Tries lingüístiques a Twitter segons el tema de conversa

Marta Lloret Llinares

Treball Final de Grau de Llengua i literatura catalanes
Universitat Oberta de Catalunya
Juny de 2024

Director: Natxo Sorolla Vidal

Agraïments

L'elaboració d'aquest treball no hauria estat possible sense el guiatge constant de Natxo Sorolla, amb qui he pogut conversar per videoconferència o correu electrònic sobre sociolingüística catalana, vitalitat etnolingüística, la teoria de l'acomodació lingüística o l'anàlisi de les dades. Les nostres converses han estat sempre un plaer i una font d'aprenentatge. Els seus comentaris i suggeriments als esborranys del treball han ajudat a millorar-lo cada vegada i espere que encara aporten molt des d'aquest moment fins el dia de la defensa. Natxo, moltíssimes gràcies per engrescar-te i acceptar dirigir aquest treball!

Les dades han estat un element essencial del treball; és imprescindible agrair molt especialment a Jordi Morales i la Xarxa CRUSCAT la cessió d'aquestes dades. Jordi ha respost sempre qualsevol pregunta sobre la recollida o el processament de les dades.

Aquest treball m'apropa al final del grau, uns estudis que vaig començar fa molt temps, sense saber si mai els acabaria. Ara que ja estic tan a prop, voldria agrair a tots els professors que he tingut en aquest grau la seua dedicació i ensenyament durant tots aquests anys, especialment a Ernest Querol, que va ser el meu tutor a la UOC durant molts anys i professor de l'assignatura Sociolingüística catalana. Sense tot el camí recorregut i els aprenentatges obtinguts, no hauria pogut realitzar aquest treball.

És imprescindible afegir un agraïment a totes les persones, amics, companys i familiars, que m'han sentit parlar aquest darrer any sobre el TFG, sense saber massa bé de què anava, sobretot a Beatriz, Inés i Dani. I molt especialment a Jose, que m'ha ajudat més del que es pensa en aquest viatge en l'ús d'RStudio, l'anàlisi de dades i la mineria de textos, un viatge en què em vaig embarcar un poc inconscientment.

Resum

Decidir quina llengua usar en cada situació és un recurs a l'abast del parlant plurilingüe. La tria depèn de diversos factors relacionats amb cada situació i amb el context social. A l'àmbit del català, s'ha vist que la tria lingüística depèn principalment de la llengua de l'interlocutor. Fins ara, la recerca s'ha centrat principalment en interaccions presencials, però a l'actualitat ens comuniquem en gran mesura a través d'Internet. Per això, hem analitzat les tries lingüístiques a Twitter, una xarxa social on els usuaris escriuen missatges curts, *tuits*, i poden interaccionar entre ells. L'estudi se situa en el marc de la teoria de l'acomodació comunicativa, que explica quan, per què i com les persones ajusten les seues interaccions amb els altres. A partir d'un corpus de tuits de 2021 escrits als territoris de llengua catalana, hem estudiat com el tema d'una conversa explica les tries lingüístiques en aquesta plataforma. Els resultats mostren que, en la majoria de casos, els usuaris s'adapten lingüísticament a l'interlocutor i trien la llengua usada per aquest, seguint el patró esperat segons el marc de la teoria de l'acomodació comunicativa. Hi ha casos, però, en què observem més divergència cap al castellà, que coincideixen amb els temes amb més interaccions iniciades en català, probablement més rellevants per a la comunitat catalanoparlant, com la política catalana.

Mots clau

Tria lingüística, teoria de l'acomodació comunicativa, català, sociolingüística, comunicació per ordinador

Índex de continguts

Índex de taules i figures

1. Introducció

Les persones plurilingües podem triar quina llengua usar en cada situació. Ho fem i ho experimentem sovint: estem amb un grup d'amics i segons a qui mirem, parlem català o castellà; un amic parla castellà amb la parella i català amb el fill; escrivim un mail en castellà i un missatge en una xarxa social en català. Si ens parem a pensar-ho, en alguns casos podem explicar per què ho fem: sempre hem parlat castellà amb aquesta persona, volem que el nostre fill domine les dues llengües, la llengua oficial de l'empresa on enviem el mail és el castellà. Aquestes raons depenen del context social i, per això, les tries lingüístiques i l'alternança de codi són objecte d'estudi de la sociolingüística (Coulmas, 2013).

En una mateixa conversa, els parlants tendeixen a usar la mateixa llengua (Coulmas, 2013, p. 187), però no sempre és així. Per explicar el comportament dels parlants, des de la psicologia social es va proposar la teoria de l'acomodació comunicativa (Giles & Ogay, 2007), que explica la manera, els motius i les conseqüències d'adaptar la nostra manera de comunicar-nos en cada situació. L'adaptació serveix per augmentar o disminuir la distància en la interacció, és a dir, podem convergir o divergir en relació a l'interlocutor. El motiu principal per convergir és provocar una percepció positiva en l'interlocutor, mentre que se sol divergir per marcar diferències, encara que també hi ha altres factors que hi intervenen.

La sociolingüística catalana (Boix & Vila, 1998), des de finals dels anys setanta, ha estudiat les tries lingüístiques en grups o contextos específics amb diverses metodologies i perspectives: enquestes demoscòpiques, estudis observacionals, anàlisi de xarxes socials, etc. A l'actualitat, ens comuniquem cada vegada més a través de la tecnologia: correu electrònic, missatgeria instantània, xarxes socials a internet, etc. La gran quantitat de dades produïdes quan ens comuniquem mitjançant els ordinadors possibiliten l'anàlisi sociolingüística a un nivell massiu, cosa que ha fet emergir la sociolingüística computacional (Nguyen et al., 2016). Una de les plataformes més usades per a l'anàlisi és Twitter (anomenada X a l'actualitat), ja que és eminentment textual i permetia descarregar les dades de manera relativament senzilla fins fa poc.

L'objectiu d'aquest treball és analitzar les tries lingüístiques a Twitter en l'àmbit del català segons el tema de conversa a partir dels principis de la teoria de l'acomodació comunicativa. Començarem fent un repàs més detallat dels treballs esmentats anteriorment i els conceptes més rellevants per a la nostra anàlisi i a continuació presentarem les preguntes de recerca i la metodologia que emprarem. Després presentarem els resultats de l'anàlisi, que

interpretarem a partir de l'actual estat de la qüestió a la discussió del treball, on també exposarem les limitacions d'aquest estudi i futures possibilitats de recerca.

2. Tries lingüístiques

Segons Coulmas (2013, p. 17), el concepte de tria és el més bàsic de la sociolingüística, la disciplina que estudia com els factors socials afecten les tries dels parlants entre les possibilitats que ofereix la llengua i el seu context lingüístic. Per estudiar aquestes tries en el context on ocorren, Dell Hymes (1972, p. 39) va proposar descriure les relacions en una comunitat des d'un punt de vista lingüístic i etnogràfic alhora. Segons Hymes les unitats bàsiques per estudiar el llenguatge en ús són els *fets de parla*, com una conversa entre dos amics. Va proposar una sèrie de huit elements per descriure'ls i analitzar-los que han estat força utilitzats per la sociolingüística: parlants, seqüència dels actes, raons, localització, agents o instruments, normes d'interacció i interpretació, to i tipus de discurs.

A partir de l'estudi de converses en el seu context, ben aviat, sociolingüistes com Gumperz i Auer van establir que l'alternança de codi i la tria lingüística són recursos a l'abast del parlant plurilingüe (Codó, 2016). Un parlant pot decidir interaccionar amb una persona o en una situació determinada en una llengua o en una altra, però també pot decidir dir una paraula o algunes paraules en una llengua diferent de la llengua principal de la conversa. Aquest darrer cas, en què els parlants trien elements en una llengua o en una altra en la mateixa conversa, és considerat alternança de codi (Coulmas, 2013, p. 155). El terme tria de llengües o tria lingüística fa referència a les eleccions que fan els parlants sobre quina llengua usar en una conversa determinada. De tota manera, les situacions reals no són sempre fàcils d'assignar a un terme o l'altre.

La majoria de catalanoparlants coneixen una altra llengua i viuen en contextos on es parlen diverses llengües, per la qual cosa triar una llengua o una altra és un fet quotidià per a bona part de la població dels territoris de llengua catalana. Per això, la sociolingüística catalana s'ha interessat tradicionalment pels usos i les tries lingüístiques. A continuació, explicarem en més detall la situació del català i alguns dels estudis principals en aquestes àrees.

2.1. La situació del català

El català és una llengua que conviu amb una llengua estatal a la majoria de territoris on es parla, amb l'excepció d'Andorra. Mentre que la llengua estatal és sempre oficial, el grau d'oficialitat del català varia: des de l'oficialitat juntament al castellà a Catalunya, les Illes

Balears i el País Valencià, a la no oficialitat a la Catalunya del Nord, l'Alguer i la Franja (Generalitat de Catalunya, s.d.). Els catalanoparlants solen ser almenys bilingües, mentre que els altres habitants dels territoris poden ser monolingües en la llengua estatal. Segons les polítiques lingüístiques i les dinàmiques socials de cada territori, el grau de coneixement del català de la població és més elevat o més baix. Aquesta convivència desigual de les llengües als territoris on es parla català ha fet que la sociolingüística catalana s'haja interessat pels usos i les tries lingüístiques.

Les enquestes sociolingüístiques que es realitzen als territoris de llengua catalana aporten dades sobre els coneixements lingüístics de la població, els usos lingüístics en diferents contextos i les actituds, representacions i opinions que tenen en relació a la llengua. L'informe més recent amb dades dels set territoris es basa en dades de les enquestes realitzades entre 2013 i 2015 (Bretxa et al., 2019). Les dades mostren diferències entre territoris on la llengua té més suport institucional i més prestigi, com Catalunya i Andorra, i altres on aquests són molt més baixos, com L'Alguer i la Catalunya del Nord. A continuació, resumirem les dades de l'informe de Bretxa et al. (2019), centrant-nos sobretot en Catalunya, les Illes Balears i la zona valencianoparlant del País Valencià, els tres territoris més poblats.

La majoria d'habitants dels territoris de llengua catalana té coneixements de català. Més d'un 90% de la població dels territoris de l'estat espanyol declara entendre'l, mentre que a l'Alguer, ho fa un 88% de la població i a la Catalunya del Nord, un 61% (Bretxa et al., 2019, p. 22). Al voltant del 80% de la població dels territoris d'Espanya sap parlar català, mentre que el percentatge és del 50% a l'Alguer i del 35% a la Catalunya del Nord (Bretxa et al., 2019, p. 23). Els percentatges per a la competència lectora són un poc més elevats als tres territoris espanyols on la llengua és cooficial, mentre que a la Franja és del 75% i a l'Alguer i a la Catalunya del Nord no arriba al 40% de la població (Bretxa et al., 2019, p. 24). Quan es pregunta sobre la capacitat per escriure, els percentatges són més baixos a tots els territoris: al voltant del 60% a Catalunya, les Illes Balears i la zona catalanoparlant del País Valencià; un 41% a la Franja; 8% a l'Alguer i 14% a la Catalunya del Nord (Bretxa et al., 2019, p. 25). Aquestes dades porten els autors de l'informe a fer una estimació del nombre de parlants de català: gairebé 12 milions de persones l'entenen; entre 9,5 i 10 milions el saben llegir i parlar; i poc més de 7 milions el saben escriure (Bretxa et al., 2019, p. 36).

Les competències en una llengua són necessàries per poder-la usar, però no són l'únic factor que intervé en la decisió dels parlants d'usar-la, especialment en entorns on conviuen diverses llengües. Per això, a l'hora de descriure la població, ens fixem en els usos lingüístics i les actituds envers la llengua. Els únics territoris on més de la meitat de la població declara

com a llengua habitual el català o el català i l'altra llengua principal són Andorra (60%) i La Franja (56%) (Bretxa et al., 2019, p. 42)[1]. A Catalunya i les Illes Balears aquest percentatge se situa entre el 40 i el 50%. No hi ha dades sobre aquesta pregunta per al País Valencià. A l'Alguer és el 17% de la població i a la Catalunya del Nord, el 6%. Les enquestes també aporten dades sobre els usos lingüístics en diferents contextos privats i institucionalitzats. En els usos amb els amics, al voltant del 60% de la població d'Andorra i de La Franja usa el català o el català i l'altra llengua principal; a Catalunya, les Illes Balears i la zona catalanoparlant del País Valencià, aquest percentatge se situa entre el 40 i el 50% (Bretxa et al., 2019, p. 71). En canvi, si ens fixem en un ús escrit, les notes personals, el percentatge no arriba al 40% en cap dels territoris (Bretxa et al., 2019, p. 70). Pel que fa als usos institucionalitzats, alguns arriben a gairebé el 80% de la població a Andorra i passen del 60% a Catalunya, mentre que a les Illes Balears i al País Valencià tenen valors similars als usos privats, entre el 30 i el 50% segons el context, però amb un major percentatge per a les opcions bilingües que l'opció "només o més català".

Les dades sobre actituds lingüístiques mostren que al voltant del 25% de la població de Catalunya i les Illes Balears inicia converses sempre en català i un percentatge similar ho fa molt sovint (Bretxa et al., 2019, p. 96). Aquests percentatges són més elevats a Andorra (42% i 27%) i a La Franja (40% i 24%). Vora el 38% de la població de Catalunya i el 32% de la població de la zona valencianoparlant del País Valencià consideren que el català s'usarà més en un futur (Bretxa et al., 2019, p. 103). A les Illes Balears, aquest percentatge és del 16%. Els percentatges de població que consideren que s'usarà igual que ara en aquests tres territoris són: 35%, 37% i 46%. Aquestes dades, juntament amb les percepcions sobre l'evolució de l'ús de la llengua, mostren que a Catalunya hi ha una vitalitat etnolingüística subjectiva positiva (Bretxa et al., 2019, p. 104), és a dir, hi ha una percepció positiva pel que fa a l'ús del català, mentre que a la resta de territoris de llengua catalana de l'Estat Espanyol aquesta percepció és d'estabilitat.

El concepte de vitalitat etnolingüística va ser proposat des de la psicologia social per Giles, Bourhis i Taylor (1977), que el van definir com allò que fa probable que un grup es comporte com un col·lectiu diferenciat en relacions intergrupals (p. 308). Així, si la vitalitat d'un grup és alta, és probable que els seus membres mostren la seua identitat diferenciada quan es relacionen amb altres grups, mentre que si la vitalitat és baixa, és més probable que els individus amaguen la seua identitat en aquestes interaccions. Giles, Bourhis i Taylor (1977)

[1] A la gràfica que apareix es mostra "País Valencià" a la segona barra, però hauria de ser "Illes Balears", segons ens han confirmat els autors i d'acord amb el text de l'apartat.

proposen que les variables que contribueixen a la vitalitat estructural pertanyen a tres grups: el demogràfic, el suport institucional i el prestigi (p. 309). Els autors apunten l'existència de dos tipus de vitalitat etnolingüística: objectiva i subjectiva (Giles et al., 1977, p. 318). L'objectiva és el resultat de mesures empíriques i la subjectiva seria la percepció que els membres del grup tenen de la vitalitat. Tant l'una com l'altra poden afectar els comportaments individuals.

2.2. Tries lingüístiques a la sociolingüística catalana

L'any 1978, poc després del final de la dictadura de Franco, Calsamiglia i Tuson (1980) van estudiar les tries lingüístiques al barri de Sant Andreu de Barcelona i van observar que en grups homogenis de catalanoparlants o castellanoparlants s'usava la llengua dels integrants del grup. En canvi, en grups heterogenis els catalanoparlants convergien al castellà del receptor, mentre que els castellanoparlants no usaven el català. Els joves que van participar en aquest estudi havien estat escolaritzats en castellà, però als anys huitanta les coses van canviar i les polítiques de normalització lingüística i la presència del català a les escoles van fer la llengua molt més habitual. Per això, Woolard i Gahng (1990) van comparar les actituds lingüístiques a Barcelona el 1980 i el 1987. El 1980 els castellanoparlants penalitzaven els "castellans" que parlaven català i als catalanoparlants els era indiferent que aquests castellanoparlants parlaren català, així que els autors consideraren que els castellanoparlants no tindrien interès a aprendre català (Woolard & Gahng, 1990, p. 80). En canvi, a l'estudi de 1987, els castellanoparlants troben més normal que algú se'ls adrece en català, ja que els passa habitualment a l'escola i ja no veuen que el català siga una llengua només per als autòctons catalanoparlants.

Pujolar (1993) va estudiar un grup d'universitaris a Barcelona i va observar que, en general, s'adaptaven a la llengua de l'interlocutor i preferien converses monolingües. Els catalanoparlants s'adaptaven amb més freqüència. En aquests universitaris, la tria no es negociava en cada conversa, sinó que «era el resultat d'un procés lligat a la història de la relació amb [cada] persona concreta» (Pujolar, 1993, p. 66). Això podia provocar que en converses on hi havia diversos interlocutors s'anés canviant d'una llengua a l'altra, ja que «aquest acord es mantindrà fins i tot en presència d'altres interlocutors mitjançant l'alternança lingüística, ja que hom assumeix que tothom entén ambdues llengües» (Pujolar, 1993, p. 66).

El fet que les tries lingüístiques a Catalunya estiguen relacionades amb la història de cada relació interpersonal és una de les raons que va portar Rosselló (2010) a estudiar les tries en una aula de parvulari, amb xiquets de 3 i 4 anys que estan començant a socialitzar amb els

companys i desenvolupant habilitats lingüístiques i comunicatives. En aquest context es produïen més converses bilingües i alternances de codi del que s'ha vist en altres grups, per la qual cosa «tot sembla indicar que a l'edat de 3 a 4 anys hi ha díades que no han consolidat encara la llengua d'interacció i que això es produirà més endavant» (Rosselló, 2010, p. 244). Per entendre millor aquest procés, Rosselló i Ginebra (2014) van estudiar els mateixos alumnes huit anys després, quan feien 6è de primària. Van observar que, en la majoria de casos, la llengua d'interacció no s'havia modificat, especialment en interaccions dins del propi grup, de catalanoparlants o castellanoparlants. En canvi, en interaccions exogrupals, hi havia hagut més canvis en la llengua d'interacció: «en general ho han fet per atorgar més pes al castellà. En el cas dels alumnes catalanoparlants inicials quan actuen com a emissors, el català ha donat pas tant als usos bilingües com a l'ús exclusiu o dominant del castellà» (Rosselló & Ginebra, 2014, p. 279). Els autors observen una consolidació de la tria lingüística que segueix els resultats d'altres treballs, però no ocorre el mateix a totes les díades analitzades, així que encara hi ha factors que cal estudiar per comprendre com es produeixen les tries.

Un altre estudi que depassa la fotografia fixa d'un grup en un moment concret és el de Pujolar et al. (2010), que se centra en «les "trajectòries" lingüístiques dels catalans i dels canvis que experimenten en el seu comportament lingüístic al llarg de la vida» (p. 65). Aquests canvis són anomenats «mudes lingüístiques» pels autors i solen coincidir amb canvis en els grups amb qui interaccionen: pas de l'escola a l'institut, arribada a la universitat, canvis de feina, etc. Les noves relacions que s'estableixen en aquests contextos permeten definir noves tries lingüístiques i canviar els usos socials de cada llengua. Normalment aquests nous usos es compaginen amb els que ja estaven establerts amb les coneixences anteriors, de manera que el parlant es troba «en un entorn social que, en certa manera, reflecteix lingüísticament el passat i el present alhora» (Pujolar et al., 2010, p. 73).

Tots aquests estudis s'han centrat en un grup limitat de gent, però també podem analitzar les tries lingüístiques de manera general a la població a partir de les enquestes d'usos lingüístics. Les dades de les enquestes provenen del que declaren les persones enquestades i poden no reflectir la realitat. Les dades de les enquestes realitzades entre 2013 i 2015 (Bretxa et al., 2019) mostren que els parlants que inicien una conversa en català tendeixen a «adoptar a la llengua de l'interlocutor, és a dir, adaptar-se lingüísticament» (p. 97) si aquest interlocutor respon en l'altra llengua oficial al territori en compte de fer-ho en català. En el cas en què inicien la conversa en l'altra llengua oficial i l'interlocutor canvia al català, la tendència majoritària també és adaptar-se. Així, el factor que més pes té a l'hora de triar una llengua o l'altra és la llengua de l'interlocutor.

Hem de tenir present que aquests estudis s'han centrat en converses i relacions presencials. En els darrers anys, amb l'augment de la comunicació per ordinadors, la sociolingüística ha començat a estudiar les tries lingüístiques en aquest context.

3. Sociolingüística computacional

A l'actualitat la comunicació es produeix en bona mesura a través d'ordinadors, tauletes i telèfons mòbils. Els correus electrònics, els missatges i les xarxes socials formen part del dia a dia. Aquestes formes de comunicació generen una gran quantitat de dades que poden ser usades per estudiar les relacions entre llengua i societat i, per tant, la sociolingüística computacional és un camp emergent (Nguyen et al., 2016). La comunicació per ordinadors permet als investigadors obtenir una gran quantitat de dades, però normalment no van acompanyades de tanta informació sobre els usuaris com la que s'obté amb la recerca sociolingüística tradicional amb mostres petites més controlades, per la qual cosa s'han d'adaptar els mètodes d'anàlisi. Per una altra banda, com que es poden aconseguir moltes dades a la vegada, s'obri la possibilitat de respondre a noves preguntes, que els sociolingüistes computacionals hauran de definir (Nguyen et al., 2016, p. 539).

Les xarxes socials digitals permeten obtenir dades massives, però no totes ho fan de la mateixa manera. Com expliquen Sorolla et al. (2017), algunes plataformes, com Facebook, tenen polítiques estrictes de privadesa; i Instagram, Snapchat i TikTok són majoritàriament visuals i no ofereixen tanta riquesa textual. Per això, Twitter ha estat una de les preferides pels investigadors: a més de ser una plataforma basada en textos, fins que Elon Musk la va adquirir a finals de 2022, era relativament senzill obtenir dades per a usos acadèmics. Els missatges a Twitter, anomenats *piulades* o *tuits*, tenen una limitació de 280 caràcters (140 abans de novembre de 2017), però com que qualsevol els pot veure i comentar, es poden produir interaccions entre els usuaris.

En el camp lingüístic, els estudis de dades provinents d'aquesta xarxa social, Twitter o X, s'han centrat en l'anàlisi de les tries lingüístiques i les seues relacions amb variables socials i en l'anàlisi de continguts (Sorolla et al., 2017, p. 42). Kim et al. (2014) i Eleta i Golbeck (2014) van estudiar el paper dels parlants multilingües a Twitter des d'una perspectiva d'anàlisi de xarxes socials, és a dir, analitzant amb qui es relacionaven i en quina llengua ho feien. Els dos estudis van mostrar que els usuaris multilingües actuen de pont entre diverses comunitats. A més, Kim et al. (2014) van analitzar si els parlants bilingües usaven llengües diferents per parlar de temes específics i van observar que era així. A les tres regions

estudiades (Qatar, Suïssa i Quebec), els usuaris empraven la llengua local per parlar de temes polítics i d'informació local per al públic local i, en canvi, usaven l'anglès per parlar d'esdeveniments, turisme i altres temes relacionats amb l'oci (Kim et al., 2014, p. 247).

Coats (2019) també va estudiar les tries lingüístiques a Twitter, en aquest cas als països nòrdics, i les va relacionar amb el gènere dels usuaris. Va trobar que els homes usaven més la llengua local que l'anglès i les dones usaven més l'anglès. Ho va interpretar com una major lleialtat dels homes cap a les identitats locals i una major rapidesa de les dones per adoptar la llengua de prestigi o certs usos lingüístics externs (Coats, 2019, p. 48). A més, va relacionar patrons de l'ús de llengües amb el gènere dels immigrants, ja que hi ha grups de determinats orígens en què la immigració és majoritàriament d'un gènere.

Pel que fa al català, és una llengua bastant usada a Twitter: segons un estudi de Mocanu et al. (2013) és la dinovena del món. Tot i això, encara no ha estat un tema molt tractat per la sociolingüística catalana. Tölke (2015) va fer un estudi sobre l'ús del català a Twitter a la Marina Alta en què analitzava els textos en català, en castellà i bilingües, però no els factors que portaven els usuaris a triar una llengua o l'altra o a canviar d'una a l'altra en el mateix missatge. Guevara (2021) es va fixar en la divulgació de la llengua catalana a Twitter a partir de comptes que en parlen. Morales i Sorolla (sense publicar) han estudiat la minorització lingüística a Twitter a partir de la mesura de les tries lingüístiques en aquesta plataforma.

Aquest treball s'insereix en el marc de la sociolingüística computacional i aporta informació sobre les tries lingüístiques a Twitter a partir de la perspectiva que ens ofereix la teoria de l'acomodació comunicativa.

4. Teoria de l'acomodació comunicativa

La teoria de l'acomodació comunicativa pretén explicar quan, per què i com les persones ajusten les seues interaccions amb els altres, i com aquestes entenen i responen a aquests ajustaments (Giles et al., 2023, p. 1). La teoria va ser proposada inicialment l'any 1973 i ha anat evolucionant per abastar un gran nombre de situacions comunicatives, per exemple: contextos institucionals o professionals com els negocis o l'ensenyament; o comunicació en relació a certes malalties, com la demència o l'autisme. La teoria se centrava en interaccions interpersonals en un principi, però també abasta les relacions intergrupals, ja que la manera com ens comportem com a individus està influenciada pel nostre context sociocultural (Giles & Ogay, 2007).

Un dels principis de la teoria de l'acomodació comunicativa és que els interlocutors usen diverses estratègies per assenyalar les seues actituds envers els altres i els grups socials a què pertanyen. Per això, segons Giles i Ogay (2007, p. 294) la interacció social és un equilibri entre la necessitat de sentir-se inclòs i la necessitat de diferenciar-se dels altres i mantenir la pròpia identitat. A més d'aquestes estratègies, hem de tenir present que cadascú té unes expectatives sobre quanta acomodació s'ha de produir. L'acompliment o no de les expectatives pot marcar la continuació de la interacció. En general, l'adaptació serveix per augmentar o disminuir la distància en la interacció, és a dir, podem convergir o divergir. La convergència o divergència es pot produir en diversos elements, verbals o no verbals: to, registre, dialecte, somriures, pauses, etc. També podem convergir en alguns elements i divergir en altres en la mateixa interacció.

La teoria de l'acomodació comunicativa no es limita a descriure els tipus d'acomodació, sinó que també n'analitza els motius. El motiu principal per convergir és provocar una percepció positiva en l'interlocutor (Giles & Ogay, 2007, p. 296). La convergència, per tant, pot semblar recomanable, però una convergència elevada pot resultar en una pèrdua d'identitat, que no sempre és adequada o desitjada. Per això, un motiu important per divergir és marcar diferències, normalment per destacar la pertinença a un grup determinat, fet que podem observar de vegades en catalanoparlants que no canvien cap a la llengua de l'interlocutor. Els parlants poden divergir per expressar una desafecció o manca de respecte cap a les característiques, comportaments o identitats dels interlocutors (Giles et al., 2006, p. 148). Un altre motiu per divergir pot ser canviar el comportament de l'altre, per exemple: un parlant no nadiu d'una llengua pot marcar més el seu accent per incitar un parlant nadiu a parlar més a poc a poc (Giles & Ogay, 2007).

L'augment de la comunicació per ordinadors va provocar que la teoria de l'acomodació comunicativa s'aplicara també a aquests contextos (Giles et al., 2023). En casos com el correu electrònic o la missatgeria instantània es transmet majoritàriament text, així que l'acomodació ha de ser a través d'elements lingüístics. Aquesta comunicació és diferent a les interaccions presencials, ja que és asíncrona, els interlocutors no hi participen alhora, però s'ha observat que l'acomodació també s'hi produeix.

Si ens centrem en el cas de Twitter, Danescu-Nicolescu-Mizil et al. (2011), van analitzar si es produïa acomodació en l'estil a les converses en aquesta plataforma. Van desenvolupar un marc probabilístic per modelar l'acomodació i el van aplicar a les converses de 7800 usuaris de Twitter. Per mesurar l'estil, van usar el mètode LIWC (Linguistic Inquiry Word Count), que

classifica els mots en categories de significat psicològic, però van eliminar les que fan referència a temes específics i es van centrar en les que fan referència a l'estil (Danescu-Niculescu-Mizil et al., 2011, p. 748). Els autors van mostrar que l'acomodació es produïa a Twitter per a totes les categories analitzades excepte per a la segona persona del plural (Danescu-Niculescu-Mizil et al., 2011, p. 750). Aquest resultat té sentit perquè la segona persona té significats diferents per als dos interlocutors. A més, el mètode emprat i la gran quantitat de dades els va permetre observar que l'acomodació és un fenomen bastant complex i que, a pesar que la tendència general era a l'acomodació, a cada parella d'usuaris es podien donar comportaments diferents: els dos convergien, un convergia i l'altre mantenia el seu comportament habitual, un convergia i l'altre divergia (Danescu-Niculescu-Mizil et al., 2011, p. 751).

La teoria de l'acomodació comunicativa també ha estat aplicada en la comunicació per ordinadors a Catalunya (Nadal, 2021). Aquest treball es va centrar en les tries lingüístiques a WhatsApp: va presentar un mètode per mesurar la convergència lingüística a converses de grups i va observar que la gent tendia a convergir, amb la qual cosa validava la teoria de l'acomodació comunicativa en aquesta plataforma (Nadal, 2021, p. 36). També va observar una tendència a convergir a la llengua majoritària de cada xat analitzat.

5. Preguntes de recerca

Com hem vist, la xarxa social Twitter ha estat poc explorada per la sociolingüística catalana per estudiar les tries lingüístiques, a pesar de la rellevància que les tries lingüístiques han tingut tradicionalment per a la sociolingüística de l'àmbit català. Per això, volem aprofundir en aquestes tries a partir de dades obtingudes d'aquesta plataforma. Ho farem des del marc que proporciona la teoria de l'acomodació comunicativa per analitzar si els usuaris s'adapten als seus interlocutors, és a dir, si convergeixen o divergeixen lingüísticament. El treball no pot abastar tots els factors que poden influir en les tries lingüístiques, així que ens haurem de centrar en algun d'aquests. Com que Twitter no proporciona informació detallada dels usuaris, ens basarem en informació que podem obtenir a partir dels tuits, com van fer Danescu-Nicolescu-Mizil et al. (2011). Així, la nostra pregunta principal de recerca és:

Com s'explica la tria lingüística a Twitter a partir dels temes de conversa?

Per abordar-la, treballarem sobre aquestes preguntes secundàries:

- Com apliquem la teoria de l'acomodació comunicativa a les interaccions a Twitter?

- Quina relació hi ha entre divergència lingüística i conflicte segons el tema de la conversa?
- Com de similars són les observacions fetes a Twitter i les observacions en converses presencials?

Aquestes preguntes secundàries ens ajudaran a contestar la pregunta principal. La teoria de l'acomodació comunicativa aporta el marc per explicar com s'adapten els usuaris i per què ho fan. Com hem vist més amunt, la divergència s'usa per marcar diferències, així que esperem trobar-ne especialment en casos de converses conflictives, on els usuaris tenen postures enfrontades. Per últim, com que les interaccions en una xarxa social com Twitter no són com les converses presencials, serà interessant comparar les nostres troballes amb els resultats d'estudis similars en converses presencials.

6. Corpus de dades

Treballem amb dades de la xarxa social Twitter cedides per la Xarxa CRUSCAT. Són dades de l'any 2021 recopilades per Jordi Morales mitjançant l'API acadèmica de la plataforma per a un projecte de recerca de la Xarxa CRUSCAT-IEC sobre la situació del català a Twitter, en el marc d'un conveni amb la Generalitat de Catalunya.

Per estudiar la convergència i la divergència lingüístiques, cal analitzar interaccions. Per això, el corpus només inclou tuits que són resposta a un altre, que anomenarem "original". Hem delimitat l'àrea als territoris de llengua catalana (Andorra, Catalunya, Catalunya del Nord, Franja de Ponent, Illes Balears, L'Alguer, País Valencià) segons la geolocalització del tuit de resposta, és a dir, el tuit original al qual respon podria provenir d'un altre lloc. En total, el corpus conté 3.412.699 tuits, distribuïts per territoris segons els percentatges que es mostren a la Taula 1. Per a cada tuit, tenim dades de 34 variables, com el número d'identificació de la conversa, de l'usuari o el moment en què s'han escrit, però ens centrarem en les variables següents: text del tuit de resposta, llengua del tuit de resposta, territori del tuit de resposta, text del tuit original, llengua del tuit original, territori del tuit original (en els casos en què hi ha dades).

Taula 1. Distribució dels tuits del corpus (en percentatge sobre el total) per territori.

Territori	Catalunya	País Valencià	Illes Balears	Catalunya del Nord	Andorra	L'Alguer	Franja de Ponent
Percentatge de tuits	64.3%	26.7%	7.8%	0.6%	0.5%	0.1%	0.1%

Les dades cedides per a la realització d'aquest treball contenen la llengua assignada a cada tuit. La detecció d'aquesta llengua va ser realitzada per Jordi Morales de manera automàtica amb l'eina de reconeixement de Google DETECTLANGUAGE (Google, 2024).

7. Metodologia

Per respondre la pregunta de recerca, *Com s'explica la tria lingüística a Twitter a partir dels temes de conversa?*, necessitem conèixer la llengua de cada tuit i establir certs temes de conversa per analitzar. A més, com que estudiem les interaccions a partir del marc de la teoria de l'acomodació comunicativa, hem de mesurar la convergència i la divergència de les interaccions.

A les dades de què partim per fer el treball, descrites a l'apartat anterior, ja tenim la llengua assignada per a cada text. Usem aquesta assignació a la nostra anàlisi i ens centrem principalment en el català i el castellà, les llengües majoritàries, amb un 59,4% de tuits originals assignats al castellà i un 21,5% assignats al català. La següent llengua en percentatge de tuits originals és l'anglès, amb un 7,2%. L'italià, el portuguès i el francès estan assignades a entre un 1% i un 2% dels tuits originals. A més, la llengua de vora un 2% d'aquests tuits està categoritzada com a "indeterminada". Pel que fa als tuits de resposta, els percentatges són similars, però una mica més baixos per a les tres llengües més habituals i lleugerament superiors per a l'italià, el portuguès i el francès. La diferència més notable per als tuits de resposta és que n'hi ha més no assignats a cap llengua: al voltant del 7% pertanyen a la categoria "indeterminat".

Les dades han estat analitzades amb el programa RStudio (versió 2023.12.1) (Posit team, 2023) amb la versió de R 4.3.2 (R Core Team, 2023) mitjançant una sèrie de paquets dissenyats per treballar en la mineria de textos i la visualització dels resultats: dplyr (Wickham et al., 2023), tidytext (Silge & Robinson, 2016), stringr (Wickham, 2023), ggplot2 (Wickham, 2016), wordcloud (Fellows, 2018). Amb aquestes eines, hem seguit un mètode artesanal per analitzar les interaccions a Twitter. A l'hora d'emprar eines que han estat desenvolupades per a l'anàlisi de textos, cal tenir present que no totes funcionen bé per als tuits, que poden usar un llenguatge informal. Per això, cal buscar les més adequades per aquest context o adaptar-les a les necessitats de l'anàlisi. El codi usat en aquest treball es troba a l'annex 1.

7.1. Tractament i anàlisi de les dades

Abans de fer l'anàlisi de les dades, les hem hagut de netejar i preparar per poder-les processar amb les eines utilitzades. Primer, hem eliminat emojis i hem convertit el text del codi UTF-8 al codi ASCII, que només codifica 128 caràcters i simplifica el text, de manera que alguns caràcters habituals a la nostra llengua, com els accents o la *ç* són eliminats i substituïts per la lletra sense accent o la *c*.

A continuació, hem eliminat les paraules buides (Annex 2), és a dir, mots que no aporten significat (Silge & Robinson, 2022), com els articles o les preposicions. Un cop fet això, hem separat les paraules de cada tuit original per tenir-les en línies separades de l'arxiu, ja que això les converteix en variables individuals i facilita el processament de les dades a partir d'aquestes paraules.

Les dades han estat filtrades de manera que només treballem amb els tuits en castellà o català. Per seleccionar paraules i temes amb què treballar, hem extret les paraules més freqüents als tuits originals. Per a cada paraula o grup de paraules seleccionades, s'ha analitzat la convergència i la divergència lingüístiques als parells de tuits on hi apareixen: hi ha divergència si un usuari contesta en català a un tuit en castellà o a l'inrevés; hi ha convergència si un usuari contesta en la mateixa llengua del tuit original, català o castellà. La divergència s'ha calculat com el percentatge de parells de tuits amb l'original en una llengua que eren contestats en l'altra:

$$\frac{número\ de\ parells\ divergents}{número\ de\ parells\ divergents + número\ de\ parells\ convergents} \times 100$$

Per fer això, hem usat les paraules seleccionades per filtrar els tuits originals que les contenen. Com que els tuits rellevants poden estar en castellà o en català, hem tingut en compte que les paraules poden estar en un idioma o l'altre. De vegades, els significants en les dues llengües són el mateix ("política"), així que, incloent una paraula a la llista podem filtrar els tuits en els dos idiomes. En canvi, hi ha casos en què el mateix significat s'expressa amb significants diferents ("felicitat" i "felicidad"). Per incloure possibles tuits en els dos idiomes, hem d'incloure els dos significants a l'anàlisi. Decidir quin és l'equivalent en l'altra llengua no sempre és senzill, ja que hi ha equivalents parcials que no s'usen exactament en els mateixos contextos ("pronto" es pot traduir per "prompte", "aviat", "prest") o significants en un idioma amb més d'un significat i, per tant, amb més d'un equivalent en l'altre idioma ("llama" en castellà pot ser una forma del verb "llamar", que podríem traduir com "cridar"; un

substantiu, traduït com “flama”; o bé un animal, que es diu igual en català). Hem de tenir en compte que la tria d’un significant o un altre pot afectar els resultats.

Després de filtrar els parells de tuits que contenen les paraules seleccionades, a partir de les dades d’assignació de la llengua, hem calculat el número de tuits originals en cada llengua i el percentatge de convergència i de divergència lingüístiques per a cada llengua.

7.2. Selecció de paraules i temes

L’anàlisi de dades descrita anteriorment s’ha realitzat per a les cinc-centes paraules més freqüents als tuits originals i per a 14 temes concrets (Taula 2). La selecció d’aquests temes s’ha realitzat de manera artesanal a partir de les cinc-centes paraules més freqüents. Per exemple, les paraules “liga”, “lliga”, i “jugadors” s’han agrupat sota el tema “esports”. Encara que hem decidit els temes a partir de les cinc-centes paraules més freqüents, als temes hem inclòs altres paraules relacionades amb aquestes, a pesar que no eren tan freqüents. Les relacions poden ser: diverses formes gramaticals d’un lema, traduccions a l’altra llengua o paraules habituals quan es parla d’un tema, com altres partits o polítics. Per exemple, “inesarrimadas” i “miqueliceta” han estat incloses al tema “politica_catalana”, a pesar que no es trobaven entre les cinc-centes paraules més freqüents.

Taula 2. Llista de temes analitzats, les paraules incloses en cadascun d’ells (noteu que les paraules estan escrites amb el codi ASCII, que és com han estat analitzades) i el número de parells de tuits filtrats per tema. Un mateix parell de tuits pot estar inclòs en més d’un tema.

tema	paraules	número de parells de tuits
dona	chica, chicas, dona, dones, madre, madres, mare, mares, mujer, mujeres, noia, noies, xica, xicona, xicones, xiques	55160
esports	equip, equipo, jugador, jugadora, jugadoras, jugadores, jugadors, jugar, liga, lliga, temporada	50002
esports_futbol	champions, equip, equipo, gol, jugador, jugadora, jugadoras, jugadores, jugadors, jugar, liga, lliga, temporada	56372
felicitat	enhorabona, enhorabuena, felic, felices, felicidad, felicidades, felicitat, felicitats, felicos, feliz	45870

geo_catala	barcelona, balear, baleares, balears, catala, catalan, catalana, catalanas, catalanes, catalans, cataluna, catalunya, mallorqui, mallorquin, mallorquina, mallorquinas, mallorquines, mallorquins, valencia, valenciana, valencianes, valencianos, valencians	105753
geo_espanyol	espanya, espana, espanol, espanola, espanolas, espanoles, espanyol, espanyola, espanyoles, espanyols	78989
geo_tot	barcelona, balear, baleares, balears, catala, catalan, catalana, catalanas, catalanes, catalans, cataluna, catalunya, espanya, espana, espanol, espanola, espanolas, espanoles, espanyol, espanyola, espanyoles, espanyols, europa, madrid, mallorqui, mallorquin, mallorquina, mallorquinas, mallorquines, mallorquins, valencia, valenciana, valencianes, valencianos, valencians	208645
negatiu	asco, fastic, mentira, miedo, odi, odio, por, problema, problemas, problemes, vergonya, verguenza, violencia	57641
pandemia	contagi, contagiados, contagiar, contagiats, contagio, contagios, contagis, coronavirus, covid, dosi, dosis, mascareta, mascarilla, pandemia, vacuna, vacunacio, virus	48639
politica	democracia, eleccions, elecciones, gobierno, govern, politics, politicos, votar	83431
politica_catalana	alejandrotgn, cataluna, catalunya, ciudadanoscs, esquerra_erc, erc, generalitat, govern, inesarrimadas, junqueras, juntsxcat, krls, lauraborras, miqueliceta, perearagones, ppcatalunya, socialistescat	87450
politica_espanyola	idiazayuso, pablocasado, pabloiglesias, psoe, sanchezcastejon, vox, yolandadiaz	45298
politica_sobiranista	esquerra_erc, erc, junqueras, juntsxcat, krls, lauraborras, perearagones	33273
positiu	abracada, abrazo, amor, anims, bonic, bonica, bonita, bonito, cor, corazon, encanta, enhorabona, enhorabuena, felic, felices, felicidad, felicidades, felicitat, felicitats, felicos, feliz, genial, guapa, guapo, precios, preciosa, precioso	99307

7.3. Particularitats de les dades

L'ús de dades provinents de Twitter permet analitzar una gran quantitat d'interaccions de nombrosos usuaris, però cal tenir en compte certes característiques de la xarxa social i de la manera de processar les dades per entendre les limitacions d'aquest tipus de dades i interpretar els resultats de manera adequada. La mostra no és representativa de tota la població, sinó que té una major representació de certs grups, principalment adults joves amb nivells formatius alts i amb més presència d'homes que de dones (Sorolla et al., 2017). A més, el fet que els missatges siguen públics pot condicionar el comportament dels usuaris. Per tot això, Sorolla et al. (2017, p. 38) afirmen que «cal tenir en compte que els estudis sobre Twitter analitzen comportaments lingüístics escrits dirigits a un públic difús i de sectors poblacionals joves amb nivells formatius superiors i masculinitzats». A més, la longitud limitada dels textos, juntament amb el llenguatge informal amb préstecs i interferències d'altres llengües, poden dificultar la detecció de la llengua del tuit. De fet, si examinem els tuits individualment, observem errors d'assignació de llengua, més presents per a tuits en català que per a tuits en castellà. Per exemple, aquest tuit està assignat al castellà: "@albertdones @helixx85 @tv3cat sr. Sánchis ja direu alguna cosa, oi ?". Aquest altre també: "@salcarria Niá un pastor ahí a Caudiel que parle valenciá".

7.4. Avaluació de la detecció de llengua

Per fer-nos una idea de l'abast d'aquest error en la detecció de llengües, hem analitzat amb atenció l'assignació d'uns quants centenars de parells de tuits. Si filtrem tuits que contenen la paraula "amor", de 500 parells convergents assignats com a català, cap té errors de detecció de llengua; de 500 parells convergents assignats al castellà, hi ha 3 respostes i 1 original en català. Dels 270 parells divergents cap al castellà, 11 missatges originals (4%) són en català i estan assignats com a castellà. De manera similar, de les respostes, n'hi ha 10 que estan marcades erròniament com a català. En canvi, entre els parells divergents cap al català, 150, hi ha 3 tuits originals (2%) assignats erròniament al català i 80 tuits de resposta (53%) assignats erròniament al castellà. En altres seleccions de tuits després de filtrar per "valencia" o per "tv3cat", hem trobat entre un quart i un terç de tuits escrits en català i assignats erròniament al castellà.

A més, hi ha tuits escrits en català assignats a altres llengües, però com que aquests han estat exclosos de la nostra anàlisi, no hem examinat si aquest fenomen és més freqüent per a tuits en català que per a tuits en castellà. També trobem tuits molt curts, alguns només amb noms, en què no és possible determinar la llengua en què estan escrits. Finalment, trobem

alguns tuits en què hi ha un canvi de codi i, per tant, les dues llengües estan presents, per exemple: “Aquesta cita d’Svetlana Alexiévich a El fin del amor. Una sociología de las relaciones negativas, d’Eva Illouz traduida per Lilia Mosconi 🙌🏽 https://t.co/zQgq2DpY3K”.

Tots aquests fenòmens que observem en el corpus de tuits analitzat han estat descrits anteriorment en regions bilingües de la península Ibèrica en un estudi per comparar sistemes de detecció de llengües (Zubiaga et al., 2016). La similitud de les llengües, com el català i el castellà, i la brevetat dels tuits són factors que dificulten la detecció de llengües (Zubiaga et al., 2016). A més, els programes de detecció de llengües solen estar basats en mètodes estadístics, que funcionen millor per a les llengües més populars, ja que aquestes constitueixen una major proporció de les dades d’entrenament; això és el que van observar Zubiaga et al. (2016) als sistemes de detecció que van comparar. Per aquest motiu, té sentit trobar més errors en l’assignació de llengua als tuits que estan escrits en català.

En principi, els errors de detecció de llengua haurien d’estar distribuïts de manera homogènia per tot el corpus i no afectar els nostres resultats; no tenim cap motiu per pensar que hi haja cap biaix que augmente els errors en relació a temes específics de conversa. Ara bé, hem de tenir present que en aquest estudi ens centrem especialment en un subconjunt minoritari de parells de tuits, els que presenten divergència lingüística. En les interaccions entre dues persones esperem convergència lingüística, segons el que s’ha observat tradicionalment en sociolingüística (Coulmas, 2013, p. 187) i el que postula la teoria de l’acomodació comunicativa (Giles & Ogay, 2007). Per tant, quan seleccionem parells divergents, probablement hi ha més errors d’assignació en tuits, com hem vist en el cas d’”amor”, que, en realitat, constitueixen un parell convergent i no divergent. De tota manera, esperem que aquests errors estiguem distribuïts homogèniament i no afecten les tendències generals dels resultats. Hem de tenir present, però, que les xifres obtingudes no són precises i ser cauts en la interpretació dels resultats.

8. Resultats

8.1. Llengües usades segons el tema

Abans d’analitzar les tries lingüístiques en una interacció, cal tenir en compte la llengua dels tuits originals, per veure si les preferències de llengua estan relacionades d’alguna manera amb els temes d’una conversa. Si analitzem només els parells de tuits en català o castellà, un 26% dels originals són en català i un 74% en castellà. Per als temes en què ens centrem al treball, el que té un menor percentatge de tuits originals en català és “política espanyola”,

amb aproximadament un 15%; i el que té el major percentatge és “política sobiranista”, amb un 86% (Fig. 1). La major part dels temes presenta un menor percentatge de tuits originals en català que en castellà, entre aquest 15% i el 39% del tema “política”. En canvi, observem un major percentatge de tuits originals en català que de tuits originals en castellà en el cas dels temes relacionats amb la política catalana (86% per al tema “política sobiranista” i 73% per al de “política catalana”) i el tema que inclou gentilicis i denominacions de territoris on es parla català (55%).

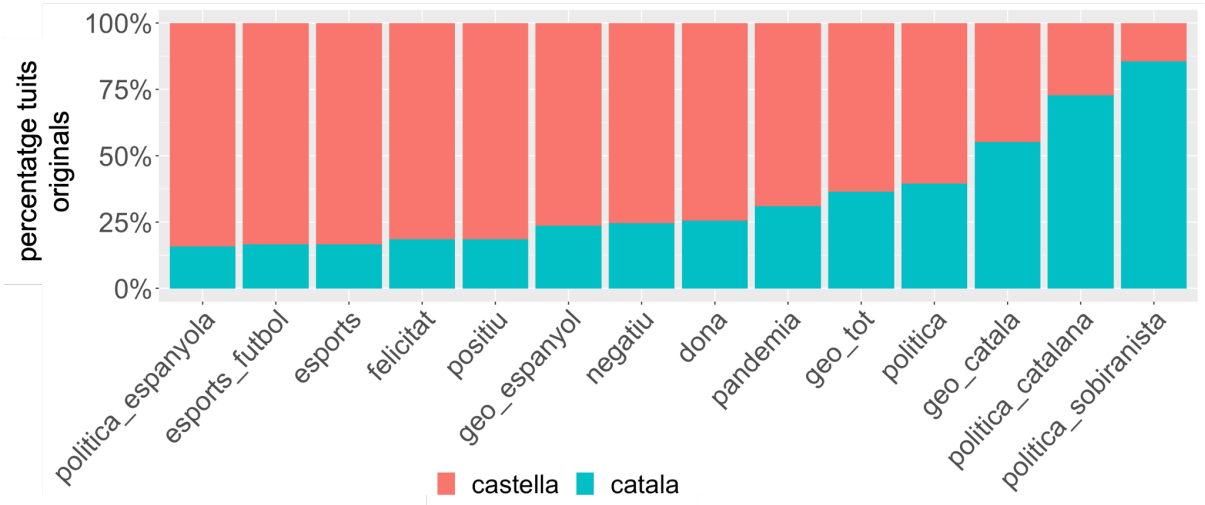

Figura 1. Percentatge de tuits originals en cada llengua per als catorze temes seleccionats, ordenats de menor a major percentatge en català.

8.2. Acomodació lingüística segons el tema

Per estudiar el nivell d’acomodació comunicativa, hem començat per analitzar el nivell general de divergència lingüística a tuits originals en català o castellà que trobem al corpus. Com ja hem vist més amunt, observem que la majoria de tuits al corpus són en castellà (59,4% del total i 74% si només tenim en compte els que són en castellà o català). Si ens fixem en el número segons la combinació de la llengua de l’original i la llengua de la resposta, les parelles en què els dos tuits són en castellà són majoria, amb més d'un milió i mig, al voltant d’un 70% del total (Fig. 2). Un 21% dels parells són en català, gairebé mig milió. Un 3,79% dels parells de tuits tenen l’original en castellà i la resposta en català i un 4,45% tenen l’original en català i la resposta en castellà. Quan calculem la divergència per a cada llengua, com a percentatge dels tuits que responen en la llengua diferent a l’original, per al castellà, la divergència total és al voltant del 5% mentre que per al català és al voltant del 17%.

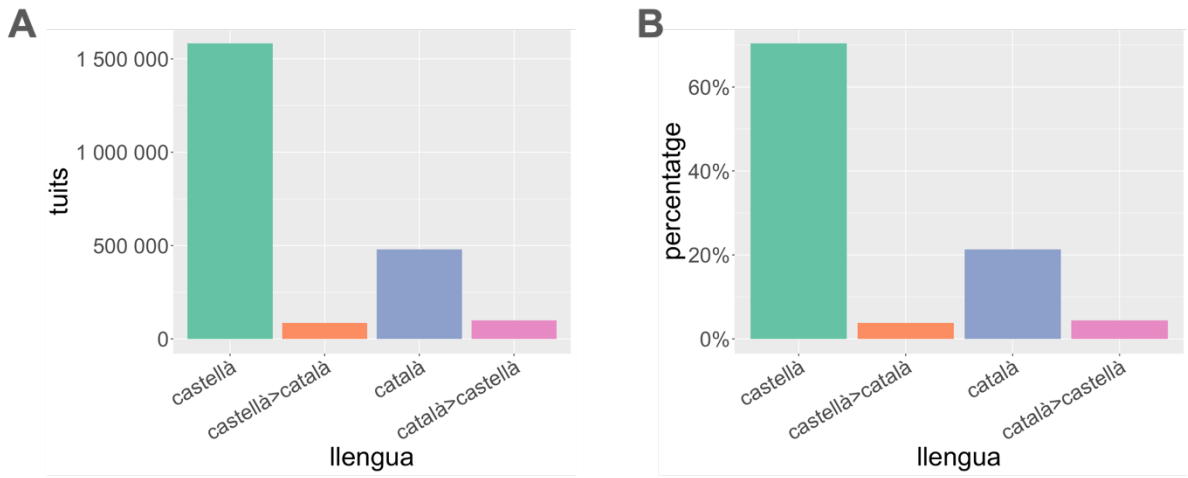

Figura 2. Acomodació lingüística als parells de tuits. Tenint en compte només les interaccions del corpus amb els dos tuits en català o castellà, es mostren: A) número i B) percentatge de parells de tuits segons la llengua de l'original i la resposta.

A continuació, per fer una aproximació al grau d'acomodació lingüística segons el tema de conversa, hem calculat la proporció de tuits en castellà o català que es responen en l'altra llengua per a les 500 paraules més freqüents entre els originals del corpus (Fig. 3B). Les 100 paraules més freqüents es mostren en un núvol de paraules a la figura 3A per donar una idea al lector de les paraules més comunes que trobem a les dades.

Per a la majoria de paraules, la divergència queda en un rang no gaire allunyat del de la divergència observada per al total de parells de tuits del corpus, amb una mediana de 16% per als originals en català i de 4% per als originals en castellà, però tant per a una llengua com per a l'altra, hi ha paraules que mostren una divergència molt elevada als parells de tuits que les contenen. La majoria d'aquestes paraules són noms d'usuaris de Twitter. Alguns exemples que tenen una divergència elevada cap al castellà són: "tv3cat", "lauraborras", "juntsxcat", "esquerraerc". D'altra banda, les paraules amb més divergència cap al català són alguns usuaris privats, paraules en castellà o paraules en altres idiomes com "vous" o "pour". En general, són paraules amb una freqüència baixa als tuits en català (menys de 25 tuits), fet que pot resultar en un percentatge elevat de divergència si hi ha una mala assignació de la llengua. Si ens fixem en les que tenen un número de tuits originals en català un poc més elevat, superior a 200, i presenten una divergència cap al català superior al 25%, trobem "valencia" (podria ser "valencià" o "València"), "ayuso" i alguns noms propis com "juan" o "jose".

Hem analitzat la divergència per a cadascun dels temes seleccionats (Taula 2) i hem observat resultats similars als obtinguts als de l'anàlisi per paraules, amb divergències en general

similars a la divergència total al corpus, però amb alguns valors més extrems. En el cas de parells amb tuits originals en castellà, els temes relacionats amb la política catalana són els que mostren resultats més extrems (Fig. 3C). Així, si els tuits originals contenen paraules relacionades amb partits o polítics catalans sobiranistes, presenten divergència cap al castellà en un 32% dels casos. Si, a més, incloem al tema paraules més generals com "govern" o "generalitat" i altres partits i polítics catalans, la divergència és de gairebé un 19%. El tercer tema pel que fa al grau de divergència cap al castellà supera el 13% i inclou gentilicis del domini del català i algun mot geogràfic de referència a aquesta àrea, com "Catalunya". Aquests tres temes són els que trobem més a la dreta a la gràfica perquè són els que tenen percentatges més elevats de tuits originals en català, d'entre el 55 i el 86% (Fig. 1).

Si ens fixem en la divergència segons el tema per a interaccions amb el tuit original en català (Fig. 3D), no hi ha diferències tan grans com en el cas anterior, però hi ha una lleugera tendència a que els parells amb un major percentatge de tuits originals en castellà (a l'esquerra a la gràfica) són els que presenten una divergència més elevada. Aquests són el de política espanyola (18%) i els dos d'esports (al voltant de 21%). El tema referent a la política sobiranista és el que té la divergència cap al català més baixa (12,7%).

Finalment, podem comparar la divergència cap al català i cap al castellà per a cada tema (Fig. 3C i 3D). Per a la majoria, observem que la divergència cap al català és major que la divergència cap al castellà, com passa per a les dades totals, però per als dos temes que tenen la divergència cap al castellà més elevada, relacionats amb la política catalana i la política sobiranista, la divergència cap al castellà és més elevada. El tema que inclou els gentilicis i mots geogràfics, tercer en el grau de divergència cap al castellà, presenta una divergència similar cap al castellà (14%) i cap al català (17%). Les dades completes de l'anàlisi per temes es poden consultar a l'annex 3.

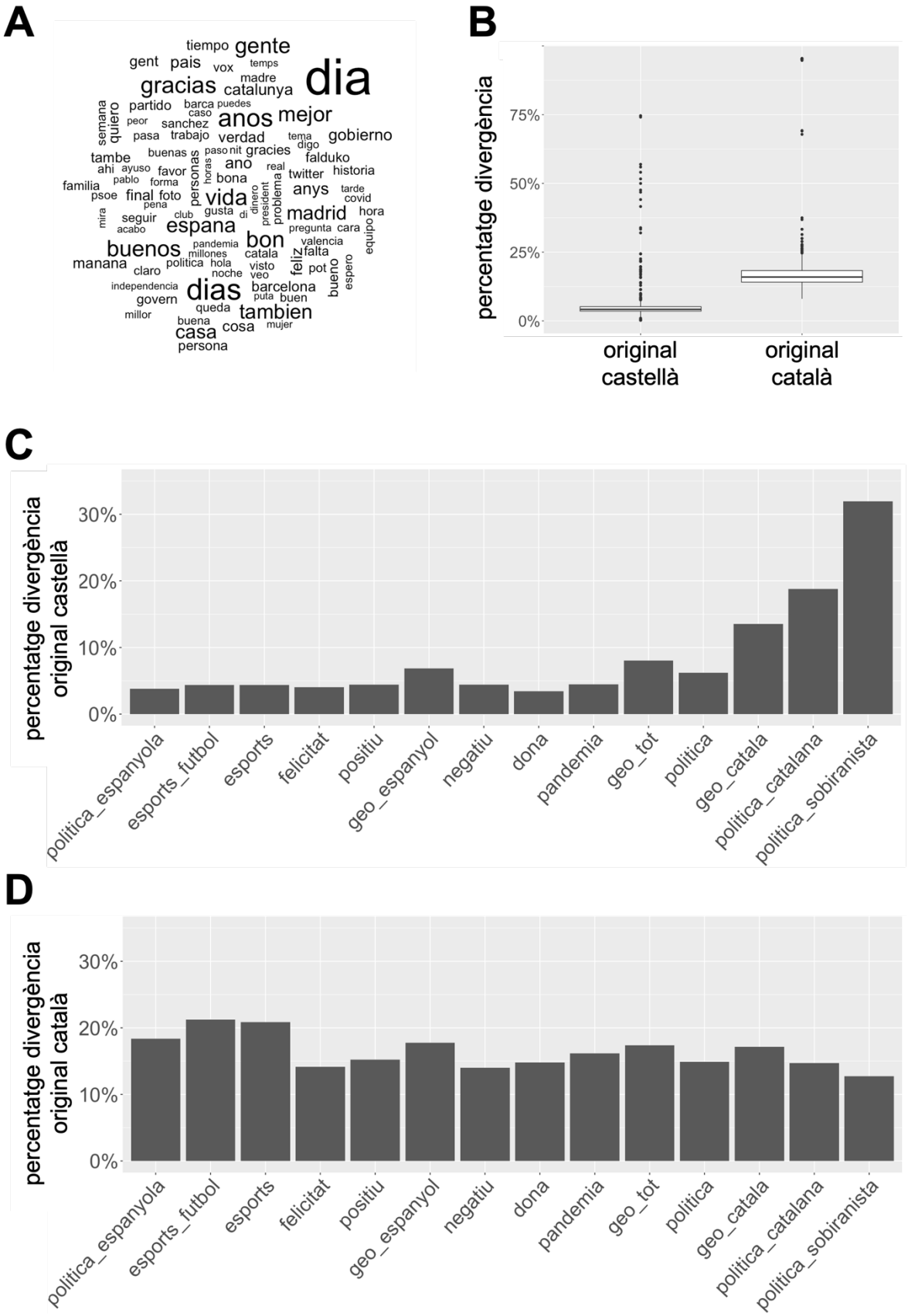

Figura 3. Divergència lingüística. A) Núvol de paraules amb les 100 paraules més freqüents entre els originals. B) Diagrames de caixa que representen els percentatges de divergència en els parells de

tuits amb l'original en la llengua indicada que contenen al tuit original una de les 500 paraules més freqüents o el seu equivalent en l'altra llengua. C) i D) Percentatge de divergència en parells de tuits amb l'original en castellà (C) i en català (D) per als catorze temes seleccionats, ordenats de menor a major percentatge d'originals en català com a la figura 1.

8.3. Anàlisi qualitativa

Després d'obtenir els resultats globals hem realitzat una anàlisi qualitativa per entendre millor què ocorre al corpus i complementar l'anàlisi quantitativa, centrant-nos especialment en casos de divergència. Com que la durada i l'extensió d'aquest treball són limitades, els exemples exposats són una mostra molt xicoteta del corpus de més de 3 milions de parells de tuits.

En examinar els parells de tuits individualment no observem un patró clar que explique per què els usuaris divergeixen lingüísticament. Així, quan ens centrem en el tema sobre la política sobiranista, el que mostra més divergència cap al castellà, trobem un to una mica tens en general i, de vegades, la divergència lingüística pot ser una manera de marcar aquesta tensió i el desacord entre usuaris:

> @govern @perearagones Que digo yo... no sería mejor defender los intereses de TODOS los catalanes, por cumplir con lo de MHP de tota Catalunya.
>
> @tatankasioux @govern @perearagones I a qui dic jo...se pot ser català desconeixent i odiant la llengua cultura i tradicions del teu pais?

També trobem casos en què dos usuaris estan d'acord en el que pensen i divergeixen lingüísticament:

> @MargaXrepublica Creo que es el momento de que todos los Consellers de @JuntsXCat presenten su dimisión en bloque y dejen a ERÑ que se revuelque en su propio fango.
>
> @COJones70182679 @XXXKW2 @MargaXrepublica @JuntsXCat Penso en el mateix. És l'única manera de desemmascarar-los

A més, hi ha alguns textos en què trobem un canvi de codi, perquè hi ha alguna citació i són assignats al castellà:

> En aquesta entrevista en @jcoscu diu: "Subió el pastelero loco y se puso al volante" referint-se al MHP @KRLS https://t.co/YdV4HS4R7o
>
> @torrents_d @jcoscu @KRLS Quin personatge esperpèntic aquest @jcoscu

Per últim, hem de tenir present que la brevetat dels tuits i el fet que els trobem fora de context impedeix entendre la intenció dels autors en alguns casos, sobretot si estan sent irònics o fan referència a altres esdeveniments o missatges:

> Hoy viernes, una magnífica tarde con sol para todos los españoles ! @KRLS @ConsellxRep @boye_g @josepcosta
>
> @tatifurri @KRLS @ConsellxRep @boye_g @josepcosta MOOOOLT SOL☀️🌞☀️🌞🎉🙌

Si ens fixem en els parells de tuits de política sobiranista amb divergència cap al català, observem casos similars, en què pot haver acord o desacord entre els usuaris:

> @KRLS Sisplau, que algú li faci arribar aquest missatge al @jordi_canyas , que a mi em té blocat!¡ "A ver cañaaaas, que no insistas, que no hay rusos, pesaooo!¡"🤣🤣🤣 https://t.co/bxnZeWl1rH
>
> @0punt0 @KRLS @jordi_canyas Cañas, que estas mintiendo, que solo teneis odio y vivis a costa de MENTIR,es por lo unico que estais en UE.como te aguantan con tanta maldad que desprendes, das ASCO.
>
> @KRLS Tots els sants tenin capbuitada,....ahir me'n vaig distreure. Per mooolts i feliços anys president!!!! 🥂 Desitjo de tot cor que els 60 els pugui cel.lebrar a casa. ||*|| 🫂🥰
>
> @12nlbel @KRLS A la casa común de todos los prófugos de la justicia

Si analitzem parells de tuits divergents del tema que conté gentilicis i denominacions geogràfiques dels territoris on es parla català, observem que, encara que els missatges poden ser conflictius, no ho són tan freqüentment com en el cas de la política sobiranista. En aquest cas, també s'observa una diversitat de casos amb divergència. Així, trobem discussions sobre l'ús de l'idioma, però també converses sobre l'aprenentatge d'aquest:

> @gerardpla @EnricGoma Ningú et pot fer canviar al castellà a Catalunya. Això que et va passar et permet posar una queixa Queixes lingüístiques: Generalitat (Direcció General de Política Lingüística) https://t.co/wIKGBIfrBO Plataforma per la llengua: https://t.co/96fN7FjaIg
>
> @EnricGoma Que te den. Si es buena profesional, ¿ a quien interesa en qué idioma hable? Gilipollas
>
> Me enseñáis... a hablar catalán? 😔👉🏽👈🏽
>
> @Ric_Courpetit Llegeix i escolta. La lectura de notícies « conegudes » o veure ràdio, tele, sèries... ajuden moltissim. I mai mai tenir por de parlar. No et preocupis si la pronùncia no és

> la « correcta » o la « normativa ». I per escriure a poc a poc, missagets al whatsapp, emails breus...

També hi ha usuaris que interactuen sobre activitats d'oci o temes de l'actualitat, com la pandèmia de COVID-19:

> 🍁A la Bassa de les Olles ens envolten els colors de tardor. Bon dilluns des de #lampolla ----> ruta natura https://t.co/9p9PB4FK6G #DeltadelEbre #Terresdelebre #Catalunya #natura @terresebretur @catexperience https://t.co/1T9nK0owZy

> @lampollaturisme @terresebretur @catexperience Precioso en media hora estaremos dando un paseo por hay espectacular 😍

> Pero cómo no se va a estabilizar si hay 35 casos. La curva de contagios de COVID en Baleares se «estabiliza» tras unas semanas en descenso https://t.co/uESyJPxcVc vía @Uhmallorca

> @MAmengualSalom @UHmallorca Fa un any teníem 11 casos actius.... 35 nous en un dia, venint d'on venim, està molt bé, però no és per aixecar la guardia del tot. I vist com ens comportam.... No sóc optimista.

L'anàlisi de parells de parells de tuits corresponents a altres temes i paraules, com "amor" o "tv3cat" també mostra una àmplia diversitat de tipus d'interaccions amb divergència.

9. Discussió

En aquest treball ens hem proposat analitzar interaccions a la xarxa social Twitter per entendre com s'explica la tria lingüística en aquesta plataforma a partir dels temes de conversa. Per fer-ho, hem dissenyat un mètode artesanal d'anàlisi a partir del marc que proporciona la teoria de l'acomodació comunicativa. Els resultats obtinguts mostren que la metodologia emprada permet respondre, almenys parcialment, les preguntes de recerca plantejades, encara que fora bo realitzar l'estudi a més gran escala de la que permet un treball de fi de grau per obtenir conclusions més generals.

En els apartats següents respondrem les preguntes de recerca a partir dels resultats obtinguts i relacionarem aquestes dades amb l'estat de la qüestió. Començarem situant-nos en la xarxa social, explicant quins són els usos lingüístics que hi trobem i com es relacionen amb altres dades sobre la població dels territoris de llengua catalana. Aquesta serà la base per endinsar-nos en les tries lingüístiques segons els temes, l'acomodació comunicativa i el conflicte. També ens sembla rellevant destacar algunes limitacions de l'estudi realitzat que permeten

al lector entendre l'abast dels resultats. Finalment, plantejarem possibles vies per ampliar els resultats obtinguts en aquest treball amb la visió de proporcionar idees per a la sociolingüística computacional catalana actual.

9.1. La tria lingüística a Twitter

Al corpus estudiat, centrant-nos només en els parells que contenen tuits en català i castellà, al voltant d'un 74% de les interaccions són iniciades en castellà, mentre que el 26% restant ho són en català. Cal situar aquestes dades en el context de la situació del català, una llengua que ha estat minoritzada al llarg de la història (Boix & Vila, 1998). Com hem vist més amunt, el català és la llengua habitual declarada de menys de la meitat de la població dels territoris de llengua catalana i els percentatges d'ús varien segons el context i el territori (Bretxa et al., 2019). Els tuits són textos escrits, que poden presentar un registre informal, sobretot quan provenen de comptes particulars i no institucionals. En aquest sentit, són similars a les notes personals, de les que tenim dades per alguns territoris a les enquestes sociolingüístiques de 2013-2015: a Catalunya, un 28,9% de la població declara que usa "només català o més català" per escriure aquestes notes; a les Illes Balears, un 22,4% (Bretxa et al., 2019, p. 70). Aquests són dos dels territoris més poblats i on la llengua és cooficial. El segon territori que més tuits aporta al nostre corpus és el País Valencià, del que no tenim dades sobre les notes personals en l'enquesta de 2014. Les dades de 2021 mostren que un 14% de la població de la zona valencianoparlant declara escriure missatges privats per internet predominantment en valencià (Generalitat valenciana, 2022, p. 31).

Al corpus de dades estudiat, si representara tota la població, esperaríem obtenir, com a molt, un percentatge de tuits equivalent al de la població que l'usa per a escriure notes personals o missatges privats. Si tenim en compte que al corpus estudiat també s'inclouen tuits de territoris no esmentats més amunt i que el País Valencià és el segon territori que més n'aporta i que la seua zona castellanoparlant no està representada a les dades anteriors, esperaríem un percentatge inferior al 26% observat. Però hem de tenir present que els usuaris de Twitter, en general, tenen un nivell formatiu alt (Sorolla et al., 2017), segment de la població que, a Catalunya i a Andorra, té el català com a llengua d'identificació en un percentatge més elevat que altres grups, per damunt del 50% (Bretxa et al., 2019, p. 48). A més, al País Valencià, les persones amb estudis superiors són les que declaren saber-lo llegir i escriure en un percentatge més alt (Generalitat valenciana, 2022, p. 12). Per últim, la vitalitat etnolingüística del català també hi juga un paper en l'ús que se'n fa en un espai públic com Twitter. Segons Giles et al. (1977) una vitalitat etnolingüística alta fa probable que els seus membres mostren la seua identitat diferenciada quan es relacionen amb altres grups. A Catalunya, aquesta

vitalitat és bona, tant pel que fa a les dades objectives, amb una bona part de la població que el coneix i l'usa, el suport institucional i el prestigi, com pel que fa a la percepció que en tenen els parlants (Bretxa et al., 2019, p. 104). A les Illes Balears i la zona valencianoparlant del País Valencià, els següents territoris amb més parlants, la situació no és tan favorable al català, però la vitalitat etnolingüística subjectiva no arriba a ser negativa (Bretxa et al., 2019, p. 104), així que és poc probable que els parlants amaguen la seua llengua.

Els resultats d'aquest treball mostren que hi ha certs temes, com la política catalana i els gentilicis o noms de llocs d'àrees catalanoparlants, en què hi ha un major percentatge de tuits originals en català que en castellà. Aquests temes es poden considerar més locals i més rellevants per als grups catalanoparlants, aquells que tenen el català com a llengua habitual o d'identificació. Aquest fet podria explicar el percentatge més elevat, de manera similar a l'ús de llengües locals que van observar Kim et al. (2014) a Quebec, Suïssa i Qatar per parlar sobre temes polítics i d'informació local. De tota manera, el nostre cas no és totalment comparable, ja que el castellà també és una llengua molt parlada als territoris de llengua catalana de l'estat espanyol, on aproximadament la meitat de la població la té com a llengua habitual (Bretxa et al., 2019, p. 42), mentre que Kim et al. comparaven les llengües locals amb una llengua estrangera i global, l'anglès.

9.2. Acomodació comunicativa a Twitter i en converses presencials

En general, al corpus de parells de tuits de 2021 observem convergència lingüística: les respostes són en la mateixa llengua que l'original i, per tant, podem dir que la teoria de l'acomodació comunicativa és vàlida per aquestes interaccions, com s'ha vist anteriorment per a grups de WhatsApp a Catalunya (Nadal, 2021). Aquests resultats segueixen la mateixa línia que el que s'observa en converses presencials. Així, al voltant del 60% de la població de Catalunya i les Illes Balears canvia de llengua si l'interlocutor usa l'altra llengua (Bretxa et al., 2019, p. 98). En interpretar aquestes dades, hem de tenir present que aproximadament un 20% de la població declara no saber català i al voltant d'un 10% no inicia mai converses en català. Diversos estudis sociolingüístics en grups limitats d'estudiants universitaris (Pujolar, 1993) o d'escolars (Rosselló & Ginebra, 2014) també van arribar a la conclusió que la tria lingüística depèn en gran mesura de la llengua de l'interlocutor.

A pesar que la majoria d'interaccions són convergents, observem divergència lingüística en un bon número de parells de tuits. Tant a nivell global com quan analitzem paraules o temes

individuals, la divergència cap al català és, en general, major que la divergència cap al castellà. Una de les raons que explica aquesta diferència és la situació de minorització del català. La comunicació a Twitter és escrita i aquesta és la competència lingüística que menys gent domina en català: aproximadament el 60% de la població de Catalunya, les Illes Balears i la zona catalanoparlant del País Valencià declara saber escriure en català (Bretxa et al., 2019, p. 25). En canvi, el percentatge de població que el sap llegir és superior al 80% (Bretxa et al., 2019, p. 24). Per tant, hi ha gent que pot interactuar amb missatges en català, però que no escriu en aquesta llengua. També hi ha gent que sap llegir i escriure en català, però que prefereix usar el castellà. Com hem vist més amunt, el percentatge que declara escriure notes personals en català a Catalunya no arriba al 30%, bastant més baix que el 60% que declara saber escriure en català (Bretxa et al., 2019). La vitalitat etnolingüística també juga un paper en la llengua usada pels usuaris i, com hem vist, aquesta no és igual a tots els territoris, així que seria interessant analitzar si els percentatges de divergència són diferents segons el territori.

Un altre factor que pot causar la divergència lingüística, en contra del que postula la teoria de l'acomodació comunicativa i del que s'ha vist en converses presencials, és que la natura de les interaccions a Twitter és diferent[2]. No és una plataforma on es mantenen converses privades amb una altra persona o un grup reduït de persones, sinó que el missatge es publica perquè qualsevol el puga veure, encara que el que veu cadascú depèn de les relacions entre seguidors, les mencions i els algorismes de la plataforma. La conversa a Twitter, si es produeix, és asíncrona i qualsevol usuari pot respondre en qualsevol moment, fins i tot anys més tard. A més, la resposta pot tenir una intenció més enllà de respondre el missatge anterior: pot ser una manera de cridar l'atenció sobre aquest a altres usuaris, una manera de mostrar la pròpia postura sobre un tema determinat o d'interactuar amb un grup específic de persones. De fet, observem que algunes de les paraules amb més divergència són noms d'usuaris, tant polítics com usuaris anònims, en molts casos perquè són mencionats per altres usuaris. Per això, l'interlocutor pot ser una persona diferent de l'autor del missatge original, cosa que pot afectar la tria lingüística. A més, la tria pot ser, simplement, una manera de marcar la pròpia identitat, independentment de la llengua del missatge original. Marcar una diferència amb l'interlocutor i mostrar-se com a membre d'un grup diferenciat pot ser un motiu per divergir segons la teoria de l'acomodació comunicativa (Giles & Ogay, 2007). En l'entorn

[2] Per saber-ne més sobre les converses a Twitter, podeu llegir el capítol 4 de *Following Searle on twitter: How words create digital institutions* (Hodgkin, 2017), o la pàgina de la plataforma sobre converses (https://help.x.com/en/using-x/x-conversations).

de Twitter, on els missatges són públics, no hem de pensar només en l'actuació d'un individu enfront d'un altre, sinó en la interacció intergrupal.

Entre els resultats obtinguts, hi ha certs temes en què la divergència cap a una llengua presenta valors més extrems, sobretot en el cas de parells de tuits amb originals en castellà. Els temes amb més divergència cap al castellà, "política sobiranista", "política catalana" i el que agrupa gentilicis i alguns noms geogràfics de l'àmbit català, coincideixen amb els temes en què s'usa més el català. Aquests dos temes de política, a més, són dels que més convergència cap al català tenen. Els resultats obtinguts van en la mateixa línia del que va observar Nadal (2021) en converses de WhatsApp: a les converses on el català era majoritari hi havia més convergència cap al català i més divergència cap al castellà (p. 36). En un estudi d'usos lingüístics en alumnes de 6è de primària dirigit per Vila a finals dels anys noranta, també es va observar una menor convergència cap al castellà en escoles on hi havia més presència de catalanoparlants familiars i una major convergència cap al castellà en aquelles amb menor percentatge d'alumnat catalanoparlant (Vila & Galindo, 2012, p. 40). Per tant, s'observa una tendència a convergir a la llengua majoritària de la conversa; aquest efecte podria ser fins i tot major en un entorn de missatges públics com Twitter, on l'audiència és major, però confirmar això depassa els límits d'aquest treball. El cas dels temes amb més divergència cap al català, els dos d'esports i el de política espanyola, podria ser similar, ja que són els temes amb una major presència del castellà, però la magnitud de l'efecte és molt menor que per als temes amb predomini del català.

9.3. La divergència lingüística a Twitter i el conflicte

La teoria de la comunicació comunicativa postula que la divergència s'usa per marcar diferències. Això pot ocórrer per expressar una desafecció o manca de respecte cap a les característiques, comportaments o identitats dels interlocutors (Giles et al., 2006, p. 148). Amb l'objectiu d'investigar si aquest és el cas als tuits del corpus analitzat, una de les preguntes de recerca plantejades és quina relació hi ha entre divergència lingüística i conflicte segons el tema de conversa.

Aquesta qüestió no és fàcil de respondre, ja que no tenim una manera senzilla de mesurar el conflicte, encara que la nostra experiència personal ens indica que temes com la política o el futbol generen més divisió a la societat que un tema com la felicitat. Una aproximació de la mineria de textos que podria contribuir a resoldre aquesta pregunta és l'anàlisi de sentiments, que classifica els textos segons si les emocions associades a certes paraules o textos són positives, negatives o neutres (Silge & Robinson, 2022). Es poden usar diversos mètodes per

realitzar aquest tipus d'anàlisi (Desai & Mehta, 2016; Silge & Robinson, 2022), normalment basats en lexicons que assignen a cada paraula un valor positiu o negatiu. A més, es pot decidir agrupar més d'una paraula per al càlcul, ja que una partícula negativa pot anul·lar el valor d'una paraula associada a un sentiment positiu: l'expressió "no estic content" indica un sentiment negatiu encara que "content" té un valor positiu. Aquests mètodes depassen l'abast del treball, però seria interessant aplicar-los al corpus de dades. De tota manera, no està clar si serien útils per detectar interaccions conflictives. Normalment s'usen per analitzar respostes a un determinat producte o una campanya política (Desai & Mehta, 2016; El Rahman et al., 2019). En el cas del conflicte i la seua relació amb la divergència, el que interessa és saber si dos usuaris tenen opinions diferents sobre un tema; però el sentiment de cada text per separat podria no donar una idea clara de l'acord o desacord entre ells. Per exemple: un text podria ser positiu cap a un polític d'un bàndol i un altre podria ser positiu cap a un polític del bàndol oposat. En un cas així, els dos presenten sentiments positius, però estan enfrontats en la discussió política.

Si alguns dels temes analitzats presentaren una divergència major que els altres com a conseqüència de ser més conflictius intrínsecament, aquesta divergència elevada s'hauria d'observar tant cap al castellà com cap al català, però no observem aquest efecte, sinó, com hem explicat més amunt, una tendència de major convergència cap a la llengua majoritària d'un tema i, per tant, major divergència cap a la llengua minoritària. Tot i així, encara podria haver una sobrerepresentació d'interaccions conflictives als parells de tuits divergents. Estudiar això de manera sistemàtica depassa els objectius d'aquest treball; només podem analitzar el que hem observat a partir d'exemples concrets, una fracció molt reduïda del corpus. Aquests exemples no indiquen una explicació única per a la divergència, sinó que mostren diverses situacions en què els usuaris divergeixen lingüísticament. És cert que hi ha alguns temes, com la política o interaccions sobre l'idioma, en què el to és més tens, però això ocorre fins i tot quan el tuit original i el de resposta estan d'acord:

> Encara que sembli mentida, aquest personatge forma part de la comissió que ha de decidir sobre si aixecar la immunitat a @KRLS https://t.co/5WheUZtOFI
>
> @DFerrerC @KRLS Tu deberías estar en la prisión de Picosen en tu tierra.Corupto lo peor del pp.Gonzalez Pons, escondido en UE.para librarse de la cárcel fraude en Formula 1.

També trobem interaccions pujades de to i que mostren desacord en parells que presenten convergència lingüística:

> @hugogijon1 Por no hablar de que si por vosotros fuera yo solo hablaría catalán en casa.
>
> @Niil95 @hugogijon1 Eso es rotundamente FALSO y lo sabes.

A més, hi ha un bon número d'interaccions amb divergència lingüística en què no s'observa cap tipus de conflicte, com les dues darreres de l'apartat 8.3 o aquesta:

> os recuerdo que estoy loca y hice esta playlist de las canciones de mitski ordenadas para que cuenten una historia de amor que acaba mal de nada :) https://t.co/44sy6GZmXw
>
> @umaflorzinha__ Ehh volia començar a escoltar mitski osigui que moltes gràcies

9.4. Limitacions de l'estudi

Fins ara hem vist que l'anàlisi realitzada i els resultats obtinguts ens permeten interpretar les dades a partir del marc teòric i extraure algunes conclusions. Cal, però, tenir presents les limitacions de l'estudi per no excedir-se en les interpretacions. En aquest apartat explicarem les limitacions més importants, que també ens serviran per reflexionar sobre possibles direccions d'aquest tipus de recerca en el futur.

El mètode artesanal emprat no permet analitzar les dades de manera exhaustiva, sinó que proporciona una manera d'aproximar-nos a la qüestió de manera exploratòria en què l'anàlisi depèn de les decisions que prenem. Hem seleccionat unes paraules determinades i uns temes determinats. Encara que ens hem guiat per les dades a l'hora de fer-ho, la selecció implica que només hem analitzat de manera sistemàtica una part de les dades, a pesar que, idealment, ho hauríem de fer amb tot el corpus, sense introduir cap biaix a priori. En tot cas, per estudiar subconjunts de les dades, s'haurien de filtrar les dades a partir de decisions empíriques.

Els textos curts i informals que s'usen a Twitter dificulten la tasca de l'assignació de llengües. Com hem vist més amunt, els errors afecten més a tuits en català. A més, els textos de les respostes, per la natura de les interaccions en aquesta xarxa social, solen ser més curts i, per tant, és més probable que hi haja errors en l'assignació de la llengua per a les respostes. Com que ens centrem en casos de divergència, que no són els més habituals, és possible que el nivell d'errors d'assignació siga major en aquests que a tot el corpus. També hem de tenir present que els usuaris que escriuen els missatges són persones: no escriuen perfectament, barregen llengües, usen enllaços, emojis i noms que no corresponen a cap llengua. Encara que fem sociolingüística computacional hem de recordar que el comportament de les persones, l'objecte d'estudi, no és sistemàtic.

Un altre element provinent de les dades de Twitter que pot afectar els resultats és la geolocalització que s'ha usat per seleccionar-los. El percentatge de tuits que estan

geolocalitzats és minoritari i, a més, és l'usuari qui escull si mostra la geolocalització, fet que pot introduir biaixos en la mostra cap a una població més connectada al lloc on viuen.

9.5. Altres possibilitats d'estudi

El mètode i l'anàlisi presentats ací obrin la porta a estudiar de manera més sistemàtica i exhaustiva la tria lingüística a les xarxes socials d'internet, usades extensament a l'actualitat. Aquest treball constitueix, més que res, una tasca exploratòria, una passa ben petita en la sociolingüística computacional catalana. El corpus analitzat es pot estudiar de moltes altres maneres per extraure dades rellevants sobre les tries lingüístiques i l'acomodació comunicativa a Internet. A més, les mateixes anàlisis es poden realitzar en altres corpus de dades. En aquest apartat, plantejarem algunes de les qüestions que es podrien adreçar d'aquesta manera.

Aquí hem analitzat tots els territoris on es parla català de manera conjunta, però la situació sociolingüística és diferent a cada territori com a conseqüència de la història i de les polítiques aplicades a territoris que constitueixen unitats polítiques i administratives diferents. Les dades de les enquestes sociolingüístiques reflecteixen clarament aquesta diversitat (Bretxa et al., 2019). L'anàlisi conjunta impedeix observar dinàmiques diferenciades que de ben segur ajudaran a entendre l'ús del català a Internet. Analitzar els territoris de llengua catalana per separat permetria veure quines similituds i diferències hi ha entre ells i com aquestes es poden relacionar amb els usos lingüístics i la vitalitat etnolingüística a cada territori i, per tant, amb la sèrie de factors demogràfics, institucionals i de prestigi que condicionen aquesta vitalitat.

Hem limitat l'anàlisi als tuits en català i castellà, però es podria ampliar a altres llengües. Al món globalitzat en què vivim, l'anglès és molt present, tant als continguts que trobem per internet com a molts llocs de treball i a les grans ciutats, en què s'usa com a llengua franca internacional. De fet, és la tercera llengua al nostre corpus. Per tant, seria interessant analitzar la convergència i la divergència cap a l'anglès. També es podria mirar què ocorre amb altres llengües.

Pel que fa a la definició dels temes, hem seleccionat paraules individuals o grups de paraules individuals relacionades. La definició de cada tema que obtenim amb aquesta aproximació és molt vaga: la paraula "Barcelona" pot estar inclosa en un tuit de temàtica turística, en un relatiu a l'equip de futbol o en un text informatiu d'un servei de transport públic, per posar tres exemples. Es podrien definir millor els temes a partir d'anàlisis per veure quines paraules apareixen juntes amb més freqüència i quines no solen aparèixer juntes, un tipus d'anàlisi

habitual a la mineria de textos (Silge & Robinson, 2022). Això permetria analitzar de manera més acurada el comportament dels usuaris segons el tema de conversa.

L'anàlisi es podria fer per usuaris, ja que també tenim les dades sobre aquests. Així podríem veure si un mateix usuari escriu sempre en la mateixa llengua, si tria una llengua diferent segons el tema de conversa, segons l'interlocutor o segons si està escrivint un tuit original o responent a un altre tuit. Les dades d'usos lingüístics mostren percentatges diferents segons el context (amb la família, amb els amics, amb l'administració, a les botigues, etc.) i, de fet, a les enquestes sociolingüístiques es mostren opcions de resposta com "català i l'altra llengua principal" (Bretxa et al., 2019), perquè la realitat és que els parlants plurilingües poden triar la llengua depenent de la situació. Així, seria esperable que alguns dels usuaris de Twitter usaren llengües diferents segons el context.

Finalment, el tipus d'anàlisi presentat ací es podria usar per estudiar altres corpus de tuits o altres corpus de textos curts a Internet, delimitats a partir d'altres paràmetres: localització, alguna característica dels usuaris, límits temporals, etc.

10. Conclusions

L'anàlisi d'un corpus de parells de tuits original-resposta amb la resposta geolocalitzada als territoris de llengua catalana ens ha permès comprovar que els usuaris se solen adaptar lingüísticament al seu interlocutor i trien la mateixa llengua usada per aquest. Així, en aquest context se segueix la mateixa tendència que s'observa a partir de la teoria de l'acomodació comunicativa en altres contextos (Giles et al., 2023; Giles & Ogay, 2007) i en estudis sobre tria lingüística realitzats per la sociolingüística catalana en converses presencials (Rosselló & Ginebra, 2014; Pujolar, 1993; Pujolar et al., 2010; Woolard & Gahng, 1990) i a WhatsApp (Nadal, 2021).

Amb la feina realitzada en aquest treball hem establert un mètode d'anàlisi que es pot usar per realitzar altres estudis sociolingüístics amb dades de la xarxa social digital Twitter, que encara no ha estat prou explorada per la sociolingüística catalana.

A més de la tendència general a la convergència lingüística observada, els resultats permeten extraure altres conclusions. En termes generals, hi ha més divergència lingüística cap al català que cap al castellà, probablement com a conseqüència de la situació del català com a llengua minoritària i minoritzada. Per una altra banda, hi ha certs temes, com la política sobiranista, la política catalana o el tema que inclou gentilicis i noms geogràfics catalans, en

què hi ha un predomini del català als tuits originals. Aquests temes són els que presenten una major divergència cap al castellà. En sentit contrari, també observem que els temes amb major predomini del castellà, el de política espanyola i els d'esports, presenten una major divergència cap al català, però la magnitud de les diferències en aquest cas és molt menor. Aquest resultats estan en consonància amb els de Nadal (2021) a WhatsApp i un estudi de Vila en escolars (Vila & Galindo, 2012), que observen més convergència cap al català en contextos on aquest és majoritari. El fet que a una xarxa global d'Internet hi haja certs temes amb predomini del català podria indicar que són temes rellevants per als catalans, però aquesta hipòtesi requereix ser comprovada. Si fos així, els usuaris que hi participen podrien estar més motivats a mostrar la seua identitat i pertinença al grup per diferenciar-se dels no catalanoparlants, una de les raons per divergir postulades per la teoria de l'acomodació comunicativa (Giles & Ogay, 2007).

Pel que fa a la relació entre el conflicte i la divergència lingüística, podem concloure que no hi ha una relació absoluta on sempre que hi ha divergència hi ha conflicte o sempre que hi ha conflicte hi ha divergència. Però l'enfocament limitat que hem pres per analitzar les dades no ens permet anar més enllà ni fer una aproximació sobre la quantitat de conflicte i la seua relació amb la divergència.

Per acabar, voldríem remarcar una de les coses que hem après en analitzar les dades: quan treballem amb textos creats per persones, no podem limitar-nos a fer operacions amb les dades, com si foren números als quals apliquem operacions matemàtiques. És important entendre bé les dades per prendre decisions adequades per a l'anàlisi i per interpretar els resultats correctament. Combinar anàlisis quantitatives globals amb anàlisis qualitatives pot proporcionar una visió més completa de les dades. En el nostre cas, cal tenir en compte que les dues llengües estudiades ací són molt properes; algunes paraules són iguals, però d'altres no ho són, així que si les analitzem de la mateixa manera, els resultats seran diferents. A més, els autors dels textos són persones: fan faltes d'ortografia, alternances de codi, mostren els seus sentiments de diverses maneres, etc. Tot això, juntament amb la brevetat dels tuits, contribueix a errors de detecció de llengua, que poden ser especialment rellevants en estudis de divergència lingüística. En conseqüència, ens sembla que establir els mètodes d'anàlisi més adients per a la sociolingüística catalana per treballar amb aquesta xarxa social o altres mètodes de comunicació per ordinador portarà un temps, però esperem que el nostre granet de sorra puga contribuir mínimament a aquesta tasca.

11. Bibliografia

Boix, E., & Vila i Moreno, F. X. (1998). *Sociolingüística de la llengua catalana*. Editorial Ariel.

Bretxa, V., Sorolla, N., Torres-Pla, J., Torrijos, A., & Vila i Moreno, F. X. (2019). *Els usos lingüístics als territoris de llengua catalana*. Generalitat de Catalunya. Departament de Cultura. Direcció General de Política Lingüística. http://llengua.gencat.cat/ca/serveis/informacio_i_difusio/publicacions_en_linia/btpl_col/usos-linguistics-territoris-llengua-catalana/index.html

Calsamiglia, H., & Tuson, E. (1980). Ús i alternança de llengües en grups de joves d'un barri de Barcelona: Sant Andreu de Palomar. *Treballs de sociolingüística catalana*, 11-82.

Coats, S. (2019). Language choice and gender in a Nordic social media corpus. *Nordic Journal of Linguistics*, *42*(01), 31-55. https://doi.org/10.1017/S0332586519000039

Codó, E. (2016). *La sociolingüística de la interacció*. Fundació Universitat Oberta de Catalunya.

Coulmas, F. (2013). *Sociolinguistics: The Study of Speakers' Choices* (2a ed.). Cambridge University Press.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark My Words! Linguistic Style Accommodation in Social Media. *Proceedings of the 20th international conference on World wide web*, 745-754. https://doi.org/10.1145/1963405.1963509

De Rosselló i Peralta, C. (2010). *Aprendre a triar. L'adquisició de les normes d'ús i alternança de codis en l'educació infantil*. Universitat de Barcelona. https://diposit.ub.edu/dspace/handle/2445/41625

De Rosselló i Peralta, C., & Ginebra Domingo, D. (2014). Tries lingüístiques: Vuit anys després. L'evolució dels usos lingüístics des de P3 fins a 6è de primària. *Treballs de sociolingüística catalana*, 267-280.

Desai, M., & Mehta, M. A. (2016). Techniques for sentiment analysis of Twitter data: A comprehensive survey. *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 149-154. https://doi.org/10.1109/CCAA.2016.7813707

El Rahman, S. A., AlOtaibi, F. A., & AlShehri, W. A. (2019). Sentiment Analysis of Twitter Data. *2019 International Conference on Computer and Information Sciences (ICCIS)*, 1-4. https://doi.org/10.1109/ICCISci.2019.8716464

Eleta, I., & Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, *41*, 424-432. https://doi.org/10.1016/j.chb.2014.05.005

Fellows, I. (2018). *wordcloud: Word Clouds*. https://CRAN.R-project.org/package=wordcloud

Generalitat de Catalunya. (s.d.). *Marc legal*. Llengua catalana. Recuperat 1 gener 2024, de http://llengua.gencat.cat/ca/el-catala/marc-legal/index.html

Generalitat valenciana. (2022). *Coneixement i ús social del valencià. Síntesi de resultats. Enquesta 2021.* Generalitat valenciana.

Giles, H., Bourhis, R. Y., & Taylor, D. (1977). Towards A Theory of Language in Ethnic Group Relations. En *Language, Ethnicity and Intergroup Relations* (p. 307-348). Academic Press London. https://www.scribd.com/document/691900898/Giles-Et-Al-

Towards-a-Theory-of-Language-in-Ethnic-Group-Relations
Giles, H., Edwards, A. L., & Walther, J. B. (2023). Communication accommodation theory: Past accomplishments, current trends, and future prospects. *Language Sciences*, *99*, 101571. https://doi.org/10.1016/j.langsci.2023.101571
Giles, H., & Ogay, T. (2007). Communication Accommodation Theory. Dins *Explaining communication: Contemporary theories and exemplars* (p. 293-310). Lawrence Erlbaum Associates Publishers.
Giles, H., Willemyns, M., Gallois, C., & Anderson, M. C. (2006). *Accommodating a New Frontier: The Context of Law Enforcement*. https://escholarship.org/uc/item/8599410c
Google. (2024). Google Docs Editors Help. *DETECTLANGUAGE*. https://support.google.com/docs/answer/3093278?hl=en
Guevara Claramunt, M. (2021). Català a Twitter, l'àgora de la llengua. *Anuari de Filologia. Estudis de Lingüística*, *11*, 125-140. https://doi.org/10.1344/AFEL2021.11.8
Hodgkin, A. (2017). *Following Searle on twitter: How words create digital institutions*. The University of Chicago Press.
Hymes, D. H. (1972). Models of the Interaction of Language and Social Life. En J. J. Gumperz & D. H. Hymes, *Direction in sociolinguistics: The ethnography of communication* (p. 35-71). Holt, Rinehart and Winston.
Kim, S., Weber, I., Wei, L., & Oh, A. (2014). Sociolinguistic analysis of Twitter in multilingual societies. *Proceedings of the 25th ACM conference on Hypertext and social media*, 243-248. https://doi.org/10.1145/2631775.2631824
Mocanu, D., Baronchelli, A., Gonçalves, B., Perra, N., & Vespignani, A. (2013). The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLoS ONE*, *8*(4), e61981. https://doi.org/10.1371/journal.pone.0061981
Nadal Ferret, M. (2021). *Parlo en la meva llengua o en la teva? Una proposta per a mesurar i analitzar l'acomodació comunicativa en converses de Whatsapp*. Universitat Oberta de Catalunya UOC.
Nguyen, D., Doğruöz, A. S., Rosé, C. P., & De Jong, F. (2016). Computational Sociolinguistics: A Survey. *Computational Linguistics*, *42*(3), 537-593. https://doi.org/10.1162/COLI_a_00258
Posit team. (2023). *RStudio: Integrated Development Environment for R*. Posit Software, PBC. http://www.posit.co/
Pujolar, J. (1993). L'Estudi de les normes d'ús des de l'anàlisi crítica del discurs. *Treballs de sociolingüística catalana*, 61-77.
Pujolar, J., González, I., & Martínez, R. (2010). Les mudes lingüístiques dels joves catalans. *Llengua i ús: revista tècnica de política lingüística*, 65-75.
R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, *1*(3). https://doi.org/10.21105/joss.00037
Silge, J., & Robinson, D. (2022). *Text Mining with R: A Tidy Approach*. https://www.tidytextmining.com/
Sorolla, N., Nobajas, À., & Gras, J. M. i. (2017). Demolingüística, internet i dades massives. *LSC– Llengua, societat i comunicació*, 36-47.

Tolke, V. (2015). L'ús de les llengües minoritàries en les xarxes socials: El valencià en Twitter. *Zeitschrift für Katalanistik*, *28*, 95-115.

Vila, F. X., & Galindo, M. (2012). Sobre la història i l'extensió de la norma de convergència lingüística a Catalunya. Dins: Vila i Moreno, F. Xavier (ed.).2012. *Posar-hi la base. Usos i aprenentatges lingüístics en el domini català*. Barcelona: Institut d'Estudis Catalans, 31-45.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr

Woolard, K. A., & Gahng, T.-J. (1990). Canvis en les avaluacions i les actituds lingüístiques a Barcelona (1980-1987). *Treballs de sociolingüística catalana*, 79-88.

Zubiaga, A., Vicente, I. S., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A., & Fresno, V. (2016). TweetLID: A benchmark for tweet language identification. *Language Resources and Evaluation*, *50*(4), 729-766. https://doi.org/10.1007/s10579-015-9317-4

12.Annexos

Annex 1. Codi usat per analitzar les dades

```
df<-read.csv(file="respostes_originals_ppcc_2021.csv", sep=",")
names(df)
head(df$created_at_resposta)
tail(df$created_at_resposta)
View(df)
saveRDS(df, "df.rds")
library(ggplot2)
ggplot(df)+geom_point(aes(x=created_at_resposta,y=territori_resposta))
library(dplyr)
head(df$territori_resposta)
table(df$territori_resposta) #summary(df$territori_resposta)
round(prop.table(table(df$territori_resposta)) * 100, 1)
table(df$lang_resposta)
round(prop.table(table(df$lang_resposta)) * 100, 1)
table(df$lang_original)
round(prop.table(table(df$lang_original)) * 100, 1)

#### Paraules buides ####
paraules_buides <- scan("paraules_buides_uniques2.csv", character(), quote
 = "")
df_tidy <- df
head(df_tidy$text_original)
df_tidy$text_original_sw <-
 iconv(df_tidy$text_original,from="UTF-8",to="ASCII//TRANSLIT", sub="")
 #Encoding(t2_stem$text_sw)
head(df_tidy$text_original_sw)
df_tidy$text_original_sw<-gsub("[^a-zA-Z 0-9]", "",
 df_tidy$text_original_sw) #això és per canviar els accents com els
 tradueix Mac a la lletra sense l'accent
library(stringr)
df_tidy$text_original_sw <- gsub(paste0("\\b(", paste0(paraules_buides,
 collapse = '|'), ")\\b"), ' ', df_tidy$text_original_sw, perl=TRUE)
df_tidy$text_original_sw <- str_replace(gsub("[[:space:]]+", " ",
 str_trim(df_tidy$text_original_sw)), "B", "b")
df_tidy$text_original_sw <- gsub('^ | $', '', df_tidy$text_original_sw,
 perl=TRUE)
###saveRDS(df_tidy, "df_tidy.rds")
###df_tidy <- readRDS("df_tidy.rds")
### df_tidy ####
library(tidytext)
library(dplyr)
df_tidy <-df_tidy %>%
    unnest_tokens(word, text_original_sw)
names(df_tidy)
names(df_tidy)
head(df_tidy)
tail(df_tidy)

df_tidy_seleccio <- subset(df_tidy, select = c(lang_resposta,
 territori_resposta, text_resposta, word, text_original, lang_original,
 territori_original))

## usar la funció anti_join entre aquest data_frame i el de paraules buides
 per eliminar-les. Ho farà amb la columna word
```

```
paraules_buides<-read.csv("paraules_buides_uniques2.csv", header=TRUE)

df_tidy_seleccio<-df_tidy_seleccio%>%
    anti_join(paraules_buides)

saveRDS(df_tidy_seleccio, "df_tidy_seleccio_filtrat.rds")

##Comptabilitzar

paraules_frequents<-df_tidy_seleccio %>%
    count(word, sort = TRUE)

write.csv(paraules_frequents, "paraules_numvegades.csv")
paraules_frequents<-read.csv(file="paraules_numvegades.csv, header=TRUE")

## per fer word cloud
library(wordcloud)
library(tm)
df_tidy_seleccio %>%
    count(word, sort=TRUE) %>%
    with(wordcloud(word,n, max.words=100))

#### o directament a partir de la llista

wordcloud(paraules_frequents$word, paraules_frequents$n, max.words=100)

###reduir el data frame, perquè pareix que hi ha massa paraules a la
 columna "word" i no el pot processar
df_tidy_seleccio<-df_tidy_seleccio%>%
    mutate(llengues=paste(lang_original, lang_resposta))%>%
    filter(llengues=="ca ca" | llengues=="ca es" | llengues=="es es" |
     llengues=="es ca")


###intentant fer un for loop amb tot això per obtenir freqüències de
 parells de tuits segons la llengua
resultats<-data.frame(paraula=vector(mode = "character", length = 500),
                      catala = vector(mode = "integer", length = 500),
                      cat_cast = vector(mode = "numeric", length = 500),
                      castella = vector(mode = "numeric", length = 500),
                      cast_cat = vector(mode = "numeric", length = 500),
                       row.names = NULL)
for (i in 1:500){
        tuits<-df_tidy_seleccio%>%
        filter(word==paraules_frequents$word[i])%>%
        distinct(text_resposta, .keep_all = TRUE)

    freq<-table(tuits$llengues)
    resultats[i, "paraula"]<-paraules_frequents$word[i]
    resultats[i, "catala"]<-freq[1]
    resultats[i, "cat_cast"]<-freq[2]
    resultats[i, "castella"]<-freq[4]
    resultats[i, "cast_cat"]<-freq[3]
}
```

```
saveRDS(resultats, "divergencia_numtuits.rds")
write.csv(resultats, "divergencia_numtuits.csv")

### calcular convergencies i divergencies a partir d'ahí

dades_divergencia<-data.frame(paraula=vector(mode = "character", length =
 length(resultats$paraula)),
                      catala_div = vector(mode = "integer", length
                       =length(resultats$paraula)),
                      catala_conv = vector(mode = "numeric", length =
                       length(resultats$paraula)),
                      castella_div = vector(mode = "numeric", length =
                       length(resultats$paraula)),
                      castella_conv = vector(mode = "numeric", length =
                       length(resultats$paraula)), row.names = NULL)

for (i in 1:length(resultats$paraula)){
    dades_divergencia[i, "paraula"]<-resultats$paraula[i]
    dades_divergencia[i,
     "catala_div"]<-resultats$cat_cast[i]/(resultats$cat_cast
     [i]+resultats$catala[i])
    dades_divergencia[i,
     "catala_conv"]<-resultats$catala[i]/(resultats$cat_cast
     [i]+resultats$catala[i])
    dades_divergencia[i,
     "castella_div"]<-resultats$cast_cat[i]/(resultats$cast_cat
     [i]+resultats$castella[i])
    dades_divergencia[i,
     "castella_conv"]<-resultats$castella[i]/(resultats$cast_cat
     [i]+resultats$castella[i])

}

### i ara ajuntem els dataframes
resultats_divergencia<-resultats%>%
    inner_join(dades_divergencia)

saveRDS(resultats_divergencia, "resultats_divergencia.rds")
write.csv(resultats_divergencia, "resultats_divergencia.csv")


### provar a tindre un número de la divergència total i fer gràfiques amb
 això
df_tidy_seleccio_divergencia_resposta<-df_tidy_seleccio%>%
    distinct(text_resposta, .keep_all = TRUE)


proporcions_total<-prop
 .table(table(df_tidy_seleccio_divergencia_resposta$llengues))
data<-data.frame(llengua=c("català", "català>castellà", "castellà>català",
 "castellà"), percentatge=c(proporcions_total))
ggplot(data, aes(x=llengua, y=percentatge, fill=llengua)) +
    geom_bar(stat = "identity")+
    scale_fill_brewer(palette = "Set2") +
    theme(legend.position="none") +
```

```
    scale_y_continuous(labels = scales::percent_format()) +
    theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1)) +
    theme(text=element_text(size=30))

nums_total<-table(df_tidy_seleccio_divergencia_resposta$llengues)
nums_tuits<-data.frame(llengua=c("català", "català>castellà",
 "castellà>català", "castellà"), tuits=c(nums_total))
ggplot(nums_tuits, aes(x=llengua, y=tuits, fill=llengua)) +
    geom_bar(stat = "identity")+
    scale_fill_brewer(palette = "Set2") +
    scale_y_continuous(labels=scales::number) +
    theme(legend.position="none") +
    theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1)) +
    theme(text=element_text(size=30))


## provar ggplot scatter plot amb número de tuits i divergencia amb escala
 logarítmica per als resultats amb les 500 paraules més freqüents
ggplot(resultats_divergencia, aes(x=catala_div, y=catala)) +
 scale_y_continuous(trans='log10')
ggplot(resultats_divergencia, aes(x=catala_div, y=catala)) +
 scale_y_continuous(trans='log10') + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)
ggplot(resultats_divergencia, aes(x=castella_div, y=castella)) +
 scale_y_continuous(trans='log10') + geom_point() +
 geom_text(label=resultats_divergencia$paraula)
ggplot(resultats_divergencia, aes(x=castella_div, y=castella)) +
 scale_y_continuous(trans='log10') + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)

ggplot(resultats_divergencia, aes(x=castella_div, y=catala)) +
 scale_y_continuous(trans='log10') + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)
ggplot(resultats_divergencia, aes(x=catala_div, y=castella)) +
 scale_y_continuous(trans='log10') + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)

### canviar x i y
ggplot(resultats_divergencia, aes(x=catala, y=catala_div)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point()
ggplot(resultats_divergencia, aes(y=catala_div, x=catala)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)
ggplot(resultats_divergencia, aes(y=castella_div, x=castella)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point()
ggplot(resultats_divergencia, aes(y=castella_div, x=castella)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)
```

```
ggplot(resultats_divergencia, aes(y=castella_div, x=catala)) +
 scale_x_continuous(trans='log10') + geom_point() +
 scale_y_continuous(labels = scales::percent_format()) +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)
ggplot(resultats_divergencia, aes(y=catala_div, x=castella)) +
 scale_x_continuous(trans='log10') + geom_point() +
 scale_y_continuous(labels = scales::percent_format())
 +geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)

### per tindre les paraules diferents en els dos idiomes amb el càlcul de
 divergència fet segons si n'apareix una o l'altra
dosidiomes<-read.csv(file="paraules_dosidiomes.csv", header=TRUE, sep=",")

resultats2<-data.frame(paraula=vector(mode = "character", length =
 length(dosidiomes$word)),
                       paraulaB=vector(mode = "character", length =
                        length(dosidiomes$word)),
                       catala2 = vector(mode = "numeric", length
                        =length(dosidiomes$word)),
                       cat_cast2 = vector(mode = "numeric", length =
                        length(dosidiomes$word)),
                       castella2 = vector(mode = "numeric", length =
                        length(dosidiomes$word)),
                       cast_cat2 = vector(mode = "numeric", length =
                        length(dosidiomes$word)),
                       catala_div2 = vector(mode = "numeric", length
                        =length(dosidiomes$word)),
                       catala_conv2 = vector(mode = "numeric", length =
                        length(dosidiomes$word)),
                       castella_div2 = vector(mode = "numeric", length =
                        length(dosidiomes$word)),
                       castella_conv2 = vector(mode = "numeric", length =
                        length(dosidiomes$word)), row.names = NULL)

for (i in 1:length(dosidiomes$word)){
    tuits2<-df_tidy_seleccio%>%
        filter(word==dosidiomes$word[i] | word==dosidiomes$word_B[i])%>%
        distinct(text_original, .keep_all = TRUE)

    freq<-table(tuits2$llengues)
    resultats2[i, "paraula"]<-dosidiomes$word[i]
    resultats2[i, "paraulaB"]<-dosidiomes$word_B[i]
    resultats2[i, "catala2"]<-freq[1]
    resultats2[i, "cat_cast2"]<-freq[2]
    resultats2[i, "castella2"]<-freq[4]
    resultats2[i, "cast_cat2"]<-freq[3]
    resultats2[i,
     "catala_div2"]<-resultats2$cat_cast2[i]/(resultats2$cat_cast2
     [i]+resultats2$catala2[i])
    resultats2[i,
     "catala_conv2"]<-resultats2$catala2[i]/(resultats2$cat_cast2
     [i]+resultats2$catala2[i])
    resultats2[i,
     "castella_div2"]<-resultats2$cast_cat2[i]/(resultats2$cast_cat2
     [i]+resultats2$castella2[i])
```

```
    resultats2[i,
     "castella_conv2"]<-resultats2$castella2[i]/(resultats2$cast_cat2
     [i]+resultats2$castella2[i])
}

saveRDS(resultats2, "dades_divergencia_dosidiomes.rds")
write.csv(resultats2, "dades_divergencia_dosidiomes.csv")

### i ara els ajunte amb els d'ahir d'un idioma
resultats_divergencia_dosidiomes<-resultats_divergencia%>%
    full_join(resultats2)

saveRDS(resultats_divergencia_dosidiomes,
 "resultats_divergencia_dosidiomes.rds")
write.csv(resultats_divergencia_dosidiomes,
 "resultats_divergencia_dosidiomes.csv")


### per no tindre gracies encara amb les dades dolentes. A més, afegim
 número de tuits originals i proporcions, per tindre-ho ja tot en el mateix
 df (en prova1, açò ho vaig afegir més endavant)

for (i in 1:length(resultats_divergencia_dosidiomes$paraula)){
    if (is.na(resultats_divergencia_dosidiomes$catala_div2[i]) &
     resultats_divergencia_dosidiomes$paraula[i] %in%
     resultats_divergencia_dosidiomes$paraulaB){
        resultats_divergencia_dosidiomes[i, "catala_divergencia"]=NA
        resultats_divergencia_dosidiomes[i, "castella_divergencia"]=NA
        resultats_divergencia_dosidiomes[i, "catala_original"]= NA
        resultats_divergencia_dosidiomes[i, "castella_original"]= NA
        resultats_divergencia_dosidiomes[i, "catala_proporcio"] = NA
        resultats_divergencia_dosidiomes[i, "castella_proporcio"] = NA
    }
    if (is.na(resultats_divergencia_dosidiomes$catala_div2[i]) &
     !resultats_divergencia_dosidiomes$paraula[i] %in%
     resultats_divergencia_dosidiomes$paraulaB){
        resultats_divergencia_dosidiomes[i,
         "catala_divergencia"]=resultats_divergencia_dosidiomes$catala_div
         [i]
        resultats_divergencia_dosidiomes[i,
         "castella_divergencia"]=resultats_divergencia_dosidiomes$castella_d
         iv[i]
        resultats_divergencia_dosidiomes[i, "catala_original"]=
         resultats_divergencia_dosidiomes$catala
         [i]+resultats_divergencia_dosidiomes$cat_cast[i]
        resultats_divergencia_dosidiomes[i, "castella_original"]=
         resultats_divergencia_dosidiomes$castella
         [i]+resultats_divergencia_dosidiomes$cast_cat[i]
        resultats_divergencia_dosidiomes[i, "catala_proporcio"] =
         resultats_divergencia_dosidiomes$catala_original
         [i]/(resultats_divergencia_dosidiomes$catala_original
         [i]+resultats_divergencia_dosidiomes$castella_original[i])
```

```
        resultats_divergencia_dosidiomes[i, "castella_proporcio"] =
         resultats_divergencia_dosidiomes$castella_original
         [i]/(resultats_divergencia_dosidiomes$catala_original
         [i]+resultats_divergencia_dosidiomes$castella_original[i])
    }
    else{resultats_divergencia_dosidiomes[i,
     "catala_divergencia"]=resultats_divergencia_dosidiomes$catala_div2[i]
    resultats_divergencia_dosidiomes[i,
     "castella_divergencia"]=resultats_divergencia_dosidiomes$castella_div2
     [i]
    resultats_divergencia_dosidiomes[i, "catala_original"]=
     resultats_divergencia_dosidiomes$catala2
     [i]+resultats_divergencia_dosidiomes$cat_cast2[i]
    resultats_divergencia_dosidiomes[i, "castella_original"]=
     resultats_divergencia_dosidiomes$castella2
     [i]+resultats_divergencia_dosidiomes$cast_cat2[i]
    resultats_divergencia_dosidiomes[i, "catala_proporcio"] =
     resultats_divergencia_dosidiomes$catala_original
     [i]/(resultats_divergencia_dosidiomes$catala_original
     [i]+resultats_divergencia_dosidiomes$castella_original[i])
    resultats_divergencia_dosidiomes[i, "castella_proporcio"] =
     resultats_divergencia_dosidiomes$castella_original
     [i]/(resultats_divergencia_dosidiomes$catala_original
     [i]+resultats_divergencia_dosidiomes$castella_original[i])}
}
saveRDS(resultats_divergencia_dosidiomes,
 "resultats_divergencia_dosidiomes.rds")
write.csv(resultats_divergencia_dosidiomes,
 "resultats_divergencia_dosidiomes.csv")


#### i ara fem uns plots amb la nova divergència
ggplot(resultats_divergencia_dosidiomes, aes(x=catala,
 y=catala_divergencia)) + scale_x_continuous(trans='log10') +
 scale_y_continuous(labels = scales::percent_format()) + geom_point()
ggplot(resultats_divergencia_dosidiomes, aes(y=catala_divergencia,
 x=catala)) + scale_x_continuous(trans='log10') + scale_y_continuous(labels
 = scales::percent_format()) + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)
ggplot(resultats_divergencia_dosidiomes, aes(y=castella_divergencia,
 x=castella)) + scale_x_continuous(trans='log10') +
 scale_y_continuous(labels = scales::percent_format()) + geom_point()
ggplot(resultats_divergencia_dosidiomes, aes(y=castella_divergencia,
 x=castella)) + scale_x_continuous(trans='log10') +
 scale_y_continuous(labels = scales::percent_format()) + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)

ggplot(resultats_divergencia_dosidiomes, aes(y=castella_divergencia,
 x=catala)) + scale_x_continuous(trans='log10') + geom_point()
ggplot(resultats_divergencia_dosidiomes, aes(y=castella_divergencia,
 x=catala)) + scale_x_continuous(trans='log10') + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)
ggplot(resultats_divergencia_dosidiomes, aes(y=catala_divergencia,
 x=castella)) + scale_x_continuous(trans='log10') + geom_point()
```

```
ggplot(resultats_divergencia_dosidiomes, aes(y=catala_divergencia,
 x=castella)) + scale_x_continuous(trans='log10') + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0)

##afegiré el número total de vegades que apareix cada paraula, per poder
 fer les gràfiques amb aquest número en compte de amb el de tuits
 convergents
paraules_numvegades<-read.csv(file="paraules_numvegades.csv", header=TRUE)
 ##canviar nom de columna de word a paraula ##(en compte d'açò, he
 important l'arxiu paraules_numvegades.csv, sense la primera columna, i
 després el passe a data.frame, que si no, no funciona la cosa)
resultats_divergencia_dosidiomes<-resultats_divergencia_dosidiomes%>%
    inner_join(paraules_numvegades)
ggplot(resultats_divergencia_dosidiomes, aes(y=castella_divergencia, x=n))
 + scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 theme(text=element_text(size=20))
ggplot(resultats_divergencia_dosidiomes, aes(y=castella_divergencia, x=n))
 + scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0) +
 theme(text=element_text(size=20))
ggplot(resultats_divergencia_dosidiomes, aes(y=catala_divergencia, x=n)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 theme(text=element_text(size=20))
ggplot(resultats_divergencia_dosidiomes, aes(y=catala_divergencia, x=n)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 geom_text(label=resultats_divergencia$paraula, hjust=0, vjust=0) +
 theme(text=element_text(size=20))



### per calcular divergències de diferents temes, ajuntant diverses
 paraules a cada tema. Vaig a vore si puc fer un data.frame per anar
 guardant-ho tot ahí. Cada tema ha estat calculat per separat i afegit al
 data.frame d'un en un.


resultats_temes<-data.frame(tema=vector(mode = "character", length = 10),
                            paraules=vector(mode = "character", length =
                             10),
                            catala = vector(mode = "numeric", length =10),
                            cat_cast = vector(mode = "numeric", length =
                             10),
                            castella = vector(mode = "numeric", length =
                             10),
                            cast_cat = vector(mode = "numeric", length =
                             10),
                            catala_div = vector(mode = "numeric", length
                             =10),
                            catala_conv = vector(mode = "numeric", length =
                             10),
```

```
                              castella_div = vector(mode = "numeric", length
                               = 10),
                              castella_conv = vector(mode = "numeric", length
                               = 10), row.names = NULL)

resultats_temes[13, "tema"]<-"geo_espanyol"
resultats_temes[13, "paraules"]<-"espanya, espana, espanol, espanola,
 espanolas, espanoles, espanyol, espanyola, espanyoles, espanyols"
resultats_temes[13, "catala"]<-freq[1]
resultats_temes[13, "cat_cast"]<-freq[2]
resultats_temes[13, "castella"]<-freq[4]
resultats_temes[13, "cast_cat"]<-freq[3]
resultats_temes[13,
 "catala_div"]<-resultats_temes$cat_cast[13]/(resultats_temes$cat_cast
 [13]+resultats_temes$catala[13])
resultats_temes[13,
 "catala_conv"]<-resultats_temes$catala[13]/(resultats_temes$cat_cast
 [13]+resultats_temes$catala[13])
resultats_temes[13,
 "castella_div"]<-resultats_temes$cast_cat[13]/(resultats_temes$cast_cat
 [13]+resultats_temes$castella[13])
resultats_temes[13,
 "castella_conv"]<-resultats_temes$castella[13]/(resultats_temes$cast_cat
 [13]+resultats_temes$castella[13])

saveRDS(resultats_temes, "resultats_temes.rds")
write.csv(resultats_temes, "resultats_temes.csv")

#### ací està cadascun dels temes
esports<-df_tidy_seleccio%>%
    filter(word=="liga" | word=="lliga" | word=="jugador" |
     word=="jugadors" | word=="jugadores" | word=="jugadoras" |
     word=="jugar" | word=="equip" | word=="equipo" | word=="temporada")%>%
    distinct(text_resposta, .keep_all = TRUE)
freq<-table(esports$llengues)

esports_futbol<-df_tidy_seleccio%>%
    filter(word=="liga" | word=="lliga" | word=="jugador" |
     word=="jugadors" | word=="jugadores" | word=="jugadoras" |
     word=="jugar" | word=="equip" | word=="equipo" | word=="temporada" |
     word=="gol" | word=="champions")%>%
      distinct(text_resposta, .keep_all = TRUE)
freq<-table(esports_futbol$llengues)

politica_sobiranista<-df_tidy_seleccio%>%
    filter(word=="lauraborras" | word=="krls" | word=="erc" |
     word=="perearagones" | word=="junqueras" | word=="juntsxcat" |
     word=="esquerra_erc")%>%
        distinct(text_resposta, .keep_all = TRUE)
freq<-table(politica_sobiranista$llengues)

politica_catalana<-df_tidy_seleccio%>%
```

```
    filter(word=="lauraborras" | word=="krls" | word=="erc" |
     word=="perearagones" | word=="junqueras" | word=="juntsxcat" |
     word=="esquerra_erc" | word=="socialistescat" | word=="ciudadanoscs" |
     word=="ppcatalunya" | word=="miqueliceta" | word=="inesarrimadas" |
     word=="alejandrotgn" | word=="govern" | word=="generalitat" |
     word=="catalunya" | word=="cataluna")%>%
    distinct(text_resposta, .keep_all = TRUE)
freq<-table(politica_catalana$llengues)

politica_espanyola<-df_tidy_seleccio%>%
    filter(word=="pablocasado" | word=="vox" | word=="sanchezcastejon" |
     word=="pabloiglesias" | word=="psoe" | word=="yolandadiaz" |
     word=="idiazayuso")%>%
    distinct(text_resposta, .keep_all = TRUE)
freq<-table(politica_espanyola$llengues)

politica<-df_tidy_seleccio%>%
    filter(word=="democracia" | word=="votar" | word=="eleccions" |
     word=="elecciones" | word=="govern" | word=="gobierno" |
     word=="politics" | word=="politicos")%>%
    distinct(text_resposta, .keep_all = TRUE)
freq<-table(politica$llengues)

dona<-df_tidy_seleccio%>%
    filter(word=="dona" | word=="dones" | word=="chica" | word=="chicas" |
     word=="mujer" | word=="mujeres" | word=="xica" | word=="xiques" |
     word=="xicona" | word=="xicones" | word=="madre" | word=="madres" |
     word=="mare" | word=="mares")%>%
    distinct(text_resposta, .keep_all = TRUE)
freq<-table(dona$llengues)

felicitat<-df_tidy_seleccio%>%
    filter(word=="feliz" | word=="felic" | word=="felicos" |
     word=="felices" | word=="felicidades" | word=="felicidad" |
     word=="felicitats" | word=="felicitat" | word=="enhorabona" |
     word=="enhorabuena")%>%
     distinct(text_resposta, .keep_all = TRUE)
freq<-table(felicitat$llengues)

positiu<-df_tidy_seleccio%>%
    filter(word=="feliz" | word=="felic" | word=="felicos" |
     word=="felices" | word=="felicidades" | word=="felicidad" |
     word=="felicitats" | word=="felicitat" | word=="enhorabona" |
     word=="enhorabuena" | word=="abrazo" | word=="abracada" | word=="cor"
     | word=="corazon" | word=="genial" | word=="guapo" | word=="guapa" |
     word=="bonito" | word=="bonita" | word=="bonic" | word=="bonica" |
     word=="amor" | word=="encanta" | word=="precios" | word=="preciosa" |
     word=="precioso" | word=="preciosos" | word=="precioses" |
     word=="anims" | word=="genial")%>%
        distinct(text_resposta, .keep_all = TRUE)
freq<-table(positiu$llengues)


negatiu<-df_tidy_seleccio%>%
```

```
    filter(word=="asco" | word=="fastic" | word=="por" | word=="miedo" |
     word=="vergonya" | word=="verguenza" | word=="odi" | word=="odio" |
     word=="problema" | word=="problemas" | word=="problemes" |
     word=="violencia" | word=="mentira")%>%
        distinct(text_resposta, .keep_all = TRUE)
freq<-table(negatiu$llengues)

pandemia<-df_tidy_seleccio%>%
    filter(word=="covid" | word=="coronavirus" | word=="virus" |
     word=="pandemia" | word=="vacuna" | word=="vacunacio" | word=="dosis"
     | word=="dosi" | word=="mascarilla" | word=="mascareta" |
     word=="contagiar" | word=="contagi" | word=="contagio" |
     word=="contagis" | word=="contagios" | word=="contagiados" |
     word=="contagiats" )%>%
    distinct(text_resposta, .keep_all = TRUE)
freq<-table(pandemia$llengues)

geo_catala<-df_tidy_seleccio%>%
    filter(word=="barcelona" | word== "valencia" | word=="catalunya" |
     word=="cataluna" | word=="catala" | word== "catalana" |
     word=="catalan" | word=="catalans" | word=="catalanes" |
     word=="catalanas" | word=="valenciano" | word=="valenciana" |
     word=="valencians" | word=="valencianos" | word=="valencianes" |
     word=="valencianas" | word=="balear" | word=="baleares" |
     word=="balears" | word=="mallorqui" | word=="mallorquin" |
     word=="mallorquina" | word=="mallorquins" | word=="mallorquines" |
     word=="mallorquinas")%>%
        distinct(text_resposta, .keep_all = TRUE)
freq<-table(geo_catala$llengues)

geo_espanyol<-df_tidy_seleccio%>%
    filter(word=="espanya" | word=="espana" | word=="espanol" |
     word=="espanola" | word=="espanoles" | word=="espanolas" |
     word=="espanyols" | word=="espanyoles" | word=="espanyol"
     |word=="espanyola")%>%
        distinct(text_resposta, .keep_all = TRUE)
freq<-table(geo_espanyol$llengues)

geo_tot<-df_tidy_seleccio%>%
    filter(word=="espanya" | word=="espana" | word=="espanol" |
     word=="espanola" | word=="espanoles" | word=="espanolas" |
     word=="espanyols" | word=="espanyoles" | word=="espanyol"
     |word=="espanyola" | word=="catalunya" | word=="cataluna" |
     word=="valencia" | word=="barcelona" | word=="madrid" | word=="europa"
     | word=="catala" | word== "catalana" | word=="catalan" |
     word=="catalans" | word=="catalanes" | word=="catalanas" |
     word=="valenciano" | word=="valenciana" | word=="valencians" |
     word=="valencianos" | word=="valencianes" | word=="valencianas" |
     word=="balear" | word=="balears" | word=="baleares" |
     word=="mallorqui" | word=="mallorquin" | word=="mallorquina" |
     word=="mallorquins" | word=="mallorquines" | word=="mallorquinas")%>%
        distinct(text_resposta, .keep_all = TRUE)
freq<-table(geo_tot$llengues)
```

```
##un bar plot amb les divergències per temes

ggplot(resultats_temes, aes(x=tema, y=castella_div)) +
    geom_bar(stat = "identity") +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1)) +
    theme(text=element_text(size=30))

ggplot(resultats_temes, aes(x=tema, y=catala_div)) +
    geom_bar(stat = "identity") +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1)) +
    theme(text=element_text(size=30))

###per reordenar això per fer gràfiques amb les barres ordenades segons el
 percentatge
###afegir primer percentatges
resultats_temes<-resultats_temes %>%
mutate(resultats_temes,
 catala_percentatge=original_catala/(original_catala+original_castella))
resultats_temes<-resultats_temes %>%
    mutate(resultats_temes,
     castella_percentatge=original_castella/
     (original_catala+original_castella))


ggplot(resultats_temes, aes(x=reorder(tema, catala_percentatge),
 y=castella_div)) +
    geom_bar(stat = "identity") +
    scale_y_continuous(limits= c(0, 0.35), labels =
     scales::percent_format()) +
    theme(axis.text.x = element_text(angle = 47, hjust = 1, vjust = 1)) +
    theme(text=element_text(size=30))

ggplot(resultats_temes, aes(x=reorder(tema, catala_percentatge),
 y=catala_div)) +
    geom_bar(stat = "identity") +
    scale_y_continuous(limits= c(0, 0.35), labels =
     scales::percent_format()) +
    theme(axis.text.x = element_text(angle = 47, hjust = 1, vjust = 1)) +
    theme(text=element_text(size=30))

## i uns amb div vs número de paraules (num en parelles ca-ca o es-es)
ggplot(resultats_temes, aes(y=catala_div, x=catala)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 theme(text=element_text(size=30))
ggplot(resultats_temes, aes(y=catala_div, x=catala)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 geom_text(label=resultats_temes$tema, hjust=0, vjust=0) +
 theme(text=element_text(size=20))
```

```
ggplot(resultats_temes, aes(y=castella_div, x=castella)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 theme(text=element_text(size=30))
ggplot(resultats_temes, aes(y=castella_div, x=castella)) +
 scale_x_continuous(trans='log10') + scale_y_continuous(labels =
 scales::percent_format()) + geom_point() +
 geom_text(label=resultats_temes$tema, hjust=0, vjust=0) +
 theme(text=element_text(size=20))
ggplot(resultats_temes, aes(y=catala_div, x=castella_div)) + geom_point() +
 scale_x_continuous(labels = scales::percent_format()) +
 scale_y_continuous(labels = scales::percent_format()) +
 geom_text(label=resultats_temes$tema, hjust=0, vjust=0) +
 theme(text=element_text(size=20))

## afegir columnes amb les sumes de ca ca i ca es per un costat, i de es es
 i es ca per un altre
resultats_temes <- mutate(resultats_temes,
 original_catala=resultats_temes$catala+resultats_temes$cat_cast)
resultats_temes <- mutate(resultats_temes,
 original_castella=resultats_temes$castella+resultats_temes$cast_cat)

### per a fer un barplot amb les barres agrupades segons l'idioma de
 l'original, cal regirar les dades, per tindre només una columna amb
 valors, separats segons tema i llengua
prova_barplot<-data.frame(tema=resultats_temes$tema,
                          original_catala=resultats_temes$original_catala,
                          original_castella=resultats_temes$original_castell
                           a)

library(tidyr)
prova_barplot %>%
    pivot_longer(col = -tema) %>%
    ggplot(aes(x = factor(tema, levels=c("politica_espanyola",
     "esports_futbol", "esports", "felicitat", "positiu", "geo_espanyol",
     "negatiu", "dona", "pandemia", "geo_tot", "politica", "geo_catala",
     "politica_catalana", "politica_sobiranista")), y = value, fill =
     name)) +
    geom_bar(stat = "identity", position = "dodge") +
    theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1)) +
     theme(text=element_text(size=20)) +
    theme(legend.position = "bottom")

##i ara amb percentatges

prova_barplot2<-mutate(prova_barplot,
 catala_percentatge=original_catala/(original_catala+original_castella))
prova_barplot2<-mutate(prova_barplot2,
 castella_percentatge=original_castella/(original_catala+original_castella))

head(prova_barplot2)

### ara eliminem les columnes amb valors que si no, hi ha massa dades

prova_percentatges<-data.frame(tema=prova_barplot2$tema,
```

```
                                catala=prova_barplot2$catala_percentatge,
                                castella=prova_barplot2$castella_percentatge)

prova_percentatges %>%
    pivot_longer(col = -tema) %>%
    ggplot(aes(x = factor(tema), y = value, fill = name)) +
    geom_bar(stat = "identity", position = "dodge")  +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1)) +
    theme(text=element_text(size=30)) +
    theme(legend.position = "bottom")


prova_percentatges %>%
    pivot_longer(col = -tema) %>%
    ggplot(aes(x = factor(name), y = value)) +
    geom_boxplot() +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(text=element_text(size=20))

prova_percentatges %>%
    pivot_longer(col = -tema) %>%
    ggplot(aes(x = factor(name), y = value)) +
    geom_violin() +
    geom_boxplot(width=0.1) +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(text=element_text(size=20))

### per fer les barres una damunt de l'altra
prova_percentatges %>%
    pivot_longer(col = -tema) %>%
    ggplot(aes(x = factor(tema, levels=c("politica_espanyola",
     "esports_futbol", "esports", "felicitat", "positiu", "geo_espanyol",
     "negatiu", "dona", "pandemia", "geo_tot", "politica", "geo_catala",
     "politica_catalana", "politica_sobiranista")), y = value, fill =
     name)) +
    geom_bar(stat = "identity", position = "stack")  +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(axis.text.x = element_text(angle = 47, hjust = 1, vjust = 1)) +
    theme(text=element_text(size=30)) +
    theme(legend.position = "bottom")

## el boxplot amb les 500 paraules més abundants. I ja voré quina és la
 millor manera per no tindre paraules que només valen en un idioma
proporcio_500paraules<-data
 .frame(paraula=resultats_divergencia_dosidiomes$paraula,
                                  proporcio_catala=
                                   (resultats_divergencia_dosidiomes$catala+
                                   resultats_divergencia_dosidiomes$cat_cast
                                   )/
                                   (resultats_divergencia_dosidiomes$catala+
                                   resultats_divergencia_dosidiomes$cat_cast
                                   +resultats_divergencia_dosidiomes$castell
                                   a+resultats_divergencia_dosidiomes$cast_c
                                   at),
```

```
                                    proporcio_castella=
                                     (resultats_divergencia_dosidiomes$castell
                                     a+resultats_divergencia_dosidiomes$cast_c
                                     at)/
                                     (resultats_divergencia_dosidiomes$catala+
                                     resultats_divergencia_dosidiomes$cat_cast
                                     +resultats_divergencia_dosidiomes$castell
                                     a+resultats_divergencia_dosidiomes$cast_c
                                     at)
                                    )

proporcio_500paraules %>%
    pivot_longer(col = -paraula) %>%
    ggplot(aes(x = factor(name), y = value)) +
    geom_boxplot() +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(text=element_text(size=20))

proporcio_500paraules %>%
    pivot_longer(col = -paraula) %>%
    ggplot(aes(x = factor(name), y = value)) +
    geom_violin() +
    geom_boxplot(width=0.1) +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(text=element_text(size=20))

### anem a fer-ho només amb les que tenen calculats resultats "2" a la
 divergencia dos idiomes, però no és el millor perquè les paraules que
 s'escriguen igual en els dos idiomes, no estaran ací
proporcio_500paraules2<-data
 .frame(paraula=resultats_divergencia_dosidiomes$paraula,
                                    proporcio_catala=
                                     (resultats_divergencia_dosidiomes$catala2
                                     +resultats_divergencia_dosidiomes$cat_cas
                                     t2)/
                                     (resultats_divergencia_dosidiomes$catala2
                                     +resultats_divergencia_dosidiomes$cat_cas
                                     t2+resultats_divergencia_dosidiomes$caste
                                     lla2+resultats_divergencia_dosidiomes$cas
                                     t_cat2),
                                    proporcio_castella=
                                     (resultats_divergencia_dosidiomes$castell
                                     a2+resultats_divergencia_dosidiomes$cast_
                                     cat2)/
                                     (resultats_divergencia_dosidiomes$catala2
                                     +resultats_divergencia_dosidiomes$cat_cas
                                     t2+resultats_divergencia_dosidiomes$caste
                                     lla2+resultats_divergencia_dosidiomes$cas
                                     t_cat2)
)

proporcio_500paraules2 %>%
    pivot_longer(col = -paraula) %>%
    ggplot(aes(x = factor(name), y = value)) +
    geom_violin() +
```

```
    geom_boxplot(width=0.1) +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(text=element_text(size=20))


### ara trac un data frame només amb les paraules i les proporcions per fer
 el boxplot
proporcions<-data.frame(paraula=resultats_divergencia_dosidiomes$paraula,
                        catala=resultats_divergencia_dosidiomes$catala_propo
                         rcio,
                        castella=resultats_divergencia_dosidiomes$castella_p
                         roporcio)

proporcions %>%
    pivot_longer(col = -paraula) %>%
    ggplot(aes(x = factor(name), y = value)) +
    geom_boxplot() +
    scale_y_continuous(labels = scales::percent_format())+
    theme(text=element_text(size=30))

proporcions %>%
    pivot_longer(col = -paraula) %>%
    ggplot(aes(x = factor(name), y = value)) +
    geom_violin() +
    geom_boxplot(width=0.1) +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(text=element_text(size=30))

#### ara anem a fer uns violin plots amb les divergències de la taula gran
 amb divergències calculades per als dos idiomes, és a dir, on "també" i
 "también" estan representades només amb una línia
divergencies <- data.frame(paraula=resultats_divergencia_dosidiomes$paraula,
                           catala_divergencia=resultats_divergencia_dosidiom
                            es$catala_divergencia,
                           castella_divergencia=resultats_divergencia_dosidi
                            omes$castella_divergencia)

divergencies %>%
    pivot_longer(col = -paraula) %>%
    ggplot(aes(x = factor(name), y = value)) +
    geom_boxplot() +
    scale_y_continuous(labels = scales::percent_format())+
    theme(text=element_text(size=30))

divergencies %>%
    pivot_longer(col = -paraula) %>%
    ggplot(aes(x = factor(name), y = value)) +
    geom_violin() +
    geom_boxplot(width=0.1) +
    scale_y_continuous(labels = scales::percent_format()) +
    theme(text=element_text(size=30))
```

Annex 2. Llistat de paraules buides

0	ante	contains	erais	gm	intentar	muchos	posible	shed	they're	vuestros
1	anterior	contra	eramos	gmt	intentas	mug	possible	shell	they've	vuit
2	antes	conyo	eran	gn	intento	mundo	possibly	shes	theyd	vuitanta
3	anybody	copy	eras	go	interest	must	potentially	should	theyll	vuite
4	anyhow	corrents	erem	goes	interested	must've	potser	should've	theyre	vuitena
5	anymore	correspondi ng	eren	going	interesting	mustn't	pp	shouldn	theyve	vuitenes
6	anyone	cosas	eres	gone	interests	mustnt	pr	shouldn't	thick	vuitens
7	anything	coses	ereu	good	into	muy	predominant ly	shouldnt	thin	w
8	anyway	could	ergo	goods	invention	mv	present	show	thing	want
9	anyways	could've	es	got	inward	mw	presented	showed	things	wanted
10	anywhere	couldn	es	gotten	io	mx	presenting	showing	think	wanting
11	ao	couldn't	esa	gov	iq	my	presents	shown	thinks	wants
12	apa	couldnt	esas	gp	ir	myse”	prest	showns	third	was
13	apart	course	escar	gq	is	myself	presumably	shows	thirty	wasn
14	apenas	cr	escriure	gr	isn	mz	previously	si	this	wasn't
15	apparently	crec	ese	gran	isn't	n	primarily	sia	thorough	wasnt
16	appear	creo	eso	grande	isnt	n'he	primer	siau	thoroughly	way
17	appreciate	cry	esos	grandes	it	n'hi	primera	sic	those	ways
18	appropriate	cs	especially	great	it'd	na	primeres	side	thou	we
19	approximately	cu	essent	greater	it'll	nada	primero	sides	though	we'd
20	apres	cual	essent	greatest	it's	nadie	primeros	sido	thoughh	we'll
21	aproximadame nte	cuales	esser	greetings	itd	name	primers	siempre	thought	we're
22	aq	cualquier	est	group	itll	namely	principalmen te	siendo	thoughts	we've

23	aqueix	cuando	esta	grouped	its	nay	pro	siete	thousand	web
24	aqueixa	cuanto	esta	grouping	itse”	nc	probably	significant	three	webpage
25	aqueixes	cuatro	estaba	groups	itself	nd	problem	significantly	throug	website
26	aqueixos	cuenta	estabais	gs	ive	ne	problems	sigue	through	wed
27	aqueixs	currently	estabamos	gt	j	near	prompte	siguem	throughout	welcome
28	aquel	cv	estaban	gu	ja	nearly	promptly	sigues	thru	well
29	aquell	cx	estabas	gueno	jamai	necessarily	prop	sigueu	thus	wells
30	aquella	cy	estad	gw	je	necessary	propi	sigui	ti	went
31	aquellas	cz	estada	gy	jm	need	propia	siguiente	tiene	were
32	aquelles	d	estadas	h	jo	needed	propias	siguin	tienen	weren
33	aquellos	danar	estades	ha	join	needing	propio	siguis	tienes	weren't
34	aquells	dhaver	estado	habeis	jp	needn't	propios	similar	til	werent
35	aquen	dun	estados	haber	junto	neednt	prou	similarly	till	weve
36	aquest	duna	estais	habia	just	needs	proud	sin	tinc	wf
37	aquesta	dunes	estais	habiais	justo	neither	provided	since	tip	what
38	aquestes	duns	estamos	habiamos	k	net	provides	sincere	tis	what'd
39	aquests	da	estan	habian	ke	netscape	proximo	sino	tj	what'll
40	aqui	dado	estan	habias	keep	never	proximos	sis	tk	what's
41	aqui	daixonses	estando	habida	keeps	neverf	pt	sise	tm	what've
42	ar	daixonses	estant	habidas	kept	neverless	puc	sisena	tn	whatever
43	ara	dallonses	estar	habido	keys	nevertheless	pudo	sisenes	to	whatll
44	are	dallonses	estaran	habidos	kg	new	pueda	sisens	toda	whats
45	area	dalt	estaras	habiendo	kh	newer	puede	site	todas	whatve
46	areas	daltabaix	estare	habra	ki	newest	pueden	six	todavia	when
47	aren	damunt	estareis	habran	kind	next	puedo	sixty	today	when'd
48	aren't	dan	estarem	habras	km	nf	pues	sj	todo	when'll
49	arent	dar	estaremos	habremos	kn	ng	puix	sk	todos	when's

50	arise	dare	estareu	habreis	knew	ni	pus	sl	together	whence
51	around	daren't	estaria	habremos	know	nine	put	slightly	ton	whenever
52	arpa	darent	estariais	habria	known	ninety	puts	sm	tons	where
53	arran	darrera	estaramos	habriais	knows	ningu	pw	small	too	where'd
54	arrera	darrere	estarian	habriamos	kp	ningun	py	smaller	took	where'll
55	arrere	date	estarias	habrian	kr	ninguna	q	smallest	top	where's
56	arreu	davall	estariem	habrias	kw	ningunas	qa	sn	tos	whereafter
57	arri	davant	estarien	hace	ky	ninguno	qual	so	tost	whereas
58	arriba	de	estaries	haceis	kz	ningunos	quals	sobre	tostemps	whereby
59	as	dear	estarieu	hacemos	l	nl	qualsevol	sobretot	tot	wherein
60	aseguro	debades	estas	hacen	l'hi	no	qualsevulla	soc	tota	wheres
61	asi	debe	estat	hacer	la	no-one	qualssevol	sois	total	whereupon
62	aside	deben	estats	hacerlo	lado	nobody	qualssevulla	sol	totes	wherever
63	aside	debido	estava	haces	large	nogensmenys	quan	sola	tothom	whether
64	ask	deca	estavem	hacia	largely	nomes	quant	solament	tothora	which
65	asked	decir	estaven	haciendo	largo	non	quanta	solamente	tots	whichever
66	asking	dedins	estaves	had	las	none	quantes	solas	toward	while
67	asks	definitely	estaveu	hadn't	last	nonetheless	quants	soles	towards	whilst
68	associated	defora	este	hadnt	lately	noone	quaranta	solo	tp	whim
69	at	deia	esteis	hagi	later	nor	quart	solos	tr	whither
70	atras	dejorn	estem	hagim	latest	noranta	quarta	sols	tras	who
71	au	dejorn	estemos	hagin	latter	normally	quartes	som	trata	who'd
72	aun	dejus	esten	hagis	latterly	nos	quarts	some	traves	who'll
73	aunque	del	estes	hagiu	lb	nosaltres	quasi	somebody	trenta	who's
74	auth	deleted	esteu	hago	lc	nosotras	quatre	someday	tres	whod
75	available	della	estic	haguda	le	nosotros	que	somehow	tret	whoever
76	avall	dels	estigue	hagudes	least	nostra	quedo	someone	tretze	whole

77	avant	demas	estiguem	hague	length	nostre	quelcom	somethan	tried	wholl
78	aviat	demasiado	estiguerem	haguerem	les	nostres	queremos	something	tries	whom
79	avui	dementre	estigueren	hagueren	less	not	qui	sometime	trillion	whomever
80	aw	demes	estigueres	hagueres	lest	noted	quickly	sometimes	truly	whos
81	away	dempeus	estiguereu	haguereu	let	nothing	quien	somewhat	try	whose
82	awfully	dentro	estigues	hagues	let's	notwithstanding	quienes	somewhere	trying	why
83	ay	des	estigues	haguessim	lets	nou	quiere	somos	ts	why'd
84	ayer	describe	estiguessis	haguessin	li	nove	quieres	son	tt	why'll
85	az	described	estigueu	haguessis	li'n	novel	quin	sons	tu	why's
86	b	desde	estigui	haguessiu	like	novena	quina	soon	tú	widely
87	ba	dese	estigui	hagui	liked	novenes	quines	sorry	turn	width
88	back	desena	estiguin	hagut	likely	novens	quins	sos	turned	will
89	backed	desenes	estiguis	haguts	likewise	now	quinze	sota	turning	willing
90	backing	desens	esto	half	line	nowhere	quisvulla	sots	turns	wish
91	backs	despite	estos	han	little	np	quite	sou	tus	with
92	backward	despres	estoy	happens	lk	nr	qv	sovint	tuve	within
93	backwards	despues	estuve	hardly	ll	ns	r	soy	tuviera	without
94	bah	dessobre	estuviera	has	lla	nu	ran	specifically	tuvierais	won
95	baix	dessota	estuvierais	hasn	llarg	nuestra	rather	specified	tuvieramos	won't
96	bajo	dessus	estuvieramos	hasn't	llavors	nuestras	rd	specify	tuvieran	wonder
97	baldament	detail	estuvieran	hasnt	llega	nuestro	re	specifying	tuvieras	wont
98	bastant	devers	estuvieras	hasta	llegó	nuestros	readily	sr	tuvieron	words
99	bastante	devora	estuvieron	haurà	lleva	nueva	realizado	st	tuviese	work
100	bastants	dice	estuviese	hauran	llevar	nuevas	realizar	state	tuvieseis	worked
2015	bb	dicen	estuvieseis	hauras	llevat	nuevo	realizo	states	tuviesemos	working
2016	bd	dicho	estuviesemos	haure	lluny	nuevos	really	status	tuviesen	works

2017	be	did	estuviesen	haurem	llur	null	reasonably	still	tuvieses	world
2018	bé	didn	estuvieses	haureu	llurs	number	rebe	stop	tuvimos	would
2019	became	didn't	estuvimos	hauria	lo	numbers	recent	strongly	tuviste	would've
2020	because	didnt	estuviste	hauríem	long	nunca	recently	su	tuvisteis	wouldn
2021	become	dieron	estuvisteis	haurien	longer	nz	ref	suara	tuvo	wouldn't
2022	becomes	diferente	estuvo	hauries	longest	o	refs	sub	tuya	wouldnt
2023	becoming	diferentes	et	haurieu	look	obtain	regarding	substantially	tuyas	ws
2024	been	diferents	et-al	have	looking	obtained	regardless	successfully	tuyo	www
2025	before	differ	etc	havem	looks	obviously	regards	such	tuyos	x
2026	beforehand	different	etcetera	haven	los	ocho	related	sufficiently	tv	xau-xau
2027	began	differently	ets	haven't	low	of	relatively	suggest	tw	xec
2028	begin	dijeron	even	havent	lower	off	renoi	sup	twas	xo
2029	beginning	dijo	evenly	haver	lr	often	rera	sure	twelve	y
2030	beginnings	dinou	ever	haveu	ls	ofthe	rere	sus	twenty	ya
20172018	begins	dins	evermore	havia	lt	oh	res	suya	twice	ye
__twitter_impression	behind	dintre	every	haviem	ltd	oi	research	suyas	two	year
'll	being	dio	everybody	havien	lu	ok	reserved	suyo	tz	years
'tis	beings	dir	everyone	havies	luego	okay	respectively	suyos	u	yes
'twas	believe	directly	everything	havieu	lugar	old	respecto	sv	ua	yet
've	below	disset	everywhere	having	lv	older	resulted	sy	uf	yo
#_TWITTER_	ben	dit	ex	hay	ly	oldest	resulting	system	ug	you
a	beside	diu	exactly	haya	m	olim	results	sz	ui	you'd
a's	besides	diuen	example	hayais	m’ha	om	retruc	t	uix	you'll
abans	best	divers	except	hayamos	m'he	omitida	right	tha	uk	you're
abans-d'ahir	better	diversa	excepte	hayan	ma	omitido	ring	t's	ultim	you've
abintestat	between	diverses	existe	hayas	made	omitted	ro	ta	ultima	youd

able	beyond	diversos	existen	he	mai	on	room	take	ultimas	youll
ableabout	bf	divuit	explicat	he'd	mainly	once	rooms	taken	ultimes	young
about	bg	dj	explico	he'll	make	one	round	taking	ultimo	younger
above	bh	dk	expreso	he's	makes	one's	rt	tal	ultimos	youngest
abroad	bi	dm	f	hecho	making	ones	ru	tals	ultims	your
abst	bien	do	fa	hed	mal	only	run	també	ultra	youre
accordance	big	does	face	hell	malgrat	onsevulga	rw	también	um	yours
according	bill	doesn	faces	hello	man	onsevulla	s	tame	un	yourself
accordingly	billion	doesn't	fact	help	mano	onto	sha	tampoc	una	yourselves
aci	biol	doesnt	facts	hem	manco	onze	shan	tampoco	unas	youve
aco	bis	doing	faig	hemos	manera	open	shavia	tan	under	yt
across	bj	don	fairly	hence	manifesto	opened	sa	tanmateix	underneath	yu
act	bm	don't	fan	her	mant	opening	sabe	tant	undoing	z
actually	bn	donar	far	here	mantinent	opens	sabeis	tanta	unes	za
actualmente	bo	donat	farther	here's	mants	opposite	sabem	tantes	unfortunately	zero
ad	both	doncs	fas	hereafter	many	or	sabemos	tanto	unic	zm
added	bottom	donde	fe	hereby	mas	ord	saben	tantost	unica	zr
adelante	br	done	felt	herein	massa	order	saber	tants	unicas	
ademas	brief	dont	fem	heres	mateix	ordered	sabes	tb	unico	
ades	briefly	dos	fent	hereupon	mateixa	ordering	sabeu	tc	unicos	
adesiara	bs	dotze	fer	hers	mateixes	orders	said	td	unics	
adeu	bt	doubtful	fet	herse”	mateixos	org	salvament	te	uniques	
adhuc	but	down	feu	herself	may	os	salvant	té	unless	
adj	buy	downed	few	hes	maybe	other	salvat	tell	unlike	
adopted	bv	downing	fewer	heu	mayn't	others	same	ten	unlikely	
ae	bw	downs	ff	hi	maynt	otherwise	sap	tendra	uno	
af	by	downwards	fi	hicieron	mayor	otra	saps	tendran	unos	

affected	bz	due	fifteen	hid	mc	otras	saw	tendras	uns	
affecting	c	dues	fifth	high	md	otro	say	tendremos	until	
affects	c'mon	durant	fifty	higher	me	otros	saying	tendreis	unto	
afirmo	c's	durante	fify	highest	mean	ought	says	tendremos	up	
after	ca	during	fin	him	means	oughtn't	sb	tendria	upa	
afterwards	cabeza	dz	find	himse”	meantime	oughtnt	sc	tendriais	upon	
ag	cada	e	finds	himself	meanwhile	our	sd	tendriamos	ups	
again	cadascu	each	fins	his	medi	ours	se	tendrian	upwards	
against	cadascun	early	fire	hither	media	ourselves	sea	tendrias	us	
ago	cadascuna	ec	first	hizo	mediante	out	seais	tends	usa	
agrego	cadascunes	ed	five	hk	medio	outside	seamos	tene	usais	
ah	cadascuns	edu	fix	hm	member	over	sean	tened	usamos	
ahead	cal	ee	fj	hn	members	overall	seas	tenen	usan	
ahir	call	effect	fk	ho	men	owing	sec	teneis	usar	
ahir	came	eg	fm	hom	menciono	own	second	tenemos	usas	
ahora	can	eh	fo	homepage	menos	p	secondly	tenen	use	
ai	can't	eight	followed	hopefully	mentre	pa	seconds	tener	used	
ain't	cannot	eighty	following	how	mentrestant	page	section	tenga	useful	
aint	cant	either	follows	how'd	menys	pages	see	tengais	usefully	
aitambe	cap	ejemplo	for	how'll	merely	para	seeing	tengamos	usefulness	
aitampoc	caption	el	fora	how's	mes	parece	seem	tengan	uses	
aitan	car	él	força	howbeit	meu	part	seemed	tengas	using	
aitant	case	ela	forem	however	meua	parte	seeming	tengo	uso	
aitantost	cases	eleven	foren	hoy	meues	parted	seems	tenia	usted	
aixa	casi	ell	fores	hr	meus	particular	seen	tenia	usually	
aixi	catorze	ella	foreu	ht	meva	particularly	sees	teniais	utm_medium	

aixo	cause	ellas	forever	htm	meves	parting	segon	teniamos	uucp	
aixo	causes	elles	former	html	mg	partir	segona	tenian	uy	
al	cc	ello	formerly	http	mh	parts	segones	tenias	uz	
al·legro	cd	ellos	forth	https	mi	pas	segons	tenida	v	
alca	cent	ells	forty	hu	mia	pasada	seguida	tenidas	va	
aleshores	centes	els	forward	hube	mias	pasado	segun	tenido	vagi	
algo	cents	else	fos	hubiera	mica	passa	segunda	tenidos	vagin	
algu	cerca	elsewhere	fossim	hubierais	microsoft	passar	segundo	teniendo	vagis	
alguien	cert	em	fossin	hubieramos	miedo	passat	seis	tenim	vaig	
algun	certa	embargo	fossis	hubieran	mientras	passim	seixanta	tenir	vair	
alguna	certain	emperò	fossiu	hubieras	mig	past	self	tens	vais	
algunas	certainly	empleais	fou	hubieron	might	pe	sembla	teniu	valor	
algunes	certes	emplean	found	hubiese	might've	pel	selves	tercer	value	
alguno	certs	emplear	four	hubieseis	mightn't	pels	semblant	tercera	vam	
algunos	cf	empleas	fr	hubiesemos	mightnt	per	semblants	terceres	vamos	
alguns	cg	empleo	fra	hubiesen	mil	per que	sempre	tercers	van	
alhora	ch	empty	free	hubieses	mill	perhaps	senalo	tes	vareig	
all	changes	en	from	hubimos	million	pero	sengles	test	varem	
alla	ci	enans	front	hubiste	mine	perq	sens	teu	vares	
allen	cierta	enant	fue	hubisteis	minus	perque	sense	teua	vareu	
alli	ciertas	enca	fuera	hubo	mio	pertot	sensible	teues	varias	
allo	cierto	encara	fuerais	hui	mios	pesar	sent	teus	varios	
allow	ciertos	encima	fueramos	hundred	mis	pf	ser	teva	various	
allows	cinc	encontinent	fueran	i	misma	pg	sera	teves	vas	
almenys	cinco	encuentra	fueras	i.e.	mismas	ph	seran	text	vau	
almost	cinquanta	end	fueron	i'd	mismo	pk	seras	tf	vaya	

alone	cinque	endalt	fuese	i'll	mismos	pl	sere	tg	vc	
along	cinquena	endarrera	fueseis	i'm	miss	place	sereis	th	ve	
alongside	cinquenes	endarrere	fuesemos	i've	missage	placed	serem	than	veces	
already	cinquens	endavant	fuesen	id	mitges	places	seremos	thank	vegada	
alrededor	ck	endebades	fueses	ídem	mitja	please	sereu	thanks	vegades	
als	cl	ended	fui	ie	mitjançant	plus	seria	thanx	vem	
also	clear	endemà	fuimos	if	mitjos	pm	seriais	that	ver	
although	clearly	endemés	fuiste	ignored	mk	pmid	seriamos	that'll	verbigracia	
alto	click	endemig	fuisteis	igual	ml	pn	serian	that's	vers	
altra	cm	ending	full	iguals	mm	poc	serias	that've	versus	
altre	cmon	endins	fully	ii	mn	poca	seriem	thatll	very	
altres	cn	endintre	further	il	mo	pocas	serien	thats	veure	
altresi	co	ends	furthered	ill	mode	poco	series	thatve	ves	
altri	co.	enfora	furthering	im	modo	pocos	serieu	the	vet	
always	com	engir	furthermore	imagen	moixoni	pocs	serious	their	veu	
am	come	enguany	furthers	immediate	mol	podeis	seriously	theirs	vez	
amb	comento	enguanyasses	fx	immediately	molt	podem	ses	them	viene	
ambdos	comes	enjus	g	importance	molta	podemos	set	themselves	vg	
ambdues	como	enlaire	ga	in	moltes	poden	setanta	then	vi	
ambos	computer	enlla	gaire	inasmuch	molts	poder	sete	thence	via	
amen	comsevulla	enlloc	gairebe	inc	moment	podeu	setena	there	video	
amid	con	enough	gaires	inc.	momento	podra	setenes	there'd	vint	
amidst	concerning	enrera	gave	inclos	mon	podran	setens	there'll	viz	
among	conocer	enrere	gb	inclusive	mons	podria	setze	there're	vn	
amongst	consegueixo	ens	gd	incluso	more	podria	seu	there's	vol	
amoungst	conseguim	ensems	ge	indeed	moreover	podriais	seua	there've	vols	
amount	conseguimos	ensota	general	index	mos	podriamos	seues	thereafter	vora	

amp.html	conseguir	ensus	generally	indicate	most	podrian	seus	thereby	vos	
ampleamos	consequently	entirely	gens	indicated	mostly	podrian	seva	thered	vosaltres	
amunt	consider	entonces	get	indicates	move	podrias	seven	therefore	vosotras	
an	considera	entorn	gets	indico	mp	point	seventy	therein	vosotros	
añadio	considering	entre	getting	information	mq	pointed	several	therell	voste	
anar	considero	entremig	gf	informo	mr	pointing	seves	thereof	vostes	
anc	consigo	entretant	gg	inner	mrs	points	sg	therere	vostra	
and	consigue	entro	gh	inside	ms	poner	sh	theres	vostre	
andante	consigueix	envers	gi	insofar	msie	poorly	shall	thereto	vostres	
andantino	consigueixen	envides	girientorn	instead	mt	poques	shan't	thereupon	voy	
anit	consigueixes	environs	give	int	mu	por	shant	thereve	vs	
announce	consiguen	environs	given	intenta	much	por que	she	these	vu	
another	consigues	ep	gives	intentais	mucha	porque	she'd	they	vuestra	
ans	contain	er	giving	intentamos	muchas	pos	she'll	they'd	vuestras	
antany	containing	era	gl	intentan	mucho	posar	she's	they'll	vuestro	

Annex 3. Resultats de l'anàlisi per temes, incloent el número de parells de tuits segons la llengua i els percentatges de convergència o divergència lingüístiques per a cada llengua

tema	català	catala>castellà	castellà	castellà>català	divergència cap al català	convergència cap al català	divergència cap al castellà	convergència cap al castellà
dona	12035	2089	39626	1410	15%	85%	3%	97%
esports	6572	1731	39870	1829	21%	79%	4%	96%
esports_futbol	7336	1978	45005	2053	21%	79%	4%	96%
felicitat	7255	1193	35902	1520	14%	86%	4%	96%
geo_catala	48365	10000	40977	6411	17%	83%	14%	86%
geo_espanyol	15371	3315	56174	4129	18%	82%	7%	93%
geo_tot	62843	13217	121941	10644	17%	83%	8%	92%
negatiu	12141	1980	41598	1922	14%	86%	4%	96%
pandemia	12629	2431	32084	1495	16%	84%	4%	96%
politica	28048	4900	47348	3135	15%	85%	6%	94%
politica_catalana	54280	9362	19337	4471	15%	85%	19%	81%
politica_espanyola	5863	1318	36670	1447	18%	82%	4%	96%
politica_sobiranista	24860	3632	3255	1526	13%	87%	32%	68%
positiu	15623	2800	77308	3576	15%	85%	4%	96%