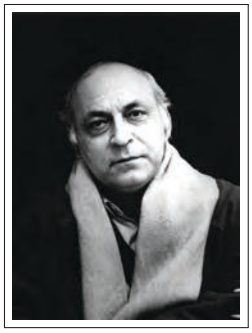

الگوریتم بازنویسی کلمات فارسی به الفبای لاتین به‌عنوان نام دامنه

سیاوش شهشهانی
دانشگاه صنعتی شریف

یادداشت (۱۳۹۹). نوشتهٔ زیر، با اندک تغییر، پیشنهادی است که نگارنده در سال ۱۳۸۳، وقتی مسئولیت بخش ثبت دامنهٔ ir. در پژوهشگاه دانش‌های بنیادی را به‌عهده داشت، تهیه کرد. هدف، ارائهٔ توصیه به ثبت‌کنندگان دامنه‌های اینترنتی بود که حتی‌المقدور از یک شیوهٔ برگرداندن نام‌های فارسی به الفبای لاتین استفاده کنند. متأسفانه روش استاندارد قابل استفاده‌ای برای بازنویسی فارسی به لاتین در مرورگرهای اینترنتی وجود نداشت، و هنوز هم وجود ندارد. استفاده از سلیقهٔ شخصی در این بازنویسی گاه منجر به آسیب‌هایی می‌شود که تبعات حقوقی ناشی از ابهام در اسامی ثبت‌شدهٔ تجاری از آن جمله است. چون برای درج نشانی‌های اینترنتی در مرورگر رایانه فقط استفاده از ۲۶ حرف لاتین به شکل استاندارد (ASCII)، ده رقم 0 تا 9 و خط تیرهٔ میانی مجاز است، نمی‌توان مثلاً از روش‌های جامع و بی‌ابهام زبان‌شناختی به این منظور استفاده کرد. الگوریتمی که در زیر مشاهده خواهید کرد با توجه به این محدودیت تدوین شده و طبعاً به روش آوانگاری نزدیک‌تر است تا حرف‌نگاری که مجموعهٔ گسترده‌تری از نمادها را می‌طلبد. البته با توجه به تنوع لهجه‌های ایران و سایر نواحی که زبان فارسی در آنها رایج است آوانگاری خالی از اشکال نیست و اتخاذ نوعی «فارسی استاندارد» اجتناب‌ناپذیر است. به این منظور تلفظ تهرانی که مورد استفادهٔ رسانه‌های رسمی صوتی/تصویری کشور است مبنا قرار گرفت.

باید اضافه کنیم که به دو دلیل کوشش گسترده‌ای برای رواج این الگوریتم صورت نگرفت. یک دلیل، ابداع و اتخاذ بین‌المللی سامانه‌ای برای استفاده از الفباهای غیرلاتین، از جمله الفبای فارسی، در مرورگرها بود که می‌توانست ضرورت استفاده از الفبای لاتین را کمرنگ‌تر کند. با اینکه مشکلات فنی این سامانه (موسوم به IDN) عملاً حل شده است و پژوهشگاه مجوز لازم برای عرضه آن را دارد، راه‌اندازی آن سامانه به‌تعویق افتاده و در هر صورت تعداد دامنه‌های لاتین ir. اکنون به بیش از یک میلیون فزونی یافته است. دلیل دوم نگرانی‌های ابرازشده به این مضمون بود که استفاده از یک شیوهٔ استاندارد فارسی‌نگاری به لاتین می‌تواند کوشش خزنده‌ای برای تغییر خط فارسی به لاتین تلقی شود. در اینجا لازم است هرگونه شبهه در این مورد را قویاً رد کنیم. کشورهای خاوردور که شیوه‌های نگارش بسیار پیچیده‌تر از فارسی دارند سال‌هاست از روش استاندارد برای لاتین‌نگاری زبان‌هایشان در شرایط مورد نیاز استفاده می‌کنند بدون آنکه نگارش سنتی‌شان را کنار بگذارند. در مقابل، نمونهٔ ترکیه که تغییر خط، گسست عمیق هویتی و تاریخی در آن پدید آورده پیش روی ماست. به‌علاوه ما معتقدیم که اشکالات خط فارسی در برابر نعمت بزرگی که همین ابهامات تلفظی در حفظ وحدت و توسعهٔ زبان فارسی ایجاد کرده است بهایی صرف‌نظرکردنی است. با توجه به تنوع تلفظها و دامنهٔ وسیع جغرافیایی استفاده از زبان فارسی که زمانی از اویغورستان و شبه‌قارهٔ هند در شرق تا امپراطوری عثمانی در غرب را فرامی‌گرفت، آویختن زبان به یک الفبای صلب و بی‌ابهام بی‌شک به تجزیه زبان منجر می‌شد. بدین ترتیب پیشنهاد ارائه شده در اینجا فقط گامی در تعدیل هرج‌ومرج ایجادشده توسط «پینگلیش»های غیراصولی است.

از خوانندگان دعوت می‌کنیم پیشنهادهای اصلاحی خود را به این نشریه یا نگارنده بفرستند.

در ضمیمهٔ این نوشته الگوریتمی برای تخصیص کُد دو حرفی به استان‌های کشور آمده است که به پیشنهاد یکی از مدیران وقت سازمان مدیریت و برنامه‌ریزی تهیه شد و به همین روال، در آن فقط از ۲۶ حرف لاتین، ده رقم ۰ تا ۹ و خط تیره استفاده می‌شود. جدول‌های تبدیل در زیر آمده‌اند.

(الف) حروف بی‌صدا

ب	b
پ	p
ت	t
ث	s
ج	j
چ	c
ح	h
خ	x
د	d
ذ	z
ر	r

ز	z
ژ	zh
س	s
ش	sh
ص	s
ض	z
ط	t
ظ	z
ع	[بحث (ب) در زیر]
غ	q
ف	f

ق	q
ک	k
گ	g
ل	l
م	m
ن	n
و	v
ه	h

(ب) حروف ساده و مرکب صدادار

آ	aa
اَ	a

اِ	e
اُ	o

او	u
ای	i

ـــ «ع» و «همزه»، اگر صدادار باشند پس از حروف صدادار حذف می‌شوند و پس از حرف بی‌صدا به صورت خط تیره نمایش داده می‌شوند:

پس از حرف صدادار: سعید (said)، شاعر (shaaer)، مؤثر (moasser)، روئید (ruid).

پس از حرف بی‌صدا: مسأله (mas-ale)، اطعام (et-aam)، تسعیر (tas-ir)، مسعود (mas-ud).

ـــ «ع» و «همزه» اگر بی‌صدا باشند با ee نمایش داده می‌شوند مگر اینکه قبل از آن صدای e ظاهر شود که در آن صورت به جای eee نوشته می‌شود ee: معروف (maeeruf)، شیء (sheyee)، صنع (sonee)، سعد (saeed)، نعمت (neemat، نه neeemat)، معراج (meeraaj، نه meeeraaj).

ـــ در بسیاری کلمات فارسی صدای «یا» هست مانند زیاد، سیاه، بیا، میان، خیال، ... در این موارد استفاده از iyaa دقیق است هر چند که iaa نزدیک‌تر به استفاده متداول (معمولاً با یک a) می‌باشد. ظاهراً خلاصه کردن iyaa به iaa نباید ابهامی ایجاد کند. تنها موردی که در حال حاضر به نظر می‌رسد خلط «میعاد» با حالت عامیانه «می‌آید» یعنی «میاد» است (هر دو miaad). اگر در قرارداد مربوط به «ع» و «همزه» پس از حرف صدادار تجدید نظر کنیم و همواره خط تیره را همه جا به کارگیریم این تشابه ایجاد نمی‌شود ولی وفور خط تیره خوش‌ظاهر نیست.

ـــ وقتی أصوات /آ/ و/ یا /اَ/ متوالیاً ظاهر شوند باید با خط تیره ابهام‌زدایی کرد، مثلاً: معاد (ma-aad)، ساعت (saa-at)، ناآرام (naa-aaraam)، معالفارق (ma-alfaareq).

ـــ صدای ضمهٔ کشیده (مانند صدای «و» در «نو» یا «مورد») در اینجا به تساهل با o کوتاه (مانند صدای «و» در «نوک») یکی گرفته شده است، ولی می‌توان قرارداد دیگری چون استفاده از ow یا w تنها را در نظر گرفت .

ـــ صداهای «اَی»، «اِی»، «اَی» و «اوی» به ترتیب به /oy/، /ey/، /ay/، و /uy/ نمایش داده می‌شوند. «ی» بی‌صدای آخر کلمه نیز به /y/ نمایش داده می‌شود مثل مشی (mashy).

(ج) ملاحظاتی در مورد حروف بی‌صدا

همهٔ حروف بی‌صدا با یک حرف لاتین نشان داده شده‌اند به استثنای «ش» و «ژ». این امر به‌خصوص در مورد صدای «ش» که در فارسی فراوان است موجب تأسف است. ولی صداهای s، h، و نیز z،و h ممکن است به طور متوالی ظاهر شوند که ایجاد ابهام در خواندن می‌کند. این پدیده در مورد لغاتی مانند تسهیل، اسحق، مذهب، اظهار، مضحک .... دیده می‌شود. می‌توان در موارد نادری که امکان ابهام باشد با خط تیره مشخص کرد که s-h و z-h نباید «ش» یا «ژ» تلفظ شوند. بعضی پیشنهاد کرده‌اند که حرف w که در تلفظ فارسی استاندارد مصداقی ندارد می‌تواند به جای «ش» به کار گرفته شود. این پیشنهاد چند اشکال دارد. یکی اینکه برای هر خوانندهٔ ایرانی که آشنایی مختصری هم با زبان‌های غربی داشته باشد تطبیق w با «ش» دور از ذهن است. در اینجا طرفداران ش=w اشاره می‌کنند که در روسی و عبری صدای (ش) با حرفی مشابه w ادا میشود که ظاهراً از حرف مشابه فنیقی آمده است (∑ یونانی نیز چنین است). اشکالی دیگر این است که هر چند w انگلیسی یا عربی در تلفظ فارسی تهرانی ظاهر نمی‌شود ولی در بسیاری لهجه‌های دیگر فارسی‌زبانان رایج است. بالاخره اینکه کلمات انگلیسی و عربی نیز بسیاری اوقات تحت دامنه ir. ثبت می‌شوند و امکان ابهام وجود دارد. مثلاً wahid را می‌توان تلفظ عربی «وحید» در نظر گرفت یا «شهید». همچنین warm به انگلیسی معنی دارد (گرم) و «شرم» نیز خوانده خواهد شد. البته به کار گرفتن c برای «چ» و x برای «خ» نیز ممکن است موارد برخوردی با کلمات غیرفارسی ایجاد کند ولی هر دوی این حروف در بعضی زبان‌های اروپایی با همین صداهای «چ» و «خ» تلفظ می‌شوند.

در مورد به کارگیری q برای هر دوی «غ» و «ق»، این در واقع تفاوتی با به کارگیری z (ذ، ز، ض، و ظ)، s (ث، س، و ص) و t (ت، و ط) به چند منظور ندارد، هر چند که در بعضی نواحی ایران این دو حرف متفاوت تلفظ می‌شوند. کاربرانی که مصر به ایجاد تمایز میان این دو حرف باشند می‌توانند از gh برای «غ» و از q برای «ق» استفاده کنند . تفکیک گسترده‌تر میان همهٔ حروف عربی که در فارسی معمولاً یکسان تلفظ می‌شوند با توجه به محدودیت علایم قابل استفاده در دامنه‌های اینترنتی چندان عملی نیست.

(د) کسرهٔ اضافه

در اینجا تصمیم‌گیری اصلی در مورد چسباندن e به آخر کلمه به‌واسطه یا بی‌واسطهٔ خط تیره است. به دلایل زیر، ما چسباندن بدون خط تیره را انتخاب کرده‌ایم:

۱. وفور کسره باعث می‌شود که تدریجاً همگان به سوی حذف خط تیره سوق داده شوند.

۲. بهتر است از خط تیره فقط در موارد واقعاً ضروری که جداسازی مانع از ابهام می‌شود استفاده شود، مثلاً در اصطلاحات مرکب مانند حدیث عشق (hadise-eshq) که hadiseeshq تلفظ دیگری را میرساند و hadis-e-eshq فقط طولانیتر از hadise-eshq است و اطلاع تازه‌ای به دست نمی‌دهد.

۳. می‌توان کسرهٔ اضافه را یک پی‌بند تلقی کرد. همچنان که در زبان‌هایی مثل آلمانی و روسی استفاده از پی‌بند کاملاً جاری است، افزودن e به آخر کلمه نهایتاً هیچ مشکلی ایجاد نمی‌کند. در مواردی که خود کلمه به حرف صدادار ختم می‌شود (مثل بُته، فضا) همچنان که در تلفظ رایج فارسی قبل از افزودن کسره، یك «ی» قرار می‌دهیم می‌توان از ye استفاده کرد (boteye-gol، fazaaye-sabz).

(هـ) استفاده از خط تیره

چون خط تیره تنها علامت غیرالفبایی و غیرعددی مجاز در دامنه‌های اینترنتی است سعی می‌کنیم در استفاده از آن زیاده‌روی نکنیم و آن را برای موارد ضروری حفظ کنیم. به غیر از مواردی که در بالا به آن اشاره شد، خط تیره در دامنه‌های اینترنتی وقتی به کار می‌رود که ثبت یك دامنهٔ مرکب و طولانی مورد نظر باشد. در اینجا نیز سلیقهٔ بعضی چسباندن کلمات بدون خط تیره است. مثلاً فرهنگستان زبان و ادب فارسی، persianacademy.ir را ثبت کرده است، نه persian-academy.ir. اگر نام شرکتی «جیغ بنفش ممتد» باشد، برای این شرکت ترکیبات متعددی از سه جزء نام امکان‌پذیر است، مانند: jiq-banafsh-momtad.ir، و jiqebanafshe-momtad.ir ....

در این موارد کوشش خواهد شد سلیقهٔ دارندهٔ هر دامنه ملحوظ شود. باید توجه داشت که دامنهٔ اینترنتی در مجموع یک واحد یا کد شناساننده محسوب میشود و ترکیب اجزای داخلی آن به‌واسطه یا بی‌واسطهٔ خط تیره تابع سلیقه ثبت‌کننده است.

(و) تشدید

تشدید با تکرار حرف نمایش داده می‌شود: مهذّب (mohazzab)، حج تمتّع (hajje-tamattoee یا hajjetamattoee).

ضمیمه

۱. الگوریتم تخصیص کد دو حرفی به استان‌های کشور

۱) نام کامل استان طبق روشی که در متن «الگوریتم بازنویسی کلمات فارسی به الفبای لاتین به عنوان نام دامنه» آمده است به حروف لاتین برگردانده می‌شود. از این پس کلیهٔ ارجاعات به نام استان به این برگردان لاتین خواهد بود. جزییات روش در بخش ۲ مرور شده است.

۲) حرف اول نام استان به عنوان حروف اول کد منظور می‌شود.

۳) در صورتی که نام استان از دو کلمه یا بیشتر تشکیل شده باشد، حرف دوم کد، حرف اول جزء دوم نام استان خواهد بود. حروف ربط نادیده گرفته می‌شوند، بنابراین مثلاً به سیستان و بلوچستان کد SB تعلق می‌گیرد.

۴) اگر نام استان فقط یک کلمه باشد حرف دوم کد، اولین حرف بی‌صدای پس از اولین حرف خواهد بود. مثلاً در مورد استان فارس (Fars) ، کد دو حرفی FR می‌شود.

روش بالا یک به یک نیست و تضادهایی پیش می‌آید. گام‌های زیر برای رفع تضاد در نظر گرفته شده‌اند.

۵) چنانچه نام یک استان جزء اولیه نام استان دیگری باشد (مثلاً کرمان جزء اولیه نام کرمانشاه است)، روش بالا برای نام کوتاهتر در نظر گرفته می‌شود و برای نام طولانی‌تر به گام (۷) مراجعه کنید.

۶) در صورت تضاد، به استثنای موردی که در (۵) اشاره شد، استان‌های مورد بحث برحسب جمعیت از جمعیت بیشتر به جمعیت کمتر اولویت‌بندی می‌شوند و تخصیص حرف دوم کد (گام ۴) برحسب اولین حرف بی‌صدای موجود خواهد بود. مثلاً برای گیلان (Gilan) و گلستان (Golestan)، چون گیلان پرجمعیت‌تر از گلستان است، کد GL برای گیلان و کد GS برای گلستان منظور می‌شود.

۷) در وضعیتی که نام یک استان جزء اولیهٔ نام استان دیگری است، حرف دوم کدِ نام طولانی‌تر، اولین حرف بی‌صدای پس از حذف نام کوتاه‌تر خواهد بود. مثلاً برای کرمانشاه (Kermanshah)، حرف S (پس از حذف Kerman) در نظر گرفته می‌شود، پس کد دو حرفی آن KS خواهد بود.

۲. شیوهٔ برگرداندن نام استان‌ها به لاتین

برای زبان عربی شیوهٔ استانداردی مبتنی بر حرف‌نگاری (transliteration) معمول است که به دو دلیل در مورد اسامی فارسی به خصوص در رابطه با محدودیت‌های نام دامنهٔ اینترنتی، مشکل می‌آفریند. این دلایل عبارت‌اند از:

۱) تلفظ فارسی و عربی در بسیاری موارد متفاوت است، مثلاً کلمه «اصفهان» در عربی به گونه‌ای تلفظ می‌شود که برگردان Isfahan را القاء کرده است در حالی‌که تلفظ فارسی آن Esfahan است.

۲) در مورد گروه‌های حروف {ذ، ز، ض و ظ}، {ث، س، ص}، {ت، ط}، {غ، ق}، و {الف، ع}، حروف درون هر گروه در اکثر نقاط ایران به یک شیوه تلفظ می‌شوند ولی در عربی تلفظ متفاوت دارند. با توجه به محدودیت ۲۶ حرفی حروف ASCII می‌توان برای همهٔ حروف یک گروه در فارسی فقط یک حرف در نظر گرفت. در حال حاضر اکثر این حروف در اسامی استان‌ها ظاهر نمی‌شوند و مشکلی ایجاد نمی‌شود. یک مورد قابل بحث دو حرف غ و ق است که ق در «قم» و «قزوین» یافت می‌شود و غ در «آذربایجان غربی». برای ق استفاده از Q که در عربی نیز مرسوم و تقریباً جا افتاده است، ولی در مورد «غ» در کلمهٔ «غربی»، می‌توان Qarbi یا Gharbi نوشت. بعضی معتقدند چون در بعضی نقاط ایران غ و ق دو گونه تلفظ می‌شوند حفظ این تمایز مناسب است.

۳) در مورد حرف «چ» چون حرف لاتین «c» در فارسی کاربرد دیگری ندارد، «c» را به جای «چ» به کار می‌گیریم. اگر شیوهٔ نمایش انگلیسی «ch» نیز برای «چ» به کار گرفته شود و مجموعهٔ «ch» یک حرف تلقی شود، تناقضی بین دو روش پدید نمی‌آید.

۴) مورد دشوارتر حرف «خ» است که معمولاً در فارسی به «kh» نمایش داده می‌شود. این امر با توجه به تعدد نام‌های استان‌هایی که با «خ» یا «ک» آغاز می‌شوند مسئله‌ساز است: کرمان، کرمانشاه، خراسان رضوی، خراسان شمالی، خراسان جنوبی، خوزستان، کهکیلویه و بویراحمد، و کردستان. بنابراین از «x» برای «خ» استفاده خواهد شد.

۵) برخلاف زبان انگلیسی که در آن استفاده از چهار مخفف N، S، E، و W برای چهار جهت خالی از ابهام است، در فارسی «شمال» و «شرق» (خراسان شمالی، آذربایجان شرقی) هر دو با حرف «ش» آغاز می‌شوند. در اینجا ما از روش دائرةالمعارف مصاحب استفاده می‌کنیم، یعنی شرق را به «ش» (SH) و شمال را به «ل» (L) نمایش می‌دهیم.

۳. جدول کدهای اختصاری استان‌ها

با توجه به مطالب بخش‌های ۱ و۲ کد دو حرفی اختصاری استان‌ها به شرح زیر خواهد بود.

آذربایجان شرقی	AS
آذربایجان غربی	AG/AQ
اردبیل	AR
اصفهان	ES
ایلام	IL
بوشهر	BS
تهران	TH
چهارمحال و بختیاری	CB
خراسان جنوبی	XJ
خراسان رضوی	XR
خراسان شمالی	XL
خوزستان	XZ
زنجان	ZN
سمنان	SM
سیستان و بلوچستان	SB

فارس	FR
قزوین	QZ
قم	QM
کردستان	KD
کرمان	KR
کرمانشاه	KS
کهگیلویه و بویراحمد	KB
گلستان	GS
گیلان	GL
لرستان	LR
مازندران	MZ
مرکزی	MR
هرمزگان	HR
همدان	HM
یزد	YZ

یادداشت ۱۳۹۹: مخفف دوحرفی استان البرز، که در زمان تهیهٔ متن بالا وجود نداشت طبق این روش به صورت AL خواهد بود.