Proposta rebuda el 15 d'octubre 2017 i acceptada per a publicació el 3 de desembre 2017.

Una eina per a una llengua en procés d'estandardització: el traductor automàtic català–sard

Machine translation from Catalan to Sardinian: a translation tool for a language in the process of standardisation

Gianfranco Fronteddu
Università degli Studi di Cagliari
gfro3d@gmail.com

Hèctor Alòs i Font
Universitat de Barcelona
hectoralos@gmail.com

Francis M. Tyers
Higher School of Economics
ftyers@hse.ru

Resum

Aquest article presenta el desenvolupament d'un sistema de traducció automàtica en codi obert basat en regles del català al sard mitjançant la plataforma Apertium, parant una atenció especial a la creació del diccionari bilingüe i de les regles de selecció lèxica i transferència estructural. Es mostren alguns problemes derivats de l'estat actual del sard estàndard. S'ha obtingut una tassa d'error per paraula (WER) del 20,5% i una tassa d'error per paraula independent de la posició (PER) del 13,9%. Mitjançant l'anàlisi qualitativa de la traducció de quatre articles enciclopèdics, s'analitzen les causes d'aquests resultats.

Paraules clau

sard, català, traducció automàtica, estandardització lingüística, Apertium, RBMT

Abstract

This article describes the development of a free/open-source rule-based machine translation system for Catalan to Sardinian based on the Apertium platform. Special attention is given to the components of the system related with transfer (structural and lexical) and lexical selection, drawing attention to issues stemming from the current state of the Sardinian written norm. The system has a word-error rate (WER) of 20.5% and a position-independent word-error rate (PER) of 13.9%. We analyse the remaining errors by doing a qualitative analysis of the translation of four articles from the encyclopaedic domain.

Keywords

Sardinian, Catalan, machine translation, language standardisation, Apertium, RBMT

1 Introducció

Aquest article presenta un sistema de traducció automàtica del català al sard basat en regles i en codi obert. Es tracta de dues llengües romàniques, la qual cosa facilita l'ús d'un sistema de transferència superficial com Apertium (Forcada et al., 2011).

L'objectiu del projecte ha estat crear un sistema de traducció que sigui capaç de traduir textos del català al sard amb una qualitat que permeti una postedició ràpida per produir un document de qualitat. Això és especialment rellevant per a una llengua com el sard, amb un nombre de recursos electrònics reduït, en particular en la varietat normativa, com es veurà més endavant.

L'objectiu bàsic del traductor és facilitar el creixement de recursos textuals en sard a Internet. Disposar d'un traductor automàtic des d'una llengua que no és la dominant (en aquest cas, l'italià) permet de posar a l'abast dels parlants de sard textos que podrien entendre només amb dificultat. Un cas paradigmàtic és la Viquipèdia, en què l'aplicació Content Translation (Laxström et al., 2015) facilita la creació de nous articles, utilitzant, si està al seu abast, traducció automàtica. En aquest cas, traduir un text de la llengua dominant a la llengua minoritzada representa, sens dubte, un enriquiment per a la llengua minoritzada en tant que incrementa els recursos que hi ha en ella. Tanmateix, per al parlant de la llengua minoritzada, que ben sovint entén bé la llengua dominant (i no poc sovint està més acostumat a llegir i escriure en ella que en la pròpia), la informació de què disposa al seu abast és pràcticament la mateixa (això sí: en la llengua que prefereixi de les dues). En canvi, poder traduir d'una altra llengua permet accedir a un contingut diferent del que ja té al seu abast.

La tria del català per a aquesta llengua diferent de la dominant es deu a diferents raons. Una és la llarga relació històrica de Catalunya i

DOI: 10.21814/lm.9.2.255

Sardenya. Això és font d'una gran quantitat de textos, testimonis i material en llengua catalana sobre la història de Sardenya que ara podrà ser disponible també en sard. Alhora, també ho estaran les nombroses publicacions i els estudis de sociolingüística i de política lingüística en català, que són de gran interès per l'estat actual de la llengua sarda. D'una manera més pragmàtica, el català és una de les llengües en què més s'ha treballat dins d'Apertium, per la qual cosa disposa d'un extens diccionari morfològic, així com d'un desambiguador morfològic força fiable. Per això, desenvolupar un traductor del català a una altra llengua romànica en Apertium resulta especialment ràpid.

Tanmateix, com es descriu més avall, el sard no es pot considerar una llengua plenament normativitzada. Això implica que els recursos lingüístics de què disposa són escassos, fins i tot havent triat de desenvolupar el traductor segons la Limba Sarda Comuna (*Llengua Sarda Comuna*), la norma aprovada com a oficial pel govern autonòmic de Sardenya el 2006. Nombrosos aspectes de la morfologia, la sintaxi o l'estil no estan encara resolts. El lèxic que pot considerar-se normatiu no arriba a les 50.000 paraules i la terminologia està molt poc desenvolupada. Això ha estat una dificultat, com es veurà més endavant.

El desenvolupament del traductor s'ha realitzat entre maig i setembre de 2017, basant-se en un prototip existent a Apertium des de 2010. S'han utilitzat els recursos preexistents a Apertium, tant per al català, com per al sard. En particular per al sard, s'han utilitzat els recursos produïts l'any anterior arran de la creació d'un traductor de l'italià al sard en la mateixa plataforma Apertium (Tyers et al., 2017).

La resta de l'article es divideix de la manera següent: a la secció 2, fem una presentació succinta del sard i de la seva situació social. A continuació, a la secció 3, expliquem la plataforma utilitzada per a construir el sistema de traducció automàtica. En la secció 4 es descriu el desenvolupament del sistema, en particular la creació del diccionari bilingüe, les regles de selecció lèxica i les de transferència estructural. Seguidament, en la secció 5 es fa una avaluació del sistema, tant quantitativa com qualitativa. Finalment, comentem possibles treballs futurs a la secció 6 i donem algunes conclusions a la 7.

2 El sard

El sard és una llengua romànica de la branca occidental parlada a Sardenya (Coròngiu, 2013, p.39), la segona illa més extensa del Mediterrani, que forma part de l'Estat italià. Sardenya té una població d'1,7 milions de persones en una superfície d'uns 24.000 quilòmetres quadrats.

El sard, amb prop d'un milió de parlants, és la més estesa de les cinc llengües parlades a Sardenya, a banda de l'italià. Les altres quatre són el cors gal·lurès (a la regió de Gal·lura), el sassarès (a la ciutat de Sàsser), el tabarquí (a l'illa de Sant Pere) i el català alguerès (a la ciutat de l'Alguer).

Està reconegut com una de les 13 llengües minoritàries de l'Estat italià i protegit com a tal per la llei 482/1999. Alhora està reconegut com a llengua cooficial per la Regió Autònoma de Sardenya en la llei regional 26/1997.

Malgrat l'aïllament geogràfic, en la qual s'ha mantingut molt de temps, ha tingut força influències d'altres idiomes. Tres són les llengües romàniques que més rastre han deixat en el sard modern en dues èpoques diferents: primerament, amb la conquesta de l'illa per part de la Corona d'Aragó, el català i el castellà des del segle XIV fins al XVIII (en un primer moment, el català i després, el castellà); a continuació l'italià, a partir que Sardenya va passar a estar sota domini piemontès fins avui, especialment en l'àmbit lèxic.

Segons la tradició i l'opinió dels primers estudiosos del sard, es poden distingir dues grans varietats: el logudorès, incloent-hi el nuorès, que cobreix una part del centre i nord de l'illa, i el campidanès, que s'estén del centre al sud. Entre els investigadors més destacats, Wagner (1951) va definir el sard com un macrosistema lingüístic constituït de dialectes diferents. Més tard, Blasco Ferrer (1986) arriba a parlar de dues llengües neosardes (el logudorès i el campidanès).

Recentment, tanmateix, acadèmics com Bolognesi (2007) i Contini (1981) afirmen que les diferències són merament fonètiques, només ocasionalment morfològiques, sense cap diferència en la sintaxi i amb un lèxic majoritàriament comú.

Segons l'Atles Interactiu de la UNESCO de les Llengües del Món en Perill (Moseley, 2010), el sard és una llengua en perill. El fet que estigui molt dialectalitzat i encara no s'hagi estès del tot una forma estàndard ha causat, en molts llocs, l'abandonament del sard en favor de la llengua de l'estat, l'italià. Avui dia, el 68% dels sards saben parlar-lo i el 29% en té una competència passiva, mentre que el 2,7% no en té cap (Oppo, 2007).

La llei 482/99 i la llei regional 26 de 1997, d'acord amb la Carta Europea de les Llengües Regionals o Minoritàries de 1992, permeten l'ensenyament del sard als alumnes de primària i se-

cundària que ho demanin, així com l'ensenyament de la història i cultura sardes, l'ús del sard en els processos penals i l'administració, inclosa la possibilitat de presentar escrits en sard a l'administració i d'obtenir documents d'identitat en sard, l'ús del sard en la toponímia i també en la televisió. Tanmateix, en l'Estatut de la Regió Autònoma no se li atorga cap reconeixement com a llengua constitucional, a diferència del que s'esdevé a la Vall d'Aosta o al Trentí-Tirol del Sud. Gràcies a un acord entre el govern autònom i la delegació a Sardenya del ministeri d'educació, a partir del curs escolar 2013/14 les famílies poden triar si fer estudiar el sard als fills com a assignatura escolar, sense que l'ensenyament *en* sard estigui previst. Tanmateix, moltes escoles no van respectar la llei i no oferien l'opció d'estudiar-lo. Per solucionar aquest problema, el govern sard, en els anys 2016 i 2017, ha patrocinat 232 projectes experimentals per al seu ensenyament en preescolar, primària i secundària. Alhora, des de 2013 hi ha hagut nombrosos intents de presentar en sard els exàmens de final del primer cicle d'ensenyament secundari ("terza media" en italià, corresponent aproximadament al segon curs d'ESO a l'Estat espanyol i al vuitè curs d'ensenyament bàsic a Portugal) i del segon cicle ("maturità" en italià, equivalent a l'examen de selectivitat espanyol i als "exàmens nacionals" portuguesos). Enlloc no hi havia ensenyament estructurat de sard fins que el 2017 la universitat de Càller n'ha creat un curs específic.

El sard està en procés d'estandardització i de fa molt temps s'està buscant un acord per establir-ne alguna forma escrita oficial. El primer intent va ser la *Limba Sarda Unificada* de 2001. Poc després, el 2003, va aparèixer la proposta de la *Limba de Mesania*. En 2006, a iniciativa del govern regional, va sortir la *Limba Sarda Comuna* (LSC), una millora de la Limba Sarda Unificada. La LSC va ser adoptada "de manera experimental" per part de la Regió Autònoma de Sardenya amb el Decret núm. 16/14 de 18 d'abril de 2006 com a llengua oficial per a les actes i documents emesos per la Regió Autònoma (tanmateix, d'acord amb l'article 8 de la llei 482/99 tenen validesa legal només els documents redactats en italià). Amb això es facultava els ciutadans a escriure a l'administració regional en qualsevol varietat del sard, alhora que instituïa l'Oficina de la Llengua Sarda.

La LSC ha estat funcionant de manera experimental fins al 2013. Aquest període ha tingut dues fases. En la primera, de 2007 a 2010, la LSC s'ha emprat només en l'administració autonòmica. En la segona, de 2011 a 2013, mitjançant el Pla Lingüístic Triennal 2011–2013, s'han dut a terme algunes accions per incentivar el seu ús més enllà de l'administració pública. Segons el "Monitoratge de l'ús experimental de la Llengua Sarda Comuna (2007-2013)" (Regione Autonoma della Sardegna, 2014), la LSC és la convenció ortogràfica sarda més freqüent en els documents de l'administració, per damunt de "grafies locals", lligades a formes dialectals de la llengua. Així, les oficines de l'administració regional han produït el 50% dels seus escrits en sard en LSC, el 9% en LSC i en una grafia local, i el 41% en una grafia local. Aquest estudi també indica que, el 2013, de les escoles on s'ensenyava sard, el 51% van preferir emprar la LSC juntament amb una grafia local, l'11% només la LSC i el 33% només una grafia local. Els projectes editorials, però, i especialment els mitjans de comunicació, es decanten més sovint per la LSC: el 35% de les publicacions en sard s'han fet en LSC, el 35% en LSC i una grafia local i el 25% només en una grafia local.

Es pot afirmar que la LSC és la convenció ortogràfica més emprada a la xarxa. El 2012 va sortir el Curretore Ortogràficu Regionale (CROS).[1] Han aparegut també revistes i un nombre significatiu d'obres literàries. El 2014 va sortir la traducció al sard, parcialment en LSC, de la xarxa social Facebook (Martín-Mor & Beccu, 2016). De fet, diferents projectes de traducció col·laborativa han utilitzat la LSC. Així, el grup d'usuaris Sardware (Martín-Mor, 2016) ha traduït als sard el programa de missatgeria Telegram i el sistema de navegació GPS uNav. En canvi, el sistema operatiu Ubuntu ha estat localitzat només parcialment.[2] En aquest context de creixement gradual de l'ús de la LSC, l'agost de 2016 va aparèixer el primer traductor automàtic al sard, que va ser el d'italià a sard sobre la plataforma Apertium.

3 Plataforma

El sistema es basa en la plataforma per desenvolupar sistemes de traducció automàtica Apertium (Forcada et al., 2011; Armentano-Oller et al., 2007). La plataforma estava inicialment orientada a les llengües romàniques de l'Estat espanyol, però ràpidament s'hi van introduir millores que permeten el tractament de parells de llengües més distants: primerament, el català i l'anglès i, més endavant, també parells sense relació genètica coneguda, com el sami septentrional i el noruec o l'èuscar i el castellà.

[1] http://www.sardegnacultura.it/cds/cros-lsc/
[2] http://wiki.ubuntu.com/Ubuntu-Sardu

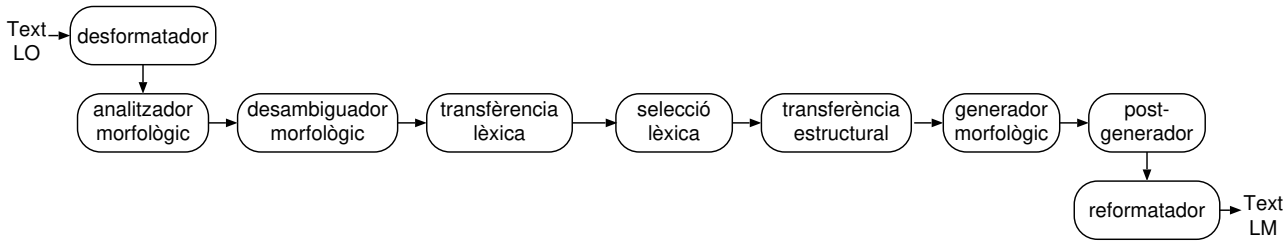

Figura 1: Arquitectura modular de la plataforma per desenvolupar sistemes de traducció automàtica Apertium. Els mòduls es comuniquen mitjançant canonades estàndard d'Unix.

Tota la plataforma, tant els programes com les dades, són de codi obert amb llicència GNU GPL.[3] El programari i les dades per als 46 parells de llengües considerats estables a data d'1.10.2017 (i molts altres que estan en desenvolupament) poden baixar-se en el web del projecte.[4]

És important assenyalar que els sistemes basats en regles són especialment adequats per a les llengües minoritzades, que típicament són també llengües que disposen de molts menys recursos, com corpus lingüístics, que les llengües dominants (Forcada, 2006). El fet que els recursos construïts a Apertium permeten no només empoderar les comunitats lingüístiques amb traductors automàtics, i no només fomentar l'estudi de les llengües en qüestió per a construir i millorar aquests traductors, sinó també extreure'n parts per produir diccionaris electrònics per a telèfons mòbils, correctors ortogràfics, etc. (Ramírez-Sánchez et al., 2006). Especialment actiu en aquesta direcció ha estat el grup de treball Giellatekno (Moshagen et al., 2014). Apertium també disposa d'un seguit d'eines que faciliten la creació de parts d'un traductor automàtic a partir de recursos lingüístics escassos, com el desambiguador morfològic (Sánchez-Martínez et al., 2006, 2007), regles de transferència (Sánchez-Martínez & Forcada, 2009), regles de selecció lèxica (Wiechetek et al., 2010; Tyers et al., 2012, 2014) o el diccionari bilingüe (Tyers & Pienaar, 2008). Això ha permès la construcció en els darrers 11 anys de traductors automàtics per a llengües minoritzades, com l'afrikaans (Otte & Tyers, 2011), l'aragonès (Martínez Cortés et al., 2012), l'asturià, el bielorús, el bretó (Tyers, 2009, 2010), el català (Armentano-Oller & Forcada, 2006; Toral et al., 2011; Ivars-Ribes & Sánchez-Cartagena, 2011), l'èuscar (Ginestí-Rosell et al., 2009; O'Regan & Forcada, 2013), el gallec, el gal·lès (Tyers & Donnelly, 2009), el kazakh (Salimzyanov et al., 2013; Sundetova et al., 2015; Balzhan et al., 2015), el maltès (Ravishankar et al., 2017), l'occità (Armentano-Oller & Forcada, 2006), el sard (Tyers et al., 2017), el tàtar (Salimzyanov et al., 2013), el tàtar de Crimea i el sami septentrional (Antonsen et al., 2017; Johnson et al., 2017). Altres llengües minoritzades amb desenvolupaments d'un volum notable a Apertium són el bengalí (Faridee & Tyers, 2009), el kurd (Gökırmak & Tyers, 2017), el marathi (Ravishankar & Tyers, 2017), el sami de Lule (Tyers et al., 2009), el sami meridional (Antonsen et al., 2017) i l'ucraïnès.

[3] http://www.gnu.org/licenses/gpl-3.0.ca.html
[4] http://wiki.apertium.org/wiki/Installation

Canonada

Típicament, un traductor construït amb Apertium consisteix en nou mòduls que es comuniquen mitjançant canonades estàndard d'Unix. Això facilita el control del procés, la inserció de mòduls nous, etc. Aquests mòduls són els següents:

- Un **desformatador**, el qual encapsula qualsevol informació de format (p.ex. etiquetes HTML o XML) a la cadena d'entrada.
- Un **analitzador morfològic**, el qual per a una forma superficial retorna una seqüència de possibles anàlisis. Cadascuna d'aquestes anàlisis consisteix en formes lèxiques amb un lema (la forma base usada habitualment en les entrades dels diccionaris), una categoria lèxica (nom, adjectiu, verb, preposició, etc.) i informació morfològica (gènere, nombre, persona, temps, etc.).
- Un **desambiguador morfològic**, el qual de la seqüència de possibles anàlisis tria la més probable. Aquest mòdul es basa o bé en el model ocult de Markov de primer nivell (Cutting et al., 1992; Sánchez-Martínez et al., 2006) o bé amb una combinació d'aquest amb Constraint Grammar (Bick & Didriksen, 2015).
- Un **mòdul de transferència lèxica**, el qual per a cada forma lèxica inambígua d'entrada retorna una o més formes lèxiques en la llengua meta.

- Un **mòdul de selecció lèxica**, el qual per a cada forma lèxica de la llengua font amb més d'una traducció possible tria una d'aquestes d'acord amb un conjunt de regles basades en el context de les paraules en la llengua font (Tyers et al., 2012).

- Un **mòdul de transferència estructural**, el qual realitza modificacions morfològiques i sintàctiques amb les formes lèxiques per convertir la representació intermèdia en llengua font en una representació intermèdia en la llengua meta. Les operacions més corrents inclouen la inserció, l'esborrament i la reubicació d'unitats lèxiques i la seva concordança (en gènere, nombre, etc.).

- Un **generador morfològic**, el qual per a cada forma lèxica en la llengua meta retorna una forma lèxica superficial (flexionada).

- Un **postgenerador**, el qual realitza transformacions ortogràfiques en la llengua meta, com per exemple apostrofacions o contraccions (*el*+*amic*=*l'amic*, *de*+*el*=*del*).

- Un **reformatador**, el qual restableix del format prèviament encapsulat.

La figura 1 dóna un exemple de canonada. Les dades utilitzades per cadascun d'aquests mòduls s'especifiquen en fitxers XML, que es compilen en fitxers binaris per a la seva execució pels mòduls.

Convé assenyalar que, tot i que cada parell de llengües era inicialment independent a Apertium, actualment els recursos específics per a una llengua es comparteixen entre els traductors a o des d'aquestes llengües (Marting & Unhammer, 2014). D'aquesta manera, per exemple, els traductors des del català comparteixen les dades de l'analitzador morfològic i el desambiguador morfològic, i els traductors al català, les del generador morfològic i el postgenerador. Un nou traductor entre dues llengües que ja estan incloses dins del sistema Apertium, en principi, només ha d'ocupar-se de la transferència lèxica (el diccionari bilingüe), la selecció lèxica i la transferència estructural. Aquest ha estat el cas del traductor de català a sard, en què ja existien traductors tant a partir del català com al sard.[5]

4 Desenvolupament

En iniciar el projecte del traductor català–sard, el maig de 2017, n'hi havia ja un prototipus a Apertium. El prototipus tenia un diccionari bilingüe de 2814 entrades, 4 regles de selecció lèxica i 33 regles de transferència estructural (algunes d'elles amb errors). El sard s'havia incorporat feia pocs mesos a les llengües amb parells estables a Apertium, després d'un projecte de quatre mesos, per la qual cosa el diccionari morfològic de què disposava (i del qual es nodreixen tant l'analitzador com el generador morfològic) era ja considerable, però també millorable.

Anàlisi

El desenvolupament va començar amb una anàlisi contrastiva entre les dues llengües. Aquesta anàlisi va utilitzar extensament la comparació prèvia entre sard i italià, sobre la base de la qual es van fer modificacions (per exemple, afegir el tractament del passat perfet perifràstic del català). Aquesta anàlisi tenia per objectiu detectar aquelles estructures que en una traducció "morfema a morfema" no resultarien correctes, per exemple:

- La meva casa. → Sa domo mea.
- Bellíssims. → Bellos a beru.
- Donar-me. → Mi dare.

Aquestes diferències van ser la base per construir subsegüentment regles de transferència estructural.

En bastir el traductor italià–sard, es va esmerçar un esforç considerable per aplegar un corpus de textos en sard. El problema no és només per la relativament poca quantitat de textos en sard disponibles en format electrònic, sinó, sobretot, per l'ús indistint de la LSC i de varietats dialectals en la Viquipèdia i publicacions periòdiques, així com les incorreccions en la norma en textos escrits, en principi, en LSC. Això representa un problema a l'hora d'esbrinar l'ús real (i autoritzat) en alguns aspectes morfològics i sintàctics que no estan prou detalladament descrits en la norma. Així, doncs, en aquesta ocasió hem confegit un petit corpus de textos literaris en LSC (206.000 mots),[6] que és el que hem utilitzat per a aquests efectes.

[5] En el moment de començar el desenvolupament del parell català–sard hi havia a Apertium vuit traductors considerats estables a partir del català: a l'anglès, aragonès, castellà, esperanto, francès, occità referencial, occità aranès i portuguès; i un traductor al sard: des de l'italià.

[6] Concretament utilitzem les següents obres: Joyce, James. *Dublinesos.* [Traducció de Sarvadore Serra.] Nùgoro: Papiros, 2011. Salgari, Emìlio. *Sas tigres de Mòmpracem.* [Traducció de Mariantonietta Piga.] Dolianova: Grafica del Parteolla, 2013. Saint-Exupéry, Antoine de. *Su printzipeddu.* [Traducció de Diegu Corràine.] Nùgoro: Papi-

Entre les vacil·lacions de la norma de la LSC per a les quals hem estat utilitzat el corpus sard, podem mencionar dues: el participi passat de verbs com *dipèndere* i *suspèndere* (respectivament, *dependre* i *suspendre*) i la posició no marcada del possessiu en els sintagmes nominals.

En el primer cas, la conjugació d'aquests verbs falta en les *Norme linguistiche di riferimento* (Regione Autonoma della Sardegna, 2006). El CROS, que és la nostra font bàsica quant a la flexió dels mots sards, admet tant *suspendidu* com *suspesu*, però per alguna raó accepta només *suspesu* en masculí singular, mentre que per a *suspendidu* també admet les formes en femení i plural. El corpus literari mostra que els dos participis s'utilitzen (i, lògicament, tenen flexió). Segons el nostre assessor lingüístic, Diegu Corràine, les dues formes són admissibles. La primera forma és característica dels dialectes septentrionals, mentre que la segona ho és dels meridionals. Així doncs, vam optar per generar la forma *suspesu*, amb la seva flexió en gènere i nombre, donat que ara mateix estem prioritzant les variants septentrionals de la LSC i, pensem, és preferible ser consegüents (una altra opció perfectament legítima seria alternar formes normatives tant del nord com del sud, però no ens considerem autoritzats per fer nosaltres la tria de l'opció a generar en cada cas que la norma permet dues formes). Cal assenyalar que hem indicat en el diccionari morfològic sard quines són septentrionals i quines meridionals per, més endavant, triar generar unes formes o altres (tal com actualment es fa amb Apertium per al català general o el valencià, i l'occità referencial o l'aranès).[7]

Una altra qüestió de vacil·lació en el sard és la posició del possessiu en els sintagmes nominals. Típicament, el possessiu va al final del sintagma, però quan hi ha adjectius darrere el nom és freqüent trobar el possessiu entre el nom i el o els adjectius. A (1) es presenten casos extrets del corpus literari.[8]

(1) a. S' istile oratòriu suo.
El seu estil oratori.

b. Sos propòsitos bonos issoro.
Els seus bons propòsits.

c. Sos ogros suos asulos.
Els seus ulls blaus.

d. Sas framas issoro debileddas.
Les seves flames una mica dèbils.

En no disposar encara d'un corpus etiquetat sard, ni d'un desambiguador morfològic, no hem pogut analitzar amb detall quina posició del possessiu és la més habitual en relació al nom i al(s) adjectiu(s). És possible que estigui influïda pel context i no només sigui estilística. Hem seguit l'opinió autoritzada del nostre informador Diegu Corràine i hem posat el possessiu sempre en posició final de sintagma (tal com ja havia estat l'opció en el traductor italià–sard).

Diccionari bilingüe

Una gran part del treball ha consistit a crear un diccionari bilingüe català–sard.

Bàsicament, això s'ha dut a terme a partir de trobar els lemes que no estaven en el diccionari bilingüe primigeni i ordenar-los per ordre de freqüència decreixent. Per a això s'ha utilitzat un corpus extret de la Viquipèdia en català de 2,6 milions de paraules. Per simplicitat en l'edició, la llista de paraules s'ha carregat en un full de càlcul de Google Docs i, a mesura que s'anaven afegint traduccions, s'ha anat carregant al diccionari bilingüe amb un programa. Es podia entrar més d'una equivalència per paraula i en una columna de comentaris s'apuntava, entre altres coses, si convenia fer una selecció lèxica en funció del significat de la paraula font. Es verificava que les paraules sardes ja estiguessin en el diccionari monolingüe (morfològic) sard. Si no, s'avaluava si la paraula podria considerar-se normativa i, si se l'hi considerava, s'entrava manualment en el diccionari sard.

D'aquesta manera s'han carregat uns 11.300 lemes catalans en el diccionari bilingüe i uns 2500 lemes nous en el diccionari sard. En arribar el diccionari bilingüe a les 8700 entrades, ja s'assolia una cobertura del 90,1%.[9]

ros, 2015. Wilde, Oscar. *Su pantasma de Canterville*. [Traducció de Sarvadore Serra.] Nùgoro: Papiros, 2013. Malauradament, de la traducció del Quixot, disponible lliurament a Internet en format PDF gràcies al suport del Govern de Sardenya, no se'n pot extreure el text. Això hagués permès incrementar considerablement el corpus i introduir textos d'un quart traductor i una tercera editorial. Cal assenyalar la manca de textos administratius o legislatius disponibles en sard, malgrat el seu estatus oficial.

[7] Per enllestir la possibilitat de generar textos en una varietat septentrional i una meridional de la LSC, caldria també estendre aquesta distinció a part del lèxic. De tota manera, el problema bàsic de posar a disposició dels usuaris la tria de varietats és sociolingüístic. És necessari que la comunitat lingüística sarda decideixi si considera més convenient per a l'arrelament d'una norma comuna normativa (i, últimament, per a la pervivència de la llengua sarda), l'ús d'una varietat única que incorpori elements tant del nord com del sud, o bé la difusió de dues subnormes estàndard dins d'un marc comú.

[8] "Suo" = "d'ell o d'ella", "issoro" = "d'ells o d'elles"

[9] Aquí s'entén com a cobertura la cobertura ingènua, és a dir, per a qualsevol forma superficial donada s'obté com a mínim una anàlisi. Pot ser que no es tinguin totes les

Una segona forma d'inclusió de paraules en el diccionari ha estat la comparació massiva dels dos diccionaris monolingües per trobar cognats. El procés de comparació tenia en compte tant diferències ortogràfiques trivials (per exemple a les formes catalanes *l·l, qua, que, qui, gue, gui, í, ú* corresponen les formes sardes *ll, cua, che, chi, ghe, ghi, ì, ù*, com una sèrie de canvis sistemàtics (per exemple, als adjectius catalans acabats en *-à*, *-ari* i *-ble* típicament corresponen adjectius sards acabats en *-anu*, *-àriu* i *-bile* i als grups consonàntics catalans *ct* i *pt* corresponen *t* en sard). Les llistes de cognats resultants s'han revisat abans de ser incloses al diccionari bilingüe. També, per evitar traduccions mecàniques, aquest mètode s'ha utilitzat només en la segona meitat del projecte, quan ja s'havien introduït milers de paraules habituals (que típicament són les més polisèmiques) mitjançant la traducció manual prèviament descrita. Gràcies al fet de disposar d'entrada de dos diccionaris monolingües extensos i a la proximitat entre les llengües, aquest mètode, enormement més ràpid que l'anterior, ha permès incloure més de 3.000 entrades en el diccionari, sense comptar noms propis.

Finalment, s'han tractat els adverbis acabats en *-ment* derivats d'adjectius. Els seus cognats en sard acaben en *-mente* i són habituals en la llengua col·loquial, però són rars en els textos literaris. El CROS n'admet només 25 i en tot el corpus literari hem trobat només 9 casos d'ús. En la fase de traducció manual i comparació massiva de diccionaris, hem traduït uns 150 adverbis catalans en *-ment* per locucions específiques sardes (per exemple, *ràpidament* → *a sa lestra*, *essencialment* → *in sustàntzia*) i 97 pels seus cognats sards. Arribats a aquest punt, hem generat automàticament la traducció dels adverbis derivats catalans de què teníem traducció de l'adjectiu segons el model *vivament* → *in manera viva*. S'han produït 1.568 traduccions automàtiques que s'han inclòs en el diccionari bilingüe. Com a simple comprovació s'ha traduït automàticament un petit corpus de prova (5.000 frases, 130.000 mots) abans i després del canvi i s'han mirat les diferències. Com a resultat, unes poques equivalències s'han canviat (i el mateix s'ha fet posteriorment, en avaluar traduccions de prova de textos reals).

També hi ha hagut una incorporació automàtica d'antropònims, tant a partir del diccionari català, com a partir del diccionari sard.

anàlisis possibles. Al llarg de tot l'article les cobertures estan calculades sobre un corpus extret de la Viquipèdia de 6,1 milions de paraules.

Categoria	**Entrades**
Substantius	7714
Adjectius	3194
Adverbis	2196
Verbs	1993
Noms propis	17303
Altres	495
Total	32895

Taula 1: Distribució de les entrades en el diccionari català-sard per categories gramaticals.

A la taula 1 es presenta la distribució de les entrades en el diccionari bilingüe per categories gramaticals.

En el projecte hem considerat important assolir una cobertura considerable per poder delimitar correctament els sintagmes nominals, cosa molt rellevant en el tractament dels possessius.

Selecció lèxica

La selecció lèxica és un element relativament nou a la canonada d'Apertium. Tradicionalment, a Apertium, la selecció d'una de les diverses traduccions possibles s'ha realitzat en el diccionari bilingüe, anul·lant totes les altres sense analitzar el context o bé afegint expressions multiparaula en els diccionaris (per exemple, "ull de bou" o "fer fora"). Només de manera excepcional la selecció lèxica es tractava en la transferència estructural, però resulta farragós.[10] En canvi, les regles de selecció lèxica permeten d'una manera molt més simple expressar contextos en què convé triar una o altra de les traduccions possibles. La figura 2 dóna un exemple de regles de selecció lèxica. Convé assenyalar que, per assegurar la rapidesa del procés, les regles només poden tenir en compte els contextos ordenats de longitud fixa, de manera que no és possible, per exemple, construir una regla que seleccioni una traducció determinada basada en si es troba una paraula donada en qualsevol posició de la frase.

S'han escrit manualment 526 regles de selecció lèxica. Han tingut característiques marcadament diferents en determinades categories gramaticals (la taula 2 desglossa les regles per categories gramaticals).

Especialment impactant quant a la qualitat de la traducció és el tractament de les preposicions catalanes *de* i *a*. La primera, bàsicament, es tradueix per *dae* per a indicar procedència o material, o, altrament, per *de*. S'han utilitzat 10 regles

[10] Per exemple, el traductor català–castellà té 6 regles de transferència estructural per triar entre *a* i *en* al traduir la preposició catalana *a*.

```
<rule weight="0.6" c="traducció per defecte">
  <match lemma="a" tags="pr"><select lemma="a" tags="pr"/></match>
</rule>
<rule weight="1.0">
  <match lemma="a" tags="pr"><select lemma="in" tags="pr"/></match>
  <or><match tags="np.loc"/><match tags="np.top.*"/>
      <match tags="np.al"/><match tags="np.al.*"/></or>
</rule>
```

Figura 2: Exemple de dues regles de selecció lèxica en què es tria la preposició sarda *a* com a opció per defecte per traduir la preposició catalana *a* i la preposició sarda *in* si *a* precedeix un topònim. (Altres regles tornen a triar *a* davant de topònim en presència de determinats verbs.)

Categoria	**Lemes catalans**	**Regles**
Preposicions	6	36
Conjuncions	3	63
Relatius	1	3
Pronoms	1	4
Verbs	12	53
Substantius	25	92
Adjectius	5	10
Noms propis	46	273
Total	99	534

Taula 2: Distribució de les regles de selecció lèxica per categories gramaticals.

de transferència lèxica, algunes de les quals contenen llistes de verbs (22) i substantius (32) que van acompanyats per *de* o *dae* en les seves traduccions al sard. Quant a la preposició catalana *a*, bàsicament es tradueix en sard com a *a*, quan es tracta de direcció o complement indirecte, i *in*, quan es tracta d'un lloc o un temps en què succeeix una acció (de forma molt semblant, si no és idèntica, al castellà). Aquí, les 11 regles de selecció lèxica, a més de contenir llistes de verbs (33) i substantius (144) típicament associats a una traducció i altra, també tenen en compte si *a* es troba davant d'un topònim. Especialment en el tractament d'*a*, certes regles competeixen entre si a l'hora de decidir la millor traducció. Per a totes dues preposicions, les llistes de verbs i substantius s'han escrit a mà a partir del coneixement de les dues llengües i de la pràctica en traduccions de prova. L'anàlisi qualitativa dels resultats (vegeu més endavant), però, mostra que aquestes regles estan lluny de resoldre els problemes.

Convé també assenyalar el problema al traduir el possessiu *seu* de la selecció de *suo* ("d'ell o ella") o *issoro* ("llur"). Això, de fet, requereix esbrinar el referent d'un pronom, qüestió per a la qual actualment la canonada d'Apertium no disposa d'eines. La nostra tria és sempre *suo*.

La selecció lèxica en els substantius presenta altra mena de dificultats. Estem parlant de casos com *tassa* (l'objecte o l'impost), *got* (l'objecte o el membre d'un poble germànic), *poble* (un vilatge o un conjunt de persones), *paper* (la substància o una funció), *pinya* (la fruita tropical o l'òrgan fructífer d'una conífera), *cop* (el que reparteix la policia antidisturbis o *vegada*), *recurs* (un mitjà o una acció judicial o administrativa), *car(a)gol* (l'animal o la vis), *taula* (el moble o la representació en forma tabular), *tret* (una característica o una descàrrega d'una arma de foc), etc. El mecanisme de selecció lèxica emprat es fixa només en les paraules immediatament anteriors o posteriors. Manualment, és sovint difícil definir contextos clars per destriar una opció o una altra i fer-ho, a més, per a desenes de paraules en un temps raonable.

Vora la meitat de les regles de selecció lèxica (273) s'han escrit per desambiguar 46 noms propis. En aquests casos es tracta de noms com *Jaume, Francesc, Isabel* o *Alexandre*. Si el context permet reconèixer que s'està parlant de reis, papes, emperadors, etc., els noms es tradueixen pels seus equivalents sards, altrament es deixen en català.

Regles de transferència estructural

Apertium, si no es diu el contrari, tradueix lemes i morfemes un per un. Òbviament, això no sempre funciona, fins i tot per a llengües genèticament molt properes. Les regles de transferència estructural són responsables de modificar la morfologia o l'ordre de les paraules per produir una sortida "correcta" en la llengua meta. En total, hem definit 93 d'aquestes regles de transferència: 48 per a construccions verbals i 45 per a nominals (incloent-hi estructures sense substantiu, però amb adjectius, numerals i/o determinants).

Convé assenyalar que, tot i que Apertium permet una jerarquia de regles per facilitar anàlisis sintàctiques més profundes i, conseqüentment, el

tractament de dependències més llunyanes, hem adoptat, per simplicitat, el model senzill, raonable per a llengües estretament emparentades. Així doncs, les regles tracten bocins de text d'esquerra a dreta. Una vegada una regla ha establert una traducció, no és possible tornar enrere, fins i tot si elements posteriors indiquen que caldria fer-ho.

A continuació es presenten els tractaments més importants que fan les regles de transferència.

Concordança dins del sintagma nominal

La majoria de regles lligades als sintagmes nominals tracten la concordança en gènere i nombre a l'interior del sintagma. Hi ha dues menes de situacions problemàtiques. Per una banda, hi ha els casos en què el substantiu català no té el mateix gènere que la seva traducció (o en alguns casos rars, no té el mateix nombre). Això fa que en la traducció calgui canviar el gènere dels determinants i adjectius associats al nom (2). Un 11% dels substantius tenen diferent gènere en català i sard en el nostre diccionari bilingüe. Una segona situació es dóna quan el substantiu en la llengua origen no té formes distintives en gènere i/o nombre, mentre que sí les té la llengua meta. Això fa que calgui assignar el gènere i/o nombre al substantiu de la llengua meta, típicament a partir de les paraules que l'acompanyen (3). Menys del 2% dels substantius presenten aquesta situació. En ambdós casos, les regles de transferència intenten solucionar el problema. I especialment en el primer cas (que és bastant freqüent, com es veu) és important delimitar correctament el sintagma nominal per canviar el gènere de tots els determinants i adjectius que acompanyen el nom.

(2) a. Una agrupació astronòmica.
Unu agrupamentu astronòmicu.

b. Les acaballes.
Sa fine.

(3) a. Un àrab marroquí.
Un' àrabu marrochinu.

b. Una àrab marroquina.
Un' àraba marrochina.

c. El temps passat.
Su tempus passadu.

d. Els temps passats.
Sos tempos passados.

Un cas especial de concordança es dóna amb el determinant *carchi* ("algun"), que sempre va en singular. Així la frase "Algunes persones mengen caragols" es tradueix en sard com "Carchi persone màndigat corrobacas", literalment, "alguna persona menja caragols", amb "persone" i el verb "màndigat" en singular.

Possessius

Com vist anteriorment, els possessius també requereixen una delimitació correcta dels sintagmes nominals, donat que han de traslladar-se des de l'inici al final (4).

(4) La seva casa natal.
Sa domo nadia sua.

En conseqüència, s'han creat força regles trivials de reordenació de l'estil de les següents:

- Possessiu[11] Adjectiu Nom → Det.Def Adjectiu Nom Possessiu
- Possessiu Adjectiu Adjectiu Nom → Det.Def Adjectiu Adjectiu Nom Possessiu
- Possessiu Nom Adjectiu → Det.Def Nom Adjectiu Possessiu
- Possessiu Adjectiu Nom Adjectiu → Det.Def Adjectiu Nom Adjectiu Possessiu
- etc.

Nombres ordinals i trencats

Els nombres ordinals i els trencats tenen una estructura inhabitual en sard.

Excepte *primu* ("primer") i *segundu* ("segon"), els ordinals en LSC no són adjectius, sinó que es construeixen mitjançant la preposició *de* i el nombre cardinal: *su de tres* ("tercer"), *su de bator* ("quart"), etc. Aquest canvi s'ha pogut tractar amb el diccionari bilingüe, sense regles de transferència.[12] No ha pogut ser així, però, amb els nombres trencats, ja que impliquen una reordenació de mots i la inserció d'un article definit (5).

(5) a. Un terç dels habitants.
Su tres unu de sos abitantes.

b. Dos terços dels habitants.
Sos tres duos de sos abitantes.

[11] En català, *el meu, el teu*, etc. s'analitzen com a una unitat.

[12] Hi ha una regla de transferència relacionada amb els ordinals, però això és degut a un canvi d'un numeral cardinal en català per un numeral ordinal en sard en parlar dels segles: *el segle XX → su de XX sèculos*.

c. En el tercer terç del
In su de tres tres unu de su
segle XX.
de XX sèculos.

Formes analítiques i ordre dels modificadors

El sard tendeix cap a l'adopció de formes analítiques. Això no només passa, com dit anteriorment, amb el sufix *-ment*, usual en altres llengües romàniques, que no s'accepta o no es recomana en LSC i se "substitueix" per locucions. El mateix succeeix amb el sufix *-íssim*, per formar superlatius d'adjectius i adverbis, que no està acceptat en LSC. En aquest cas, per exemple, *rapidíssim* es tradueix com *lestru a beru* (literalment, "ràpid de veres"). Mentre que la traducció dels adverbis es fa directament en el diccionari bilingüe, la dels superlatius es fa en regles de transferència estructural que afegeixen la locució adverbial *a beru*.

Cal assenyalar que la posició de l'adverbi *a beru* en *lestru a beru* és habitual en sard. Els adverbis tendeixen a posar-se darrere els adjectius, de la mateixa manera que els adjectius tendeixen a anar darrere els substantius encara més sovint que en català. No hem pogut, però, analitzar amb detall la posició dels adverbis en sintagmes nominals complexos (per exemple, en estructures com "una nena molt intel·ligent i decidida"), ni ens hem atrevit a posar tots els adjectius anteposats al nom darrere d'ell perquè també hi ha adjectius davant el nom en sard. En aquests casos calquem l'ordre dels mots del català.

Temps verbals

El sard també tendeix cap a formes analítiques en la conjugació. Alguns temps verbals que són sintètics en català, com en la majoria de llengües romàniques, es conjuguen en sard mitjançant perífrasis verbals, per exemple el futur (6a) i el condicional (6b). A més, la LSC no té passat perfet simple i utilitza, en canvi, el perfet compost (6c). El passat perifràstic català, el traduïm també al perfet compost sard (6d).

(6) a. Cantaré.
Apo a cantare.

b. Cantaria.
Dia cantare.

c. Cantí.
Apo cantadu.

d. Vaig cantar.
Apo cantadu.

Verbs auxiliars

Com, entre altres, l'italià, el francès i l'occità, el sard té verbs que utilitzen l'auxiliar *àere* ("haver") i *èssere* ("ser"); a més els verbs pronominals també utilitzen *èssere* (7). Això és especialment rellevant en sard, donat que l'única forma de passat perfet de l'indicatiu es construeix amb un d'aquests auxiliars. Així doncs, les regles de transferència lligades amb el pretèrit perfet sard trien un verb o un altre, segons si pertanyen a una determinada llista de verbs o bé segons si detecten una construcció pronominal.

(7) Ha permès.
At permìtidu.

(8) Ha arribat.
Est arribadu / Est arribada.

(9) S'ha permès.
S'est permìtidu / S'est permìtida.

Tanmateix, els participis darrere de l'auxiliar *èssere* concorden en gènere i nombre amb el subjecte, mentre que en els verbs catalans, en principi, no hi ha cap marca de gènere. El problema és distingir el subjecte (i el seu nucli) per assignarne el gènere. A més, aquest subjecte pot estar elidit, cosa que remet a un problema de resolució de l'anàfora semblant a l'anteriorment vist per al possessiu "seu", és a dir ara mateix inabordable a Apertium.

En general, la tria del gènere es fa per un doble mecanisme.

- En el cas dels verbs copulatius, si tenen al darrere un adjectiu, s'agafa el gènere de l'adjectiu, tal com està en català (si l'adjectiu té formes diferents per gènere, cosa que no passa, per exemple, amb *comunista*).

  (10) La reunió ha estat curta.
  Sa reunione est istada curtza.

  Aquest mètode simple, però, no l'encerta sempre:

  (11) La disfressa ha estat divertida.
  *Su disfrassu est istada ispassiosa.
  (hauria de ser:
  Su disfrassu est istadu ispassiosu.)

- En el cas dels verbs no copulatius, es tria el gènere el substantiu que precedeix el verb (com abans, si se'n pot extreure el gènere)

  (12) La directora ha vingut.
  Sa diretora est bènnida.

Com en el cas anterior, la senzillesa del mètode no permet que funcioni sempre:

(13) La directora del col·legi ha vingut.
*Sa diretora de su collègiu est bènnidu.
(hauria de ser:
Sa diretora de su collègiu est bènnida)

Existencials

Anàlogament al català *hi ha*, existeix en sard l'expressió *b'at* (literalment, "li ha").[13] A diferència amb el català estàndard, però, l'existencial sard té singular i plural (14).

(14) a. Hi ha un nen.
B' at unu pitzinnu.
b. Hi ha nens.
B' ant pitzinnos.

Això implica que hem hagut de crear regles que tracten "bocins" de text que inclouen tant l'existencial com el sintagma nominal que hi ha al darrere. Això pressuposa un gran nombre de combinacions, donat que tant l'existencial pot tenir diferents construccions (p. ex. "hi ha", "hi ha hagut", "hi va haver") com el sintagma nominal (nom, determinant + nom, determinant + adj + nom, etc.). Per falta de temps per definir cada cas, només s'han tractat les 5 construccions que s'han considerat les més habituals. Aquest tractament seria més fàcil amb la utilització de regles de dos nivells, en què el segon podria posar en plural el verb si el sintagma nominal que el segueix va en plural.

Pronoms clítics

El sistema pronominal del sard és força semblant al català, tant en les formes tòniques com febles. Hem detectat alguns usos diferents del pronom català *en*, en comparació al seu equivalent sard *nde*, però no hem sabut trobar canvis sistemàtics clars i no hem tractat aquesta qüestió en aquesta fase del treball. Un canvi important, però, és que els pronoms clítics sards van necessàriament davant del verb en infinitiu, la qual cosa ha implicat la creació de nombroses regles de transferència (15).

(15) a. Vol donar-li.
Li cheret dare.
b. Vol donar-li-ho.
Bi lu cheret dare.
c. Vaig donar-li.
L' apo dadu.
d. Vaig donar-li-ho.
Bi l' apo dadu.

5 Avaluació

El sistema s'ha avaluat tant quantitativament com qualitativament. D'una banda se n'ha analitzat la cobertura. De l'altra s'han avaluat els errors que s'han produït en la traducció de quatre textos de la Viquipèdia, comparant-los amb una versió posteditada.

Avaluació quantitativa

S'ha extret aleatòriament un corpus de 100.000 frases i 6,1 milions de paraules de la Viquipèdia en català. La cobertura "ingènua" calculada sobre aquest corpus és del 94,4%, és a dir que per a aquest percentatge de mots se n'ha obtingut com a mínim una anàlisi morfològica.

Per altra banda s'ha mesurat la qualitat de la traducció. Per a això, s'ha fet una selecció pseudoaleatòria de quatre textos de la Viquipèdia en català. Es va triar "l'article del dia" i també els dels tres dies anteriors, agafant de tots ells el resum inicial[14]. Els "articles del dia" són seleccionats pels viquipedistes segons diferents criteris, un dels més importants dels quals és la qualitat de l'article. Això garanteix, entre altres coses, la qualitat lingüística, cosa important, donat que el traductor no està pensat per tractar llengua no estàndard, amb faltes d'ortografia, barbarismes, etc. Quatre textos es consideren un nombre adequat per incloure temàtiques diferents. Els textos d'aquest corpus de prova tenien un total de 1056 paraules (34 frases). La mitjana de paraules per frase és de 31. La figura 3 presenta un fragment dels textos de prova, amb la seva traducció automàtica i la traducció posteditada.

La qualitat de la traducció s'ha mesurat mitjançant dues mètriques: la tassa d'error per paraula (*Word Error Rate*, WER) i la tassa

[13] El pronom català *hi* té unes regles de selecció lèxica justament per a ser traduït com a *bi* en aquest cas. La seva traducció habitual en sard és *nche*.

[14] "Escultura del Renaixement a la Corona d'Aragó", http://ca.wikipedia.org/wiki/Escultura_del_Renaixement_a_la_Corona_d'Aragó; "Gorgosaure", http://ca.wikipedia.org/wiki/Gorgosaure; "Vincent van Gogh", http://ca.wikipedia.org/wiki/Vincent_van_Gogh; "La gran ona de Kanagawa", http://ca.wikipedia.org/wiki/La_gran_ona_de_Kanagawa.

Català (text d'origen)	**Català→Sard** (traducció automàtica)	**Sard** (traducció posteditada)
Entre els artistes de la mateixa terra van destacar el valencià establert a Saragossa Damià Forment, Gil Morlanes el Vell, Jaume Amigó, Jeroni Xanxo, Pere Blai, Andreu Ramírez i Agustí Pujol (pare). Al segon terç del segle XVI, l'escultor d'origen basc Martín Díez de Liatzasolo va muntar un dels tallers més productius a Barcelona.	Intre sos artistas de sa matessi terra ant distacadu su valentzianu istabilidu in Zaragoza Damià *Forment,* Gil *Morlanes* su Betzu, Jaume Amigó, *Jeroni Xanxo*, Pere Blai, Andreu *Ramírez* e Agustí Pujol (babbu). A su segundu tres unu de su de XVI sèculos, s'iscultore de orìgine basca Martín Díez de *Liatzasolo* at montadu unu de sos laboratòrios prus produtivos in Bartzellona.	Intre sos artistas de sa matessi terra si sunt distinghidos su valentzianu istabilidu in Zaragoza Damià Forment, Gil Morlanes su Betzu, Jaume Amigó, Jeroni Xanxo, Pere Blai, Andreu Ramírez e Agustí Pujol (babbu). En su segundu tres unu de su de XVI sèculos, s'iscultore de orìgine basca Martín Díez de Liatzasolo at ammanniadu unu de sos laboratòrios prus produtivos in Bartzellona.

Taula 3: Part d'un dels textos del corpus de prova català amb la seva traducció automàtica i la traducció posteditada. Els segments subratllats són els que ha calgut canviar a la postedició. Les paraules en cursiva són desconegudes pel traductor però no s'han hagut de canviar. La tassa d'error d'aquest fragment és del 8,2%.

Paraules	**Desconegudes**	**WER**	**PER**
1056	8,8%	20,5%	13,9%

Taula 4: Avaluació quantitativa de la qualitat del traductor en un corpus de la Viquipèdia de 1056 paraules.

d'error per paraula independent de la posició (*Position-Independent Word Error Rate*, PER). Totes dues es basen en la distància de Levenshtein (Levenshtein, 1966) i s'han calculat amb l'eina `apertium-eval-translator`. Aquestes mètriques s'han triat, bàsicament, per dos motius. En primer lloc, volíem comparar el sistema amb sistemes basats en tecnologia similar i avaluar la utilitat del sistema en un entorn real, és a dir, traduir per a *disseminar*. En segon lloc, la traducció de referència són traduccions automàtiques editades, mentre que la majoria de les mètriques d'avaluació en traducció automàtica utilitzen referències prèviament traduïdes. Utilitzar una mètrica més habitual en traducció automàtica en un entorn poc freqüent per a les llengües amb què es treballa portaria a resultats enganyosos.

Aquests resultats són pitjors que els obtinguts en altres traductors en la plataforma Apertium per a llengües romàniques. Per exemples, el traductor castellà–portuguès va obtenir un WER del 8,3% (Armentano-Oller et al., 2006); el català–occità, del 9,6% (Armentano-Oller & Forcada, 2006); l'italià–sard, del 9,9% (Tyers et al., 2017), el castellà–aragonès, del 16,8% (Martínez Cortés et al., 2012) i el català–aragonès, del 15,5% (Juan Pablo Martínez, 9.10.2017, correu electrònic).[15]

Cal assenyalar que en tots aquests casos, excepte potser en el català–aragonès, es tracta o bé de parells de llengües molt estretament relacionats (castellà–portuguès, català–occità) o/i entre una llengua dominant i una que li està subordinada (italià–sard, castellà–aragonès). En tots dos casos, això provoca l'anivellament dels camps semàntics i de les estructures sintàctiques, per la qual cosa la traducció "paraula a paraula" resulta especialment encertada a escala estadística. En el cas del català–sard, hi ha hagut una relació històrica directa entre les dues llengües i totes dues han estat també subordinades al castellà, la qual cosa ajuda a aquest anivellament, però ja fa ben bé tres segles que el sard ha perdut la relació directa o indirecta amb el català (si no és per l'Alguer, el pes del qual difícilment pot influir significativament sobre el sard més enllà de varietats locals circumdants, o a l'inrevés, pel que fa a la influència del sard en el català normatiu).

És significativa la diferència entre el WER que obtenim (20,5%) i el PER (13,9%), quan acostumen a obtenir-se només uns dos punts percentuals de diferència entre WER i PER en traduir entre llengües romàniques. Això indica que hi ha força canvis d'ordre en l'estructura de la frase sarda, en comparació amb la de la catalana, que el traductor no ha tingut en compte.

Per entendre les causes d'aquestes tasses d'error poc satisfactòries hem estudiat les fonts dels errors en el corpus de prova traduït.

[15] Convé assenyalar que en tots els casos es tracta de tasses d'error en les primeres versions publicades dels traductors. Aquestes xifres són, per tant, comparables amb les del traductor català–sard.

Avaluació qualitativa

Paraules desconegudes

De les 93 paraules desconegudes del corpus de proves, 43 són noms propis, la gran majoria prenoms i cognoms (gairebé sempre estrangers o medievals). De les 50 altres paraules, 18 han estat copades per les paraules *albertosaure, daspletosaure, gorgosaure, hadrosaure, saurus, tiranosaure* i *tiranosàurid*.

Errors del desambiguador morfològic

Hi ha hagut tres errors atribuïbles a una mala anàlisi o desambiguació morfològica, casualment tots lligat a *va* o *van*.

En la frase "L'escultura del Renaixement a la Corona d'Aragó va lligada a la cultura humanista", la paraula *va* és morfològicament ambigua (pot ser un adjectiu o una forma del verb *anar*) i ha estat incorrectament desambiguada com a adjectiu (en canvi *va* davant d'infinitiu sempre ha estat correctament entesa com a una forma verbal).

Per altra banda, *van* és una forma verbal, però també un part de nombrosos cognoms neerlandesos i flamencs. Per evitar que *van* en aquests cognoms fos interpretat com a una forma verbal, s'havia introduït en els diccionaris *Van* com a cognom (amb majúscula inicial). En el text sobre Vincent Van Gogh tres de les quatre formes en què *Van* anava escrit en majúscules han estat ben interpretades, mentre que l'única en què estava escrit en minúscules, s'ha interpretat com un verb conjugat i "Theo van Gogh" ha esdevingut "Theo andat Gogh". L'únic cas en què *Van* iniciava una frase i, per tant, també la forma verbal hauria de portar majúscula, també ha estat incorrectament analitzat i traduït com a *Andat*.

Errors en el diccionari bilingüe

Un nombre considerable errors són atribuïbles a mancances en el traductor bilingüe (més enllà de les paraules que hi falten). Per exemple, el verb *destacar* està traduït com a *distacare*, la qual cosa és correcta quan és transitiu (en el sentit de "fer ressaltar"), però quan és intransitiu (amb el sentit de "ressaltar") la traducció hauria de ser *si distìnghere* (caldria afegir-lo a diccionari i fer regles de selecció lèxica). Qüestions semblants es troben, per exemple, en les expressions *muntar un taller*, en què caldria haver utilitzat *abèrrere* ("obrir") o *ammanire* ("desenvolupar, organitzar") en comptes de la traducció general *montare*; o bé el verb *lligar* en expressions com *estar lligat* o *anar lligat* (una cosa amb una altra en sentit figurat) cal traduir-ho més aviat com a *collegare* en comptes de la traducció general *ligare*. Aquest tipus de problemes típicament són menys freqüents quan les dues llengües tenen molt de contacte entre elles.

Català	**Trad. aut.**	**Trad. correcta**	**Ocurrències**
de	de	de	96
de	dae	dae	3
de	de	dae	5
de	dae	de	3
a	a	a	9
a	in	in	5
a	a	in	10

Taula 5: Avaluació de la traducció de les preposicions catalanes *de* i *a* en un corpus de textos de la Viquipèdia de 1056 paraules.

Quant a la inclusió automàtica d'adverbis acabats en *ment* com a locucions amb l'estructura "*in manera* + adjectiu", en el corpus de prova han aparegut quatre casos. Dos han estat posteditats i canviats per formes més fraseològiques, mentre que els altres dos, en principi, s'han mantingut. Diem "en principi" perquè "més llunyanament" s'ha traduït automàticament com a "prus in manera instrinta", que després ha estat posteditat com a "in manera prus instrinta" (mantenint l'estructura bàsica, però introduint-hi l'adverbi *prus*, "més"). El cas mostra que, si es considera vàlida aquesta generació semiautomàtica de traduccions, hauria també de tenir en compte si els adverbis van precedits per *més* o *menys*.

Errors de la selecció lèxica

La selecció lèxica té un impacte notable en el resultat final, especialment en allò referent a les preposicions. A la taula 5 es presenta l'avaluació de la traducció de les preposicions *de* i *a*. En el cas de *de*, que és molt freqüent, més del 90% dels casos (99 de 106) haurien de traduir-se per *de*. Tanmateix, la tassa d'errors és considerable: dels 8 canvis de *de* a *dae*, 5 s'han fet malament i, a més, 3 dels 99 casos de traducció a *de* tampoc han estat correctes. En total, la tassa d'encert és del 93%. La preposició *a* és molt menys freqüent (24 casos). Dels 15 casos en què hauria d'haver-se triat *in*, en 10 la tria ha estat incorrecta. La tassa d'encert és només del 58%.

L'anàlisi en detall dels errors mostra el següent:

- Si es manté, bàsicament, la mateixa estratègia, caldria ampliar molt considerablement la llista de paraules associades a una preposició o altra, incloent-hi també adjectius per al cas de *de*. Caldria també ampliar les estructures sintàctiques en què s'activen aquestes regles. Per exemple en el cas d'*a* no és suficient amb grups nominals "*a* + nom" i "*a* + det + nom", sinó que caldria incloure també, com a mínim, "*a* + det + adjectiu + nom". En el cas d'*a* caldria estudiar què fer davant de paraules desconegudes, com el topònim *Kanagawa* en els textos de la nostra anàlisi.
- En el cas d'*a*, és possible que convingui reformular les regles de manera a considerar *in* la traducció per defecte i afegint "excepcions" en què cal triar *a*. Això implicaria, entre altres coses, tenir una llista de verbs amb complement indirecte i regles capaces de reconèixer-los.

Convé també assenyalar un error també en l'única ocurrència de la conjunció *perquè*. Ha estat en la frase: "calgué esperar a la seva mort perquè els mèrits li fossin reconeguts". La raó ha estat que entre les estructures sintàctiques que es busquen per trobar el mode del verb de l'oració subordinada no hi havia: "conjunció + det + nom + pronom + verb" (sí hi havia, però, entre d'altres: "conjunció + det + nom + adverbi + verb").

Quant a la selecció lèxica dels substantius, cal assenyalar que dels 25 tractats, només hi ha hagut dues ocurrències en el corpus d'avaluació. Sense que sigui en absolut estadísticament significatiu, un cas dels dos s'ha resolt correctament i l'altre no.

Errors atribuïbles a la transferència estructural

Un cert nombre d'errors són imputables a mancances en la transferència estructural.

Per exemple, la forma verbal "s'han realitzat" es tradueix com a "si sunt realizados" gràcies a una regla que reconeix que es tracta d'una forma pronominal (de fet, una passiva reflexa) i posa l'auxiliar *èssere* en comptes d'*àere*. El problema ha estat que en el text hi havia la forma verbal "s'hi han realitzat". La presència del pronom ha fet que no es reconegués l'estructura i es traduís erròniament amb l'auxiliar *àere*.

Estructures que cal afegir en les regles de transferència són "*en* + infinitiu" (per exemple, *en arribar*) i "*tot i* + infinitiu" (per exemple, *tot i arribar*). Cal assenyalar que en el primer cas, les nostres proves en el corpus de la Viquipèdia mostren que hi ha dues traduccions possibles amb un nombre considerable de casos en cadascun d'ells: una com "després d'arribar" i l'altra com "arribant". Cal estudiar amb més compte quina opció triar. Probablement la tria mecànica d'una de les dues opcions a costa de l'altra no sigui la millor solució.

No hi ha hagut problemes de concordança dins dels sintagmes nominals, però sí en altres casos. En remarquem tres:

1. "Moltes de les quals van arribar" s'ha traduït com a "medas de sas cales sunt arribados" en comptes de "medas de sas cales sunt arribadas" perquè, per un oblit, no s'ha tingut en compte el gènere dels relatius (en aquest cas, *les quals*).
2. "El seu art fou seguit" s'ha traduït com a "s'arte sua est istadu sighidu" en comptes de "s'arte sua est istada sighida" perquè el canvi de gènere en el subjecte no s'ha traslladat als participis.
3. "Els ha unit" s'ha traduït com a "los at unidu" en comptes de "los at unidos" perquè no s'ha fet la concordança del participi amb el pronom de complement directe. El problema aquí és que, en general, el pronom català *els* pot referir-se tant al complement directe com a l'indirecte. No és possible incorporar una regla mecànica de concordança en casos del tipus "*els* + haver + participi" perquè afectaria frases com "els ha donat" en què no ha d'haver-hi aquesta concordança.

Un error recorrent és l'article definit davant l'any, que és obligatori en sard (els anys són molt freqüents en textos enciclopèdics com els del nostre corpus). Les regles reconeixen que un nombre és un any quan està precedit d'un mes i li afegeixen un article (16). En altres contextos, però, no es reconeix que es tracta d'un any, la qual cosa provoca una traducció errònia (17).

(16) 30 de març de 1853
30 de martzu de su 1853

(17) Entre 1830 i 1833
*Intre 1830 e 1833
(hauria de ser:
Intre su 1830 e su 1833)

Finalment, cal remarcar diferents canvis d'ordre de les paraules que s'han fet en la postedició, que expliquen la diferència considerable entre el WER i el PER obtinguts.

1. Diferents adjectius anteposats al nom s'han postposat, per exemple "el nou estil", traduït automàticament com a "su nou istile", s'ha corregit a "su istile nou", i "una clara influència", traduït com a "una crara influèntzia", s'ha modificat per "una influèntzia crara". Tanmateix s'han mantingut altres adjectius davant el nom com en "in sa matessi època" (cat. "a la mateixa època"). No és possible canviar mecànicament la posició de tots els adjectius de davant del nom al darrere. Cal estudiar la qüestió amb deteniment.

2. El mateix ha passat amb alguns adverbis en relació a l'adjectiu, per exemple "extremadament semblants", automàticament com a "a manera estrema sìmiles", s'ha corregit per "sìmiles meda", i "més estretament relacionat", traduït com a "prus in manera istrinta imparentadu", s'ha modificat a "imparentadu in manera prus istrinta".

3. L'expressió "ja entrat el segle XVI", traduïda automàticament com a "giai intradu su de XVI sèculos" s'ha corregit a "su de XVI sèculos giai intradu".

6 Treball futur

Hem començat a solucionar alguns dels problemes que es descriuen a la secció 5.2. Entre altres, convindria estudiar amb més detall els problemes amb l'ordre dels modificadors dins del sintagma nominal. Una possibilitat per tractar-los seria permetre que les regles de transferència siguin ambigües, i incorporar un model estadístic que tria la regla més adequada segons la combinació de lexemes.

Per altra banda, algunes de les regles de selecció lèxica es volen fer servir per millorar altres traductors de o al català i el sard. Més endavant, estaríem interessats en treballar en altres traductors per al sard, així com voldríem abordar el cors, que també es parla a Sardenya, per al qual Apertium ja té un prototipus experimental d'analitzador morfològic. Igualment, ens interessen altres llengües minoritzades de l'Estat italià, com el sicilià (per al qual Apertium ja disposa d'una versió preliminar d'un traductor a i de l'italià), el friülà i altres.

7 Conclusions

Hem presentat un traductor de català al sard. S'han discutit alguns reptes associats al desenvolupament d'una eina com aquesta per a una llengua en procés d'estandardització, com el sard. Després de presentar la feina realitzada en relació a la construcció del diccionari bilingüe i la creació de regles de selecció lèxica i transferència estructural, s'han analitzat els resultats obtinguts. El rendiment és inferior al d'altres traductors creats amb la mateixa tecnologia. Convé treballar més sobre la polisèmia dels mots i també ampliar les regles de transferència estructural. Aquestes regles haurien de reestructurar-se per facilitar el tractament de concordances més llunyanes de les que ara es tenen en compte, com les del subjecte amb l'atribut o el participi darrere de l'auxiliar *èssere*. Convé també un estudi més acurat de l'ordre d'adjectius i adverbis a l'interior dels sintagmes nominals.

El sistema està disponible com a programari de codi obert i lliure sota licència GNU GPL i es pot descarregar del servidor SVN d'Apertium.[16]

Agraïments

Voldríem agrair Diegu Corràine pels diferents aclariments que ens ha donat sobre la *Limba Sarda Comuna* al llarg de tot el projecte. Evidentment, els errors que té el traductor no són en cap manera atribuïbles a ell. El projecte ha estat parcialment finançat gràcies a una beca del programa Google Summer of Code.

Referències

Antonsen, Lene, Trond Trosterud & Francis M. Tyers. 2017. A North Saami to South Saami machine translation prototype. *Northern European Journal of Language Technology* 4. 11–27. doi:10.3384/nejlt.2000-1533.1642.

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez & Miriam A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. En *Computational Processing of the Portuguese Language (PROPOR 2006)*, 50–59.

Armentano-Oller, Carme, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Marco A. Montava, Sergio Ortiz-Rojas,

[16] http://www.apertium.org

Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez & Felipe Sánchez-Martínez. 2007. Apertium, una plataforma de código abierto para el desarrollo de sistemas de traducción automática. En *FLOSS (Free/Libre/Open Source Systems) International Conference*, 5–20.

Armentano-Oller, Carme & Mikel L. Forcada. 2006. Open-source machine translation between small languages: Catalan and Aranese Occitan. En *5th SALTMIL workshop on Minority Languages*, 51–54.

Balzhan, Abduali, Akhmadieva Zhadyra, Zholdybekova Saule, Tukeyev Ualsher & Rakhimova Diana. 2015. Study of the problem of creating structural transfer rules and lexical selection for the Kazakh–Russian machine translation system on Apertium platform. En *Turklang 2015*, 5–9.

Bick, Eckhard & Tino Didriksen. 2015. CG-3 – beyond classical constraint grammar. En *20th Nordic Conference of Computational Linguistics (NoDaLiDa'2015)*, 31–39.

Blasco Ferrer, Eduardo. 1986. *La lingua sarda contemporanea. Grammatica del logudorese e del campidanese. norma e varietà dell'uso. sintesi storica.* Della Torre.

Bolognesi, Roberto. 2007. La Limba Sarda Comuna e le varietà tradizionali del sardo. Disponible a http://www.sardegnacultura.it/documenti/7_88_20070518130841.pdf (15/10/2017).

Contini, Michel. 1981. Classificazione fonologica delle parlate sarde. *Bollettino dell'ALI* 3–4. 26–57.

Coròngiu, Giuseppe. 2013. *Il sardo: una lingua "normale"*. Condaghes.

Cutting, Doug, Julian Kupiec, Jan Pedersen & Penelope Sibun. 1992. A practical part-of-speech tagger. En *Third Conference on Applied Natural Language Processing*, 133–144.

Faridee, Abu Zaher Md. & Francis M. Tyers. 2009. Development of a morphological analyser for Bengali. En *First International Workshop on Free/Open-Source Rule-Based Machine Translation*, 43–50.

Forcada, Mikel L. 2006. Open-source machine translation: an opportunity for minor languages. En *Workshop "Strategies for developing machine translation for minority languages" (LREC'2006)*, 1–6.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez & Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation* 25(2). 127–144.

Ginestí-Rosell, Mireia, Gema Ramírez-Sánchez, Sergio Ortiz-Rojas, Francis M. Tyers & Mikel L. Forcada. 2009. Development of a free Basque to Spanish machine translation system. *Procesamiento de Lenguaje Natural* 43. 187–195.

Gökırmak, Memduh & Francis M. Tyers. 2017. A dependency treebank for Kurmanji Kurdish. En *International Conference on Dependency Linguistics (Depling'2017)*, 64–72.

Ivars-Ribes, Xavier & Victor M. Sánchez-Cartagena. 2011. A widely used machine translation service and its migration to a free/open-source solution: the case of Softcatalà. En *II International Workshop on Free/Open-Source Rule-Based Machine Translation*, 61–68.

Johnson, Ryan, Tommi Pirinen, Tiina Puolakainen, Francis. M. Tyers, Trond Trosterud, & Kevin Unhammer. 2017. North-Sámi to Finnish rule-based machine translation system. En *21st Nordic Conference on Computational Linguistics (NoDaLiDa'2017)*, 115–122.

Laxström, Niklas, Pau Giner & Santhosh Thottingal. 2015. Content translation: Computer assisted translation tool for Wikipedia articles. En *18th Annual Conference of the European Association for Machine Translation (EAMT'2015)*, 194–197.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8). 707–710.

Marting, Matthew & Kevin Brubeck Unhammer. 2014. FST trimming: Ending dictionary redundancy in Apertium. En *Proceedings of the 9th Conference on Language Resources and Evaluation (LREC'2014)*, 19–24.

Martín-Mor, Adrià. 2016. La localització de l'apli de missatgeria Telegram al sard: l'experiència de Sardware i una aplicació docent. *Tradumàtica: tecnologies de la traducció* 14. 112–127. doi:10.5565/rev/tradumatica.176.

Martín-Mor, Adrià & Alessandro Beccu. 2016. Sa localizatzione de Facebook in sardu. *Tradumàtica: tecnologies de la traducció* 14. 85–99. doi:10.5565/rev/tradumatica.179.

Martínez Cortés, Juan Pablo, Jim O'Regan & Francis Tyers. 2012. Free/open source shallow-transfer based machine translation for Spanish and Aragonese. En *Eight International Conference on Language Resources and Evaluation (LREC'2012)*, 2153–2157.

Moseley, Christopher (ed.). 2010. *Atlas of the world's languages in danger*. UNESCO Publishing 3rd ed. Disponible a http://www.unesco.org/culture/en/endangeredlanguages/atlas (15/10/2017).

Moshagen, Sjur, Jack Rueter, Tommi Pirinen, Trond Trosterud & Francis M. Tyers. 2014. Open-source infrastructure for collaborative work on under-resourced languages. En *Open-Source Infrastructure for Collaborative Work on Under-Resourced Languages (LREC'2014)*, 71–77.

Oppo, Anna. 2007. Conoscere e parlare le lingue locali. En Anna Oppo (ed.), *Le lingue dei sardi: una ricerca sociolinguistica*, capítol 1, 6–45. Regione Autonoma della Sardegna.

O'Regan, Jim & Mikel L. Forcada. 2013. Peeking through the language barrier: the development of a free/open-source gisting system for Basque to English based on apertium.org. *Procesamiento del Lenguaje Natural* 51. 15–22.

Otte, Pim & Francis M. Tyers. 2011. Rapid rule-based machine translation between Dutch and Afrikaans. En *The 15th conference of the European Association for Machine Translation (EAMT'2011)*, 153–160.

Ramírez-Sánchez, Gema, Felipe Sánchez-Martínez, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz & Mikel L. Forcada. 2006. Opentrad Apertium open-source machine translation system: an opportunity for business and research. En *Translating and the Computer 28th Conference*, n.pp.

Ravishankar, Vinit & Francis M. Tyers. 2017. Finite-state morphological analysis for Marathi. En *13th International Conference on Finite State Methods and Natural Language Processing*, 50–55.

Ravishankar, Vinit, Francis M. Tyers & Albert Gatt. 2017. A morphological analyser for Maltese. *Procedia Computer Science* 175–182.

Regione Autonoma della Sardegna. 2006. Limba Sarda Comune. Norme linguistiche di riferimento a carattere sperimentale per la lingua scritta dell'Amministrazione regionale. Disponible a http://www.regione.sardegna.it/documenti/1_72_20060418160308.pdf (15/10/2017).

Regione Autonoma della Sardegna. 2014. Monitoraggio sull'utilizzo sperimentale della Limba Sarda Comuna. Anni 2007–2013. Disponible a http://www.sardegnacultura.it/documenti/7_91_20140418114135.pdf (15/10/2017).

Salimzyanov, Ilnar, Jonathan Washington & Francis Tyers. 2013. A free/open-source Kazakh-Tatar machine translation system. En *Machine Translation Summit XIV*, 175–182.

Sundetova, Aida, Mikel Forcada & Francis Tyers. 2015. A free/open-source machine translation system for English to Kazakh. En *Turklang 2015*, 78–90.

Sánchez-Martínez, Felipe, Carme Armentano-Oller, Juan Antonio Pérez-Ortiz & Mikel L. Forcada. 2007. Training part-of-speech taggers to build machine translation systems for less-resourced language pairs. *Procesamiento del Lenguaje Natural* 39. 257–264.

Sánchez-Martínez, Felipe & Mikel L. Forcada. 2009. Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research* 34. 605–635.

Sánchez-Martínez, Felipe, Juan Antonio Pérez-Ortiz & Mikel L. Forcada. 2006. Speeding up target language driven part-of-speech tagger training for machine translation. En *5th Mexican International Conference on Artificial Intelligence (MICAI 2006)*, 844–854.

Toral, Antonio, Mireia Ginestí-Rosell & Francis M. Tyers. 2011. An Italian to Catalan RBMT system reusing data from existing language pairs. En *Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, 77–81.

Tyers, Francis M. 2009. Rule-based augmentation of training data in Breton–French statistical machine translation. En *13th Annual Conference of the European Association of Machine Translation (EAMT'2009)*, 213–218.

Tyers, Francis M. 2010. Rule-based Breton to French machine translation. En *14th Annual Conference of the European Association of Machine Translation (EAMT'2010)*, 174–181.

Tyers, Francis M. & Kevin Donnelly. 2009. apertium-cy – a collaboratively-developed free RBMT system for Welsh to English. *The Prague Bulletin of Mathematical Linguistics* 91. 57–66.

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

Tyers, Francis M. & Jacques A. Pienaar. 2008. Extracting bilingual word pairs from Wikipedia. *Collaboration: interoperability between people in the creation of language resources for less-resourced languages* 19. 19–22.

Tyers, Francis M., Felipe Sánchez-Martínez & Mikel L. Forcada. 2012. Flexible finite-state lexical selection for rule-based machine translation. En *16th Annual Conference of the European Association of Machine Translation (EAMT'2012)*, 213–220.

Tyers, Francis M., Felipe Sánchez-Martínez & Mikel L. Forcada. 2014. Unsupervised training of maximum-entropy models for lexical selection in rule-based machine translation. En *18th Annual Conference of the European Association for Machine Translation (EAMT'2014)*, 145–153.

Tyers, Francis M., Linda Wiechetek & Trond Trosterud. 2009. Developing prototypes for machine translation between two Sámi languages. En *13th Annual Conference of the European Association of Machine Translation (EAMT'2009)*, 120–128.

Wagner, Max Leopold. 1951. *La lingua sarda. storia, spirito e forma.* Ilisso.

Wiechetek, Linda, Francis M. Tyers & Thomas Omma. 2010. Shooting at flies in the dark: Rule-based lexical selection for a minority language pair. En *Advances in Natural Language Processing (NLP'2010)*, 418–429.