FACTORII CARE INFLUENȚEAZĂ MOMENTUL OPTIM DE MIGRARE LA DIACRITICELE CORECTE ÎN LIMBA ROMÂNĂ

BOGDAN STĂNCESCU
9 mai 2010, versiunea 2.0[1]

S.C. Moongate Video Production srl, București – 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 factorilor care influențează, de la caz la caz, momentul ideal de migrare.

1. Introducere

Semnele diacritice din partea de jos a caracterelor românești „ș” și „ț” 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 „ș” și „ț”, *sau orice caractere similare din punct de vedere vizual*.

Caracterul „ș” 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[4] (vezi Figura 1).

Standardul Latin-2 nu a fost niciodată asociat limbii turce[5]. Cu toate acestea, caracterele „ș” și „ț” au fost definite în acest standard, pentru limba română, drept “*s cu sedilă*” (caracterul turcesc), respectiv “*t cu sedilă*” (caracter practic nefolosit în nicio limbă).[6]

---

[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.

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

[4] Vezi și nota 12.

[5] Caracterele pentru limba turcă au fost incluse inițial în standardul ISO/IEC 8859-3 (Latin-3), creat pentru Europa de Sud; câțiva ani mai târziu a fost creat un standard special pentru limba turcă, ISO/IEC 8859-9 (Latin-5).

[6] Această discrepanță dintre uzanța autohtonă și standardele ISO a fost probabil cauzată de faptul că înșeși documentele românești care descriau semnele diacritice respective le-au numit sedile timp de aproape două secole – vezi http://www.capisci.ro/articole/Sedile_%C3%AEn_rom%C3%A2n%C4%83

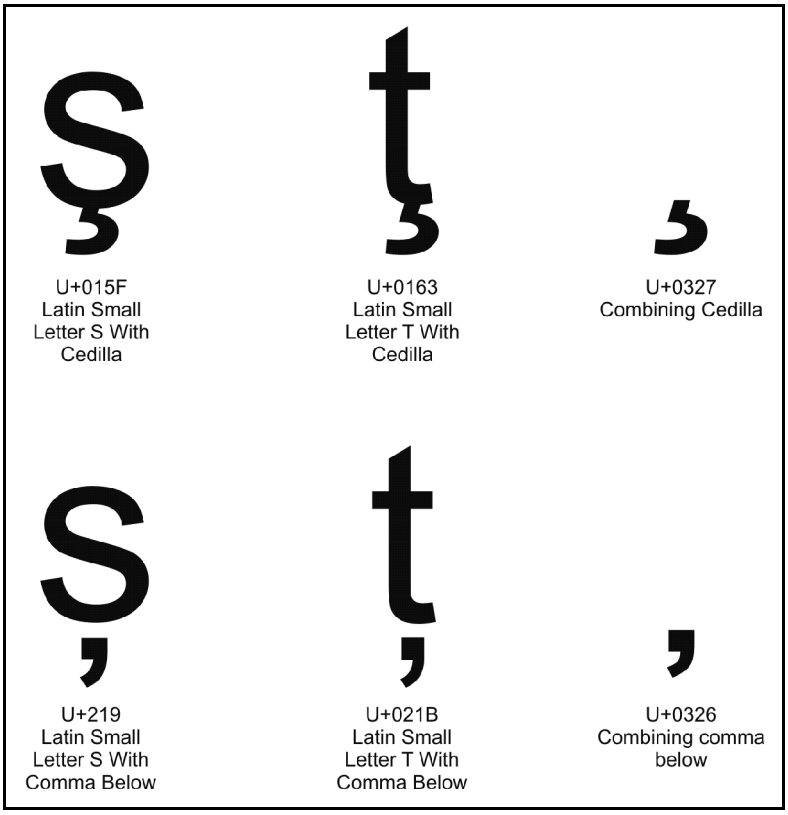

Figura 1: Caracterele în cauză, mărite pentru vizibilitatea semnelor diacritice.
Primul rând conține caracterele „*s cu sedilă*”, „*t cu sedilă*” și semnul diacritic sedilă.
Al doilea rând conține caracterele „*s cu virgulă*”, „*t cu virgulă*” și semnul diacritic virgulă.

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”.[7] 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*”, iar asta numai în măsura în care expeditorul și destinatarul mesajului se puneau cumva de acord în această privință.[8]

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.[9]

[7] `http://unicode.org/Public/MAPPINGS/VENDORS/APPLE/ROMANIAN.TXT`

[8] “*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`

[9] Caracterele cu semne diacritice corecte sunt incluse și în standardul ISO/IEC 8859-16 (Latin-10), publicat în 2001. Latin-10 este suportat în navigatoarele de Internet Firefox, Chrome și Opera și în sistemele de operare Mac OS, de la versiunea 10.4 (`http://www.opensource.apple.com/source/CF/CF-368.25/String.subproj/CFStringEncodingExt.h`), și Linux (`http://www.kernel.org/doc/man-pages/online/pages/man7/latin10.7.html`). Deși prezintă unele avantaje în fața Unicode (în special pentru că standardele ASCII extinse sunt single-byte), adoptarea pe scară largă a Unicode a făcut ca, pentru moment cel puțin, Latin-10 să nu fie adoptat pe scară largă; un alt factor decisiv în această privință este și faptul că sistemele de operare și navigatoarele de la Microsoft nu suportă acest standard.

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 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 *colecție de text* 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.[10]

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 colecțiilor de text care satisfac simultan următoarele criterii independente:

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

Dacă măcar unul dintre criteriile de deasupra nu se aplică, atunci trebuie *ignorate complet toate datele statistice concrete* din această versiune a acestui document.[11] În acest caz trebuie utilizate resursele indicate la sfârșitul acestui document (dacă se aplică), sau trebuie căutate și adaptate datele statistice concrete asociate situației concrete 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 1 sau ultimul paragraf din această lucrare).

[10] 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.

[11] 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.

3. *Alegerea momentului: de ce este important*

După cum am arătat mai sus, diferența grafică dintre cele două variante de caractere este în cea mai mare parte a timpului nesemnificativă.[12] 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*, deoarece utilizarea uneia dintre variante în defavoarea celeilalte nu aduce niciun plus de informație, indiferent de felul în care este interpretat textul.[13]

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.[14] Datorită cotei de piață uriașe a sistemelor de operare ale companiei Microsoft în rândul consumatorilor de conținut[15], *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 colecții de text de limbă română vor fi în mod inevitabil *early adopters*[16] 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 colecții de text 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 justificate pentru a amâna migrarea către diacriticele corecte[17], odată ce se atinge masa critică *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 voi arăta mai jos că independent de felul în care sunt văzuți din afară sau de corectitudinea tehnică a diacriticelor folosită în colecțiile lor de text, 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 și după momentul apariției masei critice. Scopul acestui document este tocmai acela de a identifica factorii care influențează momentul în care trebuie luate aceste decizii, de la caz la caz.

[12] 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.)

[13] 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 colecțiilor de text.

[14] Chestiunea este analizată pe larg în secțiunea 4.1. dedicată consumatorilor de conținut.

[15] Vezi Tabelul 1

[16] Persoane (fizice sau juridice) care doresc să adopte cât mai repede tehnologiile cele mai recente.

[17] Pentru brevitate voi folosi în continuare sintagmele „diacritice corecte" și „diacritice noi" pentru „caractere care folosesc semnele diacritice corecte" (virgule), respectiv „diacritice incorecte" și „diacritice vechi" pentru „caractere care folosesc semnele diacritice vechi" (sedile).

4. Factori de influență

După cum am arătat în secțiunea 3., două forțe opuse acționează simultan 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, iar rezultatul ar fi utilizarea caracterelor corecte în limba română. În 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.

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.
- *Colecțiile de text*: conținutul efectiv al revistelor, site-urilor, aplicațiilor etc.

În acest capitol voi analiza felul în care fiecare dintre acești trei piloni afectează fiecare dintre cele două forțe identificate mai sus.

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 colecțiile lor de text, 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 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 în general pe orice medii în care consumatorul are un rol pasiv din punctul de vedere al tehnologiilor implicate (cărți, jurnale, reviste, afișe, prezentări video, filme 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 chioșc[18], aplicații online sau client-side care rulează într-un mediu proprietar, conținut prezentat pe hardware dedicat precum cititoarele de cărți electronice).

[18] Terminale dedicate, instalate în locuri publice, așa cum sunt cele din aeroporturi, gări, puncte de interes turistic etc.

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:[19]

1. *Lizibilitatea* – pot consuma conținutul?[20]
2. *Interactivitatea* – pot interacționa cu interfața?

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 colecției de text* în speță.[21]

Datele statistice utilizate în acest document sunt următoarele:[22]

Sistem de operare	**Cota de piață**	**Afișează**	**Scrie[23]**
Windows XP[24]	73,14%	Lizibil (70,8%)	Simplu (34%)
		Ilizibil (29,2%)	Dificil (66%)
Windows Vista	8,91%	Perfect	Simplu
Windows 7	15,89%	Perfect	Simplu
Mac OS X	0,75%	Perfect	Simplu
Linux	0,74%	Nesigur	Simplu
Altele	0,73%	Nesigur	Dificil

Tabelul 1: Datele utilizate pentru generarea graficelor

Versiune Internet Explorer	**Cota de piață**
Internet Explorer 8	17,95%
Internet Explorer 7	9,75%
Internet Explorer 6	11,43%
Altele	0,11%

Tabelul 2: Cota de piață a diverselor versiuni de Internet Explorer

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 Linux.

[19] Cel mai reprezentativ exemplu în acest sens sunt colecțiile de text ale site-urilor și aplicațiilor Internet/intranet. În aceeași situație se află însă orice colecție de text care poate fi interpretată pe mai multe platforme software – aplicații, documente distribuite, mesaje e-mail (e.g. newsletters) și așa mai departe.

[20] Includ în această subcategorie și problemele de accesibilitate pentru persoanele cu dizabilități – de exemplu în ce măsură aplicațiile care citesc textul pentru consumatorii nevăzători sunt capabile să recunoască și să interpreteze corect texte scrise cu diacriticele noi.

[21] Această precizare este crucială pentru o analiză corectă în cazul unor situații particulare; vezi și nota 11.

[22] Datele din Tabelul 1 și Tabelul 2 sunt colectate în mai 2010 de la `http://gs.statcounter.com/`, pentru consumatorii din România. Pentru o analiză riguros exactă a consumatorilor de conținut în limba română la nivel global ar fi necesare datele statistice asociate acestui grup demografic specific, independent de poziția geografică (consumatorii de conținut în limba română nu se găsesc numai în România, ci și în Republica Moldova și pe alte meridiane; în plus, nu toți consumatorii de conținut din România preferă limba română). Totuși, așa după cum am indicat în mod repetat în acest document, fiecare creator de conținut trebuie să țină cont de capabilitățile tehnice ale propriilor consumatori. Prin urmare am considerat acceptabile aproximările rezultate din utilizarea unui criteriu geografic pentru ilustrarea orientativă a situației curente în acest document.

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

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

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:[25]

- Prin instalarea explicită a unui pachet software suplimentar de la Microsoft.[26]
- Prin instalarea Internet Explorer 7 sau Internet Explorer 8, aplicații care sunt în mod normal instalate prin mecanismul de actualizate automată al 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 dezirabilă pentru motive independente de diacritice), deci este de așteptat ca o parte semnificativă a consumatorilor să fi instalat această actualizare.

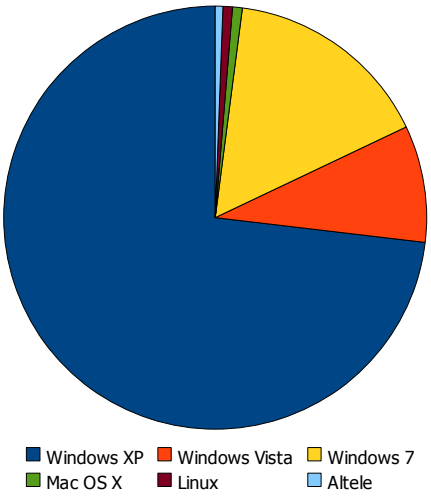

Figura 2: Gradul de utilizare al diverselor sistemelor de operare

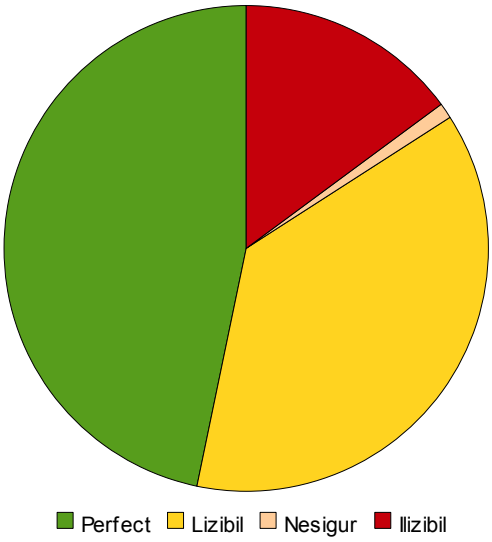

Figura 3: Capacitatea sistemelor de operare de a afișa diacriticele corecte

Totuși 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. Nu avem la dispoziție date statistice globale de încredere care să coroboreze sistemul de

[25] 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.

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

operare cu navigatorul utilizat[27], 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.

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

- Utilizatorii pentru care textul este doar lizibil (nu perfect) pot citi textul, însă, în funcție de situație, *unii 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 tabel am folosit Arial Black).
- Chiar în cazul colecțiilor de text 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.[28]

	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)

Acestea fiind spuse, trebuie menționat în mod proeminent că indiferent de platforma specifică a consumatorilor, indiferent de publicul țintă al colecției de text și indiferent de numărul lor, *fracțiunea consumatorilor care nu pot vizualiza diacriticele corecte se vor afla practic în imposibilitate de a consuma conținutul*. 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 pusă 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 colecției de text. Cea mai proeminentă funcție a interfeței în această privință este funcționalitatea de căutare: dacă colecția de text conține diacritice

[27] 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.

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

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 soluții tehnice*, însă acestea trebuie luate în considerare și rezolvate din timp.[30] Totuși *problemele de interacțiune sunt printre puținele care pot fi rezolvate înainte de începerea migrării și* ***ar trebui rezolvate cât mai curând***.[31]

4.2. Creatorii de conținut

Al doilea pilon de influență al deciziei de migrare este capacitatea creatorilor de conținut de a crea conținut folosind diacriticele corecte. *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?

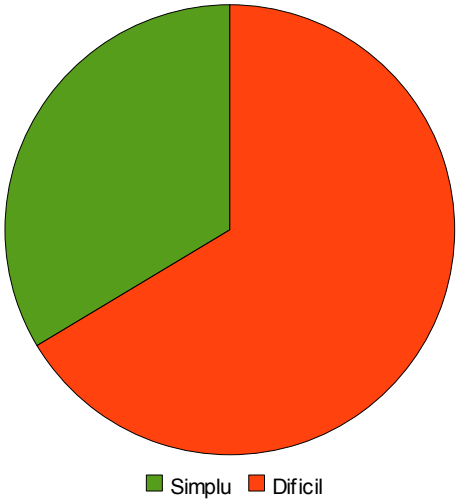

Figura 4: În ce măsură se poate scrie text cu diacriticele corecte

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

- *Redactori tradiționali* (jurnal, revistă, carte tipărită ș.a.m.d.): 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).

[29] Vezi și ultima coloană din Tabelul 1, 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 colecție de text 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 colecția de text 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 colecție de text face ca această scăpare să fie lipsită de orice efecte adverse concrete.

După cum se vede în Figura 4, *mai bine de jumătate dintre creatorii independenți de conținut au dificultăți în a scrie text utilizând diacriticele corecte*.[32] 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 factorilor decizionali de a influența mijloacele tehnice utilizate de creatorii individuali de conținut, sau de a implementa soluții tehnice pentru ameliorarea sau rezolvarea acestei probleme specifice. 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, în absența unor soluții tehnice specifice.[34]

4.3. Colecția de text

Însăși colecția de text, 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 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*.

Colecția de text influențează decizia de migrare în funcție de următorii factori:

1. *Relevanță*: este relevant să vă puneți întrebări legate de migrarea colecției de text numai în măsura în care aceasta este vizibilă. De exemplu arhiva digitală 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 colecției de text. Există însă un caz particular în care contează exclusiv mărimea: atunci când ea este nulă. Dacă un creator de conținut începe lucrul la o colecție de text complet nouă atunci trebuie cântărită cu atenție opțiunea de a utiliza de la bun început diacriticele corecte – în măsura în care aceasta este o opțiune acceptabilă din celelalte puncte de vedere atunci ar trebui adoptată, întrucât astfel se va evita mai târziu costisitorul proces de migrare retroactivă.
3. *Complexitate*: am văzut 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 voi analiza dedesubt.

[32] Datele sunt extrase de pe ultima coloană din Tabelul 1.

[33] `http://www.stefamedia.ro/diacritice-romanesti-corecte-in-windows-xp/`

[34] La Wikipedia s-a luat deja decizia migrării la diacriticele corecte, deoarece au fost identificate soluții tehnice concrete care permit vastei majorități a redactorilor individuali să genereze conținut utilizând diacriticele corecte indiferent de platforma software pe care o utilizează. Pentru detalii despre acest caz particular vezi secțiunea 5. dedicată studiului de caz Wikipedia.

Complexitatea colecției de text este o măsură a frecvenței situațiilor care necesită intervenție umană. În cazul colecției de text de la Wikipedia am identificat câteva tipologii specifice de situații problematice, probabil reprezentative pentru cazul general:

- *Texte scrise în altă limbă* (cel mai notabil în turcă) fără notație adecvată.[35] 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 acestor fragmente de text.
- *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.[36]
- *Interoperabilitatea cu alte colecții de text*, în special în condițiile unei migrări parțiale.[37]

5. *Studiu de caz: Wikipedia în limba română*

Am fost implicat în aproape toate discuțiile de la Wikipedia în limba română în privința deciziei de migrare la diacriticele corecte, iar mărimea și complexitatea colecției de text, numărul mare și eterogenitatea redactorilor și consumatorilor de conținut ai acestui proiect îl fac probabil unul dintre cele mai bune studii de caz posibile în contextul acestui document.

La Wikipedia în limba română, discuțiile despre migrarea la diacriticele corecte au început încă din anul 2007. Deși la acel moment au fost considerate premature, în urma discuției a fost creată o pagină dedicată subiectului.[38] La sfârșitul aceluiași an a fost implementată o soluție tehnică de natură să forțeze utilizarea diacriticelor *vechi*, în scopul uniformizării conținutului (unii redactori deja scriau conținut folosind diacriticele noi). La sfârșitul lui 2008 și începutul lui 2009 discuțiile au început să se orienteze către identificarea problemelor și soluțiilor tehnice concrete asociate utilizării diacriticelor corecte și s-au început câteva demersuri tehnice concrete, deși fără efecte pentru consumatori, în vederea viitoarei migrări.

Dată fiind absența unor statistici și analize concrete, comunitatea a decis la 1 martie 2010 să implementeze un proiect pilot prin care se permitea utilizarea diacriticelor corecte pe cele mai multe dintre paginile interne ale proiectului, cu scopul de a identifica numărul de redactori incapabili să le citească; nu s-a înregistrat nicio plângere.

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

[36] Cele mai la îndemână exemple sunt legate de colecția de text 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; aceasta este mai degrabă o resposabilitate morală datorată numărului relativ mare de documente care fac trimitere la enciclopedie.

[37] De exemplu atunci când colecția de text are relevanță parțială în privința diacriticelor, dar partea publică a colecției de text (cea care urmează să fie migrată) interacționează prin sisteme automate cu partea istorică/privată/confidențială care nu este migrată.

[38] `http://ro.wikipedia.org/wiki/Wikipedia:Corectarea_diacriticelor`

Pe fondul discuțiilor de la Wikipedia și în urma invitației organizatorilor ConsILR de a participa la conferință, am organizat în perioada 27 aprilie–4 mai 2010 un sondaj național[39] în scopul identificării capabilităților tehnice ale consumatorilor de limbă română în privința lizibilității diacriticelor noi. Discuțiile din timpul sondajului au dus la identificarea unui criteriu concret pe baza căruia urma să se ia decizia migrării la diacriticele corecte.[40]

Sondajul a beneficiat de un nivel adecvat de expunere și participare[41], dar am determinat ulterior că voturile exprimate au fost eronate în proporție mult prea mare pentru orice scop practic.[42] Totuși datele statistice obținute prin analizarea vizitatorilor (ignorând voturile exprimate) au confirmat faptul că distribuția capabilităților tehnice ale vizitatorilor Wikipedia în limbă română este foarte apropiată de datele statistice naționale prezentate în acest document. Coroborând această informație cu soluțiile tehnice identificate anterior, de natură să amelioreze problema lizibilității pentru segmentul consumatorilor afectați, am constatat că a fost satisfăcut criteriul convenit în timpul desfășurării sondajului. Prin urmare, pe 5 mai 2010 s-a luat decizia de migrare cât mai rapidă la diacriticele corecte.[43]

În cazul colecției de text de la Wikipedia în limba română procesul va fi cu siguranță unul de durată, însă analiza și execuția până în acest moment au urmat o traiectorie foarte apropiată de cea ideală. La Wikipedia, calitatea analizei și a deciziilor în această privință se datorează în cea mai mare măsură faptului că deciziile se iau prin consens, iar viziunile divergente ale diverșilor participanți la discuție au asigurat în cadrul discuțiilor o reprezentare optimă a celor două forțe identificate în secțiunea 4.

6. Concluzii

Toate colecțiile de text care conțin text în limba română vor urma inevitabil, mai devreme sau mai târziu, standardul corect în privința semnelor diacritice. Singurul punct delicat este alegerea momentului optim pentru migrare. Diferența dintre cele două variante este mică din punct de vedere vizual, dar există dificultăți tehnice potențiale semnificative asociate migrării.

Am căutat să identific aici factorii care influențează alegerea pragmatică a momentului optim de migrare pe baza următoarei structuri:

- Consumatorii de conținut
    - în ce măsură tehnologia se află sub controlul consumatorului
    - în ce măsură pot consuma conținutul
    - în ce măsură pot interacționa cu interfața

[39] http://www.moongate.ro/products/diacritice/sondaj/

[40] „M*ai puțin de 3% de vizitatori pe care nu îi putem ajuta (nu văd diacritice sau văd majuscule și au alt SO decât Windows sau Windows mai vechi de XP)*", la pagina menționată în nota 43.

[41] http://www.moongate.ro/products/diacritice/sondaj/date.php#statistici

[42] Aproximativ 20% dintre respondenții care au votat că nu pot citi diacriticele noi utilizau Windows Vista sau Windows 7, sisteme de operare despre care știm cu siguranță că în realitate le afișează perfect.

[43] http://ro.wikipedia.org/w/index.php?title=Discu%C5%A3ie_Wikipedia:Sfatul_B%C4%83tr%C3%A2nilor&oldid=3918861#Concluzii_finale

- Creatorii de conținut
  - în ce măsură pot controla tehnologia utilizată de redactorii individuali
  - în ce măsură redactorii individuali pot crea text folosind diacriticele corecte
- Caracteristicile colecției de text
  - relevanța
  - mărimea
  - complexitatea

Creatorii de conținut de limbă română ar trebui să ia cât mai curând următoarele măsuri concrete:

- Implementarea măsurilor tehnice necesare pentru a permite consumatorilor care folosesc deja diacriticele corecte să interacționeze cu interfața (e.g. pentru funcțiile de căutare).
- Implementarea măsurilor tehnice necesare în vederea migrării, dacă este cazul (în particular identificarea explicită a textelor scrise în limba turcă).
- Analiza situației propriului caz particular, prin prisma acestui document.
- Identificarea factorilor critici aplicabili colecției de text analizate.
- Stabilirea unui criteriu concret pentru migrare, pe baza factorilor aplicabili.
- Planificarea migrării (metodologie tehnică, estimare buget și resurse necesare).
- Monitorizarea și previzionarea evoluției situației în raport cu criteriul ales, în așa fel încât să poată aloca resursele necesare concretizării migrării în timp util.

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

Mulțumiri *organizatorilor și referenților ConsILR*, pentru că mi-au oferit imboldul inițial și încurajarea ulterioară pentru cristalizarea acestui document în formă scrisă; domnului *Cristian Secară*, pentru informațiile concrete care m-au ajutat să corectez unele statistici care altfel ar fi fost cu siguranță greșite; domnului *Cristian Adam*, pentru notele legate de Latin-10; *colectivității Wikipedia* în limba română, pentru discuțiile întotdeauna deschise; în acest context le 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, cel care a împins constant în direcția adoptării cât mai rapide a diacriticelor corecte la Wikipedia și care m-a ajutat și cu sugestii în privința acestui document), *Danutz* (care mi-a indicat site-ul `http://gs.statcounter.com/`, utilizat pentru toate statisticile din acest document; în versiuni mai vechi ale acestui document am folosit statistici globale, semnificativ diferite), *AdiJapan* (ca întotdeauna, a reușit să găsească un echilibru între interlocutorii mai tehnici și cei mai puțin tehnici, între cei avangardiști și cei conservatori); și, independent de Wikipedia, *tatălui meu și prietenilor* care m-au ajutat să dau forma curentă acestei lucrări (știți voi cine sunteți).