

Problemi tipografici di trascrizione in un dizionario

Luca Busetto

Università degli Studi di Torino

Abstract

The text introduces some typographical problems which may occur in the transcription/citation, but also in the encoding, of scientific works, such as dictionaries and grammars in Latin characters, above all of historical-comparative area. Particular reference is to the use of the Combining Diacritical Marks and some special characters. It is also pointed out the Unicode Standard of character encoding and to the OpenType fonts, which possess advance property for the text rendering (graphical arrangement): these instruments enable to go ahead in the preparation and publication of scientific articles and essays.

1 Introduzione

Solitamente, in un'opera lessicografica o grammaticale di una lingua che dispone di una norma i lemmi sono trascritti in modo più ricco dell'ortografia standard, cercando di dare informazioni sulla pronuncia, ad esempio mediante l'uso di segni diacritici di combinazione o tramite simboli speciali.¹ I dizionari della lingua italiana, in particolare, indicano per ogni lemma la sillaba che porta l'accento intensivo di parola. Allo stesso modo l'uso di un accento grave (◌̀) o di un accento acuto (◌́) indica anche il timbro aperto [e ◌̝] o chiuso [e ◌̞] delle vocali medie anteriori *e E* e posteriori *o O* accentate.²

Quando si ha a che fare con dizionari etimologici, dialettali o di lingue prive di una norma stabile, ci si trova sempre davanti a trascrizioni complesse. CANEPARI [2003: 94], nel sostenere un largo uso di trascrizioni fonetiche dettagliate, tramite la sua riforma ^{can}IPA dell'alfabeto fonemico internazionale (^{uff}IPA), porta un esempio limite di una forma come f^{h} (IPA [ˈœː]) che non fa uso di simboli speciali ma ricorre solo a diacritici. Non di rado comunque, soprattutto in ambito fonetico, da una parte, e in ambito storico-comparativo, dall'altra, ricorrono cu-

¹ In alcune vecchie edizioni di vocabolari della lingua italiana si può trovare, ad esempio, direttamente nel lemma, la *s* sonora (IPA /z/) scritta ⟨ʃ⟩ (U+0283, forma allungata di ⟨s⟩), e la *z* sonora (IPA /dz/) scritta ⟨ʒ⟩ (U+0292, variante caudata di ⟨z⟩), mescolando nell'ortografia standard l'uso del sistema di trascrizione Ascoli-Merlo. I simboli ⟨ʃ⟩ e ⟨ʒ⟩ erano in origine varianti (corsive) di ⟨s⟩ e ⟨z⟩. Per ⟨ʃ⟩ è evidente l'influsso del corsivo (cfr. nota 1), in quanto la variante di ⟨s⟩, nello stile tondo, era ⟨ſ⟩ (U+017F). In origine, nei testi a stampa, per *s* si usavano due forme tassografiche: una, ⟨ſ⟩, usata in inizio e interno di parola; l'altra, ⟨s⟩, usata, proprio come la greca ⟨ς⟩, solo in finale.

² Può essere interessante notare che in moltissime lingue africane prive di una tradizione grafica antica, e che hanno avuto una codificazione scritta in caratteri latini in tempi relativamente recenti, si usano forme distinte per le vocali medie aperte e chiuse. Per le medio-basse anteriori si usano i simboli ⟨E⟩ (U+025B, U+0190); per le medio-alte anteriori ⟨Ė⟩ (U+0065, U+0045); per le medio-basse posteriori ⟨O̝⟩ (U+0254, U+0186); per le medio-alte posteriori ⟨O̞⟩ (U+006F, U+004F).

muli di diacritici sovrascritti o sottoscritti che devono essere efficacemente distanziati tra loro per non sovrapporsi e diventare illeggibili. Ad esempio, in una trascrizione IPA, un'occlusiva dentale con fonazione cricchiata avrà bisogno di due diacritici, ̥ (U+032A, che indica il preciso carattere lamino-dentale del simbolo coronale anteriore – “dentale” in senso lato – a cui è sottoscritto [d]) e ̬ (U+0330, che indica la fonazione cricchiata, in cui le cartilagini arite-noidèe sono pressate e le pliche vocali vibrano solo anteriormente). L'unico risultato accettabile dovrà essere [d̥̬], con i diacritici centrati e adeguatamente distanziati.

2 I problemi di visualizzazione e lo standard unicode

Fino a quando un dizionario resta su carta, l'utente che lo consulterà vedrà le pagine esattamente come sono state stampate. In alcuni casi potrà avere difficoltà a riprodurre alcuni simboli ma la corretta visualizzazione del testo è salva. I problemi maggiori si hanno invece con altre forme di pubblicazione, come ad esempio quelle su supporti informatici come cd-rom o, soprattutto, quelle su internet. In questi casi i pericoli di una scorretta visualizzazione sono ancora alti (e non è raro, purtroppo, trovare documenti PDF non visualizzabili correttamente).

Prima della diffusione di tipi di carattere (*fonts*) con codifica unicode,³ e di applicazioni che li supportino appieno, i caratteri speciali usati per le pubblicazioni scientifiche, per trascrivere o traslitterare avevano necessariamente una codifica presa a prestito, ad esempio, da quella del codice ascii. Ciò significa che non avendo il carattere specifico utilizzato per scrivere il testo non è più riconoscibile, decifrabile.

I computer, in buona sostanza, non sanno far altro che trattare numeri. Per immagazzinare in memoria lettere o altri segni è necessario che a ogni carattere venga assegnato un numero. Esistono centinaia di sistemi di codifica, preesistenti a Unicode, e ognuno di questi abbina i numeri ai caratteri in modo differente. Nessuna di queste codifiche comprende un numero di caratteri sufficiente per tutte le circostanze. Per le sole lingue dell'Unione Europea, ad esempio, è necessario utilizzare parecchi sistemi di codifica distinti. Anche considerando una sola lingua, come l'italiano, non esiste una codifica unica che comprenda tutte le lettere e tutti i segni di punteggiatura e simboli tecnici di uso comune.

Questi sistemi di codifica, inoltre, sono in contraddizione l'uno con l'altro. Succede che due codifiche utilizzino lo stesso numero per due caratteri *diversi* o che, viceversa, adottino numeri diversi per lo *stesso* carattere. Qualsiasi elaboratore, e a maggior ragione un server di rete, ha bisogno di utilizzare

³ Lo standard unicode ([uni'kɔ:de]) è un sistema di codificazione dei caratteri progettato per supportare su scala mondiale lo scambio, l'elaborazione, la visualizzazione dei testi scritti nelle varie lingue e discipline tecniche del mondo moderno. In più, nelle ultime versioni supporta anche moltissime scritture antiche (gotica, italica antica, ogamica, runica, fenicia, lineare B, cuneiforme, ugaritica, persiana antica, ...). Esso costituisce l'implementazione ufficiale dello standard internazionale per la codifica dei caratteri (ISO/IEC 10646). La codifica unicode assegna un numero univoco a ogni carattere, indipendentemente dalla piattaforma, dall'applicazione, dalla lingua. (Vd.: <http://www.unicode.org/>).

codifiche diverse. Il problema è che, quando i dati passano da una codifica a un'altra, o da una piattaforma a un'altra, si corre il serio rischio di perdere informazioni.⁴

Se un testo invece è scritto in caratteri unicode, anche in assenza del tipo di carattere usato, sarà comunque visualizzato correttamente, mediante l'uso suppletivo di un altro tipo di carattere: il carattere <a> (U+0061) rimane sempre "a" indipendentemente dai *fonts* utilizzati, così come il carattere medio inglese <ɜ> (U+021D) resta sempre "yogh" (m.ingl. *daɜ*), o come la forma usata in testi e lessici del medio alto tedesco <ꝛ> (U+0225) che resta "z uncinata" (m.a.ted. *ēzzen*), o ancora come <ɝ> (U+0292) che resta sempre "z caudata, ezh", con la sua codifica specifica.⁵

3 Le scritture complesse e i caratteri OpenType

La codifica standard unicode risolve molti problemi di compatibilità, in quanto il codice univoco di ogni carattere non può essere confuso. Ciò nonostante, in sistemi di scrittura più complessi di quello latino, questo non è sufficiente per una corretta visualizzazione. Ad esempio, la scrittura araba, che va da destra a sinistra, è una scrittura di tipo corsivo (non semplicemente *italico*) e perciò le lettere all'interno di una parola devono essere legate dovunque sia possibile. A parte le sei lettere 'ا, اء, اء, اء, اء, اء' (e quelle derivate da queste nell'ampliamento e adattamento della scrittura per varie lingue nordafricane, iraniche, indiane, ...) che non legano con la lettera seguente, tutte le altre hanno quattro forme: isolata, finale, iniziale, mediana. In scritture di questo tipo è necessario che le varianti contestuali vengano selezionate automaticamente.

Esistono progetti per la creazione di caratteri "intelligenti" in grado di supportare ogni tipo di scrittura complessa. Fra questi va menzionato *Graphite*, che dispone di un elaborato linguaggio di programmazione.⁶ Tuttavia, quella che al momento sembra essere la soluzione a questi problemi è lo sviluppo su larga scala di una nuova generazione di tipi di carattere chiamati OpenType. La loro codifica è basata sullo standard unicode e quindi si possono utilizzare per qualsiasi lingua e i documenti che li utilizzano sono indipendenti dalla piattaforma. L'aspetto più significativo è che possono contenere caratteristiche tipografiche avanzate, in modo da permettere un corretto trattamento dei linguaggi complessi che non utilizzano i caratteri latini. Inoltre per i le lingue con alfabeti latini si possono ottenere effetti tipografici avanzati (legature, varianti contestuali, storiche, linguospecifiche, ...).

4 I segni diacritici di combinazione

Spesso per chi deve comporre, trascrivere, citare lessici dialettali, trascrizioni fonetiche, forme etimologiche, i segni diacritici possono rappresentare un incubo. Nei caratteri propor-

⁴ Cfr. <http://www.unicode.org/standard/translations/italian.html>.

⁵ Non di rado purtroppo questi ultimi tre simboli distinti si trovano appiattiti e resi tutti come ɜ. Nel sistema fonemico IPA ha il valore di una fricativa postalveolare sonora solcata /ɝ/, mentre in altri sistemi di trascrizione, tra cui quello Ascoli-Merlo, è usata per l'affricata dentale sonora solcata /dz/.

⁶ Cfr. <http://graphite.sil.org/>.

zionali⁷ (con o senza “grazie”, ma in misura minore anche per quelli monospaziati) un segno diacritico quale <̇> (U+0311) dovrà posizionarsi su <i> o su <W> con coordinate sensibilmente diverse (<ī, W>). Come accennato sopra il discorso si complica quando i diacritici da posizionare sono più di uno. Una soluzione può essere quella di creare un tipo di carattere con cloni di diacritici a diverse coordinate. Ciò però va a danno della codifica standard, che viene irrimediabilmente persa. Inoltre comporta una scelta manuale del clone con la posizione più appropriata per ogni singola lettera.

I tipi di carattere OpenType, invece, se debitamente compilati,⁸ possono supportare il posizionamento automatico “intelligente” del diacritico, sia rispetto ad una lettera, sia rispetto ad un altro diacritico. In questo modo si può inserire un numero illimitato di diacritici che vengono visualizzati correttamente, senza modificare l’interlinea, e che mantengono la propria codifica standard: $e + \overset{\circ}{\circ} + \overset{\circ}{\circ} + \overset{\circ}{\circ} + \overset{\circ}{\circ} + \overset{\circ}{\circ} + \overset{\circ}{\circ} \Rightarrow \overset{\circ}{\overset{\circ}{\overset{\circ}{\overset{\circ}{\overset{\circ}{\overset{\circ}{e}}}}}}$.

Un problema a parte è posto da alcune lettere – <i>, <j> e tutti i simboli derivati, quali, ad esempio, <ı> (U+0268), <ȷ> (U+0249), <ȶ> (U+029D) – che hanno la caratteristica particolare di avere, dal punto di vista sincronico, un punto diacritico esornativo (lo standard unicode attribuisce loro la proprietà di essere *softdotted*). Questo punto deve sparire – tramite sostituzione automatica di glifi senza punto – quando si pongono segni diacritici sovrascritti: $\overset{\circ}{j} + \overset{\circ}{\circ} \Rightarrow \overset{\circ}{j}, \overset{\circ}{\circ}$ + $\overset{\circ}{\circ} \Rightarrow \overset{\circ}{j}$ ⁹.

Il dizionario comparativo delle lingue indoarie di TURNER [1966] utilizza i caratteri latini e sotto ogni lemma sanscrito trascrive le varie voci di lingue e dialetti indoari in un modo che si può definire fonologico. Questo tipo di trascrizione, che in diversi casi si discosta dalla traslitterazione ufficiale, è necessariamente ricco di simboli speciali e segni diacritici. Non è infrequente trovare vocali con tre diacritici. Ad esempio abbiamo in (5027) **chāyā-** una forma kt. *čuwā́*; in (9038) ***phaṭ-1** una forma aúok. *parakáṭi*; in (9090) **phupphusa-** una forma par. *papú*.

Le occlusive “palatali”¹⁰ indoarie sono solitamente traslitterate *c j*, con l’aggiunta di *h* nel caso delle aspirate (per le corrispondenti forme indoiraniche si usano *č / ǰ*). Come visto sopra, <j> ha la proprietà di avere un punto diacritico esornativo. Nella lingua marathi le “palatali” sono realizzate come tali solo davanti a vocale anteriore e nei prestiti, mentre negli altri contesti sono realizzate come affricate dentali [ts dz]. TURNER [1966] trascrive questi

⁷ Caratteri (come quello qui usato) in cui la spaziatura di ogni lettera ha una larghezza propria, occupa un diverso numero di punti. La spaziatura proporzionale è considerata una caratteristica gradevole alla lettura.

⁸ Lo strumento, gratuito, per creare caratteri OpenType con caratteristiche tipografiche avanzate è VOLT (Visual OpenType Layout Tool): <http://www.microsoft.com/typography/developers/volt/>.

⁹ Un’eccezione è costituita da <ı> del turco, con forma maiuscola <İ> (U+0130), che si oppone a <i>/<I>, ma che non è codificata in modo autonomo. Un eventuale trascrizione ortoepica che ponga un diacritico sopra questa lettera dovrà mantenere il mantenere il punto sovrascritto funzionale: <ı̇> (*i* + $\overset{\circ}{\circ}$ + $\overset{\circ}{\circ}$) oppure <ı̇> (*i* + $\overset{\circ}{\circ}$ + $\overset{\circ}{\circ}$ + $\overset{\circ}{\circ}$), secondo lo stile usato per il lituano <ė̇>. Per questa lingua andranno anche inibite eventuali legature <fi>, che oscurerebbero l’opposizione *i ~ ı*.

¹⁰ La realizzazione neutra di questi suoni nella maggior parte delle lingue indoarie (escluso il sindhi) secondo CANEPARI [2003] non è costituita da occlusive palatali [c j], ma da affricate solcate postalveolari senza protrusione labiale (che è invece una caratteristica di [tʃ dʒ]), trascritte [tʃ dʒ].

fonemi ponendo un diacritico su *c* e *j*. Su *c* pone un punto diacritico \emptyset , usando la forma \dot{c} , ad es. mar. *khaċnē* in (3764) ***khacc-**; ma su *j* non può fare altrettanto e ne pone due \emptyset , usando quindi la forma j'' , ad es. mar. *khā□'', khā□''nē* in (3827) **kharju-1**. La coppia $\dot{c} j''$ viene usata da TURNER per trascrivere le affricate dentali anche di altre lingue e dialetti, anche quando non c'è un'origine da *c j*.

5 Problemi di trascrizione: il caso delle velari e palatali indoeuropee

Nell'*Indogermanisches etymologisches Wörterbuch* di POKORNY [1959-69], per quanto riguarda i fonemi "dorsali", accanto a lemmi con $\dot{k} k^v$, sono presenti alcuni lemmi contenenti *q* (***aleq-**, ***areq-**, ***breuq-**, ***bhleiq-**, ***enq-** / **onq-**) e q^v (***auq^v(h)-** / **uq^v(h)-**). A rigore, forse si potrebbe pensare che il redattore ponga in questi casi dei proto-fonemi diversi da * k^h * k^h . GAMKRELIDZE-IVANOV [1995] ad esempio postulano per le serie sorda (aspirata) e glottalizzata, accanto a * k^h * k^h * k^{hw} e * k^h * k^h * k^h , delle occlusive uvulari che indicano con * q^h * q^h . Nel caso del POKORNY tuttavia si ha a che fare solo con delle sviste dovute alla riabolazione di un precedente lavoro. Nel dizionario WALDE-POKORNY [1927-30], realizzato da WALDE ed edito da POKORNY, si ha una trascrizione asimmetrica delle coppie di "dorsali" sorde e sonore: le sorde sono rese con $\dot{k} y q^v$, con due grafemi diversi e con un diacritico utile ma funzionalmente pleonastico in \dot{k} ; le sonore invece con $\dot{g} y g^v$, con un'unica forma \dot{g} .¹¹ Questa asimmetria è stata eliminata nel POKORNY mediante l'uso di $k k^v$ per velari e labiovelari, ma sono rimaste alcune occorrenze residuali.

Il motivo dei due grafemi per le sorde ha origine dall'uso dei Neogrammatici (BRUGMANN, [1906]): venivano usate coppie di segni distinti per le occlusive palatali e velari sorde e sonore: * \dot{k} * \dot{kh} * \dot{g} * \dot{gh} per le palatali; * q * qh * g * gh per le velari (e * q^v * q^vh * g^v * g^vh per le labiovelari). Inoltre anche la nasale omorganica di queste consonanti era trascritta in modo differente: * \dot{n} davanti a palatale (**denk-*), * \dot{v} davanti a velare (e labiovelare, anche con funzione sillabica: **ḡq^v(i)jo-*). Una forma come **d^hég^hōm* era scritta come **dhéghōm*, metre **d^hg^hei-* era **dhq^vhei-*, con \dot{g} , simbolo purtroppo ancora sprovvisto di codifica standard unicode.

La difficoltà a riprodurre il carattere \dot{g} ne ha facilitato l'abbandono, lasciando spazio ad usi decisamente incoerenti, e per nulla episodici, come quello del WALDE-POKORNY prima del definitivo stabilizzarsi di un singolo grafema per tutti i tre ordini di "dorsali" anche per le sorde.

6 Conclusioni

Ottenere una trascrizione efficace e ben leggibile di tutti i segni di cui può avere bisogno la rappresentazione grafica di una lingua è un problema di primaria importanza, che investe

¹¹ Nei testi di quel periodo ovviamente non esisteva un grafema IPA <g> (U+0261) e g non era altro che la forma corsiva di <g> (U+0067, la forma corsiva "con coda a cappio" g non era usata, così come <a> lo è correntemente di <ā> (U+0061). Diversi simboli fonetici hanno origine tipografica da una trasposizione in tondo di forme corsive: ad. es. <ā> (U+0251) <a>; <ū> (U+028B) <u>; <š> (U+0283) <ſ> (U+017F, cfr. nota 1).

in primo luogo la fruibilità di strumenti essenziali come dizionari e grammatiche, tenendo conto delle specificità delle diverse lingue, delle loro tradizioni e delle difficoltà di resa della trascrizione fonetica ed etimologica.

Inoltre, nel caso di pubblicazioni non cartacee ci sono dei requisiti aggiuntivi imprescindibili di standardizzazione e compatibilità a cui attenersi: in questo modo anche la comunicazione fra studiosi potrà fare un passo avanti, dato l'ormai comune uso di mezzi elettronici per scrivere articoli e saggi scientifici.

Si sta finalmente iniziando ad uscire dai *secoli bui* della resa tipografica linguistica. Nessuna *soap opera* sarebbe in grado di eguagliare la serie di matrimoni e divorzi, di alleanze e conflitti che hanno caratterizzato negli ultimi vent'anni la produzione dei tipi di carattere. Sono stati introdotti, ritirati e reintrodotti ripetutamente formati incompatibili, in una vera guerra dei *fonts*. Con la diffusione dello standard unicode e con l'avvento dei caratteri OpenType molti di questi problemi sono ora risolvibili. Sebbene lo standard unicode comprenda la maggior parte degli alfabeti conosciuti, nessun tipo di carattere disponibile sui moderni computer contiene tutti i glifi. Purtroppo, non tutti i programmi di videoscrittura sono in grado di riconoscere tutti i simboli contenuti nei *fonts* o usare le varie caratteristiche avanzate OpenType (anzi, al momento nessuno è in grado di sfruttarle tutte). Questa tecnologia tipografica rappresenta, più che una soluzione già pronta per tutto, un ottimo strumento promettente in via di sviluppo.

Bibliografia

A. Dizionari

- Pokorny, J. (1959-69), *Indogermanisches etymologisches Wörterbuch*, Bern-München, Francke.
Turner, (Sir) R. L. (1966), *A Comparative Dictionary of the Indo-Aryan Languages*, London - New York - Toronto, Oxford University Press.
Turner, (Sir) R. L. (1985), *A Comparative Dictionary of the Indo-Aryan Languages - Addenda and Corrigenda* edited by J. C. Wright, London - New York - Toronto, Oxford University Press.
Walde, A., Pokorny, J. (1927-30), *Vergleichendes Wörterbuch der indogermanischen Sprachen* herausgegeben und bearbeitet von Julius Pokorny, Berlin - Leipzig, Walter de Gruyter & Co.

B. Altre opere

- Brugmann, K. (21906), *Grundriss der vergleichenden Grammatik der indogermanischen Sprachen*. 2. Aufl. (1. Aufl. 1889), Strassburg, Karl J. Trübner.
Canepari, L. (2003), *Manuale di fonetica. Fonetica naturale: articolatoria, uditiva e funzionale*, München, Lincom GmbH.
Daniels, P. T., Bright, W. (1996), *The world's writing systems*, New York - Oxford, Oxford University Press.
Gamkrelidze, T. V., Ivanov (1995), *Indo-European and the Indo-Europeans: a Reconstructon and Historical Analysis of a Proto-Language and a Proto-Culture*, Berlin - New York, Mouton de Gruyter.
The Unicode Consortium (2003), *The Unicode Standard, Version 4.0*, Boston (MA), Addison-Wesley Developers Press.