Treballs de Sociolingüística Catalana, núm. 26 (2016), p. 89-101 DOI: 10.2436/20.2504.01.110
http://revistes.iec.cat/index.php/TSC ISSN (ed. impresa): 0211-0784 ISSN (ed. electrònica): 2013-9136

Twitter com a eina per a la recerca sociolingüística: llums i ombres

Twitter as a sociolinguistic research tool: pros and cons

Alexandre Nobajas
Universitat de Keele

Data de recepció: 28 de març de 2015
Data d'acceptació: 23 de juny de 2015

Resum

La recent popularització de les xarxes socials ha provocat que una part important de la població comparteixi una quantitat ingent d'informació a través d'Internet. Bona part d'aquesta informació és textual i fàcilment accessible, per la qual cosa pot ser emprada per a l'estudi i l'anàlisi sociolingüístics d'una manera ràpida, fàcil i massiva. Tot i aquesta oportunitat per a la recerca, les dades obtingudes mitjançant aquesta metodologia tenen una sèrie de característiques que poden limitar i fins i tot esbiaixar els resultats obtinguts, i per això cal tenir-les en compte abans d'arribar a conclusions definitives. Per una banda, en aquest article s'exploren les oportunitats que les dades obtingudes a través d'una de les xarxes socials més populars, Twitter, proporcionen als investigadors de l'àmbit de la sociolingüística. Per l'altra, però, també s'expliquen quins són els riscos i les limitacions que cal tenir en compte a l'hora de realitzar investigacions sociolingüístiques utilitzant aquesta nova font de dades. Tot plegat hauria d'informar els investigadors que considerin emprar aquesta font d'informació de les llums i les ombres que es trobaran a l'hora d'utilitzar-la.

Paraules clau: Twitter, xarxes socials, metodologia sociolingüística, web 2.0.

Abstract

Due to the recent popularization of social networks, many people across the world are sharing huge amounts of information using the Internet. A large proportion of this information is in textual form and is easily accessible, so sociolinguistic research should be able to use it as means of obtaining information about how people communicate. Even though this is clearly an opportunity, data collected in this manner show some characteristics which may limit or even bias the results, so they need to be considered before starting the research process. On the one hand, this article explores the advantages of using data obtained from one of the most popular social networks, Twitter, and on the other hand it considers the problems and limitations which may be encountered by sociolinguists when using this method. In this way, researchers

Correspondència: Alexandre Nobajas. Keele University. School of Physical and Geographical Sciences. William Smith Building. Staffordshire. ST5 5BG. United Kingdom. A/e: *a.nobajas@keele.ac.uk*. Tel.: 00 44 (0) 1782 7 34962.

thinking about using Twitter as a sociolinguistic research tool will have information about the pros and the cons of using this new data gathering methodology.

KEYWORDS: Twitter, social networks, sociolinguistic methodology, Web 2.0.

1. INTRODUCCIÓ

La recollida de dades per a l'estudi sociolingüístic ha consistit tradicionalment en mètodes com ara enquestes (Martin-Jones, 1991), entrevistes (Briggs, 1986) o grups focals (Litosseliti, 2010), per citar-ne alguns dels més habituals. Aquests mètodes, tot i que encara són perfectament vàlids per a la recerca sociolingüística, pateixen d'una sèrie de limitacions que dificulten en certa mesura el progrés de la disciplina. Per començar, els mètodes esmentats solen ser cars de desenvolupar, ja que per tal de tenir quantitats suficients de participants és necessari dedicar-hi temps i recursos, cosa que implica que cal buscar finançament que en suporti els costos associats (Fowler, 2008), una tasca no sempre fàcil. En segon lloc, els mètodes tradicionals solen requerir una quantitat important de temps per poder ser completats, fet que n'incrementa no només els costos, sinó que fa que la recerca sigui menys àgil i que trobar respostes a qüestions que requereixen prestesa sigui en moltes ocasions difícil (Fowler, 2008). Finalment, com que es tracta de mètodes que en bona part depenen de contactar amb participants que acceptin participar en la recerca, l'àmbit d'actuació se sol limitar a una sèrie de d'individus o comunitats específiques, per la qual cosa la mostra sol ser més limitada del que seria desitjable (Fowler, 2008). Totes aquestes limitacions, que són ben conegudes pels sociolingüistes, no impedeixen la recerca sociolingüística, però en certes ocasions la poden dificultar més del que seria desitjable.

És en aquest context que l'ús de dades provinents de les xarxes socials es presenta com una bona oportunitat per a mitigar les dificultats que afecten la recerca sociolingüística, ja que molts dels problemes que afecten els mètodes tradicionals queden resolts, encara que sigui parcialment, en emprar aquesta nova font d'informació. L'explotació de les dades originades en les xarxes socials ofereix als sociolingüistes una nova eina que fins fa ben poc temps semblava ciència-ficció (Kim *et al.*, 2014). No és estrany doncs que l'aprofitament per part de la sociolingüística d'aquestes dades hagi despertat l'interès de molts investigadors, que hi veuen una font d'informació que pot facilitar, i fins i tot canviar, la manera que tenen de fer recerca (Nguyen, Trieschnigg i Cornips, 2015). Tot i que els potencials beneficis són molt grans, també cal saber que l'ús de dades provinents de xarxes socials no està exempt de riscos i problemes potencials de què cal ser conscients per tal de tenir-los en compte i mitigar-los en la mesura que sigui possible. És per això que, tot i que en un primer moment aquest article presenta alguns dels principals beneficis i oportunitats que les dades provinents de xarxes socials poden aportar a la recerca sociolingüística, a continuació es passen a detallar certes limitacions que cal tenir presents de cara a mitigar-les o, en tot cas, ser-ne conscients.

2. Xarxes socials i recerca sociolingüística

Des de la popularització del terme *web 2.0* a principis de la primera dècada del segle XXI, una nova generació de pàgines web va permetre que els usuaris d'Internet passessin de ser simples observadors passius a tenir la possibilitat d'esdevenir participants actius amb capacitat de crear continguts d'una manera ràpida, senzilla i interactiva (O'Reilly, 2005). Aquest canvi va representar una revolució que a mesura que s'ha anat estenent ha anant creant un nou ecosistema on els usuaris estan més interconnectats que mai, ja que poden crear i consumir continguts a la carta, centrant-se en allò que els interessi més. Aquest canvi de consumidors a creadors de continguts ha significat una evolució en molts aspectes, des dels econòmics fins als socials. Així com fins no fa gaire la creació de continguts amb potencial de difusió estava limitada a uns pocs autors que s'expressaven a través de diaris, ràdios o televisions de capçalera, avui en dia qualsevol persona pot aconseguir que les seves inquietuds arribin a tenir grans audiències (Boccia Artieri i Valeriani, 2013).

A causa de les possibilitats que aquests sistemes proporcionen, la seva popularitat ha crescut de forma exponencial i actualment alguns d'ells tenen centenars de milions d'usuaris (Champoux, Durgee i McGlynn, 2012). Aquesta ubiqüitat implica que una part important de la població interactua tot sovint amb sistemes d'intercanvi de dades socials, per la qual cosa representen una bona mostra poblacional on els usuaris interactuen ja sigui de manera escrita, pictòrica o oral, produint una quantitat de dades ingent. Tota aquesta informació és, per tant, una molt bona oportunitat per a obtenir dades que permetin realitzar estudis sociolingüístics emprant una metodologia diferent de l'habitual.

2.1. *Avantatges del web 2.0 com a mètode de recollida de dades sociolingüístiques*

Els avantatges que ofereix utilitzar dades provinents de pàgines web que encaixin dins la filosofia del web 2.0 són molts i variats, però cal destacar que excel·leixen especialment en el que els mètodes de recopilació tradicional troben dificultats: cost, velocitat i àmbit. El costos d'acumulació de dades provinents de les xarxes socials són força baixos ja que segueixen els principis de les economies d'escala. És a dir, la principal inversió és posar a punt la infraestructura de recopilació de dades, però un cop aquesta està establerta i funciona és indiferent la quantitat de dades que es recopilin: el cost no varia (Carbaugh, 2008). Això significa un canvi de paradigma respecte als mètodes anteriors, on cada enquesta o entrevista feta té un cost, cosa que limita la quantitat de dades que es poden recopilar.

Pel que fa a la velocitat de recopilació de les dades, és molt més gran si s'utilitzen els nous mètodes aquí descrits que no pas amb mètodes tradicionals. Això es deu a la gran quantitat de dades que es generen cada dia i a les quals es pot accedir. Per tant, si cal realitzar un estudi sociolingüístic en un termini breu de temps ara és possible recopi-

lar dades ràpidament, cosa que anteriorment hauria requerit augmentar la dotació de recursos o bé recopilar una quantitat d'informació inferior a la que seria desitjable. Aquesta prestesa en la recopilació de dades és també important, ja que proporciona més flexibilitat a l'investigador, que en qualsevol moment i des de qualsevol lloc pot aplegar les dades necessàries per a la seva recerca sense tenir condicionants temporals o espacials.

Finalment, la tercera característica que representa un avantatge de les dades provinents del web 2.0 respecte als mètodes tradicionals és el seu àmbit d'actuació, que és potencialment global. Així doncs, l'estudi de qualsevol àrea d'interès és en principi possible, indiferentment de la seva població o la seva extensió. Aquesta flexibilitat no va associada a una major feina o a un major cost, un tret força distintiu respecte d'altres mètodes de recopilació de dades. És per això que en aquells casos en què calgui recopilar dades d'una manera econòmica, ràpida i/o àmplia les dades provinents de mètodes que recopilin informació a partir de pàgines web 2.0 tindran un avantatge potencial respecte d'altres mètodes més detallistes.

2.2. *Twitter com a font d'informació sociolingüística*

Dins la gran varietat de tipologies de pàgina web que compleixin amb les característiques del web 2.0, potser les que més populars i les que més bé s'adeqüen a la recerca sociolingüística són les xarxes socials, ja que per definició permeten la interacció entre usuaris en un context que facilita l'intercanvi d'idees lliurement (Kaplan i Haenlein, 2010). Les xarxes socials que, de moment, són més fàcils de gestionar per a recopilar-ne dades automàticament són les que permeten principalment l'intercanvi de textos escrits. Les plataformes de compartició d'imatges i vídeos també poden ser molt útils per a l'estudi sociolingüístic, però l'anàlisi d'aquests formats és força més complex que no pas el de la paraula escrita (Laghos, Masoura i Skordi, 2012).

Dins de les xarxes socials de base principalment textual, Twitter destaca per a la recerca sociolingüística, ja que té una sèrie de característiques que faciliten l'obtenció de dades per a la recerca. En primer lloc, la informació creada dins Twitter és majoritàriament no privativa, per la qual cosa accedir-hi és relativament senzill i fins i tot els termes d'ús de la companyia permeten la descàrrega de dades sense gairebé limitacions (Twitter, 2014). Això és un gran avantatge respecte d'altres xarxes socials que no animen els seus usuaris a ser tan exhibicionistes i on, per tant, la major part de la informació només és visible per als seus creadors i aquells amb qui la vulguin compartir. Aquesta disposició de Twitter i els seus usuaris a cedir la seva informació amb tanta llibertat, encara que aquests últims molts cops no en siguin conscients (Connolly, 2015), queda palesa en la seva interfície de programació d'aplicacions —API en les seves sigles en anglès. Aquesta API permet la descàrrega de manera relativament senzilla de les dades que els usuaris de Twitter generen, de manera que es faciliten la captura i posterior anàlisi dels textos generats pels usuaris.

A més a més, Twitter és usat per una ingent quantitat d'usuaris a tot el món, cosa

que es tradueix en una producció de dades contínua i massiva. Twitter és actualment la desena xarxa social més gran del món per nombre d'usuaris actius, amb gairebé 300 milions (Griffith, 2014), i és el vuitè web més visitat del món (Alexa, 2015*a*). L'any 2013 es produïen 500 milions de piulades al dia (Twitter, 2015*a*), cosa que dóna una idea de la magnitud de la quantitat d'informació generada per aquesta xarxa social. De fet, aquesta contínua producció d'informació no ha passat desapercebuda i institucions com ara la Library of Congress americana ha decidit arxivar tots els tuits públics compartits en aquesta xarxa social a causa de la seva futura utilitat com a memòria històrica de les primeres dècades del segle XXI (Allen, 2013).

En el context dels països de parla catalana Twitter és una de les xarxes socials més populars. Encara que no existeixin estadístiques subestatals fiables, a l'Estat espanyol Twitter és el cinquè web més visitat (Alexa, 2015*c*), mentre que a França és l'onzè (Alexa, 2015*b*). Pel que fa a usuaris totals, l'any 2013 l'Estat espanyol era el sisè país del món amb més usuaris, mentre que França era el dotzè, però en termes relatius Espanya era el tercer estat del món amb més penetració de Twitter en la població, mentre que França era el dissetè (Schoonderwoerd, 2013). Per al Principat d'Andorra no s'han trobat dades disponibles.

Totes aquestes dades demostren que en el context dels Països Catalans Twitter té un grau de penetració molt elevat, fet que té com a conseqüència que és una font d'informació contínua i massiva a partir de la qual es pot treure informació rellevant. A més a més, una important quantitat de piulades duen informació sobre la ubicació des d'on s'ha fet la piulada o bé sobre el lloc d'origen de l'usuari que l'ha realitzat. És per tot això que Twitter s'està emprant cada cop més com una font de dades més a partir de la qual obtenir informació sobre els processos sociolingüístics gairebé a temps real i amb mostres de mides inimaginables si s'haguessin d'aconseguir fent servir altres mètodes més laboriosos (Graham, Hale i Gaffney, 2014).

3. Problemàtiques metodològiques

Tot i els potencials beneficis i virtuts que proporcionen les xarxes socials, hi ha tota una sèrie de reptes i limitacions que afecten la fiabilitat i la correcció de qualsevol recerca feta emprant-les, per la qual cosa cal tenir-los en compte a l'hora d'interpretar les dades i els resultats que aquestes proporcionen. En l'àmbit de la recerca sociolingüística aquestes limitacions es poden dividir en tres classes: les característiques demogràfiques dels usuaris de Twitter, els mètodes de localització dels tuits i la identificació de la llengua emprada.

3.1. *Demografia*

Tot i ser una plataforma molt popular, gratuïta i que proporciona un servei que pot ser utilitzat de manera transversal per tota la població, la demografia dels usuaris de

Twitter no sol representar de manera fefaent la realitat social dels països on és present. Com que es tracta d'una nova tecnologia basada principalment en els telèfons intel·ligents (Twitter, 2015*a*), la major part d'usuaris que l'empra és força més jove que no pas la mitjana de la població (Kiel, 2005). Globalment, el 80 % dels usuaris de Twitter tenen menys de 29 anys (Schoonderwoerd, 2013), cosa que contrasta amb el fet que tan sols el 51 % de la població mundial s'adscriu en aquesta franja d'edat (U. S. Census Bureau, 2013).

En el cas dels països de parla catalana no hi ha dades desagregades, però si s'agafa el cas de l'Estat espanyol com a aproximació demogràfica, el contrast també és important. L'any 2013 la mitjana d'edat dels usuaris de Twitter a l'Estat espanyol era d'uns 23 anys i un 43 % d'ells eren menors de 19 anys (Schoonderwoerd, 2013). En canvi, la mitjana d'edat de la població espanyola en conjunt és molt més elevada, 41,8 anys (Instituto Nacional de Estadística, 2015*b*), i el percentatge de població menor de 19 anys és de tan sols el 19,8 % del total (Instituto Nacional de Estadística, 2015*a*).

Totes aquestes dades mostren que la demografia dels usuaris de Twitter és clarament esbiaixada cap a les generacions més joves, ja que hi són sobrerepresentades d'una manera important. Aquest biaix té dues conseqüències que, en funció de l'ús que es vulgui donar a les dades obtingudes, poden ser considerades positives o negatives. En primer lloc, la no proporcionalitat entre la distribució per edats entre Twitter i la població general implica que les anàlisis efectuades a partir d'aquesta font d'informació han de tenir en compte que els resultats no seran representatius de la societat en general, només ho seran d'un segment específic. Per tant, la segona conseqüència és que si el que es vol estudiar són els usos lingüístics i sociolingüístics de les generacions més joves Twitter es presenta com una plataforma idònia. Si es té en compte que l'estudi dels usuaris d'aquesta xarxa social —o *twitizens* segons un neologisme anglosaxó— representa una mostra gens negligible de les generacions joves, les possibilitats per a l'estudi sociolingüístic són gairebé infinites.

A mesura que la tecnologia va adquirint un rol més central i les noves generacions van integrant diferents dispositius i aplicacions en el seu dia a dia (Walsh, White i Young, 2008), les fronteres entre la vida del món real i la del món virtual es van diluint. Conseqüentment, els usos i hàbits lingüístics dels nadius digitals no es poden estudiar només des de la seva vessant presencial, sinó que cal esbrinar també la manera com la població es comunica en el món virtual ja que pot tenir un impacte en com ho faci en el món real i viceversa, especialment en zones bilingües com els Països Catalans. A tall d'exemple, al Regne Unit un 20 % de les parelles heterosexuals i el 70 % de les homosexuals es coneixen per Internet (Knapton, 2014), no pas de manera presencial. En funció dels usos lingüístics d'aquestes persones quan es comuniquen a través de la Xarxa pot ser que s'estableixi una relació lingüística diferent de la que s'hauria produït si s'haguessin conegut de manera presencial, per la qual cosa també cal tenir en compte i estudiar els usos lingüístics del món virtual. Per tal de concloure aquest apartat, cal dir, doncs, que Twitter és usat principalment per gent jove i que això provoca un biaix metodològic. Aquesta característica potencialment negativa, però, es pot redreçar i interpretar com un fet positiu, ja que ens permet copsar com es comu-

niquen els joves en el món virtual, un món que cada cop és més significatiu per a les seves vides i conseqüentment per a la sociolingüística.

3.2. *Localització*

En les metodologies tradicionals de captura de dades sociolingüístiques l'investigador por establir i definir de manera ben senzilla la seva àrea d'interès, i garantir que les dades s'obtinguin d'aquells individus rellevants per a l'objectiu de la recerca, però quan es treballa amb informació generada a partir de dades obtingudes a través d'Internet tot sovint és difícil saber des de quina ubicació s'han generat. Twitter no n'és una excepció i en molts casos les piulades no tenen cap mena d'informació geogràfica associada, per la qual cosa no es pot saber des de quina ubicació s'han produït, fet que limita de manera gairebé definitiva el seu ús com a font per a la recerca sociolingüística. D'altra banda, una part dels tuits produïts pels usuaris sí que duen informació geogràfica associada. Aquesta informació pot ser de tres tipus: indirecta, la donada pel terminal o la proporcionada per l'usuari.

La major part de les piulades no du informació geogràfica associada, però tot i això en alguns casos es pot arribar a esbrinar des d'on s'han fet mitjançant l'ús de dues tècniques ben diferents: l'observació individual de cada tuit i algoritmes. El primer mètode consisteix a observar aquelles piulades que siguin de l'interès de l'investigador i a intentar deduir, si és possible, la seva ubicació d'acord amb el contingut textual del tuit. Aquest mètode és treballós i lent, i per això només és recomanable en aquells casos en què una petita quantitat de piulades sigui molt rellevant per a un estudi. L'altra opció és emprar algoritmes informàtics que permetin realitzar aquest procés de manera automàtica i, per tant, facilitar àmpliament el processament de grans quantitats d'informació. Actualment el desenvolupament d'aquests mètodes de reconeixement automàtic de les localitzacions és un camp de la recerca geoinformàtica que està generant un gran interès, per la qual cosa és d'esperar que en els propers anys hi hagi grans desenvolupaments en aquest aspecte (Cheng, Caverlee i Lee, 2010).

La segona manera d'obtenir informació geogràfica associada a una piulada és la donada per l'aparell des del qual s'ha fet, ja que, com s'ha indicat anteriorment, el 80 % de les piulades s'efectuen des de tauletes i telèfons intel·ligents que tenen la capacitat de saber les seves coordenades en tot moment. Aquesta informació és emmagatzemada i queda lligada al tuit, de manera que es pot saber des d'on s'han fet les piulades amb dos nivells de precisió. Quan les coordenades les dóna el GPS integrat en l'aparell la precisió és altíssima, amb només uns pocs metres d'error (Zandbergen i Barbeau, 2011). Tot i això, el GPS dels aparells mòbils és conegut per gastar molta bateria ràpidament (Lu *et al*., 2010), i per això la immensa majoria dels usuaris l'apaguen per evitar-ho. La segona manera d'aconseguir les coordenades és mitjançant la triangulació de les antenes de telefonia mòbil (Zhao, 2002). La resolució espacial que proporciona aquest mètode varia molt en funció de la densitat de la xarxa d'antenes: en zones urbanes molt denses es poden aconseguir resolucions d'unes poques dotze-

nes de metres, però en zones rurals aquesta xifra es pot veure espectacularment multiplicada perquè es té una xarxa d'antenes molt menys densa. De tota manera, independentment del mètode que s'hagi fet servir per a obtenir les coordenades del dispositiu, cal que l'usuari hagi donat permís a Twitter per a activar aquesta funcionalitat, i la major part dels tuitaires no el dóna. Ja sigui per estalviar bateria, ja sigui per mantenir un major grau de privadesa, tan sols un 1 % dels usuaris (Mocanu *et al.*, 2013) escullen proporcionar aquesta informació, fet que representa una limitació en la quantitat de dades que es poden recollir.

El tercer tipus d'informació geogràfica que es pot trobar associada a les piulades és aquella que l'usuari té l'opció de proporcionar voluntàriament i que incorporen una bona part dels tuits, entre el 30 % i 50 % en el cas dels Països Catalans. Els usuaris de Twitter tenen l'oportunitat d'incloure una localització associada al seu compte quan el creen o en qualsevol moment posterior. Aquesta informació sol indicar la localitat o zona on resideix l'usuari, cosa que no significa necessàriament que sigui el lloc des del qual s'ha efectuat la piulada. És, per tant, informació sobre d'on és o on resideix l'usuari habitualment, fet que pot tenir una major importància per a l'estudi sociolingüístic que no pas saber des d'on s'ha enviat el tuit. Els llocs on es creen els tuits poden ser localitzacions circumstancials, com ara unes vacances o un viatge de negocis, però saber d'on és la persona pot servir com a base per a estudiar els comportaments lingüístics de certs indrets que siguin de l'interès de l'investigador i per a, fins i tot, crear mapes lingüístics. La informació d'ubicació donada per l'usuari pot ser, per tant, la que resulti més interessant per als estudis sociolingüístics. Malauradament, com que es tracta d'un camp sense limitacions de cap mena, la informació proporcionada pels usuaris sobre el seu lloc d'origen no sempre s'adscriu a la realitat o és correcta. En aquest camp alguns usuaris no hi escriuen el lloc on viuen, hi escriuen els llocs d'origen de persones a qui admiren o indrets amb els quals tenen una relació especial. Per exemple, hi ha una quantitat important de malais i indonesis que, tot i ser d'aquests països asiàtics, indiquen en els seus comptes de Twitter que són de Cervera (Segarra). En indagar una mica més les raons que poguessin explicar aquesta casuística es va trobar que tots ells eren fans del pilot de motociclisme Marc Márquez, que és natural de Cervera i que com a homenatge ells també deien que ho eren. Aquest fenomen dels fans que canvien el seu lloc d'origen pel de la persona a qui admiren també s'ha observat amb músics, actors o altres esportistes (Hecht *et al.*, 2011).

Finalment, com que es tracta d'un camp que l'usuari pot omplir lliurement sense restriccions, els tuitaires tenen la possibilitat d'escriure la localització que els plagui sense que aquesta hagi de ser real o fidedigna. Per una banda aquesta llibertat es pot considerar positiva, ja que alguns usuaris proporcionen informació molt concreta sobre el lloc d'on són, i arriben a indicar el número, el carrer i el municipi, i la gran majoria hi inclou el municipi correctament. Fins i tot hi ha usuaris que hi inclouen dues localitzacions, la seva d'origen i la d'on resideixen, cosa que pot resultar útil per a estudiar les pràctiques lingüístiques de la gent que ha canviat de residència. Malauradament, però, tot sovint en aquest camp s'hi inclouen ubicacions que són ambigües o bé no corresponen a cap localització georeferenciable en un mapa. Les localitzacions

ambigües es poden classificar en tres tipus: 1) les que no indiquen una localitat en particular sinó un país, una regió o un continent, fet que limita l'estudi en detall; 2) les que corresponen a topònims que com que es repeteixen en diferents llocs o per manca de concreció es poden referir a indrets diferents, com ara «Vilafranca» o «Poblenou»; 3) les que, per casualitat, tenen el topònim igual que mots no toponímics, com per exemple «Bot», municipi de la Terra Alta però també terme amb el qual es coneixen els comptes automatitzats de Twitter. Per acabar, les localitzacions no georeferenciables són aquelles que no es poden definir amb les dades proporcionades, com ara «aquí» o «la meva habitació», o bé que corresponen a indrets irreals o imaginaris com, per exemple, «Nàrnia» o «Tràntor».

Quin mètode de localització de piulades es fa servir dependrà doncs, en bona part, de l'objectiu de la recerca a emprendre, però sempre caldrà tenir en compte els avantatges i inconvenients fins ara mencionats a l'hora d'inferir d'on provenen els tuits. Tot i que cada mètode de geolocalització té les seves limitacions, els seus avantatges i els seus inconvenients, hi ha maneres de mitigar els errors o imprecisions que es deriven de les característiques de la informació originada a Twitter. Per exemple, en el cas de la informació proporcionada per l'usuari en el camp «localització», es poden fer servir tècniques de proveïment participatiu (*crowdsourcing* en anglès) per tal d'agilitzar la validació de la informació i així poder treballar amb més dades amb menys esforç per a l'investigador.

3.3. *Llengua*

Tot i que Twitter ha introduït eines per a detectar les llengües en què estan escrits els tuits (Twitter, 2015*b*), actualment només detecta 29 llengües i entre elles no hi ha el català. És per tant necessari recórrer a altres mitjans per tal d'identificar fefaentment i de manera automàtica en quin idioma està escrita cada piulada. Afortunadament per als investigadors en sociolingüística, la identificació automàtica de textos és una tecnologia força avançada, fàcilment disponible i que ja és capaç de distingir entre centenars d'idiomes, entre els quals el català. Tot i això, Twitter no permet en cap cas escriure piulades més llargues de 140 caràcters, per la qual cosa, com que no es té gaire text per a identificar l'idioma, la identificació dels idiomes no sempre és correcta. Per tal d'intentar millorar-ne els resultats és preferible treballar amb piulades de com a mínim 100 caràcters, ja que en cas de treballar amb textos més curts els errors augmenten. De tota manera, la detecció idiomàtica no sempre és correcta ja que hi ha una sèrie de factors que fan que la detecció automàtica proporcioni resultats que poden ser erronis.

El primer problema és que els detectors d'idiomes funcionen mitjançant el reconeixement de paraules i estructures pròpies de cada llengua i justament és per això que és clau que el programa tingui prou text per a efectuar el reconeixement amb solvència. Tot i això, idiomes que comparteixen família lingüística, branca i/o subgrup poden arribar a ser altament intel·ligibles (Gooskens, 2007), i per tant és fàcil que el detector

idiomàtic tingui problemes per a discernir en quin idioma està escrita la piulada. Encara que cal fer més recerca en aquest àmbit, l'experiència mostra que, per exemple, hi ha grans problemes per a distingir automàticament entre gallec-portuguès-espanyol, mentre que entre català-espanyol-francès, tot i haver-hi certes dificultats, els resultats són majoritàriament satisfactoris. Per contra, la distinció idiomàtica en piulades provinents de zones on es parla l'èuscar és gairebé indefectible, amb una taxa d'èxit molt alta a causa de les grans diferències que hi ha entre aquest idioma i els altres parlats al seu voltant. D'altra banda, també cal destacar que no hi ha algoritmes per a detectar tots els idiomes, ja que no s'han desenvolupat o bé no són fàcilment disponibles. Això implica que idiomes com ara l'asturià o l'aragonès no siguin detectables automàticament per alguns dels identificadors d'idioma més populars (Detect Language, 2015), de manera que es limiten en certa mesura les possibilitats de recerca sociolingüística.

L'altre problema per a la detecció automàtica dels idiomes en què estan escrits els tuits és en el contingut, ja que no sempre facilita la identificació de l'idioma, fins i tot quan aquesta tasca es realitza de manera manual. Com que estan escrits principalment mitjançant dispositius mòbils (Twitter, 2015*a*), és fàcil que els usuaris cometin errors tipogràfics o que no siguin gaire curosos en l'ortografia, de manera que s'ocasionen en alguns casos dificultats per al programa de detecció idiomàtica. En altres casos l'ús de barbarismes i neologismes fa que l'algoritme no sigui capaç de saber correctament en quina llengua està escrita una piulada. Finalment, hi ha casos en què les piulades estan compostes de frases escrites en més d'una llengua, i per això el programa normalment identificarà l'idioma preponderant en el tuit, encara que la llengua vehicular sigui una altra.

Totes aquestes dificultats no anul·len l'estudi automatitzat dels idiomes emprats a Twitter, però cal que els investigadors siguin conscients d'aquestes limitacions abans d'arribar a determinades conclusions. Una de les possibles solucions és realitzar un estudi estadístic de la taxa d'error que els detectors d'idioma tenen quan es fan servir juntament amb Twitter. Això permetria ponderar i avaluar si els resultats són adients en cada cas i compensar els resultats en cas que fos necessari.

4. Conclusió

Twitter es presenta com una font d'informació que té un enorme potencial per a la recerca en tots els àmbits de les ciències humanes i socials, però, tot i això, com cada cop que una nova font d'informació apareix, cal tenir en compte que hi ha una sèrie de reptes i riscos que no es poden ignorar per a poder fer-ne un ús efectiu. En el cas de la recerca sociolingüística aquests riscos, però també les virtuts, han estat explorats al llarg de l'article, i s'ha demostrat que, tot i ser possible, cal anar amb cura a l'hora de fer servir les dades provinents de Twitter. Ja sigui per la limitació en la longitud de les piulades que la plataforma imposa, les dificultats per a localitzar els tuits o la demografia dels usuaris, treballar amb dades provinents de Twitter implica ser conscient que no tot seran flors i violes, i per això, inicialment, i fins que els mètodes no siguin

millorats i provats, caldrà emprar aquesta font d'informació amb cura. Inicialment caldrà contrastar que les conclusions obtingudes amb aquesta font d'informació són comparables a les que es puguin obtenir mitjançant els mètodes tradicionals en la sociolingüística, fins i tot combinant tots dos sistemes per a obtenir una visió holística de la realitat.

Tot i aquests reptes i precaucions, l'estudi de diferents disciplines de les ciències humanes i socials mitjançant les xarxes socials, i especialment Twitter, obre un ventall gairebé infinit de possibilitats. Per primer cop a la història es pot obtenir informació de bona part de la població en temps real, de manera extremament econòmica i sense limitacions més enllà de les ja esmentades. Encara més, un cop s'ha emmagatzemat un corpus de piulades significatiu, es poden «fer preguntes» a mida i segons les necessitats de cada moment a quantitats ingents de dades que poden ajudar a obtenir una visió més profunda del món. Tot això hauria de ser motiu d'optimisme entre els investigadors i hauria d'iniciar una nova manera de fer recerca sociolingüística; això sí, sempre tenint en compte els reptes que encara queden per resoldre.

Bibliografia de referència

Alexa (actual. 1 març 2015*a*). «The top 500 sites on the web». A: *Alexa Internet* [en línia]. <http://www.alexa.com/topsites> [Consulta: 25 març 2015].

— (actual. 1 març 2015*b*). «Top sites in France». A: *Alexa Internet* [en línia]. <http://www.alexa.com/topsites/countries/FR> [Consulta: 1 març 2015].

— (actual. 1 març 2015*c*). «Top sites in Spain». A: *Alexa Internet* [en línia]. <http://www.alexa.com/topsites/countries/ES> [Consulta: 25 març 2015].

Allen, E. (actual. 4 gener 2013). «Update on the Twitter Archive at the Library of Congress». A: *Library of Congress* [en línia]. <http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/> [Consulta: 25 març 2015].

Boccia Artieri, G.; Valeriani, A. (2013). «"Racconto le rivoluzioni. Dal basso". Il caso di @tigella tra giornalismo, attivismo, socialmedia curation e celebrità online». *Mediascapes Journal*, núm. 1, p. 11-26.

Briggs, C. L. (1986). *Learning how to ask*: *A sociolinguistic appraisal of the role of the interview in social science research.* Cambridge: Cambridge University Press.

Carbaugh, R. (2008). *International economics*. Mason, Ohio: Cengage Learning.

Champoux, V.; Durgee, J.; McGlynn, L. (2012). «Corporate Facebook pages: when "fans" attack». *Journal of Business Strategy*, vol. 33, núm. 2, p. 22-30.

Cheng, Z.; Caverlee, J.; Lee, K. (2010). «You are where you tweet: a content-based approach to geo-locating Twitter users». A: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management 2010*. Nova York: ACM, p. 759-768.

Connolly, K. (2015). «Germany's Pegida leader steps down over Adolf Hitler photo». *The Guardian* [en línia] (21 gener). <http://www.theguardian.com/world/2015/jan/21/germany-pegida-adolf-hitler-lutz-bachmann> [Consulta: 24 març 2015].

Detect Language (actual. 2015). «Detected languages». A: *Detect Language* [en línia]. <https://detectlanguage.com/languages> [Consulta: 25 març 2015].

FOWLER, F. J. (2008). *Survey research methods*. Los Angeles: Sage.

GOOSKENS, C. (2007). «The contribution of linguistic factors to the intelligibility of closely related languages». *Journal of Multilingual and Multicultural Development*, vol. 28, núm. 6, p. 445-467.

GRAHAM, M.; HALE, S. A.; GAFFNEY, D. (2014). «Where in the world are you? Geolocation and language identification in Twitter». *The Professional Geographer*, vol. 66, núm. 4, p. 568-578.

GRIFFITH, E. (2014). «Twitter co-founder Evan Williams: "I don't give a shit" if Instagram has more users». *Fortune* [en línia] (11 desembre). <http://fortune.com/2014/12/11/twitter-evan-williams-instagram/> [Consulta: 25 març 2015].

HECHT, B.; HONG, L.; SUH, B.; CHI, E. H. (2011). «Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles». A: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2011*. Nova York: ACM, p. 237-246.

INSTITUTO NACIONAL DE ESTADÍSTICA (actual. 2015*a*). «Censos de Población y Viviendas 2011. Pirámide población españoles/extranjeros». A: *Instituto Nacional de Estadística* [en línia]. <http://www.ine.es/censos2011_datos/cen11_datos_inicio.htm> [Consulta: 25 març 2015].

— (actual. 2015*b*). «Indicadores de Estructura de la Población. Edad media de la población según sexo». A: *Instituto Nacional de Estadística* [en línia]. <http://www.ine.es/jaxiT3/Datos.htm?t=3197&L=0> [Consulta: 25 març 2015].

KAPLAN, A. M.; HAENLEIN, M. (2010). «Users of the world, unite! The challenges and opportunities of social media». *Business Horizons*, vol. 53, núm. 1, p. 59-68.

KIEL, J. M. (2005). «The digital divide: Internet and e-mail use by the elderly». *Informatics for Health and Social Care*, vol. 30, núm. 1, p. 19-23.

KIM, S.; WEBER, I.; WEI, L.; OH, A. (2014). «Sociolinguistic analysis of Twitter in multilingual societies». A: *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. Nova York: ACM, p. 243-248.

KNAPTON, S. (2014). «Couples who met online three times more likely to divorce». *The Telegraph* [en línia] (26 setembre). <http://www.telegraph.co.uk/news/science/science-news/11124140/Couples-who-met-online-three-times-more-likely-to-divorce.html> [Consulta: 25 març 2015].

LAGHOS, A.; MASOURA, S.; SKORDI, A. (2012). «Linguistics in social networks». *American International Journal of Contemporary Research*, vol. 2, núm. 1, p. 1-5.

LITOSSELITI, L. (2010). *Research methods in linguistics.* Londres: Continuum.

LU, H.; YANG, J.; LIU, Z.; LANE, N. D.; CHOUDHURY, T.; CAMPBELL, A. T. (2010). «The Jigsaw continuous sensing engine for mobile phone applications». A: *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems 2010*. Nova York: ACM, p. 71-84.

MARTIN-JONES, M. (1991). «Sociolinguistic surveys as a source of evidence in the study of bilingualism: a critical assessment of survey work conducted among linguistic minorities in three British cities». *International Journal of the Sociology of Language*, vol. 90, núm. 1, p. 37-56.

MOCANU, D.; BARONCHELLI, A.; PERRA, N.; GONÇALVES, B.; ZHANG, Q.; VESPIGANI, A. (actual. 2013). «The Twitter of Babel». A: *MOBS Lab* [en línia]. <http://www.ccs.neu.edu/home/qianz/MapTwitterLanguage/v1/index.html> [Consulta: 18 març 2015].

NGUYEN, D.; TRIESCHNIGG, D.; CORNIPS, L. (2015). «Audience and the use of minority languages on Twitter». A: *Proceedings of the Ninth International AAAI Conference on Web and Social Media*. Palo Alto, Calif.: Association for the Advancement of Artificial Intelligence, p. 666-669.

O'REILLY, T. (2005). «What is web 2.0: design patterns and business models for the next generation of software». *O'Reilly* [en línia] (30 setembre). <http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html> [Consulta: 27 febrer 2015].

SCHOONDERWOERD, N. (2013). «4 ways how Twitter can keep growing». *Peerreach* [en línia] (7 novembre). <http://blog.peerreach.com/2013/11/4-ways-how-twitter-can-keep-growing/> [Consulta: 25 març 2015].

TWITTER (actual. 22 octubre 2014). «Developer Agreement & Policy. Twitter Developer Agreement». A: *Twitter* [en línia]. <https://dev.twitter.com/overview/terms/agreement-and-policy> [Consulta: 24 març 2015].

— (actual. 2015*a*). «About Twitter». A: *Twitter* [en línia]. <https://about.twitter.com/company> [Consulta: 25 març 2015].

— (actual. 2015*b*). «GET help/languages». A: *Twitter* [en línia]. <https://dev.twitter.com/rest/reference/get/help/languages> [Consulta: 25 març 2015].

U. S. CENSUS BUREAU (2013). «World midyear population by age and sex for 2013». *U. S. Census Bureau* [en línia] (19 desembre). <https://www.census.gov/population/international/data/idb/worldpop.php> [Consulta: 25 març 2015].

WALSH, S. P.; WHITE, K. M.; YOUNG, R. M. (2008). «Over-connected? A qualitative exploration of the relationship between Australian youth and their mobile phones». *Journal of Adolescence*, vol. 31, núm. 1, p. 77-92.

ZANDBERGEN, P. A.; BARBEAU, S. J. (2011). «Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones». *The Journal of Navigation*, vol. 64, núm. 3, p. 381-399.

ZHAO, Y. (2002). «Standardization of mobile phone positioning for 3G systems». *Communications Magazine, IEEE*, vol. 40, núm. 7, p. 108-116.