CÂND MIGRĂM LA DIACRITICELE CORECTE?

BOGDAN STĂNCESCU
18 aprilie 2010, versiunea 1.04[1]

S.C. Moongate Video Production srl, Bucureşti – România

bogdan@moongate.ro

Rezumat

Preconizez că în perioada 2010-2015 va avea loc în cea mai mare parte migrarea conținutului de limbă română de la utilizarea caracterelor cu diacritice cu sedilă (conform ISO-8859-2) la caracterele cu diacritice corecte cu virgulă (conform Unicode 3.0). Acest document încearcă o sinteză a elementelor care influențează, de la caz la caz, momentul ideal de migrare.

1. Introducere

Semnele diacritice din partea de jos a caracterelor românești „ş" și „ţ" sunt virgule. Acest detaliu este de la sine înțeles pentru orice vorbitor nativ de limbă română: este un fapt nedisputat care se învață în clasele primare și nu mai trebuie repetat niciodată în mod explicit. Iar asta într-o asemenea măsură încât însăși Academia Română nu a simțit nevoia să se pronunțe în această privință decât în anul 2003, și chiar și atunci numai pentru că a răspuns unei întrebări explicite în acest sens.[2]

Atunci când a fost creată prima codare a caracterelor pentru Europa de Est, în 1987 (Latin-2[3]), caracterele pentru limba română au fost comasate cu cele pentru alte limbi din această zonă geografică scrise în mod uzual cu grafie latină, precum ceha, maghiara, poloneza și altele. Între limbile asociate acestui standard, limba română este singura care folosește caracterele „ş" și „ţ", *sau orice caractere similare din punct de vedere vizual*.

Caracterul „ş" din limba română („*s cu virgulă*") este foarte similar din punct de vedere vizual cu litera „ş" din limba turcă („*s cu sedilă*"). Diferența grafică dintre cele două semne diacritice este aproape insesizabilă pentru mărimi mici de text (vezi Figura 1).

Standardul Latin-2 nu a fost niciodată asociat limbii turce[4]. Cu toate acestea, caracterele „ş" și „ţ" au fost definite în acest standard, pentru uzul în limba română, drept "*s cu sedilă*" (caracter specific turcesc), respectiv "*t cu sedilă*" (caracter ce nu este folosit în nicio limbă).

---

[1] Cea mai actualizată versiune a acestui document, împreună cu alte resurse conexe, se vor găsi întotdeauna la adresa http://www.moongate.ro/products/diacritice/

[2] Vezi http://www.secarica.ro/html/s-uri_si_t-uri.html. Ce-i drept, semnele respective au fost numite sedile încă de la Titu Maiorescu, iar uneori chiar au fost folosite ca atare (http://decatorevista.ro/DecatORevista.pdf, p. 78)

[3] Standardul ISO/IEC 8859-2, cunoscut și ca Latin-2

[4] Caracterele pentru limba turcă au fost incluse inițial în standardul ISO/IEC 8859-3, numit și Latin-3, standard creat pentru Europa de Sud; câțiva ani mai târziu limbii turce i-a fost asociat standardul ISO/IEC 8859-9, numit și Latin-5, standard creat în mod specific pentru limba turcă.

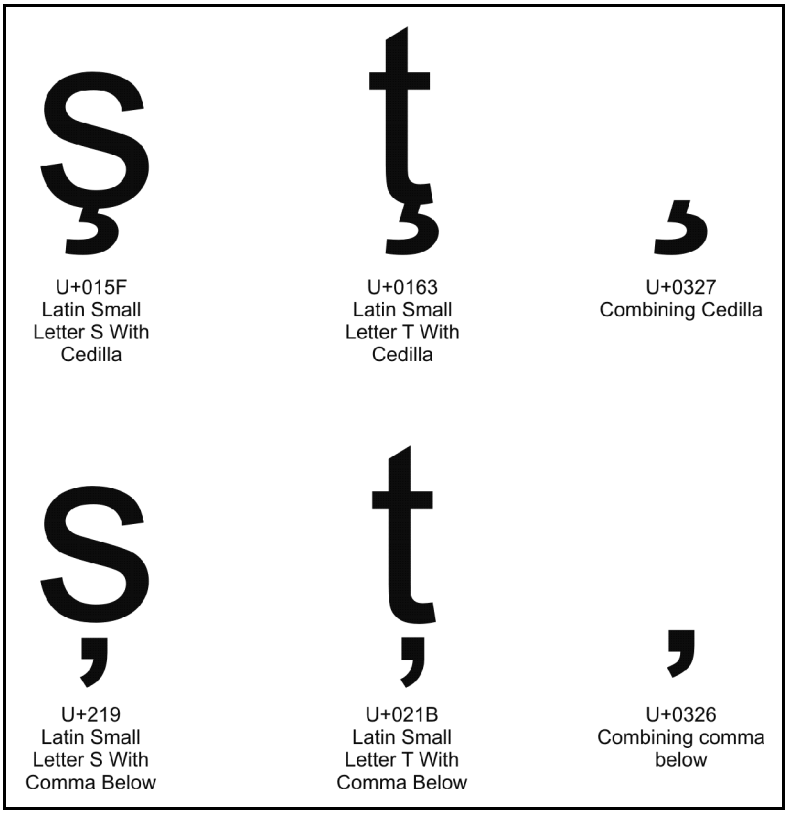

Figura 1: Caracterele în cauză, mărite pentru vizibilitatea semnelor diacritice.
Primul rând conține caracterele „*s cu sedilă*”, „*t cu sedilă*” și semnul diacritic sedilă.
Al doilea rând conține caracterele „*s cu virgulă*”, „*t cu virgulă*” și semnul diacritic virgulă.

În anul 1997 compania Apple a schimbat caracterele din standardul proprietar MacOS Romanian în așa fel încât să utilizeze semnele diacritice corecte pentru „ş” și „ţ”[5]. În același an Asociația de Standardizare din România a protestat pe lângă ISO în privința standardului Latin-2, însă singura modificare făcută un an mai târziu de către ISO a fost adăugarea unei note care permitea interpretarea semnelor respective drept „*s cu virgulă*”, respectiv „*t cu virgulă*” numai în măsura în care expeditorul și destinatarul mesajului se puneau de acord în această privință.[6]

În anul 1999 a apărut prima versiune a standardului Unicode care să conțină caracterele corecte românești „*s cu virgulă*” și „*t cu virgulă*”. Dificultățile de adoptare a standardului Unicode în formă completă au cauzat însă întârzieri de mai bine de zece ani în adoptarea sa pe scară largă.

Așa se face că timp de douăzeci de ani aproape toate textele scrise în limba română în medii informatice au fost scrise fie fără semne diacritice, fie cu semne diacritice greșite. Abia sistemul de operare Windows Vista, apărut la sfârșitul anului 2006, a fost primul sistem de operare utilizat pe scară largă de consumatorii de conținut care să folosească în mod nativ caracterele corecte pentru limba română. Chiar și așa, penetrarea lentă a acestui sistem de operare și a celor ulterioare face ca migrarea către utilizarea în practică

[5] `http://unicode.org/Public/MAPPINGS/VENDORS/APPLE/ROMANIAN.TXT`
[6] “*Note - Subject to the agreement of originator and receiver, in information interchange the letters S and T WITH CEDILLA BELOW may be used to substitute for the letters S and T WITH COMMA BELOW*”, `http://www.secarica.ro/html/s-uri_si_t-uri.html`

a semnelor diacritice corecte să fie încă și astăzi, în 2010, mai mult un subiect de discuție sporadică decât obiectul unor acțiuni concrete.

2. Contextul de aplicabilitate al acestei lucrări

În această lucrare voi utiliza termenul de *bază de date* în sens cât se poate de abstract și de cuprinzător: orice colecție de texte în limba română stocate electronic, indiferent de formatul concret de stocare sau de modul de prezentare. De la jurnal, revistă, carte sau enciclopedie tipărită până la mesaje e-mail, site-uri Internet sau etichete de text din cadrul aplicațiilor software, toate vor migra mai devreme sau mai târziu de la caractere cu sedile la caractere cu virgule.[7]

În ceea ce privește *factorii abstracți* care influențează alegerea momentului optim de migrare, am căutat să identific o structură suficient de generică încât să fie aplicabilă oricărei situații practice, în contextul definiției cuprinzătoare din paragraful anterior.

Pe de altă parte, *datele statistice concrete* prezentate în această lucrare sunt specifice numai bazelor de date care satisfac simultan următoarele criterii independente:

1. sunt consultate prin intermediul unui navigator de Internet (*web browser*);
2. sunt consultate de consumatori foarte eterogeni în privința platformei software.

Dacă măcar unul dintre criteriile de deasupra nu se aplică bazei de date pe care o gestionați, atunci trebuie să *ignorați complet toate datele statistice concrete* din această versiune a acestui document.[8] În acest caz va trebui fie să utilizați resursele indicate la sfârșitul acestui document (dacă se aplică), fie să căutați și să adaptați datele statistice concrete asociate situației dumneavoastră la structura prezentată aici.

În privința dimensiunii temporale a deciziei de migrare am decis să nu includ niciun fel de date concrete, întrucât nivelul estimat de eroare al oricărei predicții de această natură ar fi prea mare pentru orice scop practic. Am ales în schimb să actualizez acest document și resursele conexe pe măsură ce evoluează situația (vezi nota prima notă de subsol sau ultimul paragraf din această lucrare).

3. Alegerea momentului: de ce este important

După cum am văzut mai sus, diferența grafică dintre cele două variante de caractere este în cea mai mare parte a timpului nesemnificativă.[9] Din acest motiv *nu există nicio presiune naturală considerabilă pentru adoptarea semnelor corecte* – practic toți consumatorii de conținut pot interpreta corect semnele diacritice „vechi”, iar majoritatea acestora nu sunt oricum la curent cu această problemă grafică minoră. Chiar și într-un sens mai profund chestiunea este la fel de neimportantă: *într-un text scris în limba română distincția dintre cele două tipuri de semne diacritice nu are valoare semantică*,

[7] Eu personal am întâlnit această problemă în contextul discuțiilor de la Wikipedia în limba română; acolo mi-am și format și sintetizat în mare parte argumentele expuse în această lucrare, interacționând cu comunitatea de voluntari din cadrul proiectului respectiv.

[8] De exemplu dacă (1) este vorba despre o aplicație client-side, (2) este o aplicație online dedicată clienților care folosesc terminale mobile, (2) este o aplicație (online sau offline) care rulează numai pe o platformă anume, sau (1+2) este o aplicație client-side pentru terminale mobile.

[9] Există totuși situații în care diferența este ușor de sesizat chiar și pentru un consumator neavizat, mai ales atunci când se folosesc mărimi mari de literă (titlul unei cărți pe copertă, titlurile de pe afișe sau materiale publicitare etc.)

deoarece utilizarea uneia dintre variante în defavoarea celeilalte nu aduce niciun plus de informație, indiferent de felul în care este interpretat textul.[10]

Prin urmare avem de-a face cu o *problemă semnificativă de interoperabilitate cauzată de rezolvarea unei probleme minore de prezentare*. Situația pare absurdă, însă faptul că nu există (și nu poate exista) nicio soluție alternativă pentru problema de prezentare legitimează problema de interoperabilitate.

În sistemele de operare ale companiei Microsoft anterioare Windows Vista caracterele cu diacritice corecte sunt vizibile numai în Windows XP, și asta numai în anumite condiții.[11] Datorită cotei de piață uriașe a sistemelor de operare ale companiei Microsoft în rândul consumatorilor de conținut, *aceste considerente fac migrarea către caracterele cu diacritice corecte o chestiune discutabilă în absența penetrării masive a sistemelor de operare Windows Vista sau mai noi* pe piață.

Pe de altă parte, unii factori de decizie ai diverselor baze de date de limbă română vor fi în mod inevitabil *early adopters*[12] ai noilor caractere cu diacritice corecte. Pe măsură ce trece timpul, pe măsură ce penetrează Windows Vista și sisteme de operare mai recente și pe măsură ce diverse baze de date migrează la noile diacritice, va crește masa de conținut și de consumatori de conținut axați pe noile caractere. Odată ce se atinge o masă critică, *acei creatori de conținut care vor mai oferi text cu diacriticele incorecte vor fi văzuți ca depășiți*. Dacă în prezent există motive absolut justificate pentru a amâna migrarea către diacriticele corecte[13], odată ce se atinge masa critică (în special în ceea ce îi privește pe consumatorii de conținut), *nu va mai exista nicio scuză pentru întârzierea migrării*: în ultimă instanță, diacriticele noi sunt cele corecte, iar cele vechi sunt pur și simplu incorecte în limba română!

Totuși vom vedea mai jos că independent de felul în care sunt văzuți din afară sau de corectitudinea tehnică a diacriticelor folosită în bazele lor de date, unii dintre creatorii de conținut de limbă română vor avea *motive întemeiate pentru a adopta noile diacritice înaintea celorlalți*, iar alții vor avea *motive întemeiate pentru a întârzia migrarea pentru o perioadă semnificativă* chiar după momentul apariției masei critice despre care am vorbit mai sus. Scopul acestui document este tocmai acela de a identifica factorii care duc la aceste decizii, de la caz la caz.

4. Factori de influență

După cum am văzut în secțiunea 3., două forțe opuse acționează asupra deciziei de migrare la diacriticele corecte:

- *Pentru migrare cât mai rapidă*: tehnologia este deja disponibilă, conținutul ar putea fi deja vizualizat de majoritatea consumatorilor (dar vezi secțiunile următoare), iar rezultatul ar fi utilizarea caracterelor corecte în limba română. În

[10] Mai puțin cazul în care un text scris în română conține citate sau nume turcești. Chestiunea este discutată mai pe larg în secțiunea 4.3. dedicată bazelor de date.
[11] Chestiunea este analizată pe larg în secțiunea 4.1. dedicată consumatorilor de conținut.
[12] Persoane (fizice sau juridice) care doresc să adopte cât mai repede tehnologiile cele mai recente.
[13] Pentru brevitate voi folosi în continuare sintagmele „diacritice corecte” pentru „caractere care folosesc semnele diacritice corecte” (virgule), respectiv „diacritice vechi” pentru „caractere care folosesc semnele diacritice vechi” (sedile).

plus, imaginea ultimilor creatori de conținut care să migreze va avea probabil de suferit într-o oarecare măsură.

- *Pentru amânarea migrării*: diverse probleme de lizibilitate și interoperabilitate, dintre care unele foarte semnificative (vezi secțiunile următoare).

Aceste două forțe vor avea o evoluție dinamică de-a lungul timpului, în sensul că prima va crește în defavoarea celei de-a doua, până la eliminarea completă a acesteia din urmă.

După o analiză îndelungată a structurii diverșilor factori care influențează aceste două forțe contrare, am identificat trei piloni pe care se sprijină întregul raționament:

- *Consumatorii de conținut:* cititorii, utilizatorii produselor software etc.
- *Creatorii de conținut*: edituri, deținători de site-uri, producători de software etc.
- *Bazele de date*: conținutul efectiv al revistelor, site-urilor, aplicațiilor etc.

Voi analiza felul în care fiecare dintre acești trei piloni afectează fiecare dintre cele două forțe și în ce fel.

4.1. Consumatorii de conținut

Este de la sine înțeles că cel mai important dintre cei trei piloni este cel reprezentat de consumatorii de conținut: indiferent de capabilitățile tehnice ale creatorilor de conținut și de bazele lor de date, orice demers este inutil în măsura în care conținutul nu poate fi consumat sau, în cazul aplicațiilor interactive, consumatorul nu poate interacționa cu interfața în așa fel încât să se obțină acces la conținutul propriu-zis.

Prima întrebare în ceea ce-l privește pe consumatorul de conținut este legată de modalitatea prin care acesta consumă în mod concret conținutul. *Dacă mediul final de consum nu implică tehnologii aflate sub controlul consumatorului, atunci consumatorul nu este un factor semnificativ* în luarea deciziei de migrare. În această situație se află conținutul prezentat exclusiv pe medii tipărite, sau folosind orice alte medii în care consumatorul are un rol pasiv din punctul de vedere al tehnologiilor implicate (cărți, jurnale, reviste, afișe, prezentări video și așa mai departe). În plus, conținutul consumat în condiții controlate de creatorii de conținut beneficiază în mare măsură de aceleași derogări (de exemplu aplicații care rulează în mod kiosk[14], aplicații online sau client-side care rulează într-un mediu proprietar, conținut prezentat pe hardware dedicat precum cititoarele de cărți electronice).

Dacă însă mediul de consum este dependent de tehnologii controlate de consumator, atunci devin relevante două subcategorii de factori care influențează capacitatea consumatorului de a utiliza conținutul:[15]

1. *Lizibilitatea* – pot consuma conținutul?[16]
2. *Interactivitatea* – pot interacționa cu interfața?

[14] Terminale dedicate, instalate în locuri publice, așa cum sunt cele din aeroporturi, gări, puncte de interes turistic etc.
[15] Cel mai reprezentativ exemplu în acest sens sunt bazele de date ale site-urilor și aplicațiilor Internet/intranet. În aceeași situație se află însă orice bază de date (în sensul larg utilizat în această lucrare) care poate fi interpretată pe mai multe platforme software – aplicații, documente distribuite, mesaje e-mail (e.g. newsletters) și așa mai departe.
[16] Includ în această subcategorie și problemele de accesibilitate pentru persoanele cu dizabilități.

Pentru a cântări capacitatea de lizibilitate a consumatorilor în contextul diacriticelor corecte trebuie analizat nivelul de utilizare al platformelor care suportă diacriticele corecte *în contextul consumatorilor bazei de date* în speță. Acest aspect este foarte important, deoarece aplicațiile care rulează pe o platformă anume depind de capacitatea utilizatorilor acelei platforme specifice de a citi text care conține diacriticele corecte.

În martie 2010, nivelul global de utilizare al platformelor software era următorul:[17]

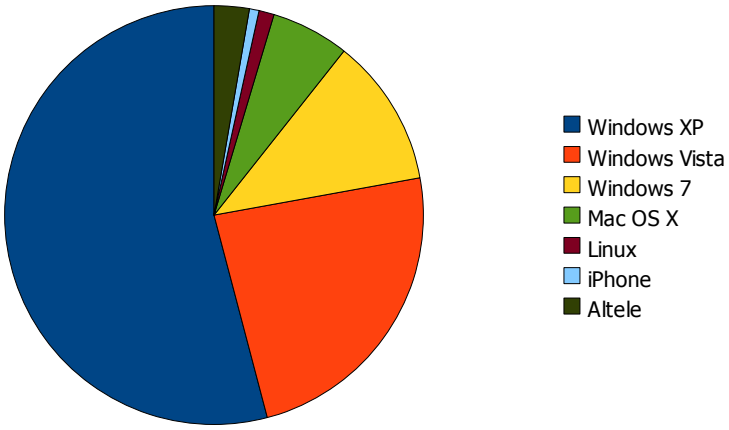

Figura 2: Gradul de utilizare al diverselor sistemelor de operare

În privința capacității de a citi conținut scris cu diacriticele noi, următoarele platforme suportă în mod nativ diacriticele corecte: Windows Vista, Windows 7, Mac OS X și iPhone. Platformele Linux și cele cu cotă minoră de piață sunt incerte.[18]

În privința Windows XP, care deocamdată rămâne lider detașat (nu numai că este cea mai utilizată platformă, dar este mai utilizată decât toate celelalte la un loc), situația este la fel de incertă, însă merită investigată. La instalare, Windows XP nu suportă deloc diacriticele corecte – pur și simplu pe ecran apar niște „pătrățele" în locul caracterelor respective (ultima coloană din Tabelul 3). Există două modalități principale de a obține compatibilitate cu diacriticele corecte în Windows XP:[19]

- Prin instalarea explicită a unui pachet software suplimentar de la Microsoft.[20]
- Prin instalarea Internet Explorer 7 sau Internet Explorer 8, aplicații care sunt în mod normal actualizate automat de către Windows.

Pentru prima opțiune nu avem la dispoziție statistici, însă este destul de puțin probabil că un consumator oarecare de conținut a făcut efortul să descarce și să instaleze un astfel de pachet software. Pe de altă parte actualizarea navigatorului Internet Explorer este una automată (și una dezirabilă, independent de diacritice), deci ne așteptăm că o parte semnificativă a consumatorilor au instalat această actualizare.

[17] Datele sunt extrase din Tabelul 2.

[18] Practic toate versiunile curente de Linux suportă în mod nativ diacriticele corecte. Însă nivelul scăzut de utilizare al sistemelor de operare Linux între consumatorii de conținut, coroborat cu numărul mare de combinații de distribuții și versiuni de Linux fac orice statistică din această categorie să fie nepractică. Totuși vezi și nota 22.

[19] `http://www.microsoft.com/Romania/Diacritice.aspx` – în realitate atât Internet Explorer 8 cât și Internet Explorer 7 sunt capabile să opereze substituția de corp de literă pentru a obține rezultatele de pe coloana a doua din Tabelul 3.

[20] „Actualizare de fonturi corespunzătoare extinderii Uniunii Europene" de la `http://www.microsoft.com/downloads/details.aspx?FamilyID=0ec6f335-c3de-44c5-a13d-a1e7cea5ddea&DisplayLang=ro`

Într-adevăr, cota de piață acestor versiuni de Internet Explorer este semnificativă:[21]

Versiune Internet Explorer	**Cota de piață** (relativă, între utilizatorii de IE)
Internet Explorer 8	42,59%
Internet Explorer 7	22,91%
Internet Explorer 6	34,39%
Altele	0,11%

Tabelul 1: Cota de piață a diverselor versiuni de Internet Explorer

Totuşi asta înseamnă că mai bine de o treime dintre utilizatorii de Internet Explorer nu pot vedea text scris cu diacriticele corecte, atâta timp cât utilizează Windows XP.[22] Nu avem la dispoziție date statistice globale de încredere care să coroboreze sistemul de operare cu navigatorul utilizat[23], deci vom fi nevoiți să presupunem că toți utilizatorii de Windows XP se supun aceleiași proporții identificate pentru Internet Explorer, indiferent dacă utilizează acest navigator sau altul.

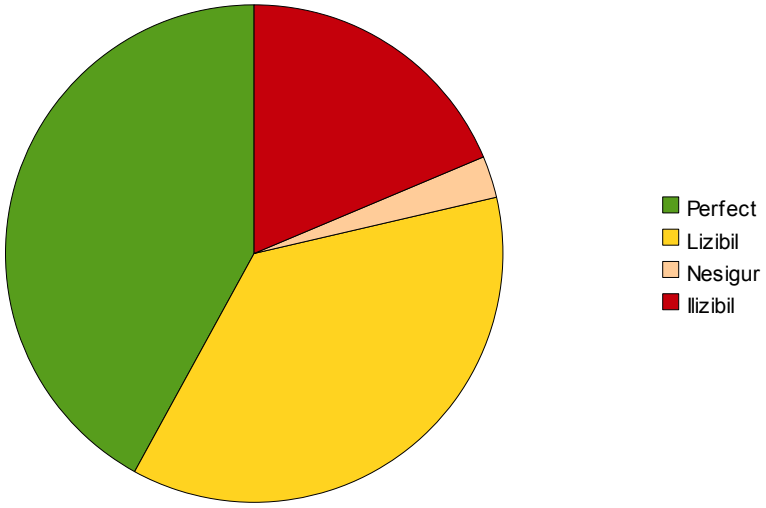

Figura 3: Capacitatea sistemelor de operare de a afișa diacriticele corecte

[21] `http://en.wikipedia.org/wiki/Internet_Explorer#Market_adoption_and_usage_share` [en], date estimate pentru februarie 2010

[22] Mulți (dar nu toți) dintre utilizatorii altor navigatoare, precum Mozilla Firefox, Google Chrome sau Opera, beneficiază de funcționalitatea de substituție a corpului de literă pe care o implementează IE începând cu versiunea 7. În statisticile din acest document am făcut presupunerea oarecum optimistă că numai Internet Explorer 6 executat în Windows XP este incapabil să afișeze diacriticele noi. Am estimat că această aproximare optimistă compensează aproximarea pesimistă pentru sistemele de operare „nesigure”.

[23] De fapt statistica ideală ar fi chiar cea pe care încerc să o estimez aici (lizibilitate perfectă/lizibilitate limitată/ilizibilitate); în lipsa ei, cea mai bună aproximare ar fi o coroborare a lizibilității limitate (gradul de instalare al EUupdate.EXE sau IE7 sau IE8) cu sistemul de operare Windows XP. Totuși procentele sunt deocamdată suficient de mari în toate categoriile încât aproximările din text să nu afecteze în mod semnificativ calitatea analizei.

Datele utilizate în graficele de mai sus sunt următoarele:[24]

Sistem de operare	**Cota de piață**	**Afișează**	**Scrie**[25]
Windows XP[26]	52,94%	Lizibil (65,5%)	Simplu (10%)
		Ilizibil (34,5%)	Dificil (90%)
Windows Vista	23,25%	Perfect	Simplu
Windows 7	11,24%	Perfect	Simplu
Mac OS X	5,90%	Perfect	Simplu
Linux	1,14%	Nesigur	Dificil
iPhone	0,73%	Perfect	Simplu
Altele	2,65%	Nesigur	Dificil

Tabelul 2: Datele utilizate pentru generarea graficelor

	Windows 7, Vista	Windows XP (nou)	Windows XP (vechi)
Diacritice corecte	arșiță	arșiță	ar□i□ă
Diacritice vechi	arşiţă	arşiță	arşiță

Tabelul 3: Afișarea celor două tipuri de diacritice în funcție de gradul de lizibilitate al diacriticelor noi (de notat că diacriticele vechi sunt perfect lizibile indiferent de context)

Deja în 2010 *majoritatea consumatorilor de conținut pot să citească text care folosește diacriticele corecte*. Există totuși câteva rezerve întru totul semnificative:

- Utilizatorii pentru care textul este doar lizibil (nu perfect) pot citi textul, însă, în funcție de situație, *vor observa o diferență sesizabilă, inestetică și deranjantă de afișare a caracterelor respective* (a doua coloană din Tabelul 3). Suportul pentru diacriticele corecte se limitează în Windows XP la numai câteva corpuri de literă – pentru celelalte, sistemul de operare substituie pur și simplu caracterele respective cu aceleași caractere din cele mai apropiate corpuri de literă pentru care dispune de caracterele în speță. Rezultatul este cel așteptat: textul este lizibil, însă în funcție de diferența vizuală dintre corpul de literă utilizat în text și cel disponibil rezultatele pot fi inestetice (în figură am folosit Arial Black).
- Chiar în cazul bazelor de date accesibile via Internet este posibil ca unele aplicații specifice să se adreseze în mod particular consumatorilor care utilizează o paletă relativ îngustă de sisteme de operare. Un exemplu relevant sunt aplicațiile sau subdomeniile dedicate pentru platforme mobile.[27] Dacă în cazul general contează în mod covârșitor cele peste 96 de procente ale sistemelor de

[24] `http://en.wikipedia.org/wiki/Usage_share_of_operating_systems` [en] Procentele adunate nu dau 100% din cauză că la Wikipedia procentele respective sunt mediana valorilor agregate din mai multe surse; am preferat să le păstrăm ca atare pentru a permite verificarea sursei. Indiferent de asta, proporțiile din grafice sunt corecte.

[25] Statistică relevantă în special pentru secțiunea 4.2. legată de creatorii de conținut.

[26] Aproximări semnificative în ambele subcategorii. Fracțiile din subcategorii sunt procente din cota Windows XP.

[27] Multe site-uri publice oferă alternative pentru platforme mobile. De pildă un consumator care accesează pentru prima dată domeniul `http://www.moongate.ro/` folosind un dispozitiv mobil este redirecționat automat către `http://m.moongate.ro/`; în cazul acestui al doilea domeniu sunt relevante numai statisticile legate de platformele mobile.

operare desktop, în cazul site-urilor orientate către dispozitive mobile trebuie analizate numai capabilitățile celor mai puțin de 4% dintre platformele analizate mai sus, reprezentând platformele software utilizate pentru consultarea acelor baze de date (ultimele două rânduri din Tabelul 2).

- Toate datele statistice utilizate aici sunt cele globale. Este posibil ca proporțiile specifice publicului românesc să difere semnificativ, iar dacă diferă atunci diferența este aproape sigur în defavoarea lizibilității perfecte.[28]

Acestea fiind spuse, trebuie menționat în mod proeminent că indiferent de platforma specifică a consumatorilor, indiferent de publicul țintă al bazei de date și indiferent de numărul lor, *fracțiunea consumatorilor care nu pot vizualiza diacriticele corecte se vor afla practic în imposibilitate de a consuma conținutul pe care îl oferiți*. Iar alternativa este afișarea aceluiași text, utilizând caractere cu semne diacritice tehnic incorecte, dar care sunt aproape identice din punct de vedere vizual și pe care le poate citi oricine (vezi al doilea rând din Tabelul 3). În ultimă instanță trebuie să *puneți în balanță corectitudinea academică față de pierderea de facto a unei fracțiuni a cititorilor*.

Celălalt factor care trebuie luat în considerare este capacitatea consumatorilor de a interacționa cu interfața bazei de date. Cea mai proeminentă funcție a interfeței în această privință este funcționalitatea de căutare: dacă baza de date conține diacritice corecte iar consumatorul operează o căutare utilizând diacriticele vechi (sau viceversa)[29] atunci consumatorul nu va obține rezultatele dorite. *Practic toate problemele de interacțiune pot fi rezolvate prin software*, însă acestea trebuie luate în considerare și rezolvate din timp.[30] Totuși *problemele de interacțiune sunt printre puținele care pot fi rezolvate încă înainte de începerea migrării și* ***ar trebui rezolvate cât mai curând***.[31]

4.2. Creatorii de conținut

Dacă în privința consumatorilor de conținut ați determinat că factorii ce influențează decizia de adoptare a diacriticelor corecte au ajuns să avantajeze migrarea, trebuie acum să vă întrebați în ce măsură puteți crea conținut care să utilizeze noile diacritice. *Este prea puțin important dacă cititorii pot citi, atâta vreme cât scriitorii nu pot scrie.*

Prin urmare factorii care influențează creatorii de conținut sunt dictați în primul rând de o logică similară celei legate de consumatori:

> Poate autoritatea sub egida căreia se generează conținut să controleze mijloacele tehnice ale creatorilor individuali de conținut?

[28] Estimăm că aceasta este situația în prezent când există relativ puține baze de date publice care să exploateze diacriticele corecte. Credem că pe măsură ce ne apropiem de masa critică (și cu atât mai mult după aceea) diferența dintre consumatorii români și media globală va înclina în favoarea compatibilității consumatorilor de limbă română.

[29] Vezi și ultima coloană din Tabelul 2, reprezentată grafic în Figura 3.

[30] Aproape orice problemă de acest fel poate fi rezolvată prin interpretarea flexibilă a datelor de intrare, așa cum procedează Google sau DEX online.

[31] O bază de date proeminentă de limbă română care a migrat foarte curând la diacriticele corecte este DEX online (`http://dexonline.ro/`). Au fost însă luate în calcul aproape toate problemele identificate în acest document: consumatorii care nu pot citi diacriticele corecte pot alege să consulte dicționarul folosind diacriticele vechi, iar interacțiunile cu baza de date ignoră în mod implicit semnele diacritice din textul de intrare. Singura scăpare este legată de limba turcă, în sensul că și cuvintele care ar trebui scrise cu sedilă au fost convertite automat la forma cu virgulă (e.g. `http://dexonline.ro/definitie/siret` versus `http://tr.wiktionary.org/wiki/%C5%9Ferit`). Totuși absența unei necesități de interoperabilitate ulterioară internă dintre dicționar și orice altă bază de date face ca această scăpare să fie lipsită de orice efecte adverse concrete.

Răspunsurile posibile la această întrebare sunt cu mult mai variate decât în cazul consumatorilor. Am ales să identific numai cazurile extreme aici, pentru exemplificare:

- *Redactori tradiționali* (jurnal, revistă, carte tipărită șamd): acești redactori lucrează la sediul societății care i-a angajat, iar societatea respectivă are control complet asupra platformei software utilizate la sediu. Chiar și redactorii care preferă să folosească mijloace tehnice proprii pot fi constrânși să urmeze standardele societății (e.g. atunci când redactează text utilizând laptopul sau computerul propriu). În acest caz creatorii de conținut sunt un factor neglijabil, deoarece autoritatea decizională va opera migrarea pe baza celorlalți doi piloni.
- *Redactori voluntari, independenți* (de exemplu redactori la Wikipedia sau alte proiecte similare, persoanele care contribuie cu comentarii în diverse site-uri, site-uri de socializare șamd): acesta este unul dintre cei mai puternici factori pentru amânarea adoptării diacriticelor corecte (vezi dedesubt).

Statisticile legate de capacitatea creatorilor independenți de conținut de a genera text folosind diacriticele corecte arată astfel:[32]

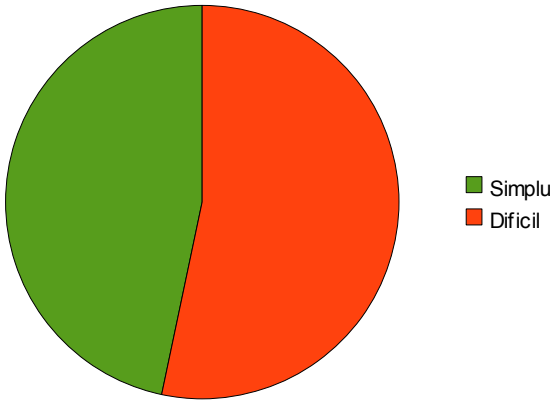

Figura 4: În ce măsură se poate scrie text cu diacriticele corecte

Așadar *mai bine de jumătate dintre creatorii independenți de conținut au dificultăți în a scrie text utilizând diacriticele corecte*. Motivul central pentru această limitare este faptul că utilizatorii de Windows XP, deocamdată majoritari în statisticile globale, nu pot genera conținut utilizând diacriticele corecte decât în condiții destul de stricte.[33] În aceste condiții decizia trebuie luată în concordanță cu capacitatea dumneavoastră de a influența mijloacele tehnice utilizate de creatorii individuali de conținut. În cazul proiectelor bazate pe voluntariat, așa cum este Wikipedia, este evident că migrarea în condițiile actuale nu ar avea sorți de izbândă, chiar lăsând la o parte ceilalți piloni.

4.3. Baza de date

Însăși baza de date, cel mai puțin important dintre cei trei piloni identificați aici, va fi pentru mulți creatori de conținut impedimentul major în decizia de migrare, chiar și atunci când migrarea ar fi dezirabilă din celelalte puncte de vedere. Orice analiză a deciziei de migrare către diacriticele corecte este în mod firesc axată în primul rând pe utilizabilitate din punctul de vedere al consumatorului de conținut, în al doilea rând pe

[32] Sunt reprezentate datele de pe ultima coloană din Tabelul 2
[33] http://www.stefamedia.ro/diacritice-romanesti-corecte-in-windows-xp/

capacitatea creatorului de conținut de a genera conținut și abia în ultimul rând pe *disponibilitatea creatorului de conținut de a migra în mod retroactiv conținutul existent la diacriticele corecte*.

Baza de date influențează decizia de migrare în funcție de următorii factori:

1. *Relevanță*: este relevant să vă puneți întrebări legate de migrarea bazei de date numai în măsura în care aceasta este vizibilă. De exemplu baza de date istorică a unui periodic publicat exclusiv în formă tipărită nu face în general obiectul migrării. Evident, dacă există o parte accesibilă în format electronic și una stocată intern atunci numai partea accesibilă este relevantă.
2. *Mărime*: semnificația acestui factor trebuie coroborată în general cu următorul, complexitatea bazei de date. Există însă un caz particular în care contează exclusiv mărimea: atunci când ea este nulă. Dacă începeți în perioada următoare lucrul la o bază de date complet nouă atunci cântăriți cu atenție opțiunea de a utiliza de la bun început cu diacriticele corecte – în măsura în care aceasta este o opțiune acceptabilă din celelalte puncte de vedere atunci adoptați-o, fiindcă veți evita mai târziu costisitorul proces de migrare retroactivă.
3. *Complexitate*: am spus mai sus că distincția dintre caracterele cu virgulă și cele cu sedilă nu are valoare semantică în limba română. Altfel spus, un text scris în exclusivitate în limba română ar putea fi convertit cu ușurință la diacriticele corecte printr-o simplă operațiune de căutare și înlocuire automată. În general acest lucru este adevărat, însă cu unele rezerve pe care le vom analiza dedesubt.

Complexitatea bazei de date este deci o măsură a frecvenței situațiilor care necesită intervenție umană. În cazul bazei de date de la Wikipedia am identificat următoarele tipologii specifice de situații problematice:

- *Texte scrise în altă limbă* (cel mai notabil în turcă) fără notație adecvată.[34] Dacă textele (inclusiv numele de persoane, locuri, evenimente ș.a.m.d.) în turcă sunt marcate explicit ca fiind scrise în turcă există posibilitatea de a automatiza procesul prin evitarea schimbării semnelor diacritice în cazul acestora.
- *Identificatori de resurse care conțin semne diacritice*. Pe lângă cazul general (URI) mai pot exista o sumedenie de identificatori interni sau externi a căror integritate structurală trebuie menținută de-a lungul procesului de migrare, în ambele sensuri.[35]
- *Interoperabilitatea cu alte baze de date*, în special în condițiile unei migrări parțiale.[36]

[34] De exemplu în HTML: `http://www.w3.org/TR/WCAG10-HTML-TECHS/#language` [en]

[35] Cele mai la îndemână exemple sunt legate de baza de date de la Wikipedia în limba română. Printre identificatorii interni de resurse se numără legăturile interne între articole și cele care leagă articole despre același subiect în mai multe limbi. Identificatorii externi conținuți în Wikipedia sunt legăturile (URI) către pagini de pe alte site-uri și care pot conține caractere cu diacritice. Identificatorii externi pe care trebuie să-i gestioneze Wikipedia sunt legăturile (URI) dinspre alte site-uri către articolele din Wikipedia care pot conține caractere cu diacritice (e drept, aceasta este mai degrabă o resposabilitate morală, în condițiile numărului relativ mare de documente care fac trimitere la enciclopedie).

[36] De exemplu dacă baza de date are relevanță parțială în privința diacriticelor, dar partea publică a bazei de date (cea care urmează să fie migrată) interacționează prin sisteme automate cu partea istorică/privată/confidențială care nu este migrată.

5. Concluzii

Toate bazele de date care conțin text în limba română vor urma în mod inevitabil standardul corect în privința semnelor diacritice. Singurul punct delicat este alegerea momentului optim pentru migrare. Diferența dintre cele două variante este foarte mică din punct de vedere vizual, dar există dificultăți tehnice potențiale semnificative asociate migrării. Am încercat să identific aici factorii care influențează alegerea pragmatică a momentului optim de migrare pe baza a trei piloni: consumatorii de conținut, creatorii de conținut și caracteristicile bazei de date.

Ultima versiune a acestui document, precum și alte resurse suplimentare în această materie se găsesc la adresa `http://www.moongate.ro/products/diacritice/`

Mulțumiri domnului Cristian Secară pentru informațiile concrete care m-au ajutat să corectez unele statistici care altfel ar fi fost cu siguranță greșite; tatălui meu și prietenilor care m-au ajutat să dau forma curentă acestei lucrări (știți voi cine sunteți); colectivității Wikipedia în limba română pentru discuțiile întotdeauna deschise; pentru inspirația acestui material mulțumesc în particular redactorilor Cezarika1 (primul care a ridicat această problemă și ne-a arătat de ce merită discutată), Strainu (un early adopter prin definiție, foarte capabil pe partea tehnică) și nu în ultimul rând lui AdiJapan (ca întotdeauna, a reușit să creeze un echilibru între interlocutorii mai tehnici și cei mai puțin tehnici, între cei avangardiști și cei conservatori).