Unelte și metode digitale în cercetarea comunicării mediate de calculator. Importanța corpusului online

Andrei STIPIUC[*]

Key-words: *Digital Humanities, Internet Linguistics, CMC, User-Generated Content, Social Networks*

1. Introducere – Un milion contează

Robert Wilensky, celebru profesor de matematică și informatică de la Universitatea din Berkeley, afirma cândva că „deși am auzit cu toții, de atâtea ori, că un milion de maimuțe care tastează la un milion de mașini de scris ar putea recrea întreaga operă a lui Shakespeare, știm cu siguranță acum, uitându-ne în Internet, că acest lucru nu este posibil”[1]. Wilensky parafraza teorema maimuțelor infinite care definește, ca obiect al studiului probabilităților, în funcție de diversele sale variante, că, la un moment dat, opera scriitorului englez va fi creată aleatoriu, grație puterii de calcul. În același timp, intervențiile scrise ale utilizatorilor de Internet, în calitatea lor de *prosumeri*, erau luate în derizoriu.

23 septembrie 2011 este data la care Robert Wilensky a fost contrazis. Maimuțele digitale ale inginerului creativ Jesse Anderson au produs, în laborator, *A Lover's Complaint* de William Shakespeare, pe baza unui model computațional de tip *cloud* oferit de Amazon (EC2-Amazon Cloud System) și a unor operații de lucru cu text ASCII (cu literele alfabetului latin), a comparării textelor produse, în grupuri de 9 caractere, cu baza de date a operei literare (Gross 2011). De fiecare dată când se producea o potrivire, fragmentul din baza de date era bifat, iar maimuța digitală primea, ca recompensă, o banană digitală.

Au fost contrazișsi astfel toți cei care desconsiderau rolul datelor în societatea contemporană, cei care luau în derizoriu importanța puterii de calcul și a inginerilor creativi, și s-a reconsiderat faptul că nu există date generate de utilizatori lipsite de semnificație științifică, un aspect extrem de important din punctul de vedere al științelor umaniste și mai ales al umanismului informatic.

[*] Universitatea „Alexandru Ioan Cuza”, Iași, România.

[1] „We've all heard that a million monkeys banging on a million typewriters will eventually reproduce the entire works of Shakespeare. Now, thanks to the Internet, we know this is not true”, http://www.eecs.berkeley.edu/Faculty/Homepages/wilensky.html.

2. Importanța datelor în societatea contemporană și în cercetarea modernă

Trăim în plină revoluție industrială a datelor (Hellerstein 2008), în care nu doar guvernele și corporațiile dețin și utilizează metode de culegere și prelucrare a unui număr impresionant de date, ci, de cele mai multe ori, şi utilizatorul comun, individul obișnuit în mediul căruia există o suită de echipamente digitale (de la banalul laptop și până la smartphone și ceas inteligent) cu care se pot produce și prelucra date. Aproape orice activitate devine cuantificabilă: câte pagini citește un individ într-o singură sesiune de lectură (dispozitivele și aplicațiile Kindle), câte cărți (rețeaua socială GoodReads), câte calorii se ard la o singură activitate fizică (aplicațiile Endomondo). Reiterate, aceste utilizări generează, doar de la câțiva indivizi, un număr impresionant de date. Prin multiplicarea datelor cu numărul utilizatorilor din întreaga lume, putem să ne imaginăm anvergura pe care o presupune operarea cu date. Proiectul astronomic Sloan Digital Sky Survey (http://www.sdss.org) a strâns în prima săptămână de funcționare a telescopului din New Mexico mai multe date decât s-au strâns în întreaga istorie a astronomiei. Supermarketul Wal-Mart înregistrează 1 milion de tranzacții pe oră, în timp ce rețeaua socială Facebook găzduiește 40 miliarde de fotografii (McCarthy 2008). Operarea cu date a devenit esențială pentru activități socio-administrative vitale, de la combaterea crimei la identificarea tendințelor în afaceri, de la urmărirea indivizilor (inițial doar cei ce ridicau suspiciuni de siguranță, acum aproape pe toată lumea) la prevenția și combaterea epidemiilor.

Câmpul de date digitale a reunit, în cercetare, echipe de cercetători din domenii variate, ceea ce conduce la sporirea activității de cercetare inderdisciplinară, precum și la creșterea numărului de discipline care pot fi conectate acum, în premieră. În funcție de metodele de prelucrare, volumul de date poate fi alcătuit din date care altădată erau doar la dispoziția specialiștilor dintr-un singur domeniu (Bartscherer, Coover 2011: 24). Cercetarea de date a continuat tradiția analizelor cantitative, dar a lărgit orizonturile științifice: se pot trasa relații și se pot urmări fenomene imposibil de identificat și de abordat până atunci. Conceptul consacrat, The Big Data[2], a fost formulat chiar de către echipele multiversate, formate din cercetători din câmpuri de studiu diferite, iar de foarte puțină vreme se discută deja despre o meserie emergentă: cea de savant de date (*data scientist*) (Patil 2012), în care titularul trebuie să combine aptitudinile programatorului, ale statisticianului, ale povestitorului și ale artistului, pentru a scoate la lumină diamantul, partea semnificativă, din cantitatea imensă de date.

Îmbinarea metodelor digitale și a altor mijloace informatice, a informaticii, în sine, cu științele umaniste a dus la crearea, încă de la mijlocul secolului trecut, a umanismului informatic. În cadrul acestui câmp de cercetare interdisciplinar prin definiție există patru etape fundamentale, pe care ar trebui să le parcurgă, într-o măsură sau alta, orice studiu modern: investigarea datelor; analizarea datelor; sinteza datelor; prezentarea datelor.

Urmând aceste etape, în lingvistică și în CMC se pot măsura: numărul de vorbitori ai unei anumite limbi în funcție de numărul de utilizatori, poate fi analizat

[2] Webopedia, http://www.webopedia.com/TERM/B/big_data.html.

statistic conținutul generat, numărul substantivelor proprii și al altor părți de vorbire (Metoda POS); se pot identifica tendințe într-o dezbatere publică, se pot realiza hărți lingvistice interactive (Gold 2012), se pot conduce analize lexicale și studii ale conversației sincrone și asincrone etc.

3. Debutul în cercetarea din cadrul umanismului digital

Pentru un cercetător debutant cu buget subvenționat inexistent, dar cu buget personal care să permită o autofinanțare suficientă astfel încât să nu ne situăm în sfera excluşilor digitali[3] (McIntosh, Varoglu 2005), apelarea la mediul online a reprezentat o gură de oxigen. Cercetarea în mediul online presupune accesibilitatea corpusului, costuri reduse de cercetare, multiple posibilități de selecție a metodelor și a uneltelor digitale de operare asupra corpusului.

În studiul doctoral în derulare, *Particularităţi lingvistice ale textului din cadrul conţinutului generat de utilizatorii români pe platformele sociale online*, am încercat să ne înscriem în rândul exerciţiilor, deloc numeroase în literatura românească de specialitate, care apelează la sincronie în spaţiul online românesc pentru a studia manifestările limbii utilizate de vorbitorii nativi de limbă română activi în spaţiul Web.

În prezent, World Wide Web-ul se confundă cu cele mai populare aplicaţii ale sale. Caracterul social al spațiului Web (Berners-Lee, Fischetti 2001: 12), acolo unde numărul utilizatorilor români creşte de la an la an, este indiscutabil, în reţelele sociale sau în cadrul altor platforme ori lumi virtuale, oamenii interacţionând între ei şi comunicând zilnic. Web-ul nu mai este doar un loc de lectură, de consum, este un loc al interactivităţii şi convergenţei activităţilor umane cotidiene (Morris, Ogan 1996). Dat fiind acest context social fundamental, limba joacă în continuare un rol esenţial în interacţiunea utilizatorilor în cadrul comunităţilor locale, naţionale sau chiar globale.

Lucrarea noastră se plasează, în mod tradițional, în cadrul comunicării mediate de calculator, deși, în perspectiva umanismului digital despre care vorbim, o putem apropia de ceea ce David Crystal (2004) numea, cu sintagma *Internet Linguistics*, studiul limbii pe Internet. Lucrarea se orientează către inovaţiile din spaţiul Web ale ultimilor ani – platformele UCG (*User Generated Content*) – și asupra utilizatorilor români ai acestor platforme, în principal către membrii rețelelor sociale (Gumperz 1966: 27-38)[4]. Deși există o întreagă dezbatere terminologică, am preferat sintagma de „rețea socială”, pe care l-am adaptat din raționamente ce țin de abordarea sociolingvistică a lucrării noastre, păstrând termenul așa cum a fost conceput de Gumperz.

Am urmărit modul în care se manifestă variaţiile limbii scrise în mediul online în permanenta sa adaptare la caracteristicile acestor platforme. Am avut în vedere trăsăturile variantei lingvistice pentru limba scrisă (text): caracteristicile grafice (aspectele legate de tipografie, aşezare în pagină, design, layout, ilustraţii, culoare) și ortografice (sistemul scris al limbii române pe aceste platforme: alfabet, majuscule,

[3] În original, *digital divide*. Este vorba de persoanele care au puterea financiară de a se dota cu calculatoare și dispozitive mobile pentru a se conecta la noua societate bazată pe cunoaștere.

[4] Perspectiva sociolingvistică pe care am ales-o pentru studiu ține cont de definițiile date de Gumperz pentru a identifica și califica relațiile existente între membrii comunităților lingvistice.

diacritice, punctuație, emfază – cursiv, aldin, subliniat), caracteristicile lexico-semantice (elemente de vocabular, termeni existenţi în urma pătrunderii activităţilor online în viaţa cotidiană, expresii idiomatice) şi morfo-sintactice, caracteristicile discursului prin care se clădește, pe Facebook, în particular, narațiunea personală, modalitățile de manifestare a politeții online și modurile de încălcare a acesteia, jocurile lingvistice, marcările emoțiilor pozitive și negative, comutarea de cod în funcție de afilierea sau izolarea culturală, imigrația unor utilizatori în alte țări și adoptarea limbii locale, continua expansiune a „limbilor engleze globale" (Schneider 2011: 336), căile prin care argoul informatic reușește să influențeze norma comună din mediul offline. O serie de caracteristici lingvistice sunt comune ambelor moduri de comunicare – sincronă și asincronă (utilizarea emoticonului, abrevierile, scrierea prescurtată, dar și o înclinație spre informalitatea verbală) –, ceea ce apropie și mai mult limba scrisă de trăsăturile limbii vorbite.

Am orientat cercetarea în jurul actualizărilor (*status updates*) din Facebook, deoarece există un consens că aceste actualizări, indiferent de natura și tematica lor, oferă indicații asupra stării psihologice a fiecărui individ care le distribuie. Motivele distribuirii (*share*) sunt de natură utilitară, informațională, militantă și, mai ales, autobiografică, iar utilizatorii care trimit informația o înțeleg mai bine atunci când o retransmit. S-a constatat că aproape 80% dintre mesajele publicate pe Facebook au un conţinut autobiografic. Din perspectivă lingvistică, este clară intertextualitatea mediilor, asemănarea cu publicistica în ceea ce privește locuțiunile sau expresiile idiomatice, ca parte a discursului repetat. În discurs apar întrebările, ca particularitate de interacțiune, presecvențele, cu care se verifică disponibilitatea receptorului, scrierea în stil memorialistic.

Am intuit și o evoluție a comunicării prin mijloace lingvistico-imagistice, prin utilizarea *meme*-lor (combinarea voit stupidă de text și imagine, care prezintă un grad mare de *viralitate*), dar și prin replicile acordate cu hypertexte care trimit la imagini – ca răspuns ce marchează propria opinie, în urma selecției dintr-o bază de date uriașă de răspunsuri posibile.

Particularităţile lingvistice ale conţinutului scris generat de utilizatorii români pe care le-am observat în cadrul rețelelor sociale pot fi grupate astfel:

1. Particularități ortografice:
   Scrierea prescurtată; scrierea cu acronime; scrierea alfanumerică, prin substituția unor litere, asemănătoare ca formă grafică cu cifre; scrierea neconvențională, cu caracterul „@", dedicat adreselor de email, cu scopul de a marca apartenența organizației la spațiul online;
   Scrierea telegrafică (enunțuri scurte, tastate rapid, cu omisiuni legate de punctuație sau de ortografie, scrierea fără spații după semnele de punctuație, scrierea cu litere mici după semnele de punctuație sau la începutul propoziției, scrierea fără diacritice);
   Scrierea sau apelarea la simboluri și emoticoane pentru a putea reproduce, în scris, efectele paralingvistice (emoții, expresivitate) care se regăsesc, altfel, doar în cadrul comunicării față în față;
   Scrierea cu repetarea consoanelor sau a vocalelor pentru marcarea unui sentiment.

Figura 1. *Scrierea cu majuscule, utilizarea emoticoanelor*

2. Câteva particularități lexicale de argou tehnic sau care definesc elemente de interfață ale unor platfome online, care au părăsit mediul online și au intrat în uzul curent.

3. Sintaxa interfețelor, care include modul în care este utilizată limba pentru a pune la dispoziția utilizatorului mijloacele de comunicare în scris (pentru scrierea propriilor texte, pentru distribuția unor materiale online – hyperlink, imagini). Utilizarea semnului „#" pentru crearea de etichete (*hashtags*) care să înscrie intervenția într-o serie cu același subiect. Utilizarea uneltelor din cadrul Facebook pentru a completa câmpuri predefinite prin care un utilizator își poate înregistra anumite momente din viața personală.

Figura 2. *Exemplu de sintaxă Twitter*

4. Comutarea de cod, în care se alternează, după caz, limba română cu o limbă străină (de obicei engleza, dar se întâlnesc și limbile franceză sau italiană) sau în care se selectează o altă limbă decât limba română pentru a marca, de obicei în cazul românilor emigrați, apartenența la noua comunitate socio-culturală. Pentru majoritatea utilizatorilor români, limba engleză este *lingua franca* în cadrul mediului online.

Figura 3. *Exemplu de comutare de cod inter-propozițională*

5. Limbajul text + imagine (*LOLcats*, *meme*), deturnat de la forma și sensul de bază și utilizat pentru exprimarea plastică și ludic-parodică a unor anxietăți cotidiene sau pentru a răspunde, în mod ironic, unor comentarii.

Figura 4. *Exemplu de* meme *studențesc în preajma examenului de licență*

6. Respectarea sau încălcarea politeții online, ale cărei reguli se constituie într-o *netichetă* de care unii utilizatori țin cont, în timp ce alții nu.

7. Jocurile lingvistice, în cadrul cărora se remarcă lexicul inventat, parodierea unor greşeli de gramatică, unități frazeologice ca parte a discursului repetat, prin care utilizatorii își manifestă inventivitatea și creativitatea lingvistică.

Figura 5. *Joc lingvistic de scriere fonetică*

8. Particularități metalingvistice, în cadrul cărora se discută despre limbă, în general, și în care se ia atitudine față de utilizarea greșită a unor termeni și față de cei care nu cunosc normele gramaticale.

Figura 6. *Intervenție metalingvistică de corectare a titlurilor administrative în funcție de sex*

9. Particularități pragma-lingvistice ale narațiunii personale, eul definit cotidian prin intervenții memorialistice sau prin textele care însoțesc ritualurile sociale, anuale sau cotidiene: sărbători românești și internaționale, sărbători personale etc.

Deși am propus un studiu descriptiv, am ținut cont, în analiza noastră, de materiale care să nu adune neapărat particularități de exprimare neglijentă sau

neconformă cu standardele gramaticale, din care să nu reiasă utilizarea defectuoasă a limbii, supunând analizei doar materialele relevante din punctul de vedere al modificărilor impuse de mediu.

Cercetarea digitală a CMC (în cadrul căreia se înscrie și studiul numit anterior) poate debuta, cu o serie de avantaje clare:

- accesarea bibliotecilor de date, a revistelor științifice online de tip *open acces* (Burdick *et alii* 2012: 113); a cărților de referință digitalizate (prin cumpărare sau împrumut);
- apelarea la dicționare online – prestigioase (Cambridge Dictionaries Online) sau agregatoare (Dexonline);
- eliminarea unor deplasări, gestionarea eficientă a timpului;
- accesul la corpusul online. Exemplele pentru corpusul studiului doctoral au fost culese pe parcursul a doi ani (2012-2014). În urma unei selecții riguroase, am publicat exemplele care ilustrau, fără niciun dubiu, particularitățile lingvistice ale conținutului scris generat de utilizatorii români în mediul online. În afara materialelor publice, corpusul de față este constituit în mare parte din texte culese de pe două conturi de Facebook diferite, de unde a fost observat, săptămânal, un număr de aproximativ 1.000 de utilizatori unici. Am optat pentru cenzurarea numelui de familie sau al celui de-al doilea nume de utilizator, pentru a asigura anonimatul autorilor textelor publicate. Exemplele alese sunt texte scrise de utilizatori români, majoritatea stabiliți în România (din Iași, București, din alte orașe din regiunea Moldovei), dar și din Republica Moldova. Un număr important de utilizatori, minor, totuși, ca pondere, din totalitatea celor observați, s-a stabilit peste granițe (Anglia, Italia, Franța), și am specificat acest lucru la orice exemplu unde era necesară mențiunea. Aproximativ 45% dintre utilizatorii urmăriți sunt studenți (20-25 de ani) la școli superioare (studii de licență, master, doctorat), 30% sunt tineri cu vârste cuprinse între 25 și 30 de ani, iar restul de 25% au vârste mai mari de 30 de ani.

Dar cercetarea digitală și, mai ales, mediul online românesc prezintă și suficiente dezavantaje și greutăți pentru cercetătorul umanist:

- lipsa bibliotecilor online românești, a cataloagelor online românești cu reviste și lucrări de doctorat, indexate corespunzător și actualizate anual;
- dificultatea lecturii de pe ecrane electronice (retroluminarea, norme de design al textului nerespectate: lungimea rândurilor, mărimea și lizibilitatea fonturilor etc.);
- o parte dintre programe sunt contra cost, iar accesarea lor favorizează pirateria digitală;
- cea mai mare parte a programelor nu sunt optimizate pentru limba română (diacritice, forme flexionare, localizare).

4. Metode în cercetarea digitală a CMC

a) *Investigarea datelor* (are rolul de a strânge corpusul de text necesar studiului, sub formă de text sau imagine):

- Capturi grafice – *HoverSnap*, capturi de text – *Text Capture*, combinație între cele două (de tip OCR) – *SolidCapture*.
- Procesarea imaginilor – *HeadsUp Digitisation* (când obiectul supus capturii este deja un raster).
- Extragerea algoritmizată a textului din imagini – *Project Naphta*.
- Scanarea cărților și articolelor de pe hârtie (sau fotografiere) și convertirea în format digital.
- OCR pentru documente mai scurte; supus erorilor, care depind de calitatea sursei.
- Formulare online cu centralizarea răspunsurilor în baze de date (*Google Docs*).
- *Data Mining* (*Text Data Mining*) = extragerea automată a datelor care îndeplinesc anumite condiții.

b) *Analiza și sinteza datelor* (cu rolul de a prelucra datele în funcție de scopul urmărit):

- Procesarea imaginilor – *FocusOPEN Digital Asset Manager*. Extragerea textului pentru procesarea ulterioară.
- *LitStats* – unealtă pentru analiza limbajului natural din orice fișier text ASCII. Generează frecvența cuvintelor, lungimea cuvintelor, frecvența literei capitale, frecvența lungimii frazei și frecvența secvențelor verbale.
- Analiza *POS* (*Parts of Speech*) (Muñoz-García, Navarro 2012) – se etichetează părțile principale de vorbire (substantive, verbe, adverbe, pronume) și părți secundare de vorbire (substantive comune, nume proprii, pronume personale) pentru a determina anumite proprietăți ale unui text sau mediu: Twitter: substantivele comune și numele proprii sunt constante; prevalența adjectivelor de cantitate, prevalența numeralelor. Facebook: prevalența pronumelui personal de persoana I (mediu autobiografic).
- Analiza *MEM* (*Meaning Extraction Method*) (Kramer, Chung 2011: 169) – metodă semiautomată, împrumutată din psihologia personalității. Aplicată pe *status updates*, urmărește modul în care se manifestă exprimarea sinelui. Pe baza numărării ocurențelor și a coocurențelor cuvintelor, apoi a unei ordonări matriceale, ilustrează modul în care oamenii discută despre ei înșiși sau despre un eveniment. Aplicată pe *status updates*: evenimente cu impact emoțional pozitiv: Crăciunul, Paștele, Valentine's Day; evenimente cu impact emoțional negativ: proiectul Roșia Montană, moartea actorului Philip Seymour Hoffman, tăirea teilor din Iași, conflictul din Ucraina.

c) *Strategie și managemet de proiect* (cu rolul de a asigura organizarea spațiului de lucru, a programelor, precum și siguranța datelor):

- *Scriber* – procesor de texte special conceput (și) pentru lucrări științifice.
- Back-up automat (*Dropbox*).
- Managementul resurselor informatice (stabilitate, securitate, interconectare dispozitive, rețele personale, rețele locale).

d) *Colaborare* (pentru menținerea legăturii între colaboratori, a lucrului în comun cu echipamente interconectate):

- Colaborare cu colegi și utilizatori (*sharing*) – *Evernote*, *Pushbullet*.
- Forumuri online: *HASTAC* – forum pentru colaborare interdisciplinară și între instituții.
- Sisteme de videoconferință – *Google Hangouts*, *AccessGrid*.
- Prezentarea datelor.
- Hărți interactive și hărți mentale – *MindJet*.
- Infografice – *infogr.am*.
- Wordclouds – *Wordle* (generare de nor de cuvinte după input de text).

5. Concluzii

Pentru umanistul informatician, corpusul online poate constitui un câmp de cercetare extrem de fertil, având la dispoziție diferite instrumente digitale și metode, începând de la cele de tip *mainstream* (utilizate în mod obișnuit pentru activitate de birou, pentru organizarea datelor personale sau în scop educațional) ce pot fi deturnate către cercetare și terminând cu aplicațiile specializate, create de alți cercetători și puse la dispoziția comunității oamenilor de știință. Textul generat de către utilizatori în mediul online oferă noi posibilități de analiză și poate fi abordat din direcții diferite, din punct de vedere lingvistic, al teoriilor comunicării sau al pragmaticii.

Bibliografie

Bartscherer, Coover 2011: Thomas Bartscherer, Roderick Coover (eds.), *Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts*, Chicago, University of Chicago Press.

Berners-Lee, Fischetti 2001: Tim Berners-Lee, Max Fischetti, *Weaving the Web*, New York, Barnes & Noble.

Burdick *et alii* 2012: Anne Burdick *et alii*, *Digital Humanities*, Cambridge, MIT Press.

Crystal 2003: David Crystal, *The Scope of Internet Linguistics*, http://www.davidcrystal.com/DC_articles/Internet2.pdf.

Crystal 2004: David Crystal, *Language and the Internet*, Cambridge, Cambridge University Press.

Gold 2012: Matthew K. Gold (ed.), *Debates in the Digital Humanities*, Minneapolis, University of Minnesota Press.

Gross 2011: Doug Gross, *Digital Monkeys with Typewriters Recreate Shakespeare*, http://edition.cnn.com/2011/09/26/tech/web/monkeys-typewriters-shakespeare/.
Gumperz 1966: John Gumperz, *On the Ethnology of Linguistic Change*, în W. Bright (ed.), *Sociolinguistics*, The Hague, Paris, Mouton.
Hellerstein 2008: Joseph Hellerstein, *The Commoditization of Massive Data Analysis*, O'Reilly Radar, http://radar.oreilly.com/2008/11/the-commoditization-of-massive.html.
Kramer, Chung 2011: Adam D.I. Kramer, K. Chung, *Dimensions of Self-Expression in Facebook Status Updates*, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. Barcelona, Catalonia, Spain, July 17-21, 169.
McCarthy 2008: Caroline McCarthy, *Facebook Hosts 10 billion Photos*, CNET, 15 octombrie, http://www.cnet.com/news/facebook-hosts-10-billion-photos/.
McIntosh, Varoglu 2005: C. McIntosh, Z. Varoglu (eds.), *Perspectives on Distance Education: Lifelong Learning & Distance Higher Education*, Paris, UNESCO.
Morris, Ogan 1996: Merrill Morris, Christine Ogan, *The Internet as Mass Medium*, în „Journal of Communication", 46, 1, martie.
Muñoz-García, Navarro 2012: Óscar Muñoz-García, Carlos Navarro, *Comparing User Generated Content Published in Different Social Media Sources*, lrec-conf.com.
Patil 2012: D.J. Patil, *Data Jujitsu: the Art of Turning Data into Product*, O'Reilly Media.
Schneider 2011: Edgar W. Schneider, *Colonization, Globalization, and the Sociolinguistics of World Englishes*, în Rajend Mesthrie (ed.), *The Cambridge Handbook of Sociolinguistics*, Cambridge, Cambridge University Press.

Digital Methods and Tools for CMC Research. Online Corpus Analysis

It is well acknowledged that the contemporary research in social studies is increasingly required to respond to new efficiency criteria, in the process of adapting to the industrial and capital-intensive scale of scholarly work. To an equal extent, humanities research must closely study new types of data (big data, online corpora) and to succeed, researchers will need to resort to the use of digital methods and tools for collecting, analyzing and representing data and results. It is essential for theoretical Computer Mediated Communication research to enfold digital methods of analysis in order to illustrate new linguistic and cultural directions in an increasingly technology dominated society. Our paper will illustrate the modern methods employed in researching the linguistic aspects of user-generated content (UGC) regarding online social networks, from data mining and gathering techniques, to data processing practices such as MEM (Meaning Extraction Methods) or POS (Parts of Speech Analysis), and exemplifying the results through diagrams and infographics for a better representation to a wider public.