the consonants of a Hebrew source without troubles we haven't seen before with other scripts. Even the surprising fact that the Hebrew characters run from right to left while the roman ones run from left to right is no problem for the computer. At the touch of a key line 2 here can become line 3.

[Figure 10. Reversibility]

But no romanization system could ever enable the computer, or even a person who doesn't know Hebrew very well, to find the right roman vowels to fit an *unvocalized* text. We seldom see vocalized Hebrew script in our work, and yet for purposes of collocation in our catalogs, we need to record those vowels.

Now just a few words about Arabic, a language which poses very similar problems to the romanizer--maybe a few more than Hebrew, because Arabic has even more essential vowels which don't show than Hebrew does.

[Figure 11. Arabic]

First, It's clear that the people who dreamed up the ALA-LC Arabic romanization table were different from those who did the Hebrew one, because they did a much more Lambdin-like job on it. If a consonant is marked with the "doubling" indicator, we romanize it as doubled in Arabic. If a vowel is written as long, we put a macron over the roman vowel, and we don't if it's written as short.

The trouble is, of course, that in most Arabic text, as in Hebrew, these discriminations are *not* marked--the vowel points *aren't* written. And much of the grammar is hiding in the diacritics. The consonants say "w-l-d-t"--is that "waladat" = "she gave birth," walladat = "she assisted someone else to give birth," or "wulidat" = "she was born"? So, even more than in romanizing Hebrew, you really have to be able to deduce the meaning--to put it another way, you really have to know the language--before you can romanize Arabic correctly. This is not work for computers.

Now let's consider Ottoman Turkish.

[Figure 12. Ottoman]

Here's a language which looks, most people would say, just like Arabic--maybe a bit messier, as if written with more verve and nonchalance. But if you really try to read it, it's impossible. I think so, anyway. It's *Turkish* written in Arabic script, and the Arabic script fits it very badly. For one thing, Turkish is famous for its numerous mellifluous vowel sounds, and unvocalized Arabic script can discriminate only three--a kind of a, a kind of i, and a kind of u. Most of those vowels you see in romanized Turkish with umlauts and so on have to be deduced from the romanizer's knowledge of the language. Arabic script also provides an abundance of consonants which aren't needed in Turkish, and doesn't provide some that are required, like "ch."

Another trouble comes from the fact that Ottoman text often contains words borrowed from Arabic and Persian, and these are spelled as they are spelled in the source language, not as they're pronounced in Turkish. So even when you get past assigning the proper umlauts in the etymologically Turkish words, you run up against words in the same script that use the vowels and consonants in completely different ways.

But to the Ottoman reader, these are not "foreign" words--he pronounces them as Turkish when he reads aloud, umlauts and all, rather as an Akkadian speaker reads "ma'dûtu" when he sees "didli." How on earth does he know what to do, and how can we teach a computer to do the same? A one-for-one, reversible romanization system can't be devised for this sort of thing.

So here's the second principle of romanization, "conversion." In the words of the ALA-LC table for Ottoman itself:

"The principle of conversion is applied as far as possible, i.e., the word, phrase, name, or title being romanized is represented, if possible, by the form it has in modern Turkish orthography, even if that means converting some letters to their modern equivalents. Foreign words, or words of non-Turkish origin that have become loan-words in Turkish, are converted like Ottoman Turkish."

Suppose you find a word in an Ottoman text that looks exactly like the Arabic "qismah." You transcribe it according to its form in modern Turkish orthography, namely "kismet." You make this substitution automatically if you know Ottoman well. But can we teach a computer to do it?

For sure you can't teach a computer to discriminate between character strings just on the basis of whether they're etymologically Turkish or Arabic. Computers don't care. The best you can do is give the machine a list of character strings and tell it, "When you see this string, do not romanize character by character, but substitute this other string as a block." When you see "qāf-sīn-mīm-tā-marbūtah," with spaces on both sides--i.e., "qismah"--write "k-i-s-m-e-t."

A machine can easily accept this instruction--but there may be some problems in formulating the list of character strings you'll have to give it. New strings that belong on the list will keep turning up, and you'll have to keep track of them and add them. If you are kind, you'll create or authorize a source (a dictionary, perhaps) where other catalogers can look them up. And sometimes you'll have to make corrections to the list. It's a bother, but it can be done. If the string you want isn't in the list, you can ask a speaker of Turkish how to pronounce the thing, or romanize it by analogy with similar strings that *are* in the list. *You* can do this--the computer can't--or anyway, can't yet. The computers we have at the Library of Congress can't do it, so Ottoman catalogers still have plenty of job security.

Now for some examples from my recent experience romanizing languages written in Hebrew script which are not Hebrew.

First, Yiddish**.**

[**Figure 13. Yiddish**]

Yiddish in itself, pure Yiddish, is very easy to romanize. Generally speaking, each character in Yiddish text matches to a single character in the Yiddish romanization system. The few Yiddish characters which have more than one value, such as alef and tsvey-yudn, are usually marked in text with diacritics which disambiguate them. The ALA-LC romanization scheme for Yiddish, though of course it doesn't please everyone in every detail, is a piece of cake to use. It can be correctly applied by people (or machines) that don't even understand Yiddish.

The main trouble comes from the fact that Yiddish text often contains Hebrew words, and these words are not pronounced à la Hebrew but à la Yiddish--just as Arabic words in Ottoman are pronounced à la Turkish, and in Akkadian "didli" is pronounced "ma'dûtu." You see "ma'aśiyot," you pronounce-- and romanize--"mayśes." So in order to romanize Yiddish, we need either an experienced Yiddish speaker, or an authorized reference source--which may be a person or a look-up list for the computer like the one needed for Ottoman.

Are there still problems? Well, there's the problem that not everyone may like the reference source you authorize. Perhaps it represents a dialect they really hate or are offended by! A more or less arbitrary choice has to be made, and a reference source intended for one dialect has to do duty for several similar ones.

After studying all these languages and their problems, there's not much to worry about when we come to Ladino, a Spanish dialect with some etymologically Hebrew vocabulary, written in Hebrew script.

[Figure 14. Ladino]

As with Yiddish, the one-to-one match-up of consonants is normally not a problem. There are a couple of ambiguous characters, and these are normally *not* disambiguated in printed text. But there's a good reference source, Bunis's *Leshon G'udezmo*, which has a dictionary that distinguishes them for you. There are dialects--different styles, so to speak, of reading the same text--but this problem has to be dealt with by fiat, as was the case in Yiddish. And to deal with those etymologically Hebrew words we'll need another look-up list.

At present there is no ALA-LC romanization table for Ladino, but AJL colleagues of ours have recently created one for which we're now seeking authorization from ALA and LC. Until it's authorized, we'll continue to do what we've always done--which pleased nobody--follow an unwritten rule to romanize "consonants as in Hebrew, vowels as in Spanish." What an improvement we have to look forward to!

Lastly for today, there's Judeo-Arabic,

[Figure 15. Judeo-Arabic]

an Arabic dialect once spoken and published in many corners of the former Ottoman Empire. It's written in Hebrew characters, of course, but it differs from both Yiddish and Ladino in that (at least in the works I've studied closely) it *doesn't* import many Hebrew words wholesale, but uses Arabic equivalents or cognates. "Torah" is "Tōrat," "Yiśra'el" is "Yisrā'īl." So the problem of look-up lists doesn't arise. Thank goodness.

Does that mean there are no problems in romanizing Judeo-Arabic? Ha ha, indeed there are! But to discuss them in detail I'll have to write another paper.

My proposed Judeo-Arabic romanization table is based on vocalized texts I've encountered as a cataloger, all printed in Baghdad (so maybe all the same dialect). From these I have abstracted the consonants and vowels that seem appropriate to use in romanization. I've begun a sort of dictionary and a series of verb-conjugation tables. I think these studies, with input from scholars and speakers, may eventually become the "authorized reference sources" for Judeo-Arabic romanization, because as yet I've found no Bunis-like dictionary with enough material for our needs. I wonder how many years it will take to gather up enough to be helpful? Would any of you like to help?

As for this piece of text on the screen, it says (you can see my romanization): "Lam ilēk el an ah li-tamimhā"--"It is not required of you to complete the work." But we do intend to keep on with it, and some day it will be finished to everyone's satisfaction.

P.S. I also believe that cats can paint.

BIBLIOGRAPHY

1. *ALA-LC romanization tables* (Washington : Library of Congress, c1997)
2. Bunis, David M. *Leshon Gʾudezmo* (Jerusalem : Magnes, c1999)
3. Busch, Heather. *Why cats paint* (Berkeley, Calif. : Ten Speed Press, c1994)
4. Labat, René. *Manuel d'épigraphie akkadienne* (Paris : Imprimerie nationale, 1963)
5. Lambdin, Thomas O. *Introduction to Biblical Hebrew* (New York : Scribner's, c1971)
6. *Seder Hagadah shel Pesa*   (Baghdad : Shelomoh Bekhor   utsin, 669 [1908 or 1909])
7. *Sefer Pir   e Avot* (Baghdad : 'Ezra Re'uven Dangur, 5669 [1908 or 1909])