REVISTA DE LLENGUA I DRET // JOURNAL OF LANGUAGE AND LAW #73

TECNOLOGIES LINGÜÍSTIQUES PER A LLENGÜES MINORITZADES: EL CAS DE L'ALGUERÈS

Adrià Martín Mor*

Francesc Ballone**

Resum***

La tecnologia pot jugar un rol decisiu en els processos de normalització lingüística. La creació de recursos lingüístics —amb el potencial formatiu o de disseminació que comporten, especialment en llengües en procés d'estandardització— és una possibilitat que cal tenir present en dissenyar estratègies per a la normalització. Aquest article se proposa contribuir al procés de normalització de l'alguerès, varietat parlada a l'Alguer (Sardenya) per unes dotze mil persones, mitjançant una anàlisi de les obres de consulta digitals i dels recursos lingüístics existents. En la primera part se proporcionen dades sobre el context sociolingüístic i se fa un estat de la qüestió sobre el procés d'estandardització de l'alguerès. La segona part mira d'identificar, amb referències a altres comunitats lingüístiques en situacions similars, accions en l'àmbit tecnològic que podrien dur-se a terme en paral·lel al procés d'estandardització de l'alguerès.

Paraules clau: alguerès; tecnologies lingüístiques; llengües minoritzades; llengües en perill; estandardització lingüística.

LINGUISTIC TECHNOLOGIES FOR MINORITY LANGUAGES: THE CASE OF ALGHERESE

Abstract

Technology can play a decisive role in linguistic normalisation processes. The creation of linguistic resources (with the potential they have for education or to encourage the spread of languages, especially those in the process of standardisation) is a possibility that should be taken into account in designing normalisation strategies. This article proposes contributing to the process of normalising Algherese, a variety spoken in Alghero (Sardinia) by around twelve thousand people, through an analysis of digital reference works and existing linguistic resources. The first part provides data about the sociolinguistic context and establishes the current situation regarding the process of standardising Algherese. The second part seeks to identify, with references to other language communities in similar situations, actions in the technological sphere that could be carried out in parallel with the process of standardising Algherese.

Keywords: Algherese; linguistic technologies; minority languages; endangered languages; linguistic standardisation.

---

*Adrià Martín Mor, professor Serra-Húnter del Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental (Universitat Autònoma de Barcelona). Les seves línies de recerca són les tecnologies de la traducció i les llengües minoritzades. adria.martin@uab.cat 0000-0003-0842-3190

** Francesc Ballone, doctor en lingüística aplicada i membre corresponent de l'Alguer per la Secció Filològica de l'IEC. Els seus camps de recerca són la fonètica segmental del català i del sard, i també la prosòdia i la sociolingüística catalana. fballone@iec.cat

*** Els autors d'aquest article signen com a ciutadans de la República catalana proclamada pel Govern legítim de Catalunya, en protesta per l'empresonament i l'exili d'activistes i membres del Govern i en solidaritat amb la ciutadania que va patir la repressió de l'Estat espanyol arran del referèndum d'autodeterminació de l'1 d'octubre del 2017.

Citació recomanada: Martín Mor, Adrià, i Ballone, Francesc. (2020). Tecnologies lingüístiques per a llengües minoritzades: el cas de l'alguerès. *Revista de Llengua i Dret, Journal of Language and Law*, *73*, 82-93. https://doi.org/10.2436/rld.i73.2020.3397

Sumari

1 El context sociolingüístic de l'alguerès

1.1 La varietat no dominant algueresa en el context pluricèntric del català

1.2 Procés de normativització de la llengua escrita

1.3 Obres digitals de consulta disponibles

2 Recursos de suport lingüístic per a llengües minoritzades

2.1 Implementació i recursos lingüístics per a l'alguerès

3 Conclusions

Referències bibliogràfiques

1 El context sociolingüístic de l'alguerès

L'*Atlas of the world's languages in danger* de la UNESCO (Moseley, 2010) classifica les llengües en un contínuum que va des de *safe* fins a *extinct*, passant per *vulnerable* i *definitely*, *severely* i *critically endangered*.[1] Aquesta classificació és el resultat de l'aplicació de nou criteris d'anàlisi de la vulnerabilitat de cada llengua (UNESCO, 2003). Sis d'aquestos criteris són agrupats en una primera subcategoria que fa referència a la vitalitat de la llengua, en què s'inclouen, per exemple, la transmissió intergeneracional o l'adaptabilitat a nous dominis i mitjans (*response to new domains and media*). Dos criteris més fan referència a les actituds lingüístiques a diversos nivells (en el pla institucional i en un altre de més individual), mentre que l'últim factor fa referència a la urgència de la necessitat de documentació. El document adverteix que cap d'aquestos factors, per ell mateix, no pot ésser utilitzat per a avaluar la vitalitat d'una llengua, sinó que és mitjançant la combinació de tots ells i, sobretot, l'aplicació a cada context, que se'n pot obtenir una fotografia més o menys precisa. En el cas del català alguerès (l'única varietat del català inclosa en l'*Atlas*), la UNESCO li atribueix l'estatus de llengua *definitely endangered*.

En el cas d'*Ethnologue*,[2] una altra publicació que té per objectiu catalogar l'estatus de les llengües del món pel que fa al seu risc d'extinció, l'escala de vulnerabilitat té tretze nivells, que van del 0 (llengua "internacional") al 10 (extingida).[3] A diferència de la publicació anterior, *Ethnologue* recull les varietats del català que se parlen a l'Estat espanyol, a la república francesa i a l'Estat italià, i ne proporciona els corresponents nivells de vulnerabilitat. En el cas del català alguerès, la publicació se basa en Argenter (2008) per atribuir-li 7.480 parlants i l'estatus de 4, *educational* (Eberhard, Simons i Fennig, 2019), categoria reservada als idiomes amb una certa presència en l'àmbit de la formació bàsica.[4]

Segons l'*Enquesta d'usos lingüístics de l'Alguer, 2015* (EULA), elaborada per la Direcció General de Política Lingüística de la Generalitat de Catalunya (Ballone, 2017), els algueresos de major edat (més de divuit anys) que tenen una bona competència oral activa en català són el 30,7%[5] de la població (c. 12.000, sobre un total de 44.000 residents), xifra que puja al 72,1% en termes de bona comprensió oral.[6] Pel que fa a la llengua escrita, solament el 3,0% dels algueresos declara saber produir un text de manera autònoma,[7] mentre que poc més d'una quarta part de la població (26,1%) declara saber llegir textos sense massa dificultats (2017: 11).

El mateix estudi reporta valors dramàtics pel que fa a la transmissió intergeneracional de la llengua, com que solament el 3,6% dels genitors més joves (18-44 anys) declara utilitzar aquest idioma com a llengua única o principal amb els fills (2017: 18); al contrari, els valors relatius al nivell d'atractivitat de l'alguerès són decisivament més positius, vist que quasi tres algueresos de cada quatre (72,1%) mostren el desig teòric de voler conèixer i utilitzar la pròpia varietat de català (2017: 28). Aquesta actitud positiva vers la llengua històrica de l'Alguer és evident també en el percentatge de persones que voldrien la introducció de les llengües locals —alguerès inclòs— en les escoles de Sardenya (2017: 27), percentatge que abraça la quasi totalitat de la població resident, és a dir, el 92,3%. Considerant aquesta darrera xifra, és evident la distància entre la voluntat popular de poder tenir una instrucció pública també en català i la falta d'aquesta llengua en l'ensenyament curricular.[8] Aquesta absència resulta encara més greu si considerem que el català de l'Alguer és tutelat, juntament amb altres llengües de l'estat, per la constitució italiana (art. 6), per una llei ordinària estatal (Llei 482/1999) i per dues lleis regionals sardes (Llei 26/1997 i Llei 22/2018).

1 La publicació també preveu una categoria per a les llengües revitalitzades.

2 Es pot consultar l'obra aquí.

3 Alguns d'aquestos nivells tenen subseccions (com ara 8a, *moribund*, i 8b, *nearly* extinct). La categorització completa, juntament amb la metodologia utilitzada, se pot trobar al web.

4 En el cas de l'alguerès, aquesta presència s'havia d'atribuir, en l'època de la publicació d'Argenter (2008), al Projecte Palomba, relatiu a l'escola *elementare* (de sis a deu anys) i *media* (d'onze a tretze anys), i al Projecte Costura (cf. Bosch i Rodoreda, 2007), reservat a una classe d'alumnes d'educació infantil de segon cicle (de tres a sis anys). De moment, cap d'aquestos projectes està actiu.

5 Aquest percentatge inclou els usuaris que tenen com a mínim la "capacitat de portar a terme una conversa sencera sense massa problemes" (Ballone, 2017: 11).

6 Aquest percentatge inclou els usuaris que tenen poc o cap problema de comprensió, independentment de la rapiditat de parla de l'interlocutor (Ballone, 2017: 10).

7 La xifra puja al 6,5% si s'hi inclouen els usuaris que declaren que estan aprenent a escriure en alguerès normatiu, tot i que encara no dominen aquest codi (Ballone, 2017: 12).

8 Caria (2006) oferia una panoràmica exhaustiva de la situació sociolingüística del català alguerès, en la qual destacava una primera secció sobre la presència de la llengua en els mitjans de comunicació, en l'administració i en l'ensenyament.

1.1 La varietat no dominant algueresa en el context pluricèntric del català

Ammon (2005: 1540) posa en relleu el paper mediador de la noció de *pluricentrisme* (Clyne, 1992) en la resolució de les tensions originades entre diversos *centres culturals o polítics* d'una mateixa llengua en processos d'estandardització i de normativització. La dimensió política d'aquestos processos és evident; Ammon (2005: 1536), de fet, proposa diverses categories en funció de la situació política de les comunitats implicades (pluriestatal, pluriregional i plurinacional). Les tensions esmentades, a més, poden ésser agreujades en el cas de les llengües minoritzades o en perill d'extinció, a causa de la urgència d'assolir un estàndard escrit com a primer pas per a la supervivència.

En el cas del català, la bibliografia relativament extensa ne mostra inequívocament la condició pluricèntrica. No solament la llengua és parlada en diversos estats i regions, sinó que, tal com fa notar Mas (2019: 211), "[l]a simple existència física de més d'una institució oficial creada amb la missió de treballar sobre la normativa de la llengua porta a haver de qualificar-la com a policèntrica".

Tenint en compte el context sociolingüístic de l'alguerès descrit a l'apartat 1 (El context sociolingüístic de l'alguerès) —especialment el fet que és l'única varietat del català considerada en perill d'extinció per la UNESCO—, juntament amb la condició d'ésser "la més perifèrica de les comunitats dins l'àmbit catalanoparlant" (Argenter, 2010: 133) i també el seu pes demogràfic, és evident que els desacords en el procés d'estandardització de l'alguerès poden ésser analitzats des de la perspectiva de la teoria dels centres. En aquest sentit, en termes de Muhr (2012), caldria considerar l'alguerès una varietat no dominant (VND) del sistema lingüístic català. Són, d'altra banda, aquestos mateixos motius els que justifiquen la urgència d'emprendre accions, seguint l'exemple d'altres VND, com ara la valenciana (Mas, 2019).

1.2 Procés de normativització de la llengua escrita

Els dos estudis considerats tradicionalment com a primers exemples de gramàtica de l'alguerès, és a dir, la *Grammatica del dialetto alghерese odierno* (1906) de Joan Palomba i la *Gramàtica algueresa* de Joan Pais, publicada pòstuma el 1970, no poden ésser inserits en realitat plenament en el procés de normativització de l'alguerès modern, vist que en el primer cas, més que d'una gramàtica, se tracta d'una aproximació descriptiva (molt parcial i insegura) a la fonètica de l'alguerès, i, en el segon cas, d'un treball normatiu de tipus prefabrià.[9] Els primers documents dels quals tenim notícia i que eren finalitzats a la introducció en l'alguerès de l'ortografia catalana moderna són estats redactats per l'intel·lectual alguerès Antoni Simon Mossa; aquestos constituïen el material didàctic dels cursos d'alguerès de l'Escoleta del Bon Pescador, començats el 1959 i acabats el 1970.

La primera publicació que ha permès a centenars d'algueresos de familiaritzar amb l'alguerès normatiu és la *Santa Missa* (Nughes i Sanna, 1980), seguida, alguns anys després, per l'obra que probablement ha tengut més impacte sobre el procés de normativització d'aquesta varietat de català, és a dir, el *Diccionari català de l'Alguer* (a partir d'ara, DCA), a cura de Josep Sanna. Un altre text fonamental dins aquest procés és *El català de l'Alguer: un model d'àmbit restringit* (a partir d'ara, MÀR), curat per Luca Scala (2003) i publicat per l'Institut d'Estudis Catalans. Aquestos documents, integrats sovent per diccionaris i gramàtiques del català general, han constituït la base que ha permès en els darrers decennis la producció d'una gran quantitat de textos i publicacions en alguerès normatiu. En diversos casos, però, en aquesta producció escrita són presents variants formals que van més enllà del simple àmbit estilístic, i, de fet, són el resultat de l'adopció de criteris ortogràfics en part diferents entre un autor i l'altre. Per citar-ne alguns exemples:

a) Transcripció de la [u] final àtona en paraules no catalanes o amb la fonètica afectada per altres llengües: *sard[u]*, *dueny[u]* ("amo"), *porqued[u]* ("porc petit"), etc. Alguns autors han mostrat preferència pel criteri etimològic (per ex., R. Caria), segons el qual la [u] en qüestió s'hauria d'ortografiar *u* en sardismes (*porquedu*) i *o* en castellanismes i/o italianismes (*duenyo*), mentre altres han preferit el criteri contextual utilitzat al DCA, segons el qual s'escriu *u* si la vocal final és precedida d'una altra vocal (*botaiu*, "boter") i *o* si és precedida de consonant (*porquedo*). Com a resultat, en literatura trobem paraules com *sard*

9 Per exemple, s'hi troben indicacions per ortografiar amb el dígraf *ch* les oclusives velars finals (*dich*), i s'utilitza el grafema y per indicar la conjunció *i*.

escrites en tres maneres diferents, és a dir, *sardu* (amb [u] atribuïda a la influència del sard), *sardo* (criteri contextual), *sard* (model del català general).

b) Distribució de les formes sil·làbica i asil·làbica de l'article definit *lo*, pel qual són disponibles en literatura almenys dues propostes: la primera, proposada per Pais (1970) i utilitzada en la *Missa* en alguerès, segons la qual s'ha d'utilitzar la forma *el* després de paraula acabada amb: vocal, -[k], -[p], -[t], -[t͡ʃ] i -[t͡s] (*M'agrada més la carn que el peix*, *M'acab el còmpito*), sempre que el mot que segueix l'article no comenci amb [d͡ʒ], [j], [l], [ʎ], [r], [s], [t͡ʃ], [ʃ], [t͡s], cas en el qual el mateix article conserva la forma *lo(s)* (*veig los relotges*); la segona, present a MÀR, preveu l'ús de la forma *el* després de paraula que acaba amb vocal (*talla-te els cabells*), *r* muda (*obrir els ulls*), preposició *amb* (*amb el temps*), i consent l'ús de *lo* en els altres casos (*lo pare*).

c) Ús de l'accentuació general (*alguerès*) o local (*alguerés*) en els casos on el timbre de *e* tònica correspon a la pronúncia valenciana i no a aquella del català oriental.

d) Variació en l'ús d'elisió gràfica en alguns casos característics de l'alguerès (*de amic* / *d'amic*, *se hi vol comprar* / *s'hi vol comprar*, etc.).

e) Variació en l'ús de formes ortogràfiques més representatives de la fonètica local (*nostro*, *vostro*, *algua*, etc.) o més adherents a la llengua general (*nostre*, *vostre*, *aigua*, etc.).

Aquestes (i altres) diferències formals han sovent alimentat la idea en una part de la població que l'alguerès no disposa d'un model ortogràfic local compartit (si més no) per la majoria de les associacions culturals i els operadors lingüístics. Per aquesta raó, la Consulta Cívica per les Polítiques Lingüístiques del Català de l'Alguer[10] (a partir d'ara, Consulta) ha creat un grup de treball per a la normativització del català de l'Alguer que té la finalitat d'elaborar propostes normatives que valorin alhora convergències internes, entre associacions i operadors lingüístics locals, i externes, tenint com a punt de referència les normes del català general, a partir de l'*Ortografia catalana* i de la *Gramàtica catalana* de l'IEC.

1.3 Obres digitals de consulta disponibles

Si considerem les taxonomies dialectals més comunes en la dialectologia catalana (per ex., Veny, 1982[2002]), l'alguerès resulta la varietat amb el nombre més reduït de parlants. La situació no canvia gaire en el vessant de l'Estat italià, del qual l'Alguer fa part, considerant que els catalanoparlants històrics constitueixen una entre les minories quantitativament menys representatives de la República.[11] Tot i això, la supervivència del català en un racó aïllat de Sardenya ha suscitat l'interès de nombrosos estudiosos, que s'ha manifestat —entre d'altres— en la producció d'un nombre impressionant d'articles i recerques de tipus especialment lingüístic i sociolingüístic.[12]

Aquest interès és visible també en termes de recursos digitals presents en la xarxa en pàgines web d'accés lliure. Ne mostrem aquí baix alguns exemples:

a) Corpus orals: és possible trobar en la xarxa —en conjunt— centenars d'hores de producció oral, a voltes amb algun tipus d'estructuració i informació basilar sobre context, informants i temes tractats,[13] i altres voltes amb material penjat amb informació complementària escassa o inexistent. Un exemple de corpus oral ben estructurat és el Corpus Oral de l'Alguerès, que conté diverses hores de parla espontània i semiespontània produïda per parlants natius; la pàgina inclou també la transcripció ortogràfica dels textos orals recollits.

b) Corpus lèxics tematitzats: fan part d'aquesta categoria els reculls de mots pertanyents a camps semàntics específics, com ara l'*Atles lingüístic del domini català*, que reporta milers de termes tradicionals per cadascuna de les dues-centes varietats de català seleccionades, alguerès inclòs.

10 La Consulta és un òrgan consultiu del municipi de l'Alguer instituït el 2018 per participar a les activitats de tutela i promoció de l'alguerès, en una òptica unitària de la llengua catalana.

11 Cf. Orioles (2003).

12 Cf., a títol d'exemple, Ibba (2004).

13 Per exemple, els materials produïts pel mitjà de comunicació local Catalan TV i que se poden trobar al seu canal de YouTube.

c) Corpus lèxics: en aquest àmbit, el treball que de moment ha tengut l'impacte major sobre la comunitat dels usuaris de l'alguerès és el *Diccionari de alguerés*, un projecte col·laboratiu de diccionari en línia que naix a partir de la inserció en una plataforma web del text del DCA, integrat per material provinent d'altres estudis, com Corbera Pou (2000), i també de l'activitat de recerca en àmbit del lèxic tradicional portada a terme per alguns voluntaris del grup de treball. En la xarxa són disponibles també altres indrets on s'han recollit porcions de lèxic alguerès, com el *Diccionari català-valencià-balear* (a partir d'ara, DCVB) i algunes breus videolliçons que tenen com a tema paraules o maneres de dir tradicionals.[14]

d) Recollides de textos escrits: aquesta categoria inclou documents històrics,[15] cançoners,[16] estudis científics,[17] articles a la Viquipèdia[18] i altres tipologies de textos.

Aquesta llista és exemplificativa i en cap manera exhaustiva, i podria enriquir-se ulteriorment amb seccions relatives al teatre (Associació per a la Salvaguarda del Patrimoni Historicocultural de l'Alguer), pàgines institucionals (Comune Alghero), música (Pino Piras), etc.

Recordem que tot el material citat és disponible en pàgines d'accés lliure, però fem notar que, en molts casos, l'accés efectiu és subordinat al coneixement directe de les adreces en què se troba el corpus o a la inserció d'una sèrie de paraules clau molt específiques dins els motors de recerca, condicions aquestes que podrien limitar la consulta del mateix material a usuaris no especialistes. A partir d'aquestes consideracions, el grup de treball de l'IEC a l'Alguer està portant a terme el projecte www.llenguamia.cat —actualment (desembre de 2019) en la seva versió beta—, que té la finalitat de facilitar també a un públic no especialista la consulta de diferents tipologies de material digital present a la xarxa, a través d'un treball d'indexació de contingut i l'atribució a cada objecte indexat d'una sèrie de paraules clau lligades, per exemple, al tema tractat, a la tipologia de mitjà que caracteritza el contingut (vídeo, àudio, text), a l'autor, etc.[19]

2 Recursos de suport lingüístic per a llengües minoritzades

En la secció precedent ens hem centrat en objectes digitals definits com a obres "de consulta" (diccionaris i reculls de textos), mentre que en la present tractarem recursos i aplicacions de suport lingüístic, com ara correctors i traductors.

La pregunta a la qual intentarem donar resposta en els paràgrafs següents és si una llengua minoritzada se pot permetre avui de no tenir en la justa consideració la possibilitat de servir-se, per reforçar-se o fins i tot per sobreviure, també d'eines de suport lingüístic digital.

Tal com recorda el relator especial de les Nacions Unides en qüestions de minories al seu informe, "[i] nnovations such as using new technologies and the Internet offer encouraging approaches to reaching small groups or widely dispersed minorities" (United Nations Special Rapporteur on minority issues, 2017). És probablement per aquest potencial que hi ha diversos exemples de comunitats lingüístiques minoritzades que s'han organitzat al voltant de grups d'usuaris de tecnologia (Martín-Mor, 2017). També des de l'àmbit acadèmic són innombrables les iniciatives adreçades a generar tecnologies per a llengües minoritzades. En aquest sentit, experiències com la de la comunitat bascòfona resulten d'un alt interès, pel fet que contribueixen a establir un seguit de bones pràctiques (de manera força estructurada) que poden ésser preses per altres comunitats com a full de ruta. El grup IXA, ja el 2001, posava en relleu la importància de la col·laboració entre acadèmia i associacions de voluntaris, i proposava una metodologia de cinc fases per al desenvolupament de tecnologies (Agirre et al., 2001), començant per simples corpus textuals i acabant per eines per al processament automàtic de traduccions.[20]

14 Vegeu-ne un exemple aquí.

15 Vegeu-ne un exemple aquí

16 Vegeu-ne un exemple aquí.

17 Vegeu-ne un exemple aquí.

18 Vegeu-ne un exemple aquí.

19 El projecte Els sons del català és una plataforma didàctica que permet als usuaris dels diferents territoris del domini de familiaritzar amb la fonètica de les varietats del català. El projecte té previst inserir, per a la primera part del 2020, també l'apartat alguerès.

20 Atès que l'article referenciat és del 2001, creiem que el grup se referia a la traducció automàtica quan esmentava "translation of noun phrases and simple sentences" (Agirre et al., 2001: 5).

No és casual que entre les tecnologies que el grup IXA proposa en una primera fase hi hagi els corpus textuals. És, de fet, sobre la base de corpus textuals que se poden construir recursos lingüístics com ara correctors ortotipogràfics o traductors automàtics. Els corpus paral·lels, per exemple, permeten entrenar motors de traducció automàtica estadística (TAE) o neuronal (TAN) sense grans inversions de recursos (ni humans ni econòmics). Per contra, l'absència d'un corpus digital consistent —una situació ben habitual en l'àmbit de les llengües minoritzades—, si bé impedeix el desenvolupament de tecnologies com les esmentades, pot ésser suplida amb estratègies alternatives. En qualsevol cas, la cerca i compilació de corpus paral·lels solen concentrar una part considerable dels esforços dels projectes de recerca en llengües minoritzades. Entre les estratègies habituals per a augmentar la mida dels corpus per a llengües minoritzades, hi ha l'ús de tècniques de *web scraping* per a descarregar i alinear llocs web multilingües (Doğru, Martín-Mor, i Aguilar-Amat, 2018), o la combinació de corpus especialitzats amb corpus generals. Més recentment, algunes línies de recerca apunten cap a una metodologia per suplir la manca de corpus paral·lels amb retraducció (*back-translation*) per a la creació de motors de TAE i de TAN (Artetxe, Labaka, i Agirre, 2018; Artetxe, Labaka, Agirre, i Cho, 2018). En el camp de les tecnologies de reconeixement de la parla, de manera paral·lela, la disponibilitat de corpus orals facilita el desenvolupament de tecnologies de base estadística.[21]

En aquest sentit, la UNESCO (2008) suggereix que el programari lliure "can play an important role as a practical instrument for development as its free and aspirations make it a natural component of development efforts in the context of the Millennium Development Goals". De fet, tal com fan notar Paricio-Martín i Martínez Cortés (2010: 7) per a la llengua aragonesa, la possibilitat d'unir esforços amb altres comunitats mitjançant l'ús o l'adaptació de plataformes existents a les pròpies necessitats contribueix a l'alta presència de llengües minoritzades en els programes lliures i de codi obert.[22] En el cas del català, això és especialment cert en casos com el dels correctors gramaticals i de la traducció automàtica (TA): LanguageTool, un dels correctors gramaticals més utilitzats i amb llicència lliure, incorpora el català amb un nombre de regles molt superior al de llengües amb molts més parlants (com ara l'anglès, l'italià o el castellà);[23] Apertium, en el camp de la traducció automàtica basada en regles (TABR), incorpora, a més, llengües tan poc avesades a tenir presència en aquestes tecnologies com el sard (Tyers, Alòs i Font, Fronteddu, i Martín-Mor, 2017) o l'occità.

A banda de l'avantatge que suposa per a les llengües minoritzades la possibilitat de fer servir plataformes lliures, és menys conegut que aquestes se poden permetre reflectir la variació lingüística de manera diferent al programari de propietat. Mas (2019) analitza diverses eines de suport lingüístic (entre les quals hi ha els esmentats LanguageTool i Apertium, i també programari de propietat) i la manera com presenten les variants del català, ja que:

> "[a] diferència d'un text escrit, que haurà d'utilitzar una variant o una altra indefectiblement, el corrector o traductor automàtic pot contindre els diversos geosinònims (lèxics, morfosintàctics, ortogràfics...) admesos per la normativa i deixar-ne la tria al criteri de l'usuari, siga en peu d'igualtat, siga recomanant-ne l'ús d'unes en detriment d'altres." (Mas, 2019: 213.)

L'estudi posa en relleu el fet que sovent el programari lliure és adoptat per les institucions, i confirma en bona part (i amb algun matís) que el programari lliure "reflecteix la variació lingüística i nominal d'una manera més pròxima a l'acadèmia que el programari comercial, que tendiria més a l'isolament territorial" (Mas, 2019: 220). En concret, aquest reflex té lloc tant en l'àmbit de la presentació de la llengua (com ara la denominació *català/valencià*) com en el de la representació de la variació.

21 Resulten d'interès en aquest sentit projectes com ara Common Voice, una plataforma col·laborativa impulsada per Mozilla amb l'objectiu de generar tecnologies de reconeixement lingüístic per a qualsevol llengua del món. La plataforma ofereix als usuaris la possibilitat de carregar corpus textuals i, en un segon moment, enregistrar directament a la plataforma la lectura de segments del corpus. Se genera així un corpus oral prou consistent (amb variació dialectal i idiolectes) per a entrenar sistemes de base estadística (Ardila et al., 2019).

22 Diversos programes han optat els últims anys per metodologies de traducció col•laborativa, entre els quals n'hi ha alguns de difusió mundial com Facebook. Si bé les llengües minoritzades poden aprofitar aquest fet per incrementar la seva visibilitat, també és cert que, en funció de la llicència del programa, les traduccions no se puguin reaprofitar per a altres finalitats (Martín-Mor i Beccu, 2016).

23 Segons l'apartat Languages del web de l'eina LanguageTool.

En qualsevol cas, malgrat aquesta tendència cap al pluricentrisme del programari lliure en català, cal fer ressaltar que cap dels programes analitzats per Mas no inclou l'alguerès, una de les varietats de la llengua amb una necessitat més urgent de tecnologies lingüístiques (vegeu l'apartat 2.1, Implementació i recursos lingüístics per a l'alguerès).

2.1 Implementació i recursos lingüístics per a l'alguerès

En termes generals, és evident el gran desequilibri quantitatiu entre les nombroses obres de consulta digital en alguerès i els pocs recursos de suport lingüístic. Sobretot, fa reflexionar l'absència de correctors ortogràfics o gramaticals pensats per a aquesta varietat. Ja hem citat la importància dels corpus textuals digitalitzats que poden constituir la base per a la creació de correctors automàtics (ortotipogràfics, gramaticals, etc.) o de traductors automàtics. Si considerem la gran quantitat i riquesa d'obres digitals, fins i tot de tipus textual, citada a 1.3 (Obres digitals de consulta disponibles), se podria pensar que ja avui hi ha una base suficient per reflectir el pluricentrisme del programari lliure en català adaptant a l'alguerès la base de dades ja existent (tant per fer un exemple) a LanguageTool. En realitat, com hem comentat a 1.2 (Procés de normativització de la llengua escrita), els textos locals de referència normativa (especialment DCA i MÀR) estan actualment en fase de revisió per part de la Consulta; altres reculls lèxics, com el del projecte Diccionari de alguerés, tenen una funció més divulgativa que normativa, i, per tant, no poden constituir —de moment— una base de dades fiable. Un altre possible text de referència, el DCVB, no facilita la cerca sistemàtica de les entrades algueresas, com que el portal que allotja el diccionari no permet recuperar les entrades més que a partir del lema de les paraules.[24]

Un cas anàleg és el de la Viquipèdia catalana, que inclou articles escrits en alguerès, però que no permet —a diferència del que passa en altres versions de la Viquipèdia, com ara la mateixa Viquipèdia sarda— l'etiquetatge lingüístic en funció de la varietat, cosa que, a la pràctica i fins on sabem, no dona la possibilitat de recuperar de manera automatitzada les entrades escrites en català de l'Alguer (Martín-Mor, 2017: 378).

Un altre obstacle a una eventual immediata possibilitat d'estructurar una base de dades per un programari lliure de correcció automàtica és que no hi ha encara cap estudi general i sistemàtic sobre la morfologia verbal algueresa, la qual se discosta sovent dels models de les altres varietats (inclosa la balear) i que, per tant, no pot sempre utilitzar models de referència normatius ja aplicats en altres varietats catalanes.

Tot això significa que no se pot treballar ja avui en la perspectiva de crear recursos de suport lingüístic? La pregunta és retòrica i la resposta és, òbviament, negativa. El mateix Grup per a la Normativització de l'Alguerès està produint, juntament amb propostes de revisió normativa, documents detallats que permeten començar la digitalització i indexació de les entrades del DCA sobre les quals s'aplica la revisió. Igualment, ja a partir d'ara seria possible introduir tota la part del DCA que representa l'àmbit tecnicoespecialístic modern, per al qual no és prevista cap diferència formal entre varietat local i general.[25]

Portar a terme aquest procés significarà poder inserir l'alguerès als programes lingüístics (traductors automàtics i correctors ortogràfics i gramaticals), com aqueixos ressenyats per Mas (2019: 215). També se podrà inserir l'alguerès en la majoria dels aparells de butxaca (telèfons, tauletes, lectors de llibres electrònics, etc.),[26] amb beneficis per l'ús normal de la llengua fàcilment previsibles en casos com els correctors ortogràfics amb text predictiu.

En aquest context, les línies que segueixen se proposen d'identificar àmbits per a incrementar la presència de l'alguerès en la tecnologia. Aquesta proposta se basa en l'ús de programari amb llicència lliure, ja que ne permet la modificació sense restriccions. En primer lloc, i tenint en compte que el procés d'estandardització de l'alguerès encara és en curs, és evident que manquen els fonaments sobre els quals se podria construir la majoria de les tecnologies esmentades a l'apartat 2 (Recursos de suport lingüístic per a llengües

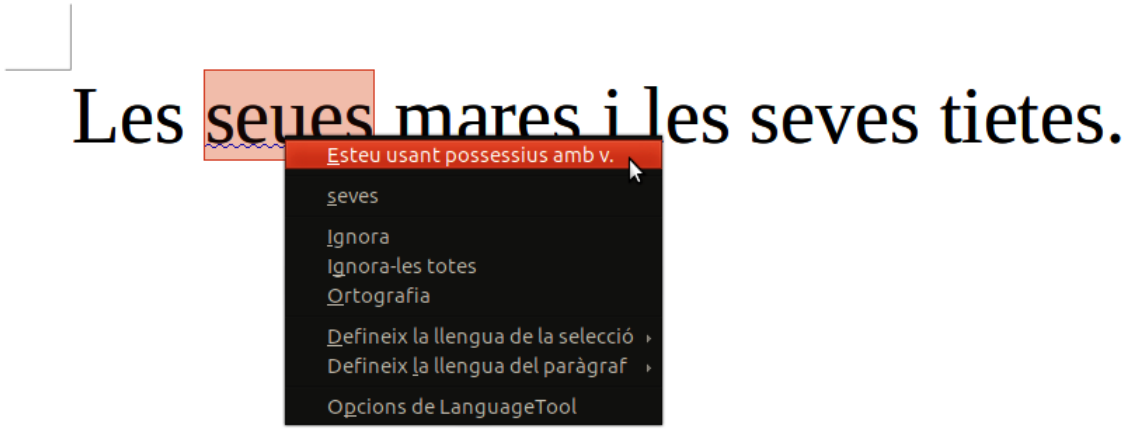
24 Se poden utilitzar, però, buscadors genèrics d'internet (com Google o DuckDuckGo) per recuperar les entrades marcades amb les etiquetes *alg*. i *Alg*. amb una cerca com: "'alg.' site:dcvb.iec.cat".

25 També en els pocs casos en què alguna diferència és possible, com en la formació del plural dels mots acabats en consonant sibilant seguida per una altra consonant (*asterisc*/*asteriscos*), és suficient adaptar a l'alguerès les regles ja utilitzades en les varietats compatibles (*asteriscos*).

26 Segons la guia elaborada per Softcatalà, un 75% dels dispositius mòbils (telèfons, tauletes, lectors de llibres electrònics, etc.) se pot configurar en català, i un percentatge similar dels aparells té un corrector ortogràfic en català per mitjà d'un teclat virtual.

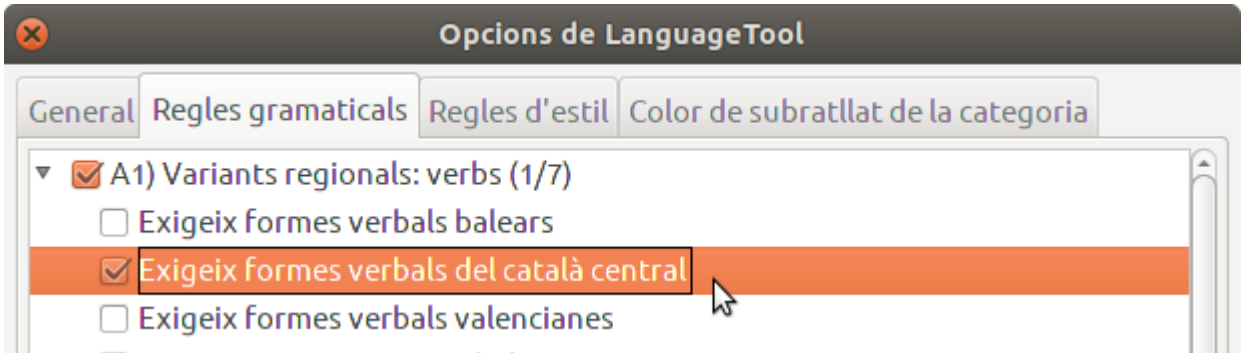
minoritzades). Amb tot, les obres de referències normatives, degudament digitalitzades, ja podrien servir de base per a desenvolupar diccionaris de sinònims i de partició de mots en format de fitxer de text, a partir de les definicions i les equivalències entre variants.

En l'àmbit dels correctors ortogràfics i gramaticals, el treball de Softvalencià ofereix un bon exemple que podria ésser adaptat al cas de l'alguerès: el diccionari Hunspell de valencià[27] conté les entrades del diccionari de català general i les amplia amb les formes pròpies del valencià. Quant al programa per a la correcció gramatical LanguageTool, incorpora tres varietats del català ("central", valencià i balear). Un exemple il·lustratiu és el dels possessius: LanguageTool mira de garantir la coherència en un mateix document, tal com mostra la imatge següent.



Imatge 1. Incoherència en l'ús de possessius detectada per LanguageTool.

Una anàlisi de la normativa per a l'alguerès permetria convertir en regles informàtiques les regles lingüístiques per tal d'informar sobre l'adequació de les produccions textuals a l'alguerès normatiu. Qüestions com la triada de l'accentuació local (*alguerés*) o general (*alguerès*) podrien ésser incorporades entre les opcions de personalització:



Imatge 2. Opcions de configuració de LanguageTool en LibreOffice

Idealment, l'existència d'aquestos primers recursos contribuiria a un increment de la producció de textos digitals i, per conseqüència, a la disponibilitat de corpus sobre els quals eventualment se podrien desenvolupar recursos basats en tecnologies estadístiques. Atesa una grandària suficient, aquestos corpus podrien servir de base per a traductors estadístics o motors de reconeixement de veu (vegeu la secció 2, Recursos de suport lingüístic per a llengües minoritzades). El procés de creació de corpus digitals se pot agilitzar per mitjà de l'adaptació de manera automatitzada de textos escrits en altres varietats del català. En aquest sentit, la de

27 Hunspell és un corrector ortogràfic lliure per a diverses llengües, entre les quals hi ha el català.

l'adaptador de variants de Softvalencià oferiria, un cop més, una via per explorar: se tracta d'un conjunt de macros de Python per a adaptar textos del català central al valencià. Mentre això no és possible, se pot recórrer a la TABR. Apertium, de fet, ja ofereix algunes combinacions lingüístiques que inclouen variants, com ara en el cas del català i el valencià, el portuguès europeu i del Brasil, o l'occità general i l'aranès. El fet d'afegir regles específiques per a l'alguerès a un motor d'Apertium, a banda d'acostar el model de llengua a aquesta varietat, permetria reduir l'esforç de postedició per a finalitats de disseminació (Forcada, 2009), de manera que se'n facilitaria l'adopció.

3 Conclusions

Les dades sociolingüístiques presentades revelen una situació crítica, amb una alta predisposició de la ciutadania algueresa cap a l'ús de la llengua que no se correspon amb l'ús social. És en aquest context d'emergència que les iniciatives empreses —també des de l'àmbit polític— per revertir les xifres topen amb la necessitat imperiosa de disposar d'un marc normatiu de referència. Per elaborar aquest marc, però, és urgent disposar de recursos lingüístics bàsics. La tecnologia pot ésser un element que contribueixi a la disrupció d'aquest cercle viciós, propi de les situacions de llengües en perill. Més encara, el sol fet de desenvolupar tecnologies lingüístiques pot enriquir el procés d'estandardització, alhora que el procés d'estandardització facilita la tasca del desenvolupament. És convenient, doncs, per tal d'afavorir aquesta retroalimentació, que els progressos en tots dos àmbits avancin de manera paral·lela i coordinada.

Com recordàvem ara, la població algueresa demanda en bona part instruments que li permetin d'ampliar els coneixements i, per tant, incrementar l'ús social de la llengua. També en aquest pla formatiu la tecnologia és un agent col·laborador irrenunciable: un cop generats els instruments lingüístics, convé emprendre accions específiques per a facilitar-ne l'adopció a tots els àmbits (institucional, educatiu, social, etc.). Les sinergies entre l'àmbit acadèmic, el sector públic i l'associacionisme, habituals en comunitats de llengües minoritzades, solen portar resultats positius. Atès que ja hi ha exemples d'associacions de la resta del domini lingüístic que han reforçat la seva presència al territori de l'Alguer, una línia de treball paral·lela amb associacions que treballin amb les tecnologies lingüístiques (Softcatalà, Amical Wikimedia, etc.) comportaria, ben segur, un augment de la presència de l'alguerès en la tecnologia.

Per últim, tal com s'ha esmentat a l'apartat 1.2 (Procés de normativització de la llengua escrita), cal considerar que els avenços assolits en termes de normativització en el marc de la Consulta fan menys arbitrària (en tant que més consensuada) la presa de decisions, la qual cosa agilitzarà el procés de normalització de l'alguerès en la tecnologia. Serà indispensable, doncs, que se publiquin en format digital i obert les obres de referència de caràcter normatiu de l'alguerès. És justament aquesta digitalització el que determinarà la viabilitat de la resta de tecnologies que encara no existeixen per a l'alguerès.

Referències bibliogràfiques

Agirre, Eneko, Aldezabal, Izaskun, Alegria, Iñaki, Arregi, Xabier, Arriola, Jose Mari, Artola, Xabier, i Soroa, Aitor, et al. (2001). Developing language technology for a minority language: progress and strategy. *ELSNews*, *10*(1. Special issue on minority languages), 4-5.

Ammon, Ulrich. (2005). Pluricentric and divided languages. *Sociolinguistics. An International Handbook of the Science of Language and Society*, *2*, 1536-1543.

Ardila, Rosana, Branson, Megan, Davis, Kelly, Henretty, Michael, Kohler, Michael, Meyer, Josh, i Weber, Gregor, et al. (2019). Common Voice: a massively-multilingual speech corpus. *ArXiv:1912.06670 [Cs]*.

Argenter, Joan A. (2008). L'Alguer (Alghero), a Catalan linguistic enclave in Sardinia. *International Journal of the Sociology of Language*, *2008*(193-194).

Argenter, Joan A. (2010). Comunitat perifèrica local, llengua tradicional i invisibilitat del centre. *Revista de l'Alguer*, *9*(9), 127-136.

Artetxe, Mikel, Labaka, Gorka, i Agirre, Eneko. (2018). Unsupervised statistical machine translation. *ArXiv:1809.01272 [Cs]*.

Artetxe, Mikel, Labaka, Gorka, Agirre, Eneko, i Cho, Kyunghyun. (2018). Unsupervised neural machine translation. *Proceedings of the Sixth International Conference on Learning Representations*.

Ballone, Francesc (ed.). (2017). *Els usos lingüístics a l'Alguer 2015*. Generalitat de Catalunya, Departament de Cultura, Direcció General de Política Lingüística.

Bosch i Rodoreda, Manuel. (2007). El català de l'Alguer, entre la desaparició i la dissolució. Dins Germà Colón Domènech i Lluís Gimeno Betí (ed.), *Ecologia lingüística i desaparició de llengües* (p. 35-52). Castelló de la Plana: Universitat Jaume I.

Caria, Rafael. (2006). El català a l'Alguer: Apunts per a un llibre blanc. *Revista de Llengua i Dret*, *46*.

Clyne, Michael. (1992). Pluricentric languages – Introduction. Dins Michael Clyne (ed.), *Pluricentric languages* (p. 1-10). https://doi.org/10.1515/9783110888140.1

Corbera Pou, Jaume. (2000). *Caracterització del lèxic alguerès*. Palma: Universitat de les Illes Balears.

Doğru, Gökhan, Martín-Mor, Adrià, i Aguilar-Amat, Anna. (2018). Parallel corpora preparation for machine translation of low-resource languages: Turkish to English cardiology corpora. Dins Maite Melero, Martin Krallinger, i Aitor Gonzalez-Agirre (ed.), *Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing"* (p. 12-15).

Eberhard, David M., Simons, Gary F., i Fennig, Charles D. (ed.). (2019). *Ethnologue: Languages of the World* (22a ed.).

Forcada, Mikel L. (2009). Apertium: Traducció automàtica de codi obert per a les llengües romàniques. *Linguamática*, *1*(1), 13-23.

Ibba, Joan. (2004). *Bibliografia algueresa*. L'Alguer: Edizioni del Sole.

Martín-Mor, Adrià, i Beccu, Alessandro. (2016). Sa tradutzione de Facebook in sardu. *Revista Tradumàtica: Tecnologies de la Traducció*, *14*, 85-99. https://doi.org/10.5565/rev/tradumatica.179

Martín-Mor, Adrià. (2017). Technologies for endangered languages: The languages of Sardinia as a case in point. *MTm*, *9*, 365-386.

Mas, Josep Àngel. (2019). El pluricentrisme de la llengua catalana en els principals correctors i traductors automàtics. *Revista de Llengua i Dret*, *71*, 208-222. https://doi.org/10.2436/rld.i71.2019.3229

Moseley, Christopher (ed.). (2010). *Atlas of the world's languages in danger* (3a ed.). UNESCO.

Muhr, Rudolf (ed.). (2012). *Non-dominant varieties of pluricentric languages. Getting the picture*. https://doi.org/10.3726/978-3-653-01621-5

Nughes, Antoni, i Sanna, Josep. (1980). *Santa Missa pels fidels de l'Alguer*. L'Alguer: La Poligràfica Peana.

Orioles, Vincenzo. (2003). *Le minoranze linguistiche. Profili sociolinguistici e quadro dei documenti di tutela*. Roma: Il Calamo.

Pais, Joan. (1970, a cura de Pasqual Scanu). *Gramàtica algueresa*. Barcelona: Editorial Barcino.

Palomba, Joan. (1906 [2001], a cura de Francesco Bertino). *Grammatica del dialetto algherese odierno*. L'Alguer: Edicions Obra Cultural de l'Alguer.

Paricio-Martín, Santiago Jorge, i Martínez-Cortés, Juan Pablo. (2010). New ways to revitalise minority languages: the impact of the internet in the case of Aragonese. *Digithum*, *12*, 1-11.

Sampson, Geoffrey. (2001). What is a minority language? *ELSNews*, *10*(1. Special issue on minority languages), 1-2.

Sanna, Josep. (1988). *Diccionari català de l'Alguer*. L'Alguer, Barcelona: Fundació del II Congrés de la Llengua Catalana.

Scala, Luca. (2003). *El català de l'Alguer: un model d'àmbit restringit*. Barcelona: Institut d'Estudis Catalans.

Tyers, Francis M., Alòs i Font, Hèctor, Fronteddu, Gianfranco, i Martín-Mor, Adrià. (2017). Rule-based machine translation for the Italian–Sardinian language pair. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 221-232. https://doi.org/10.1515/pralin-2017-0022.

UNESCO. (2003). *Language vitality and endangerment*.

UNESCO. (2008). *Proprietary and free and open source software*.

United Nations Special Rapporteur on minority issues. (2017). *Language rights of linguistic minorities: A practical guide for implementation*.

Veny, Joan. (1982[2002]). *Els parlars catalans. Síntesi de dialectologia*. Palma: Moll.