Breaking Arabic: the creative inventiveness of Uyghur script reforms

Yannis Haralambous

DESIGN REGRESSION

ARTICLE | YANNIS HARALAMBOUS

Breaking Arabic: the creative inventiveness of Uyghur script reforms

15 MIN READ | 12 OCT 2021

Abjad or not?

The Arabic writing system is notoriously an *abjad*①. Nevertheless, let us not confuse writing system and script! A script is neither abjad nor *not* abjad per se—it is its use in the frame of a particular language that can be qualified as such (for further discussion see Meletis, 2020, p. 21). In theory, for languages other than Arabic, the Arabic script could be used in manifold ways.

However, the rule is that most languages that use Arabic script use it as an abjad. The reasons are historical: the Arabic script has been spread as the vehicle of Islam and, consequently, Arabic-script languages contain a substantial amount of Arabic words, in their original Arabic spelling. If Arabic loan words are written and read in abjad mode, then it is not surprising that words in native languages are written and read in the same way.

This paper deals with a renowned exception to this: a language that uses the Arabic script in a strictly *phonographic*[2] way. This language takes the liberty of respelling Arabic loan words and names. And furthermore, this language dares to transgress the most sacrosanct rule of Arabic calligraphy/typography by actually breaking words at the end of the line.

This language is Uyghur[3].

A glance at history

Uyghur is a Turkic language and, as such, it historically faced the same problems as Ottoman, Kirghiz, Kazakh, and other Turkic languages. The abjad approach used for Arabic loan words in their original spelling was extended to native words, even though these languages are morphologically quite different from Arabic.

In Turkey, Kemal Atatürk solved the problem in 1929 by imposing a purely phonographic Latin-script based writing system, and by removing most Arabic and Persian loan words from the language (Lewis, 1999).

The Xinjiang region (Northwest China), where most of the Uyghurs live, has not been affected by Atatürk's reform—at the time it was politically part of the Republic of China and culturally under the Soviet sphere of influence (Wei, 1993, p. 263). Uyghurs chose a different path: from 1937 to 1983 (and despite short periods of use of the Cyrillic and the Latin script[4]) they reformed their writing system (Wei, 1993; Reheman & Guo, 2019), and progressively transformed the use of Arabic script in Uyghur from abjad mode to phonographic mode.

Here is how they did it.

How to move from abjad to phonography

First of all the Uyghurs needed to solve the problem of Arabic loan words that would need to be read in abjad mode. They solved it by respelling the words (in a way similar to the way the Hungarians spell the name of the French capital as “Párizs”). For example, the Arabic word سلطان (sultan)—that uses a non-Uyghur emphatic letter ط and does not mark the short /u/—becomes سۇلتان in Uyghur; ط has become ت (of which the medial form is ـتـ) and the phoneme /u/ is represented by grapheme ۇ.

Next, they cleaned up their consonants. They kept Arabic consonants belonging to their phonology, without changing their grapheme-phoneme correspondence. All other consonants (ث, ح, ذ, ص, ض, ط, ظ, ع, ة and the stand-alone ء) had to go. And as the Arabic writing system was not sufficient to represent all Uyghur consonantic phonemes, they added graphemes پ, چ, ژ, گ and ڭ, which were already used in Ottoman (Buğday, 2009).

Now comes the difficult part: how to deal with vowels. Disregarding vocalic length, Arabic language has only three vowels, while the Uyghur phonology requires eight vowels. The principle of strict phonography requires all vowels to be denoted by distinct graphemes—Uyghurs were keen on applying this principle and, after 47 years of reforms, they achieved it.

The main constraint was to do it in a way that is natural to the Arabic script, i.e. to use variations of original Arabic graphemes in order to obtain new graphemes for Uyghur. Can this be done in a simple and straightforward way? Not really, but a smart way to do it is to keep frequent graphemes simple and allow more visual complexity in the rarer ones.

Let us therefore consider vowels in decreasing order of frequency, to discover the inventiveness of Uyghur reformers.

The most frequent Uyghur vowel phoneme is /i/; it appears in 42.64% of Uyghur words. Arabic uses ي for the long vowel /i/. The medial form of this Arabic letter is ـيـ, which is the standard “tooth”⑤ with

two dots underneath. The simple solution would be to use ﻳ for phoneme /i/, but that would be a tedious solution since Uyghur texts would be flooded with ﻳ letters, and strongly-opinionated Arabs might have said: “See, this is why we prefer abjad: it allows us to avoid repeating the obvious and to show only pertinent information”.

Uyghur reformers were definitely not tedious. They cut the Gordian

knot by changing the rules of the Arabic script: they used a *non-letter*, namely a tooth without dots or diacritics. Using an undotted tooth ﺒ is both radical and ingenuous—indeed, the most unobtrusive grapheme you can get is the one you wouldn’t notice in the first place. A non-Uyghur reader of the Arabic script may not even detect its presence. As a test, try to locate the four occurrences ⑥ of the letter ﺒ in Figure 1.

دىزاين رېگرەسسون

Figure 1: An (approximate) Uyghur transcription of “Design Regression”, typeset in the typeface Amiri.

The second most frequent Uyghur vowel phoneme is /a/, which is non-problematic since it is perfectly compatible with the Arabic letter ا. So let us turn to the third most frequent Uyghur vowel phoneme, namely /ɛ/. For this, the Ottomans used the letter ە. The problem is that in the Arabic writing system, ە stands for /h/. Uyghur reformers had a second rule-breaking idea: since the Arabic ە letter has four quite distinct contextual forms, namely ھ (initial), ﻫ (medial), ﻪ (final) and ە (isolated), why not use its initial contextual form as a grapheme representing /h/ and its final and isolated forms as a distinct grapheme representing /ɛ/?

For an average (non-Uyghur) Arabic-script reader هه is just a sequence of two identical /h/ graphemes—in Uyghur it is an /h/ followed by an /ɛ/. In grapholinguistic jargon one would say that the contextual allograph ھ changed status and became a distinct

grapheme.

Last problem: how to deal with vowel phonemes /o/, /u/, /ø/, /y/ and semi-vowel /w/. These phonemes were, incidentally, all represented by a single grapheme in Ottoman, namely و. This time, Uyghurs reformers used a more conventional approach: diacritization.

This leads us to the one and only case where Uyghurs broke the original grapheme-phoneme correspondence of the Arabic language. They decided that و (a long /u/ or a semi-vocalic /w/ in Arabic) would be used for /o/, a phoneme that does not exist in the Arabic language. But if و is /o/, how do you represent the original /u/? Once again Uyghurs found a smart solution by saying: “if you really want an /u/, then write it twice”, namely as ۇ, that is a و letter combined with a diacritic looking like ۇ. This diacritic is actually the Arabic-language short vowel *damma* /u/.

As for the relatively rare phonemes /ø/ and /y/, Uyghurs used diacritics absent from the Arabic writing system: ۆ and ۈ.

And last but not least, for the semi-vowel /w/, they chose a triple-dotted vowel⑦ ۋ —also a clever choice since triple dots are exclusively used for consonants in the Arabic and Ottoman writing systems, so that they carry the “consonant” connotation.

And this is how the Arabic script joined the world of phonography.

Hyphenation of Arabic script

What is common to Semitic languages is the fact that the morphology of a graphemic word is identified by the reader through a specific mental process. This process consists of detecting a *root* combined with a *scheme*, a potential prefix, and a potential suffix. The root is a set of three or four consonants carrying semantic information. The scheme is a pattern of vowels or consonants carrying morphosyntactic information (such as: is it a noun? a verb?

what number? what tense? what person?, etc.). Root and scheme are intertwined. A frequently used example of this phenomenon is the root *k t b* ك ت ب that provides words such as /kātib/ كاتب "author" (scheme /ā-i-∅/), /kitābun/ كتاب "a book" (scheme /i-ā-un/), /kutub/ كتب "books" (scheme /u-u-∅/), /katabtu/ كتبت "I wrote" (scheme /a-a-tu/), etc.

To be able to carry out this process, the reader of the Arabic language needs a simultaneous visualization of the complete Arabic graphemic word[8]. Hyphenation would break this visualization and make it more difficult to identify its parts. Consequently the Arabic language has never been hyphenated. Furthermore, the connected nature of the Arabic script allows the calligrapher (and, to a lesser extent, the typographer) to elongate or shrink words to fit them on lines of equal width, so that there is no need for hyphenation in the first place.

The choice of not hyphenating Arabic language has been inherited by all other languages using the Arabic script, even though languages such as Turkish[9] or Persian could very well benefit from such a method. This would apply especially in printed media, in which columns may be narrow. Furthermore, in typography, lengthening of connecting strokes between Arabic letters in order to compensate

for excessive whitespace is not as easy as in calligraphy.

Once again Uyghur broke the rule. Reformers probably considered hyphenation as a feature inherent in all phonographic writing systems and decided to take the plunge and be the first (and only, for the moment) Arabic-script language in the world adopting it.

Of course this decision was taken at a time when typesetting was mechanical and such an operation was still possible. In digital typography, hyphenation of Arabic script is quite a challenge since it requires redefining the basic contextual behaviour of letters, which is handled at a very low level by the operating system. Systems such as XeTeX can do it in a rather tricky way (cf. Haralambous, 2021) but not (yet) word processing software, at least not automatically. And hyphenation, by its nature, should be automatic.

Font design of hyphenated Arabic script

A natural question to ask is: do the features of the Uyghur language affect font design?

In Figures 2, 3 and 4 the reader can see excerpts from three Uyghur printed documents, covering the period 1979–2021 (documents from the Web site https://elkitab.org).

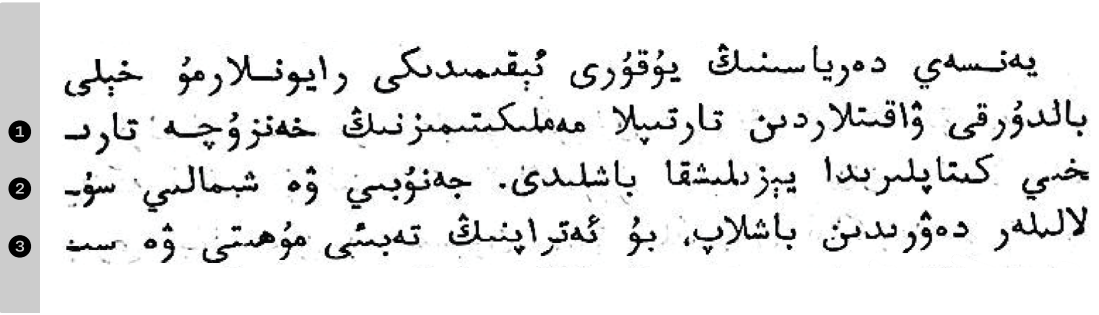


Figure 2: From شىنجاڭنىڭ قىسقىچە تارىخى (*Shinjangning qisqiche tarixi*), 1979. Hyphenation with hyphens (located at the base line) for biform letters as in (2), and without hyphens for quadriform letters, as in (1) and (3).

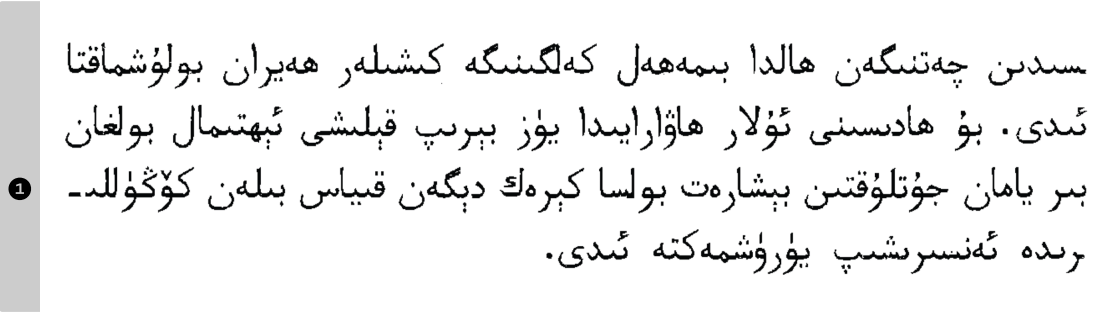
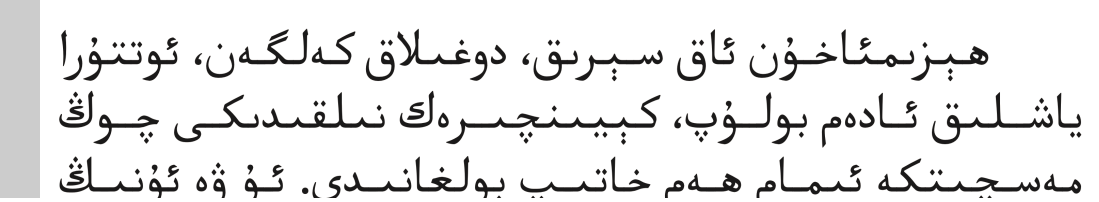


Figure 3: From لالە-قۇربان (*Lale-qurban*), 1997. Standard hyphenation of a quadriform letter, with base line hyphen (1).



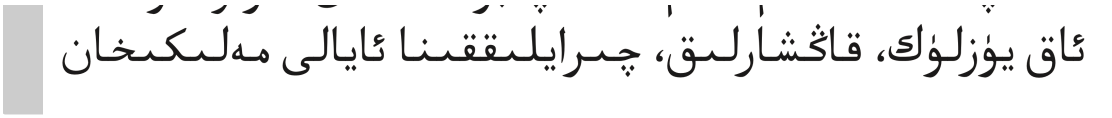


Figure 4: From لۇتۇن (*Lutun*), 2021, typeset in InDesign. No hyphenation. Connections between letters have been elongated to compensate for the lack of hyphenation. This is a standard practice in Arabic language but does not comply with Uyghur typographic tradition, as can be seen in Figures 2 and 3.

Observing these documents, one immediately notices that Uyghur typesetters avoid Arabic aesthetic ligatures (see Haralambous & Dürst, 2019, p. 152) and favour a *flat design*.[10] Maybe the purpose of this communicative strategy is to emphasize the phonographic nature of the writing system, where graphemes are directly matched to phonemes and we have a tendency to perceive the order of phonemes as inherently linear since our speech organs can only emit a single phoneme at a time.

The preference of non-ligatured text is easy to achieve; it suffices to deactivate aesthetic ligatures in the font. What is more challenging for designers is to cope with hyphenation.

Splitting after a biform letter involves no special design. After all, being biform,[11] the letter is already necessarily in final or isolated form. Thus, the hyphen is simply appended to a final or isolated form, something perfectly natural in all the Arabic-script languages (see Table 1).

HYPHENATING AT	NON-HYPHENATED	UPPER LINE	LOWER LINE
biform letter	مەھى	مە-	ھى
quadriform letter	مىھى	مى-	ھى

Table 1: Hyphenating after biform and quadriform letters in Uyghur.

The difficulty is to hyphenate quadriform letters, which are connected to the following letter by a calligraphic stroke and hyphenation occurs between forms that are normally meant to be connected.

Indeed, for centuries Arabic type designers have connected shapes in such a way that the connecting curve seems as natural and elegant as possible. The reader is not supposed to notice the junction point between the letters. The illusion of an unbroken natural calligraphic stroke between the two letters is crucial to obtain the fluidity of a high-quality Arabic font. By splitting the connecting stroke between two letters, the entrails of the font are laid bare—compare the fluidity of a Uyghur word and the awkwardness of its individual segments in Figure 5.

Figure 5: An Uyghur word typeset in Decotype Naskh and its individual segments laid bare.

One way to solve the problems inherent in Uyghur typography would be to design additional contextual forms for letters involved in hyphenation and to develop systems that would use these forms automatically whenever hyphenation of quadriform letters occurred.

Since the Unicode standard has adequately encoded the various uses of the Arabic script, the internationalization of Arabic fonts has been straightforward: the font designer only had to add some additional dots and diacritics and to build ligatures involving these additions. In the light of what we said above, adding special forms for hyphenation may be a real internationalization challenge for designers of Arabic fonts, since they may need to re-invent the very basics of the Arabic script. This is an amazing opportunity for designers to contribute with an elegant and effective solution, and will require the

collaboration of programmers and designers.

Conclusion

Uyghurs reformed their writing system to use the Arabic script phonographically rather than as an abjad. Their decision to break with

tradition was remarkable, but even more remarkable was the way they actually achieved it. Instead of switching to a different script, as did the Turks, they chose to stay as close as possible to their cultural heritage of the Arabic script, but redesigned its use to adhere to strict phonography. Their methods were highly innovative. Today, the decision to hyphenate, contradicting the fluid connections between letters, calls for new letter forms and constitutes the main challenge for programmers and designers. We hope that in the years to come font designers will accept this challenge and produce Arabic font masterpieces rendering equally well Uyghur and all other Arabic-script languages.

Solution to the exercise

Figure 6: Solution to the test of identification of the letter ـىـ.

❶ An abjad is a writing system in which primary graphemes represent only consonants or long vowels. Typical examples are the writing systems of Arabic,

Hebrew and Syriac languages.

❷ A script is used in a phonographic mode when graphemes represent all types of phonemes, whether consonant, long vowels, or short vowels.

❸ While the venerable *Oxford English Dictionary* and *Encyclopaedia Britannica* write “Uighur” (with an “i”), Wikipedia and mass media use the “Uyghur” spelling. We kept the latter because it coincides with the official Uyghur Latin-alphabet transcription of the Uyghur word for Uyghur, namely ئۇيغۇر.

❹ According to Zhou (2003, p. 137–138), Uyghur may return once again to the Latin script since “information technology clearly favors romanized writing systems”, an extremely Eurocentric point of view.

❺ The tooth form (cf. Smitshuijzen, 2001, p. 181, for the name) is used in six Arabic-language letters, namely ـبـ, ـتـ, ـثـ, ـنـ, ـيـ and ـئـ, always accompanied by dots or other forms.

❻ For the solution, see Figure 6.

❼ Indeed, all triple-dotted graphemes represent consonants: ث /θ/, پ /p/, ش /ʃ/, ڭ /ŋ/, چ /ʧ/.

❽ See (Meletis, 2020, §2.5) for the definition of “graphemic word”.

❾ Turkish is a so-called *agglutinative* language where graphemic words may contain entire phrases and therefore can be quite long.

❿ At least in the Naskh style (see Osborn, 2017, p. 55 and Haralambous, 1994), aesthetic Arabic ligatures often correspond to a vertical stacking of shapes. In the absence of aesthetic ligatures, the base line is visually omnipresent and the general impression is one of a flat design. Compare the ligatured (left) and unligatured (right) sequence of the three letters in the following figure:

⓫ In the Arabic script, *biform* letters are not connected to the following letter; *quadriform* letters are connected to the following letter.

References

Buğday, K. (2009). *The Routledge introduction to literary Ottoman*. Routledge. https://www.routledge.com/The-Routledge-Introduction-to-Literary-Ottoman/Bugday/p/book/9780415494380 accessed 18 September 2021.

Haralambous, Y. (1994). The traditional Arabic typecase extended to the Unicode set of glyphs. *Electronic Publishing—Origination, Dissemination, and Design*, *8*(2/3), 125–138. http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume8/issue2/2point10.png accessed 18 September 2021.

Haralambous, Y. (2021). Implementing Uyghur hyphenation in XeTeX. (To appear in *TUGboat*.)

Haralambous, Y. & Dürst, M. (2019). Unicode from a linguistic point of view. In *Proceedings of Graphemics in the 21st Century. Brest, June 13–15, 2018*. *Grapholinguistics and its Applications*, Vol. 1. Fluxus Editions, pp. 127–166. https://doi.org/10.36824/2018-graf-hara1

accessed 1 October 2021.

Lewis, G. (1999). *The Turkish language reform. A catastrophic success*. Oxford University Press. https://global.oup.com/academic/product/the-turkish-language-reform-9780199256693?lang=en accessed 18 September 2021.

Meletis, D. (2020). *The nature of writing. A theory of grapholinguistics*. *Grapholinguistics and its Applications*, Vol. 3. Fluxus Editions. https://doi.org/10.36824/2020-meletis accessed 18 September 2021.

Osborn,J. (2017). *Letters of light*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674971127 accessed 18 September 2021.

Reheman, A. & Guo, Y. (2019). A field research of Chinese Uyghur people's writing reforms and influences in the 20th century. In *3rd International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2019)*, pp. 173–178. Atlantis Press. https://dx.doi.org/10.2991/iccese-19.2019.40 accessed 18 September 2021.

Smitshuijzen Abi Farès, H. (2001). *Arabic typography: A Comprehensive Sourcebook*. SaqiBooks.

Wei, C. (1993). An historical survey of modern Uighur writing since the 1950s in Xinjiang, China. *Central Asiatic Journal*, *37*(3–4), 249–322. https://www.jstor.org/stable/24467845?refreqid=excelsior%3Afcd2252a88eee2d971a96655780114dc accessed 18 September 2021.

Zhou, M. (2003). *Multilingualism in China: The politics of writing reforms for minority languages, 1949-2002*. Mouton De Gruyter. https://www.degruyter.com/document/doi/10.1515/9783110924596/html accessed 18 September 2021.

Yannis Haralambous

After a PhD in Algebraic Topology at the University of Lille 1, Yannis Haralambous is doing research in Digital Typography, Grapholinguistics and Natural Language Processing. He is currently Professor at the Computer Science Department of IMT Atlantique, a French grande école in Brest, Brittany, and member of the DECIDE team of the CNRS Laboratory Lab-STICC. He has written a book on *Fonts & Encodings* (O'Reilly, 2007). He is the organizer of the biennial *Grapholinguistics in the 21st Century* conference and the editor-in-chief of the book series *Grapholinguistics and Its Applications* at Fluxus Editions.

Yannis Haralambous at IMT Atlantique