Multilingual Typesetting with Ω, a Case Study: Arabic

Yannis Haralambous[*] John Plaice[†]

Abstract

In this paper we describe the internal structure of the Arabic script package for the Ω typesetting system, as well as the techniques and tools used for its development. This package allows typesetting using regular LaTeX styles, in all Arabic alphabet languages: Arabic, Berber, Farsi, Urdu, Pashto, Sindhi, Uighur, etc.

We also give a description of the character codes added to Unicode, to obtain the Unicode++ encoding, used by the Ω system for typesetting purposes.

1 Overview of the Ω Arabic Script Package

Typesetting with Ω is a process similar to typesetting with TeX: the user prepares a "source" file, containing the text of $\substack{\text{his}\\\text{her}}$ document and a certain number of macro-commands for attribute changes of the text (font characteristics, language, case, etc.), references to figures (included in graphical format files on disk) and other material included in or accompanying the text.

Once this source file prepared, Ω is launched: it reads the file, expands the commands and typesets the text accordingly. To perform this task, Ω loads and executes several ΩTPs (Ω Translation Processes), which take care of low level properties of the document (contextual analysis of the script, case switching according to script and language, etc.). It also uses different fonts, most of which are *virtual*, in the sense that they themselves call other fonts. On a higher level, such a document uses LaTeX packages, some of them modified to take advantage of the additional features of Ω vs. TeX.

The leading idea of the Ω Arabic Script Package (as of any Ω language package) is that the low level properties of the script have to be separated from higher level typesetting commands. For example, contextual analysis of the Arabic script has to be completely independent of the LaTeX command level, so that one can use Arabic text in any context (inside a table or a formula, or deeply nested inside several LaTeX environments and commands, etc.) and under any circumstances, as in

[*] Atelier Fluxus Virus, 187, rue Nationale, 59800 Lille, France, yannis@fluxus-virus.com

[†] School of Computer Science and Engineering, The University of New South Wales, Sydney 2052 Australia, plaice@cse.unsw.edu.au

the following example, which has been typeset with ordinary LaTeX environments and macros:

حَياة	مَيِّت
$\int_{\text{صِفْر}}^{\text{غير مَحْدُود}} f(x)\,dx$	أَنا

There are two key aspects to Arabic script typesetting, unfortunately of unequal complexity: the first one is contextual analysis, that is the fact that Arabic letters change shape according to their position in a word, or according to the fact that they are part of an abbreviation, etc. This aspect can be handled easily and efficiently by ΩTPs. The second aspect is more global: it is the fact that Arabic script is written from right to left.

Two methods can be applied: the first one is to change the default direction of the whole document. This method is extremely efficient when the document is entirely in Arabic, or if left-to-right text excerpts are exceptional. Being global, this method applies also to page-level typesetting methods, such as the order of columns in a multicolumn environment, etc. Of course, mathematical formulas are not affected by this global direction change.

The second method is to keep left-to-right as default direction and to temporarily switch to right-to-left for every Arabic script sentence. This can be practical for a document where Arabic excerpts are exceptional.

2 Parts of the Ω Arabic Script Package

This package consists of the following elements:

1. The `OmegaSerifArabic` PostScript fonts: files `omsea1.pfb`, `omsea2.pfb`, `omsea3.pfb` and the corresponding AFM files. A Sans-serif font (`OmegaSansArabic`), as well as additional styles of the Serif font are under development.

2. The virtual font `omrl`: files `omrl.ovf`, `omrl.ofm`, `omsea1.tfm`, `omsea2.tfm`, `omsea3.tfm`.

3. The configuration file `omrl.cfg`, which is used by the PERL utility MakeOVP to create the virtual font out of the AFM files and other information.

4. A certain number of ΩTPs:

   (a) `7arb2uni.otp`, 7-bit Arabic/Farsi transcription to Unicode;

   (b) `7ber2uni.otp`, 7-bit Berber transcription to Unicode;

   (c) `7urd2uni.otp`, 7-bit Urdu transcription to Unicode;

   (d) `7pas2uni.otp`, 7-bit Afghanistani Pashto transcription to Unicode;

   (e) `7pap2uni.otp`, 7-bit Pakistani Pashto transcription to Unicode;
   (f) `7snd2uni.otp`, 7-bit Sindhi transcription to Unicode;
   (g) `uni2cuni.otp`, contextual analysis, sending Unicode++ to cUnicode++ ('c' for 'contextual');
   (h) `cuni2oar.otp`, cUnicode++ to `omrl` font.

   These ΩTPs are available in human-readable and compiled binary format (OCP), the latter being loaded by Ω on runtime.

5. A LaTeX style (`arabic.sty`) defining a command that will activate and deactivate the ΩTPs.

6. Documentation and test files (`testarab.tex`, `testsind.tex`).

3 Installation of the Ω Arabic Script Package

To use the Ω Arabic Script Package you must have Ω version 1.45 or higher installed on your machine. Place OFM, OVF, TFM and OCP files where the system expects to find them (if in doubt, consult the `texmf.conf` file). Keep the `arabic.sty` file somewhere where it can be found by Ω. Finally add the following few lines to the `psfonts.map` configuration file of `odvips`:

```
omsea1 OmegaSerifArabicOne </foo/omsea1.pfb
omsea2 OmegaSerifArabicTwo </foo/omsea2.pfb
omsea3 OmegaSerifArabicThree </foo/omsea3.pfb
```

where `/foo` stands for the absolute path of the directory containing the PFB files.

This is all you need to do: you can start already by launching Ω on files `testarab.tex` and `testsind.tex`.

In the following sections we will describe the use of the package, from the end users' point of view. We will assume that the user is familiar with the TeX typesetting system and the LaTeX macro package.

4 Basic Macros

Before starting a new document one has to choose if the "background language" is going to be an Arabic alphabet language, in other terms, if we expect pages and columns to be typeset from right to left, and the whole global page design to be right-to-left oriented.

If this the case, then the macro `\GlobalArabic[`*`language`*`]` has to be used in the document header, where the optional argument *`language`* is one of the following: `arabic` (by default), `farsi`, `urdu`, `pashto`, `sindhi`, `custom`.

This macro will switch the global typesetting direction of the document to right-to-left and will launch the ΩTPs necessary for the language chosen.

Inside the document, independently of the choice of background language, one can use LaTeX environments `arabic`, `berber`, `farsi`, `urdu`, `pashto`, `pashtop`, `sindhi` to switch to the corresponding language, and `latin` or `greek` to switch to a Latin alphabet language or some flavour of Greek. It should be noted that these macros are only temporary and will be adapted to a more global language-switching scheme, currently being elaborated by the LaTeX3 and Ω working groups.

5 Input of Arabic Alphabet Text

5.1 You Have an Arabic Keyboard

If you have an Arabic Keyboard, containing sufficiently many keys for the language you want to typeset (for example, with a standard Arabic keyboard one can perhaps typeset Farsi, possibly Urdu but not Pashto and certainly not Sindhi), you need to configure Ω to your *input encoding*, by providing the appropriate input ΩTP by use of the `\ArabicInputEncoding` macro, which you have to place in the header of your document. We have already written such ΩTPs for three input encodings: Macintosh Arabic (`applemac`, covering Arabic, Farsi, Urdu), Windows Arabic (`1256`, covering Arabic and Farsi), MS-DOS Arabic ASMO (`708`, covering Arabic only) and ISO 8859-6 (`iso8859-6`, covering only Arabic). If your equipment is not in this list, go to section 6 to see how to write your own ΩTPs.

5.2 You Don't Have an Arabic Keyboard

In that case you can use a Latin transcription: we have prepared ASCII Latin transcriptions for each of the main Arabic-alphabet languages: Arabic, Berber, Farsi, Urdu, Pashto (Afghanistani and Pakistani), Sindhi. Here they are:

5.2.1 Arabic/Farsi Transcription

A	ا	p	پ	z	ز	`	ع	m	م	I	ى
'a	أ	j	ج	zh	ژ	gh	غ	n	ن	y	ي
'i	إ	H	ح	s	س	f	ف	'n	ں	'y	ئ
'A	آ	kh	خ	sh	ش	q	ق	-h	ه	||	ء
"A	ٱ	ch	چ	S	ص	v	ڤ	"h	ة	E	ے
b	ب	d	د	D	ض	k	ك	e	ۀ		
t	ت	dh	ذ	T	ط	g	گ	U	و	LLah	ﷲ
th	ث	r	ر	Z	ظ	l	ل	'u	ؤ	SLh	ﷺ

Remarks:

1. The *tah marbutah* ة can be written in two ways: `"h` or `"t`.

2. The *waw* و can be written in two ways: `w` or `U`.

3. The hyphen in front of the transcription of ه is only necessary to prevent confusion between cases such as `kh` (خ) and `k-h` (كه). We suggest you use it all the time.

4. VERY IMPORTANT: the duplication of consonants (*shaddah*) is obtained by writing the consonants twice. So for example, `Dmm"h` will produce ضمّة and not ضممة; to obtain the latter, type `Dm-m"h`, as for example in the word تتحرّك, which presents both cases, and which is typed `t-tHrrk`.

Vowels and other diacritics are obtained in the following way: (they are typed after the consonant to which they belong)

fatha	`a`
kasra	`i`
damma	`u`
soukoun	`<>`
vertical fatha	`a|`
fathatan	`aN`
kasratan	`iN`
dammatan	`uN`

Example: it is a trivial task now to welcome you to this system of Arabic input, by saying

```
\begin{arab}
\Huge
'aahlAaN wa sahlAaN!
\end{arab}
```

أَهلاً وَ سَهلاً!

Example of vowelized Arabic:

لِأَنَّهَا ٱلْآنَ لَا تُفَكِّرُ فِي نَفْسِهَا، وَلَكِنَّهَا تُفَكِّرُ فِى
أَخَوَيْهَا وَفِى ٱلْخَطَرِ ٱلَّذِى لَحِقَهُمَا.

transcribed:

```
li'aannahaA "Al<>'Ana laA tufakkiru fiI naf<>sihaA,
walakinnahaA tufakkiru fiI 'aakhaway<>haA
wafiI "Al<>khaTari "AlladhiI laHiqahumaA.
```

5.2.2 Urdu Transcription

The Urdu transcription is similar to the Arabic/Farsi one described above, with a few additional characters, and one exception.

The additional characters are ٹ, ڈ and ڑ, transcribed by `'t`, `'d`, `'r`. The exception concerns the two different uses of the *hah* glyph ہ. In Urdu it can be used as the second part of a digraph, such as for example جھ, in which case we transcribe it as `-h`; it can also be the standard consonant *hah*, in which case we transcribe it by `x`. Notice the four forms of the latter in Urdu: ہ ہہہ, while in Arabic the same letter is written ه ههه.

Example:

> ہماری طرف پرانے زمانے میں دستور تھا کہ اگر کسی شخص کو کاغذ پر
> کچھ لکھا ہوا گرا پڑا مل جاتا تو وہ اس پرزے کو احتیاط سے اٹھا کر
> کہیں رکھ دیتا یا پانی میں بہا دیتا تاکہ لکھے ہوئے حروف کی بے حرمتی
> نہ ہو۔

transcribed:

```
xmArI Trf prAnE zmAnE my'n dstUr t-hA kx Agr ksI
shkhS kU kAghdh pr kchh lk-hA xUA grA p'rA ml jAtA tU Uh
As przE kU AHtyAT sE A't-hA kr kxy'n rk-h dytA yA pAnI mI'n
bxA dytA tAkx lk-hE xU'yE HrUf kI bE HrmtI nx xU.
```

5.2.3 Pashto Transcription

The Pashto transcription is similar to the Arabic/Farsi one described beyond, with a few additional characters and some exceptions. We are proposing two ΩTPs, using the same transcription, for the two flavors of written Pashto: Afganistani and Pakistani.

1. Afghanistani Pashto

A	ا	'z	ځ	'r	ړ	D	ض	g	گ	-y	ي
b	ب	c	څ	z	ز	T	ط	l	ل	e	ې
p	پ	H	ح	zh	ژ	Z	ظ	m	م	ay	ئ
t	ت	kh	خ	'g	ږ	`	ع	n	ن	ey	ۍ
't	ټ	d	د	s	س	gh	غ	'n	ڼ	||	ء
's	ث	'd	ډ	sh	ش	f	ف	w	و		
j	ج	dh	ذ	x	ښ	q	ق	-h	ه	LLah	ﷲ
ch	چ	r	ر	S	ص	k	ک	L		SLh	ﷺ

2. Pakistani Pashto

A	ا	'z	ځ	'r	ړ	D	ض	g	ګ	-y	ے
b	ب	c	څ	z	ز	T	ط	l	ل	e	ے
p	پ	H	ح	zh	ژ	Z	ظ	m	م	ay	ے
t	ت	kh	خ	'g	گ	`	ع	n	ن	ey	ے
't	ټ	d	د	s	س	gh	غ	'n	ڼ	||	ء
's	ث	'd	ډ	sh	ش	f	ف	w	و		
j	ج	dh	ذ	x	ښ	q	ق	-h	ه	LLah	ﷲ
ch	چ	r	ر	S	ص	k	ك	L		SLh	ﷺ

Nevertheless, one should be aware that an automatic transcription from one glyph set to the other is not possible because, for example, a letter such as ښ is not used in Pakistani Pashto and can be replaced by خ or ش, depending on its pronunciation in a given word.
Example of Afghanistani Pashto:

که غواړئ چه دعقل یه زیان اوضرّپوه شئ دا و منی ء چه عقل هغه.
قوتونه پهسړی کښی وژنی ژه ژوندي ولاړدی. ژوندون پهعمل اواراده
ولاړدی. غوښتنه لواراده دهریشر فت اصل اواساس دی. څومره چه عقل
زیاتبز هغومره اراده ضعیفه کبزی.

and the same in Pakistani Pashto:

که غواڑے چه دعقل ےه زےان اوضرّپوه شے دا و منے ء چه عقل هغه.
قوتونه پهسڑے کخے وژنے ژه ژوندے ولاڑدے. ژوندون پهعمل اواراده ولاڑدے.
غوختنه لواراده دهرےےشر فت اصل اواساس دے. څومره چه عقل زےاتےبز
هغومره اراده ضعےفه کےبزے.

transcribed:

```
k-h ghUA'ray chh d`ql yh zyAn AUDrrpUh shay dA
U mnI || chh `ql hghh. qUtUnh p-hs'rI kxI wzhnI zhh zhUnde
wlA'rdI. zhUndUn p-h`ml AUArAd-h wlA'rdI. ghUxtnh lUArAd-h
d-hreyshr ft ASl AUAsAs dI. cUmrh chh `ql zyAtebz hghUmrh
ArAd-h D`yf-h kebzI.
```

A variant form ړ of ړ is provided in the font. The user can change the ΩTPs (see 6) so that the former is used instead of the latter.

5.2.4 Sindhi Transcription

Sindhi being a language with many more letters than Arabic, and using Arabic letters in a way quite different than Arabic, it is not surprising that the Sindhi transcription is fundamentally different from the Arabic, Farsi, Urdu and Pashto ones. As a matter of fact we have tried to use as few non-alphabetic characters as possible, following a more-or-less rational scheme loosely based on the correspondence between Sindhi written in Arabic and in Devanagari script and the standard transcription of the latter. Since shadda is much more seldom in Sindhi than in Arabic, the “double consonant = consonant + shadda” convention is not valid in this transcription; instead we propose a transcription of the shadda diacritic: `+`.

`A`	ا	`p`	پ	`dh`	ڌ	`sh`	ش	`kh`	ک	`y`	ي
`'A`	آ	`ph`	ڦ	`.=d`	ڏ	`.s`	ص	`.n`	ڱ	`'y`	ئ
`b`	ب	`j`	ج	`.d`	ڊ	`.z`	ض	`g`	گ	`meN`	۾
`=b`	ٻ	`=j`	ڄ	`.dh`	ڍ	`..t`	ط	`=g`	ڳ	`||eN`	۽
`bh`	ڀ	`=n`	ڃ	`=z`	ذ	`..z`	ظ	`l`	ل	`||`	ء
`t`	ت	`c`	چ	`r`	ر	`` ` ``	ع	`m`	م		
`th`	ٿ	`ch`	ڇ	`.r`	ڙ	`gh`	غ	`n`	ن		
`.t`	ٽ	`.h`	ح	`z`	ز	`f`	ف	`'n`	ڻ		
`.th`	ٺ	`=kh`	خ	`zh`	ژ	`q`	ق	`U`	و	`LLah`	ﷲ
`=s`	ث	`d`	د	`s`	س	`k`	ڪ	`-h`	ه	`SLh`	ﷺ

Remarks:

1. The transcription `/` is used for constructions such as بہ (`b/`), تہ (`t/`), کہ (`kh/`), etc.
2. The *waw* و can be written in two ways: `w` or `U`.

Example:

تنهن ڪري اسان کي پنهنجي ذهن کي سجاڳ رکڻو پوندو ۽ پنهنجي
جدوجهد ۾ ڏاهپ پيدا ڪرڻي. اهو بہ معلوم ڪرڻو پوندو تہ سنڌ ۾ هر
آئي وقت چا چا ٿي رهيو آهي ۽ دشمن اسان جي ۽ اسان جي جدوجهد
جي ڪلاف ڪهڙا ڪهڙا گهاٽ گهڙي رهيو آهي.

transcribed:

```
tn-hn kry AsAn khy pn-hnjy =z-hnn khy sjA=g rkh'nU
pUndU ||eN pn-hnjy jdUj-hd meN .=dA-hp pydA kr'ny. AhU b/
m`lUm kr'nU pUndU t/ sndh meN hr 'A'yy wqt chA chA thy r-hyU
'Ahy ||eN dshmn AsAn jy ||eN AsAn jy jdUj-hd jy khlAf
k-h.rA
k-h.rA g-hA.t g-h.ry r-hyU 'Ahy.
```

5.2.5 Berber Transcription

The Berber transcription is different from the previous ones because it is based on a tri-alphabetic system (Tifinagh, Latin and Arabic alphabets).[1] The goal of this

[1] The reader can find more information in *Un système TEX berbère*, Études et Documents Berbères, 11 (1994), La boîte à Documents/Édisud, Paris (France).

transcription is to enable output in the three alphabets, out of the same code. In particular, since Latin alphabet has upper and lower case, it should be possible to distinguish these (and of course ignore the distinction when typesetting in Arabic or Tifinagh). In the table below, all transcribed letters are in lowercase ASCII, but can very well be written also in uppercase, producing the same result: `Tifinagh`, `tifinagh` or `TIFINAGH` will all three produce تيفيناغ.

Tr.	Lat.	Ar.	Tif.	Tr.	Lat.	Ar.	Tif.	Tr.	Lat.	Ar.	Tif.
`a`	a	ا	·	`.h`	ḥ	ح	⋯	`.s`	ṣ	ص	Ⱶ
`b`	b	ب	Θ	`i`	i	ي	ϟ	`t`	t	ت	X
`c`	c	ش	C	`j`	j	چ	ⵊ	`.t`	ṭ	ط	Ɇ
`gh`	ɣ	غ	⋮	`k`	k	ك	⇛	`u`	u	و	:
`d`	d	د	Λ	`l`	l	ل	Ⅱ	`x`	x	خ	::
`.d`	ḍ	ض	E	`m`	m	م	ⵎ	`z`	z	ز	Ж
`.e`	ɛ	ع		`n`	n	ن	I	`.z`	ẓ	ز	ⵥ
`f`	f	ف	ⴼ	`.n`	ñ	.ن	ǂ	`.i`	i	ئ	ϟ
`g`	g	ڨ	X	`q`	q	ق	Ξ	`--`	_	ـ	_
`.g`	ğ	ج	İ	`r`	r	ر	O				
`h`	h	ه	Ξ	`s`	s	س	N				

Remarks:

1. Letter و can also be transcribed `w`.
2. Letter ى can also be transcribed `y`.
3. The stroke ـ is not to be confused with the graphical connecting stroke *keshideh*. It is placed between words and plays a grammatical role.
4. Duplication of consonants (*shaddah*) again is transcribed by writing the corresponding consonant twice.

Example:

تيفيناغ، د_تيرا تيمزوّرا ن ئمازيغن. لّانت دي تمورت_نّغ دات تيرا ن
تاعرابت د_تلاطينيت. نّولفانت_دّ دات ئمير ن وڤلّيد ماسينيسن. ئمازيغن
ن ئمير_ن، تّارون_تنت غفيزرا، دڤّ ئفران، غف ئڤدورن، ماشا تيڤّتي غف
يزكوان : تّارون فلّ_اسن ئسم ن ومتّين، د وي_ت_يلان، د واين يخدم دي
تودرت_يس اكّن ور ت تتّون ئناطفارن.

transcribed:

```
Tifinagh, d--tira timezwura n .imazighen.
Llant di tmurt--nnegh dat tira n ta.erabt d--tla.tinit.
Nnulfant--edd dat .imir n ugellid Masinisen. .Imazighen n
```

```
.imir--en, ttarun--tent ghefi.zra, degg .ifran, ghef .igduren,
maca tiggti ghef i.zekwan : ttarun fell--asen .isem n umettin,
d wi--t--ilan, d wayen yexdem di tudert--is akken ur t ttettun
.ina.tfaren.
```

The same code will produce the following output in the Tifinagh alphabet:

> XϟƆƐϟI·⁝, Λ_XϟO· XϟƐЖ≕:O· I ϟƐ·Жϟ⁝I. ⵍ\\·IX Λϟ XƐ:ⵥ_I\⁝ Λ·X XϟO·
> I X·O·ⵥ Λ_Xⵍ·Eϟ IϟX. I\:ⵍƆƐ·IX_ΛΛ Λ·X ϟƐϟO I :Xⵍ\\ϟΛ Ɛ·ⵅϟIϟⵅI.
> ϟƐ·Жϟ⁝I I ϟƐϟO_I, XX·O:I_XIX ⁝ƆƐϟ#O·, ΛXX ϟƆƐO·I, ⁝ƆƐ ϟXΛ:OI, Ɛ·G·
> XϟXⵅϟ ⁝ƆƐ ϟ#⇒≕·I : XX·O:I ƆƐⵍ_·ⵅI ϟⵅƐ I :EXϟI, Λ ≕ϟ_X_ϟⵍ·I, Λ
> ≕·ΣI Σ∷ΛƐ Λϟ X:Λⵥ_ϟⵅ ·⇒⇒I :O X XXXX:I ϟI·EƆƐ·OI.

and the following one in the Latin alphabet:

> Tifinaɣ, d_tira timezwura n imaziɣen. Llant di tmurt_nneɣ dat tira n taɛrabt d_tlaṭinit. Nnulfant_edd dat imir n ugellid Masinisen. Imaziɣen n imir_en, ttarun_tent ɣefiẓra, degg ifran, ɣef igduren, maca tiggti ɣef iẓekwan : ttarun fell_asen isem n umettin, d wi_t_ilan, d wayen yexdem di tudert_is akken ur t ttettun inaṭfaren.

6 Writing Your Own Transcription

We have developed and presented in this paper a certain number of Arabic alphabet language transcriptions for two reasons: first, to show the possibilities and power of Ω, and second, to give a starting point for the user to create $^{\text{his}}_{\text{her}}$ own transcriptions.

The process of creating a new transcription is twofold: the first part, which can be very difficult and painful, consists of finding the combination of letters, digits and ASCII symbols which will transcribe each character; the second one, which is straightforward (modulo some precautions) is to implement this in Ω by writing the appropriate ΩTP.

6.1 A Good Transcription: Is it Possible?

There are (at least) two goals for a good transcription:

1. *It has to be readable and easily memorizable.* In other words, `AHmd` is better than `'.hmd`, for denoting احمد : although an apostrophe can be considered a logical choice for transcribing an alif and the period in front of the h may denote that it is an emphatic ‘h’ sound, taking an A for alif and a capital H for the emphatic h is more readable; also using rules such as “uppercase ASCII characters transcribe emphatic letters” is an easy way to memorize the transcriptions of ح, ط, ض, ص, ظ.

2. *It has to be complete and avoid ambiguities*. Of course all letters of the target language have to be covered, but having many letters to transcribe leads sometimes to ambiguities: for example taking `h` for ه, `k` for ك and `kh` for خ are perfectly logical choices; nevertheless there is a hitch: when you need to transcribe كه you are tempted to write simply `kh` and this will of course produce خ instead. The solution we have given to this problem is to type a hyphen between the letters which are not considered as a 'digraph', but this is only a compromise solution: the user must constantly be aware of this problem, and this is hardly the case when you are concentrated in your text...

It is clear that these two goals are contradictory: an accurate and unambiguous transcription has to be complicated and will be difficult to read and memorize; a friendly and easily readable transcription will be full of ambiguities.

An additional problem when making a transcription is to choose between *(etymo)logical*, *phonetic* and *graphical* representations of characters. A typical example is the standard Ω transcription of Greek: `w` is chosen for letter ω, this is a purely *graphical* choice: the 'w' looks like an omega, but has absolutely no other relation with, neither historical nor phonetic (the letter omega represents the sound 'o' in modern Greek); `b` is chosen for letter β, this is an *etymological* choice: the Latin 'B' derives from the ancient Greek 'B', otherwise β looks quite different than 'b' and is pronounced 'v' in modern Greek; finally, `x` is a *phonetic* transcription of letter ξ; clearly they do not bear any resemblance, and historically it is not clear (at least to the author) why 'x' should be derived from ξ (their positions in the alphabet is quite different as well, and this is an argument speaking against an etymological relation between the letters).

The reader may object that this distinction between etymological, phonetic and graphical representations is not relevant for Arabic alphabet transcriptions; actually this is only partly true: take for example `bh` for ڀ, this is an *etymological* transcription in the sense that it reflects the standard transcription of the Indic alphabet letter which corresponds to that Sindhi letter. Also `` ` `` for ayn is in some sense a *graphical* representation: it has been chosen because it resembles the IPA transcription of the ayn, which is ʕ. For the same reason, `'` has been chosen for the hamza with carrier (in أ, ؤ, etc.): the hamza's IPA transcription is ʔ.

We hope to have convinced the reader that the making of a transcription is a difficult task, needing a lot of thought, compromises and tests. Once again, we would like to emphasize the fact that our transcriptions are only temptative proposals and should not be taken as standards of any kind; after all the power of Ω is that it can work with any input transcription without affecting further processing, be it contextual analysis, diacritic placement or esthetical ligaturing.

In the next section we will see how to implement a new transcription or change an existing one by writing/modifying an ΩTP file. But first some generalities on the ΩTPs used by the Arabic Ω system.

6.1.1 The ΩTPs used by the Arabic Ω system

When Ω reads the text flow it places letters, digits and punctuation (whatever is not an escape or special character) into a buffer. When it encounters a special character it stops buffering and executes one after the other all currently active ΩTPs on the buffer. In theory, ΩTPs could be used to arbitrarily send character combinations to other combinations: one could very well imagine an ΩTP sending the string "Yannis" to "John" and "John" to "Yannis", or "Microsoft Word" to "ʧǫξⱡⱥɞ"; nevertheless, such an ΩTP would not be of general use...

Our development has mainly been focused in building ΩTPs in accordance to the following scheme:

$$\boxed{\text{Input text}} \xrightarrow{\texttt{foo2uni}} \boxed{\text{Unicode++}} \xrightarrow{\texttt{uni2foo}} \boxed{\text{DVI output}}$$

where `foo2uni` sends text encoded in an arbitrary encoding into Unicode++ (Unicode++ is Unicode extended for the needs of Ω and typography), and `uni2foo` converts Unicode++-encoded data into the encoding of the output font. By this method we are able to keep completely separate input encoding and font encoding.

In the case of Arabic things are slightly more complicated since an additional step is needed: contextual analysis. This is where our scheme proves to be extremely efficient: by performing contextual analysis on the level of Unicode++, and hence obtaining the following new scheme:

$$\boxed{\text{Input text}} \xrightarrow{\texttt{foo2uni}} \boxed{\text{Unicode++}} \xrightarrow{\texttt{uni2cuni}} \boxed{\text{cUnicode++}} \xrightarrow{\texttt{cuni2oar}} \boxed{\text{DVI output}}$$

we still remain independent of both the input and the font encoding. This means that if we need to adapt Ω to a new Arabic encoding we only need to indicate which code position corresponds to which Unicode character, and, on the other hand, if we want to adapt a new font to Ω, we only need to indicate which font position corresponds to which contextual form of which character, in cUnicode++.

In the next section we will partly describe the syntax of ΩTP files by giving examples of `foo2uni` cases.

6.2 Implementing a Transcription

The ΩTP files we will need for input encoding → Unicode++ transformations use only part of the syntax of ΩTP files.[2] Such an ΩTP file is of the following form:

```
input: 1;
output: 2;
expressions:
...
...
```

[2] The `uni2cuni` ΩTP file already needs more complicated constructions.

where `input: 1; output: 2;` means that input is 8-bit while output is 16-bit, and `...` are lines of the following form:

```
before => after ;
```

where `before` is an expression before the transformation, and `after` after it. For example,

```
`a' => "o" ;
```

will transform all 'a's in the file into 'o's.

How do we describe characters and strings? On the left side of `=>` we can only put separate characters: they can be written either as "grave accent+ASCII character+apostrophe" or as `@"XYZT` where `XYZT` are hexadecimal digits: in this case we are not restricted to ASCII characters. The latter syntax can also be used on the right side. For example,

```
`i'`j' => @"0133 ;
@"008E => @"00E9 ;
```

will send the string 'ij' to the Unicode++ character representing the Dutch ij ligature, and the 8-bit code 8E (a Macintosh 'e' with acute accent) to the Unicode++ character 00E9 (which is the Unicode 'e' with acute accent).

On the right side of `=>` we can also write complete strings, possibly containing Ω commands, which will be forwarded to the next ΩTP or to the typesetting engine of Ω. For example,

```
`~' => "\penalty10000" ;
```

sends the tilde character to the TEX command of infinite penalty.[3] We can also use ranges on the left side: for example, `` `a'-`k' `` means "all characters between a and k".

By using parentheses and the vertical bar on the left side, we obtain the Boolean 'or' operator:

```
(`E'|`e') => ;
```

for example, will send both uppercase and lowercase letters 'e' to nothing (a transformation which would leave Perec's book *La disparition* unchanged[4]).

This operator becomes even more useful by the fact that we can use on the right side the exact character matched on the left side: the commands `\1`, `\2`, ... , `\9` used on the right side stand for the first, second, ..., ninth character matched on the left side. For example:

[3] By this we obtain the same result as in TEX but without turning tilde into an active character, a fact that TEX users will surely appreciate.

[4] Although there are rumors that there is a single 'e' in that book... The authors were not able to find it yet.

```
`c'(`a'|`e'|`i'|`o'|`u')`t' => "m" \1 "p" ;
```

will send cat, cet, cit, cot, cut respectively to map, mep, mip, mop, mup.

We can go even further: ΩTP syntax allows us to add or substract a fixed offset to the characters matched on the left side. For example:

```
`a'-`z' => #(\1 - @"0020) ;
```

will substract 20 from the code position of the character found on the left side. The characters on the left side being precisely lowercase letters, this offset will turn them into uppercase ones.

6.2.1 Examples

The beginning of the ΩTP 7arb2uni, used to send the ASCII transcription of Arabic to Unicode++, described in 5.2.1, to Unicode++, looks like this:

```
input: 1;
output: 2;

expressions:

`L'`L'`a'`h'    => @"FDF2 ;
`S'`L'`h'      => @"FDFA ;

`|'`|'`|'`|'    => @"0621 @"0651 ;
`|'`|'        => @"0621 ;
`z'`h'`z'`h'    => @"0698 @"0651 ;
`z'`h'        => @"0698 ;
`z'`z'        => @"0632 @"0651 ;
`z'          => @"0632 ;
`y'`y'        => @"064A @"0651 ;
`y'          => @"064A ;
`v'`v'        => @"06A4 @"0651 ;
`v'          => @"06A4 ;
`u'`N'        => @"064C ;
`u'          => @"064F ;
```

Let us take a closer look at these lines. The left sides `L'`L'`a'`h' and `S'`L'`h' correspond to the (religious) ligatures ﷲ and ﷺ which appear in the *Arabic Presentation Forms* part of Unicode, that's why the code positions we send them to are so high. The line `|'`|'`|'`|' corresponds to a double hamza; according to our transcription rules, by writing a letter's transcription twice without intermediate hyphen, we get the letter followed by a *shaddah* diacritic. On the right side of `|'`|'`|'`|' you see two codes: 0621 stands for the stand-alone hamza in Unicode++, and 0651 for the *shaddah*. The next line will send `||` to the stand-alone hamza.

WARNING: the order of these lines is very important: transformations are matched in the order lines are read. By putting the double hamza before the single one, Ω will first look for a double hamza and *only if it does not find any* will then proceed to transforming a single one.

For the same reason digraphs such as `zh` must appear before their first letter in the ΩTP file (and trigraphs before the starting digraph, etc.). That's why the order of lines starting with a 'z' is 'zhzh', 'zh', 'zz', 'z'.[5]

Our sample file ends like this:

```
`h'`h'   => #(@"0647) #(@"0651) ;
`h'      => #(@"0647) ;
`-'`-'`-' => @"2014;
`-'      => ;
.        => #(\1) ;
```

This means that after having entered all digraphs using 'h' as second character, we enter the stand-alone 'h', first as a double letter, and secondly as a single letter. Finally we send the triple hyphen to an m-dash '—' and the single hyphen to nothing: its purpose is to prevent combinations of letters to be interpreted as digraphs: when reading `k-h`, Ω will not match it with `kh`: it will first match `k` with letter kaf, then send the hyphen to the vacuum of non-existence and when arriving to the `h` the `k` will already be matched so that it is too late to construct a `kh` digraph.

The period at the beginning of the last line is part of the ΩTP syntax we have not seen yet: it means 'any character'. Since this is the last line of the file, we can interpret it rather like 'any still not matched character'. This line simply sends any character not yet matched to itself.

6.3 Wrapping it up

Once the ΩTP file has been written or modified, one only needs to compile it (by using the `otp2ocp` utility) and place it where Ω expects to find it. On the LaTeX command level, ΩTPs are loaded via the `\ocp` command, in a way similar to fonts: to load the file `foo2uni` one will write

```
\ocp\FooUni=foo2uni
```

Of course this is preferably done inside a LaTeX package or style file: the final user should not need to deal with or understand this kind of code. Once the ΩTPs are loaded they are combined into *lists*. In this way we can push or pop simultaneously ΩTPs on/from a stack. This is useful because a language switch usually

[5]There is a simple way of avoiding ordering problems: after having written this part of the ΩTP file, run a line sorting program on it so that lines are sorted in *inverse* lexicographical order. This will automatically place trigraphs before digraphs before singletons, etc.

requires several ΩTPs to be changed at once. To define ΩTP lists we use the following syntax:

```
\ocplist\ArabicOCP=
\addbeforeocplist 100 \ArabUni
\addbeforeocplist 200 \UniCUni
\addbeforeocplist 300 \CUniArab
\nullocplist
```

The numbers (100, 200, 300) allow us to introduce additional ΩTPs, if necessary, between the already defined ones. Finally, to activate/desactivate an ΩTP list, we use the commands `\pushocplist` (followed by the name of the ΩTP list) and `\popocplist`. To take a real life example,

```
\ocp\ArabUni=7arb2uni
\ocp\UniCUni=uni2cuni
\ocp\CUniArab=cuni2oar
\ocplist\ArabicOCP=
\addbeforeocplist 100 \ArabUni
\addbeforeocplist 200 \UniCUni
\addbeforeocplist 300 \CUniArab
\nullocplist
\pushocplist\ArabicOCP
```

is sufficient to load all ΩTPs necessary for typesetting in the Arabic language.

7 Availability and Further Information

The Ω system is entirely in the public domain. It can be obtained from any CTAN server. The latest information on Ω and its Arabic system can be found on the Ω server:

`http://www.ens.fr/omega`

courtesy of the École Normale Supérieure de Paris.

8 Samples

Starting from next page, a few samples (Arabic, Berber, Sindhi). For these examples we have switched the background language to Arabic, so that even page numbers are in Arabic.

٨.١ أطفال الغابة

كان لأحد الملوك القدماء أخت تعيش معه فى قصره، بعد أن ماتت زوجته، وتركت له من الأولاد ثلاثة: أميرين وأميرة. وقد ازداد حبّ الملك لأولاده، بعد وفاة والدتهم الملكة، وأحبّهم حبّا كثيرا؛ ليعوّضهم ما فقدوه من عطف أمّهم وحبّها لهم، وتفكير ها فيهم؛ فكان يسأل عنهم كلّما حضر، ويفكّر فيهم كلّما دخل، ويوصى بهم كلّما خرج، ويطلبهم كلّما جلس لتناول طعام الإفطار أو الغداء أو الشّاى أو العشاء.

محّمة أخيها لأولاده، وصمّمت فيما بينها وبين نفسها أن تعمل سرّا كلّ وسيلة ممكنة لإبعادهم عن أبيهم والتّخلّص منهم.

وفى يوم من الأيّام كان الأميران يلعبان مع أختهما الأميرة فى حدائق القصر بعد خروج الملك، فشوّقتهم عمّتهم وحبّبت إليهم الذّهاب معها إلى الغابة للّعاب فيها، ووعدتهم أن تريهم أشياء جميلة وألعابا لذيذة سارّة تحت الأشجار هناك.

فصدّق الأميران والأميرة ما قالته عمّتهم، ولم يعرفوا ما تخفيه عنهم من الشّرّ، وذهبوا معها للّعب والرّياضة فى ألغابة، ومشاهدة الأشاء الجميلة فيها، ورؤية الألعاب الغريبة تحت أشجارها.

وقد شعر الأطفال بسرور كثير عند ماخرجوا مع عمّتهم لهذه الرّحلة. وأخذوا يمشون معها فى الغابة حتّى وصلوا إلى وسطها، فأحسّوا بالتّعب الشّديد، وطهرت علاماته فى مشيتهم، وعلى وجوههم بعد هذه الرّحلة الطّويلة المتعبة الّتى لم يجرّبوها من قبل. ولما شعرت العمّة بشدّة تعبهم، قالت لهم: ناموا هنا تحت هذه الشّجرة حتّى تحضر الحوريّات لتلعب أمامكم ألعابا لم تروها، وستجدون فى مشاهدتها كلّ لذّة وسرور.

٨.٢ الّال ي وسقدش ن يضريسن Ω دتامازيغت

ا دّ نسّكن س وايس يّف وسقدش ن Ω ي تيرا س توتلايت تامازيغت، اما س تيفيناغ، اما س يسكّيلن يلاطانيّن. نويّدّ تامازيغت ام، توتلايت يدّرن (يتّوارون س تيفيناغ تيّنايّن) : يزمر ومدان اد يسدّو يالّ تيغورا ن وسودّس ن تيرا، ي واراتن وسنّانن، يتكنيكن نغ ي ويد ن تسيكّلا، ام ويد سّخدامن ي وسمسارو ن تفرانسيست.

Ω، د امسلاي ن وسميهل ي وسودّس ن تيرا. امواكّن نزرا، د اين ي دّ يتّاكّن يزوبّاز وار تاڨّارا ي وسقدش د وسيهرو، ماشا يسّفك اد يلمد وقدّاش كرا توسّنيوين. نونزاس، تازّايت ن ولمادا، نزمر ا تّ نسيفسّ س وسقدش ن يناڨراون ن ورماس ن تيرا، يسغزانن ن وسميهشل ن واراتن، يتّواسّنن اطاس (ويد يتّنوزون، سراين غف ومدان، ويد يزمرن اد سّخدمن تازمرت تاسمسيراوت ن كرا يناڨراون يمهلانن ام ويد ن Macintosh, Windows, Unix.

تانضا تامزواروت ن Ω — غاس تين اي يتّالاسن يسم ن Ω —، وس تلي اڨرودم ي وقدّاش. ام ڨّ يمسلاين ن وسميهل اكّ، اد يارو ومدان اهيل، دڨّير، ا ت يسّفسو اكّن ا ت يسّوغال س انقال ن تماشينت. دي Ω، اهيل د ارا ن وضريس (نطّضن غورس كرا ن تسونضيوين ي وسبوني د تغسّا تامزّولت). اسفسو، د اسلكم ن واهيل Ω ؛ انڨال ن تماشينت ارا دّ يفّغن، د وين، ي د اڨلام ن وسبتر اي يتّوسودّسن، يقّيمدّ يمين وسمسارو.

اكالايا، يزمر ا ت ياف يفرغ وين ينومن يسقداش ن يضريسن غف Macintosh, Windows، د ويّيض، ي دقّ اضريس ا دّ يفّغ دي تسمساروت اكّن يلّا ڨّ وقديل [اناڨراوا يتّواسّنن س يّسميس يميوزيل، س تڨليزيت « wysiwyg » ، يشسّغلاض كرا : اضريس ارا دّ يسوفّغ وقدّاش، اد يلّي غاس س تسدّي ومي يسّاوض وڨديل ؛ اسڨموض ارا دّ تسوفّغ تسمساروت، يسمر اد يلّي يوسر كرا.]

يواكّن اد يقّيم وسقدش سراي ف ومدان، يزمر اد يسّخدم اسغزان يتّواسّن د الاّلن ي ورماس. تاغسّا تامزّولت ن وارا (يغفاون، تيفولا، تيسدّارين، تيزميلين تينادّاين، تيميتار تينمو‌داق، اسمل ن تكتابين)، ا تّ يسّيغال سي تبونيت ن وسغزانّي غر تسونضيوين ن Ω. يمير، Ω، اد يسّفسو انڤالنّي ا دّ يسّوفّغ اضريس يوقزن تاغسّا تامزّولت تامزواروت، ماشا تيراينس اد يلينت ولاغنت وڤار.

٨.٣ كتين كر موڙيا جڏهن

تنهن ڪري اسان کي پنهنجي ذهن کي سجاڳ رکڻو پوندو ۽ پنهنجي جدوجهد ۾ ڏاهپ پيدا ڪرڻي. اهو بہ معلوم ڪرڻو پوندو تہ سنڌ ۾ هر آئي وقت ڇا ڇا ٿي رهيو آهي ۽ دشمن اسان جي ۽ اسان جي جدوجهد جي ڪلاف ڪهڙا ڪهڙا گهاٽ گهڙي رهيو آهي.

اسان کي اها بہ خبر هجڻ گهرجي تہ اسان جي آس پاس ۽ يسگردائي ۾ ڇا ڇا ٿي رهيو آهي. هندستان ۾ ڇا ٿي رهيو آهي، افغانستان ۾ ڇا ٿي رهيو آهي. عراق ۽ ايران ۾ ڇا ٿي رهيو آهي ۽ آمريڪا ۽ سوويت يونين ڇا ڇا سوچي رهيا آهن. جڏهن اسان سڄي دنيا جي سياست تي ۽ سڄي دنيا جي جدوجهد تي ۽ سڄي دنيا جي تنديلين تي نظر رکنداسون ۽ انهن تبديلين جي اثرن کي پنهنجي ملڪ، قوم ۽ عرام تي پوندي ڏسنداسون تہ انهن تبديلين مان ڪهڙا منفي ۽ ڪهڙا مثبت اثر آهن. تڏهن ئي اسان پنهنجي جدوجهد کي بهتر بہ ڪري سگهنداسون تہ چوٽڪاري وارو حل بہ ڳولي ويندا‌سون.

رڳو ڳالهيون ڪندي ۽ نعرن هڻندي اسان جي قوم چاليھ سال پيڙائون ۽ عذاب پوڳيا آهن ۽ انهن نعرن اسان جي قوم لاءِ وڌيڪ پيڙائون ۽ عذاب نازل ڪيا آهن . جيڪڏهن اسان ۾ اڄ قوم جي اميد پيدا ٿي آهي تہ اها اسان جي عمل ۽ اسان جي بي لوث جدوجهد جي ڪري پيدا ٿي آهي ۽ مائهو اسان ڏانهن واجهائي رهيا آهن. تہ اسان ئي آهيون جيڪي ڪجھ نہ ڪجھ ڪنداسون. پر اسان کي ڏسڻو آهي تہ دنيا جي اندر ڇا ٿي رهيو آهي ۽ اسان جو دشمن ڪيئن حالتن کي پنهنجن مفادن ۾ ڪتب آڻڻ جي ڪوشش ڪري رهيو آهي، انهيءَ جي لاءِ ضروري آهي تہ اسان پاڻ ۾ ڏاهپ پيدا ڪريون ۽ پاڻ ۾ ڄاڻ جو هڪ وسيع خزانو پيدا ڪريون. ۽ اسان موجوده صورت حال کي سمجهڻ لاءِ روزمره جي ميڊيا ۽ دنيا جي اندر ٿيندڙ ڪاروائين تي گهري نظر رکون تہ دشمن ۽ جارحيت پسند قوتون ۽ اسان تي قابض قوتون دنيا جي اندر ٿيندڙ تبديلين کي سندن حق ۾ ۽ سندن مفادن جي حق ۾، سنڌ تي جارحيت قائم رکڻ جي حق ۾، سنڌ کي مستقل قبضي ۾ ڪرڻ جي حق ۾ ڪيئن ڪتب آڻي رهيون آهن.