Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 9, Número 2- Dezembro 2017

ISSN: 1647-0818

Volume 9, Número 2 – Dezembro 2017

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Editorial

Amb aquest volum, Linguamàtica arriba als seus nou anys, amb 19 exemplars publicats des del 2009 i a una cadència de dos números a l'any (el 2010, excepcionalment, se'n van publicar tres). Una collita selecta de més de 100 articles científics sobre les tecnologies de les llengües peninsulars, divulgant i analitzant la recerca duta a terme a les seves institucions per més de 200 autores i autors que han volgut compartir amb nosaltres la seva feina en alguna de les nostres llengües.

En aquests nou anys, Linguamàtica s'ha anat consolidant com una revista científica de prestigi gràcies a l'esforç sostingut d'autors, revisors i editors. I, sens dubte, aquest esforç ha valgut i val la pena. Encetem el 2018, doncs, amb l'orgull de la feina feita i amb la il·lusió de la feina per fer, sabent que formem part d'una comunitat que vol seguir treballant en l'àmbit tecnològic per a les nostres llengües i en les nostres llengües, sense renunciar ni a la qualitat científica ni a la nostra identitat. Per molts anys.

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís,
Universidade de Vigo

Alberto Simões,
Universidade do Minho

Aline Villavicencio,
Universidade Federal do Rio Grande do Sul

Álvaro Iriarte Sanroman,
Universidade do Minho

Ana Frankenberg-Garcia,
University of Surrey

Anselmo Peñas,
Univers. Nac. de Educación a Distancia

Antón Santamarina,
Universidade de Santiago de Compostela

Antoni Oliver González,
Universitat Oberta de Catalunya,

Antonio Moreno Sandoval,
Universidad Autónoma de Madrid

António Teixeira,
Universidade de Aveiro

Arantza Díaz de Ilarraza,
Euskal Herriko Unibertsitatea

Arkaitz Zubiaga,
Dublin Institute of Technology

Belinda Maia,
Universidade do Porto

Carmen García Mateo,
Universidade de Vigo

Diana Santos,
Linguateca/Universidade de Oslo

Ferran Pla,
Universitat Politècnica de València

Gael Harry Dias,
Université de Caen Basse-Normandie

Gerardo Sierra,
Univers. Nacional Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Iñaki Alegria,
Euskal Herriko Unibertsitatea

Irene Castellón Masalles,
Universitat de Barcelona

Joaquim Llisterri,
Universitat Autònoma de Barcelona

José João Almeida,
Universidade do Minho

José Paulo Leal,
Universidade do Porto

Joseba Abaitua,
Universidad de Deusto

Juan-Manuel Torres-Moreno,
Lab. Informatique d'Avignon - UAPV

Kepa Sarasola,
Euskal Herriko Unibertsitatea

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Marcos Garcia,
Universidade da Corunha

María Inés Torres,
Euskal Herriko Unibertsitatea

Maria das Graças Volpe Nunes,
Universidade de São Paulo

Mercè Lorente Casafont,
Universitat Pompeu Fabra

Mikel Forcada,
Universitat d'Alacant

Pablo Gamallo Otero,
Universidade de Santiago de Compostela

Patrícia Cunha França,
Universidade do Minho

Rui Pedro Marques,
Universidade de Lisboa

Susana Afonso Cavadas,
University of Sheffield

Tony Berber Sardinha,
Pontifícia Univ. Católica de São Paulo

Xavier Gómez Guinovart,
Universidade de Vigo

Revisora Convidada

Nora Aranberri, Euskal Herriko Unibertsitatea

Artigos de Investigação

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

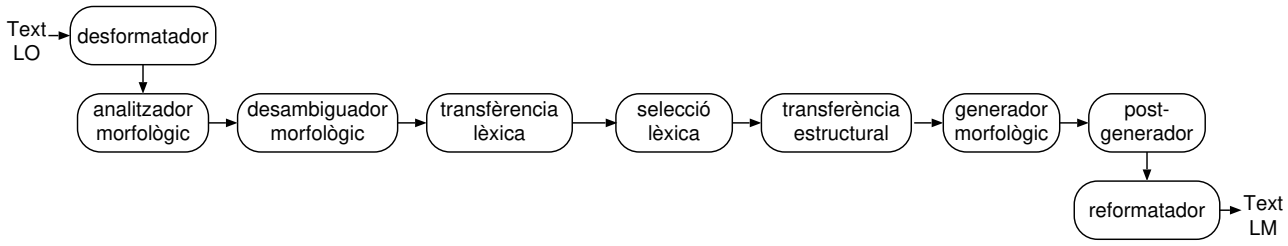

Figura 1: Arquitectura modular de la plataforma per desenvolupar sistemes de traducció automàtica Apertium. Els mòduls es comuniquen mitjançant canonades estàndard d'Unix.

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.

Propuesta recibida el 15 de octubre 2017 y aceptada para publicación el 3 de diciembre 2017.

Detección automática de nombres eventivos no deverbales en castellano: un enfoque cuantitativo basado en corpus

Automatic detection of non-deverbal eventive nouns in Spanish: a quantitative, corpus-based approach

Rogelio Nazar
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Rebeca Soto
Pontificia Universidad Católica de Valparaíso
rebecasotoriveros@gmail.com

Karen Urrejola
Pontificia Universidad Católica de Chile
kuc@uc.cl

Resumen

Presentamos un estudio en el campo de la detección de nombres eventivos no deverbales, que son aquellos nombres que designan eventos pero que no han pasado por un proceso de derivación a partir de verbos, como *fiesta* o *cóctel*, y no presentan por ello las pistas morfológicas típicas de los nombres deverbales, como los afijos *-ción*, *-miento*, etc., por lo que son justamente los más difíciles de detectar.

En el presente artículo continuamos y extendemos el trabajo iniciado por Resnik (2010), quien ya ofrece pistas para la detección automática de este tipo de unidades léxicas. A las sugerencias de Resnik añadimos otras, entre ellas el análisis inductivo de corpus, analizando con qué tipo de palabras suele coocurrir el nombre eventivo, y utilizándolas como predictores de esta condición. Además, simplificamos considerablemente el algoritmo de detección y aplicamos los experimentos a un corpus de mayor tamaño, el EsTenTen (Kilgarriff & Renau, 2013), de más de 9 mil millones de palabras. Finalmente, presentamos los primeros resultados de nombres eventivos extraídos automáticamente, incluyendo numerosos no deverbales.

Palabras clave

análisis inductivo de corpus, lexicografía computacional, sustantivos eventivos no deverbales

Abstract

We present a study in the field of the automatic detection of non-deverbal eventive nouns, which are those nouns that designate events but have not experienced a process of derivation from verbs, such as *fiesta* ('party') or *cóctel* ('cocktail') and, for this reason, do not present the typical morphological features of deverbal nouns, such as *-ción*, *-miento*, and are therefore more difficult to detect.

In the present research we continue and extend the work initiated by Resnik (2010), who offers a number of cues for the detection of this type of lexical unit. We apply Resnik's ideas and we also add new ones, among them, the inductive analysis of the words that tend to co-occur with eventive nouns in corpora, in order to use them as predictors of this condition. Furthermore, we simplify the classification algorithm considerably, and we apply the experiments to a larger corpus, the EsTenTen (Kilgarriff & Renau, 2013), comprising more than 9 billion running words. Finally, we present the first results of the automatic extraction of eventive nouns from the corpus, among which we find plenty non-deverbal nouns.

Keywords

computacional lexicography, inductive corpus analysis, non-deverbal eventive nouns

1 Introducción

En el contexto de la clasificación de los tipos de nombres o sustantivos, en los últimos años ha resultado de interés la distinción entre los nombres que designan eventos (por ej. *ceremonia*) de aquellos que hacen referencia a entidades en lugar de eventos (ej. *silla*). Sin embargo, el examen de la bibliografía revela que el estudio de este fenómeno ha sido abordado principalmente desde una mirada teórica e introspectiva, y los criterios para la caracterización de los nombres eventivos que se han establecido en la investigación actual se han basado predominantemente en una lógica deductiva (Graña López, 1993; Bosque, 1999; De Miguel, 2006; Real Academia Española, 2010; Fábregas, 2010). En otras palabras, los investigadores, a partir de su conocimiento de la lengua,

DOI: 10.21814/lm.9.2.253

dan cuenta de las particularidades del comportamiento sintáctico-semántico de este tipo de sustantivos y establecen criterios para su identificación. Son todavía pocos los estudios que adoptan una mirada empírica o que intentan contrastar con el corpus las propiedades establecidas deductivamente, como es el caso de Resnik (2010).

A partir de esta constatación, el presente trabajo busca ampliar el estudio pionero de Resnik para aportar a la caracterización de nombres eventivos por medio del análisis de su comportamiento en un corpus, con el fin de establecer criterios que permitan su identificación de forma objetiva y sistemática. Creemos que avanzar desde la teoría y los métodos puramente introspectivos hacia un análisis empírico es un paso fundamental, ya que el corpus representa el uso efectivo que los hablantes hacen de la lengua.

Concretamente, aplicamos los criterios de detección aportados por Resnik e incluimos otros que obtuvimos de manera inductiva; es decir, planteamos un método mixto, combinando pistas inductivas con aquellas encontradas en la bibliografía. Además, hemos conseguido simplificar considerablemente el algoritmo de clasificación, desde un método computacionalmente intensivo como el aprendizaje automático a uno basado en simples cálculos de coocurrencia. Esta simplificación metodológica permite la aplicación a un mayor volumen de datos de manera más rápida, lo que nos permite aplicar el método al corpus EsTenTen, que supera los 9 mil millones de palabras, y obtener grandes cantidades de nombres eventivos con precisión suficiente para ser de utilidad práctica en el campo del análisis lexicográfico.

El artículo se estructura de la siguiente manera: en la sección 2 revisamos la caracterización de los nombres eventivos en español que se ha realizado en los últimos años. En la sección 3 presentamos nuestra metodología de trabajo, que consiste en un algoritmo de clasificación de nombres eventivos a partir del estudio de sus contextos de coocurrencia. En la sección 4, en tanto, presentamos los resultados de la aplicación de este método primero con un listado de 100 nombres eventivos (no deverbales) y 100 nombres no eventivos compilados previamente por Resnik, para luego ofrecer también los resultados de la aplicación del método al corpus EsTenTen. El resultado es evaluado de manera manual examinando una muestra de 400 candidatos.

Este artículo es acompañado además de un sitio web[1] en el que se ofrecen los resultados del análisis y todo el código fuente del proyecto. Pensamos que este algoritmo de clasificación y su implementación pueden ser reaprovechados para realizar otro tipo de clasificaciones en el campo de la lexicología y lexicografía computacional.

[1] http://www.tecling.com/neven

2 Marco teórico

Las clasificaciones principales aportadas por la gramática

La gramática tradicional ha clasificado las palabras en diferentes clases sintácticas: artículo, sustantivo, pronombre, verbo, adverbio, preposiciones y conjunciones (Real Academia Española, 2010). Dentro de los denominados sustantivos o nombres, se han distinguido diferentes tipos a partir de criterios diversos: contables e incontables, abstractos y concretos, comunes y propios, individuales o colectivos, etc. La clase de los sustantivos es heterogénea en cuanto al comportamiento sintáctico-semántico de las unidades léxicas que la conforman y, de este modo, constituye un área de interés para los investigadores caracterizarlas desde diversos enfoques.

Esta investigación en particular se centra en la diferencia de los nombres eventivos respecto de los no eventivos. Los primeros corresponden a "un tipo de sustantivos individuales (por tanto, contables) que no designan objetos físicos, sino acontecimientos o sucesos" (Bosque, 1999, p. 55), por ejemplo: *fiesta*. Los no eventivos, al contrario, designan entidades, contables y no contables, que no se corresponden con sucesos o acontecimientos, por ejemplo: *gato*.

Dentro de la categoría de los nombres eventivos es posible diferenciar, por un lado, entre aquellos que derivan de verbos, proceso que puede ser acusado por la presencia de un morfema nominal (ej. *inaugura-ción*) o puede no presentar dicha marca (ej. *desfile*), y, por otro lado, los que no provienen de un verbo (ej. *boda*) (Fábregas, 2010). A su vez, dentro de la clase de los sustantivos deverbales, algunos pueden denotar un acontecimiento (eventivos) o bien el resultado del proceso implicado en el mismo (resultativos) (Grimshaw, 1990; Pustejovsky, 1995; Picallo, 1999; Alexiadou, 2001; Alonso Ramos, 2004). Grimshaw (1990) señala que las nominalizaciones eventivas no son contables pero las resultativas sí, como se verá en detalle más adelante (Cuadro 1).

La investigación en este ámbito ha establecido diferencias en el comportamiento sintáctico de los nombres eventivos con el fin de caracterizar esta clase de sustantivos, que aunque no derive de un verbo, de todos modos expresa un evento (Fábregas, 2010). Esto los distingue de los nombres pu-

ramente designativos, ya que los eventivos tienen una capacidad predicativa (De Miguel, 2006) o estructura argumental (Grimshaw, 1990). Al respecto, De Miguel (2006) señala que los nombres eventivos suelen aparecer con verbos de soporte o de escaso contenido léxico como en *dar una cena* o *hacer una fiesta*, aunque no de forma exclusiva, ya que también son comunes construcciones como *dar un golpe* o *hacer un pastel*.

Bosque (1999) destaca que los nombres eventivos pueden ser sujeto de predicados como *tener lugar* y también complemento directo de verbos como *presenciar*. Por otro lado, y dado que poseen límites temporales, también se acompañan de verbos como *empezar* y *concluir* y aparecen como complemento preposicional de *durante*: *durante la clase/el eclipse/la ocupación alemana*; *antes* y *después* (o *tras*): *después de la cena, antes de la conferencia*.

En este contexto, sin embargo, pueden también producir una lectura eventiva nombres que en principio no son eventivos (como *después del último autobús* o *antes del cigarrillo*). De hecho, la lectura eventiva de nombres no eventivos no es infrecuente. El nombre *libro* también puede resultar ambiguo en enunciados que pueden admitir una interpretación eventiva (como en *empecé el libro esta tarde*) y una interpretación objetual (como *el libro está sobre la mesa*). Esta ambigüedad entre una lectura eventiva y una objetual en los sustantivos eventivos es sistemática y observable en muchos otros casos, tales como *cena* o *concierto*, y entraría en el ámbito de lo que Apresjan (1974) y Pustejovsky (1995) denominan polisemia sistemática o regular. En este caso particular, podríamos hablar de coerción de tipos, o bien de tipos complejos (dotted types) en la terminología de Pustejovsky (1995). En castellano, el estudio de estos casos ha sido desarrollado por Adelstein et al. (2012), quienes contrastan los usos locativos y eventivos que puede mostrar un mismo sustantivo.

El sustantivo eventivo no deverbal como una clase autónoma

Un precedente importante en el estudio de los nombres eventivos no deverbales en lengua castellana es el estudio de Resnik (2010). En este se presenta un panorama completo de la investigación en el ámbito hasta esa fecha y, además, ofrece una propuesta de caracterización que, si bien se basa en el método introspectivo, es contrastada con un análisis de corpus.

La autora propone que los nombres eventivos no deverbales serían una clase autónoma, con un comportamiento sintáctico distinto al de las otras categorías, como los eventivos deverbales, de proceso y de resultado, y, por supuesto, de los sustantivos no eventivos. Así, si bien los eventivos no deverbales se suelen clasificar dentro de las nominalizaciones resultativas (Grimshaw, 1990), la autora propone, a partir de distintas pruebas, que los no deverbales serían una clase diferente.

Resnik parte de las propuestas de Grimshaw (1990), quien distingue entre nominalizaciones eventivas y nominalizaciones resultativas en función de la presencia o ausencia de una estructura eventiva compleja, y Picallo (1999), que lleva la distinción de Grimshaw al plano sintáctico y afirma que las nominalizaciones eventivas se realizan con una construcción pasiva, mientras que las nominalizaciones resultativas lo hacen con una construcción activa. Picallo sostiene que las nominalizaciones eventivas/pasivas se distinguen de las resultativas/activas por la presencia de elementos como la expresión del agente, que aparece en un sintagma preposicional introducido con *(por parte) de* en las nominalizaciones eventivas pero con un nombre genitivo con *de* en las resultativas. Las nominalizaciones eventivas aparecerían así en función de sujeto de predicados como *tener lugar, durar u ocurrir*; las resultativas, en cambio, serán sujeto de predicados tales como *ser inconsistente, ser considerado incorrecto*, *ser publicado*, etc. En respuesta a esta idea, Resnik sostiene que "la interpretación de la diferencia entre lectura eventiva y lectura resultativa de las nominalizaciones en términos de construcción pasiva y construcción activa se vuelve extraña al incorporar el caso de la nominalización creada a partir de una base inacusativa: es cierto que se trata de una construcción sin argumento externo, con un tema como sujeto, pero está claro que no es una construcción pasiva (no se puede incluir un agente como adjunto) y en ese sentido sería más adecuado, en todo caso, hablar de las nominalizaciones eventivas en general como construcciones ergativas" (Resnik, 2010, p. 75).

Para Alexiadou (2001), en tanto, se distingue entre nominalizaciones y nombres no deverbales en función de la estructura morfológica. Los últimos, a su vez, se diferencian de los nombres resultativos en que no se interpretan como eventivos. Así, las nominalizaciones eventivas tendrían parte de la estructura funcional de los verbos, a diferencia de las nominalizaciones resultativas, que carecen de estas proyecciones verbales. Resnik (2010) adapta esta propuesta para los nombres eventivos simples que, si bien carecen de morfología verbal, sí tienen propiedades aspectuales.

Teniendo en cuenta las clases aspectuales de Vendler (1967) (estados, actividades, logros, realizaciones), Resnik también distingue clases aspectuales de nombres no deverbales en función de los modificadores que admiten. El caso del nombre *clase*, por ejemplo, en tanto que actividad, corresponde a un evento durativo, por tanto atélico, mientras que otros, como *accidente*, ya son un evento puntual y, por tanto, no admiten el modificador durativo. Esto, en definitiva, lleva a pensar que la categoría de aspecto léxico no es una propiedad intrínseca de las raíces verbales, sino una categoría funcional que puede aparecer tanto con núcleos verbales como nominales.

Así, a partir de la clasificación de los nombres eventivos en español en nominalizaciones eventivas, nominalizaciones resultativas y nombres eventivos no deverbales, Resnik se centra en el análisis de las propiedades de los eventivos no deverbales y demuestra que estos no son equiparables a las nominalizaciones resultativas, ya que tienen una estructura funcional específica que incluye propiedades aspectuales.

El Cuadro 1, adaptado de Bel et al. (2010), resume las propiedades léxico semánticas de los diferentes sustantivos descritos en este apartado. En este cuadro se muestran, por un lado, las clases distinguidas por Grimshaw (1990): los sustantivos no eventivos, los sustantivos eventivos de proceso y los sustantivos resultativos; y por otro, los sustantivos eventivos no deverbales como una clase autónoma, tal como propone Resnik (2010). De esta manera, es posible observar cómo los no deverbales presentan propiedades específicas en relación con los demás.

La detección automática de sustantivos eventivos

La dificultad principal de la detección automática de sustantivos eventivos es, como se ha explicado, que el nombre eventivo no deverbal no ofrece las marcas morfológicas que son propias de los eventivos deverbales, como *accidente* o *guerra*, y por tanto no pueden ser detectados con afijos como *-ción*, *-miento*, etc[2].

Resnik (2010) no recurre a pistas morfológicas porque trabaja específicamente con nombres eventivos no deverbales, y es, en cambio, el comportamiento sintáctico-semántico de estos nombres lo que permite su detección automática. Entre los rasgos predictores para esta clasificación, la autora destaca los siguientes:

- Los nombres son complementos de *durante, hasta el final de, desde el principio de*, entre otras
- Son argumento de verbos tales como *ocurrir, producir, celebrar* y verbos aspectuales como *empezar* o *durar*
- Admiten cuantificadores aspectuales como *dos días/semanas de* o *una etapa/un período de*, entre otros
- Predicados aspectuales como *ocurrir, producir, desatar, desencadenar, celebrar*
- Cláusulas sustantivas: proceso/hecho/actividad/evento de + construcción nominal
- Referencia anafórica con *esto*
- Selección de ser/estar: el verbo ser selecciona un evento como sujeto cuando presenta complementos locativos o temporales
- Paráfrasis con *hecho, actividad, evento*
- Argumento del verbo *presenciar*
- Complemento de preposiciones aspectuales (*en medio de, durante*)
- Modificación con adjetivos aspectuales
- Modificación con cláusula temporal al + infinitivo

Utilizando pistas de este tipo, Resnik (2010) intenta la clasificación de los nombres en las listas que aparecen en el Cuadro 2.

El corpus utilizado en su experimento está conformado por textos de prensa de El País y La Vanguardia con un total de 21 millones de palabras, parte del corpus técnico del IULA o CT-IULA, (Cabré et al., 2006). Utilizando este corpus y el software Weka (Hall et al., 2009), entrenó un clasificador del tipo árbol de decisión (Decision Tree) C4.5 (Quinlan, 1993). La clasificación se llevó a cabo por medio del "10-fold cross-validation", que implica repetir el experimento diez veces utilizando cada vez un 90 % del set como entrenamiento y un 10 % como test, garantizando que cada elemento haya estado en ambos conjuntos, es decir, en el entrenamiento y en el test.

Con este método la autora reporta una tasa de éxito importante para ser un estudio pionero en el área: una precisión de 0.84 y 0.82 clasificando los nombres eventivos y no eventivos, y una

[2] Esto no quiere decir que la detección de nombres eventivos mediante pistas morfológicas sea una tarea sencilla, como señalan Balvet et al. (2011): no todo nombre con *-ción* será eventivo: no lo es *población*, aunque sí *poblamiento*.

	Sustantivo Eventivo no-deverbal	**Sustantivos de Proceso**	**Sustantivos Resultativos**	**Sustantivos no-eventivos**
Ejemplo	*guerra*	*construcción = evento*	*construcción = objeto resultativo*	*mapa*
Argumento interno obligatorio	No	Sí	No	No
Realización del argumento externo	Genitivo frase determinante	Frase preposicional 'por'	genitivo frase determinante	genitivo frase determinante
Sujeto de verbo aspectual (comenzar, terminar)	Sí	Sí	No	No
Cuantificador aspectual (un periodo de)	Sí	Sí	No	No
Complemento de durante...	Sí	Sí	No	No
Contable/ /no-contable (determinantes, formas plurales)	contable/ /no-contable	no-contable	contable	contable/ /no-contable

Cuadro 1: Clasificación de sustantivos adaptado de Bel et al. (2010, p. 47)

cobertura de 0.82 y 0.84, respectivamente. La limitación está en el aspecto empírico, ya que se opera con un listado de 100 nombres eventivos y 100 no eventivos, y en un corpus de tamaño limitado.

En un estudio posterior, Bel et al. (2010) presentan el análisis de los sustantivos eventivos no deverbales y un experimento de detección automática para el inglés y el español. Allí reproducen los experimentos y se habla de un "accuracy" de 80 % para el castellano. Se presume que con ese término se refieren en realidad a la precisión, definida como (1) junto con la cobertura (2) (Baeza-Yates & Ribeiro-Neto, 1999), y no a la "accuracy" definida en (3), donde *tp* o *true positive* sería el sustantivo correctamente detectado como eventivo, el *fp* o *false positive* el sustantivo no eventivo incorrectamente seleccionado como eventivo, el *fn* o *false negative* el sustantivo eventivo no detectado y *tn* o *true negative*, el sustantivo no eventivo correctamente descartado.

$$precision = \frac{tp}{fp + tp} \tag{1}$$

$$recall = \frac{tp}{fn + tp} \tag{2}$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

En este segundo estudio, Bel et al. (2010) intentan elevar la precisión clasificando solo los elementos más seguros, y consiguen llegar a un 95 %, aunque entonces la cobertura disminuye a 43 %. Más allá de la tasa de éxito en la detección, estos dos trabajos son de importancia porque se propone una caracterización que, aunque basada en el método introspectivo, es contrastada con un análisis de corpus. Las limitaciones, sin embargo, son el tamaño de la muestra de sustantivos (Cuadro 2) y que el corpus con el que trabajan es de un tamaño muy reducido.

El trabajo que presentamos ahora comparte el objetivo de Resnik (2010) y Bel et al. (2010) en cuanto el relevamiento de características contextuales que permitan identificar de manera automática en un corpus sustantivos eventivos no deverbales y diferenciarlos de sustantivos no eventivos. Sin embargo, la metodología seguida es ahora un enfoque mixto con elementos que provienen del análisis de corpus basado en la observación del comportamiento de los sustantivos eventivos no deverbales en un corpus. En esta investigación sistematizamos y simplificamos el esquema de elementos predictores y utilizamos un

Categoría	Sustantivos
Eventivos	*fiesta, feria, festival, boda, funeral, velorio, velatorio, ceremonia, evento, picnic, cóctel, té, banquete, festín, ágape, tertulia, campaña, cónclave, cumbre, asamblea, sesión, misa, vacaciones, receso, excursión, trayecto, travesía, clase, conferencia, curso, taller, workshop, congreso, simposio, jornadas, tumulto, coloquio, entrevista, audiencia, concierto, ópera, serenata, espectáculo, show, programa, película, ciclo, discurso, sermón, torneo, campeonato, carrera, rally, tormenta, tempestad, temporal, borrasca, terremoto, sismo, huracán, maremoto, sequía, catástrofe, cataclismo, desastre, tragedia, holocausto, drama, incendio, accidente, impacto, siniestro, caos, crisis, guerra, batalla, conflicto, paz, silencio, ruido, escándalo, lío, follón, problema, motín, huelga, incidente, boicot, pánico, miedo, pasión, furor, rabia, siesta, frío, calor, hambre, pereza, dolor, fiebre, gripe*
No eventivos	*mapa, antología, característica, droga, plasma, teléfono, montaña, tubo, estética, cliente, escena, colectividad, canal, arquitectura, cara, levedad, estadio, batuta, súbdito, ciudad, madera, cifra, habitación, fotocopia, vivienda, gas, literatura, especie, paisaje, diferencia, carretera, seguridad, red, contraseña, rodilla, virus, cantidad, provincia, detalle, público, garganta, maqueta, dato, volcán, cárcel, familia, dinero, estereotipo, tarifa, compañía, justicia, humo, balneario, paquete, prensa, vehículo, dueño, prejuicio, banda, consorcio, economía, figura, mar, pancarta, grupo, arma, informe, diario, trama, zona, misterio, facultad, cadáver, nivel, pista, columna, combustible, estructura, ruta, alimento, herramienta, factura, miembro, forma, tema, fuente, temperatura, euro, ilusión, punto, batería, silueta, unidad, organismo, norma, vía, planta, autobús, perspectiva, antena*

Cuadro 2: Listados de sustantivos eventivos y no eventivos compilados por Resnik (2010)

algoritmo más sencillo que el árbol de decisión, lo que nos permite trabajar con un corpus de tamaño mucho mayor.

3 Materiales y Métodos

Para conseguir el objetivo de establecer criterios que permitan identificar de forma objetiva y sistemática los nombres eventivos, planteamos el problema como una tarea de clasificación y aplicamos para ello un método basado en corpus. En nuestro caso, este corpus es el EsTenTen (Kilgarriff & Renau, 2013), un corpus de gran tamaño (9 mil millones de palabras), constituido por páginas web de distintos países de habla castellana. Entendemos que este corpus cumple con la definición de Sinclair (1991): una colección de textos que manifiestan ocurrencias de lenguaje natural y que han sido escogidos para caracterizar un estado o variedad de lenguaje.

La idea principal

El examen de las concordancias de un sustantivo eventivo no deverbal como *fiesta* en el corpus EsTenTen revela rápidamente una serie de pistas sistemáticas ofrecidas por el contexto inmediato, infraoracional. El Cuadro 3 ofrece algunos ejemplos extraídos de este corpus.

Ya en el examen visual de estas concordancias, antes de aplicar ningún criterio de contabilización de frecuencia de aparición de las palabras del contexto (la metodología básica para este tipo de problema), podemos encontrar con facilidad una serie de elementos que son consistentes con la condición de eventivo del sustantivo analizado.

Como primera aproximación, lo que salta a la vista es que el sustantivo tiene más de un significado, ya que se utiliza para designar un tipo particular de automóvil, el *Ford Fiesta*. Difícilmente se puede hablar de un caso de polisemia en este caso, ya que no se puede confundir el sustantivo con la condición de nombre propio que tiene cuando se usa para designar el coche. Las concordancias con este uso aparecen, en el Cuadro 3, agrupadas en las líneas 1 a 8. En el resto, sin embargo, advertimos la interpretación eventiva y encontramos elementos como días de la semana, especificadores como *durante*, sustantivos utilizados para las medidas de tiempo: *hora, día, semana, año* y los adverbios utilizados para el orden secuencial: *antes* y *después*.

Estos elementos específicos aparecen con una alta frecuencia en los contextos de los nombres eventivos, lo que sugiere que puede ser

1	ejemplo , asique no se si Ford permitirá que el	Fiesta	supere al Focus en ese aspecto , o mejorarà el
2	ese aspecto , o mejorará el Focus cuando venga el	Fiesta	Un autazo A ver si entendi bien por
3	y pensar que van a hacerle un restyling al pedorrisimo	Fiesta	que venden aca no El fiesta actual es un
4	restyling al pedorrisimo Fiesta que venden aca no El	fiesta	actual es un desastre mi novia tenia uno y los
5	Hace meses que estoy esperando este	Fiesta	porque me estaba por comprar un Agile y despues de
6	jamas compraría un 207 c y si compraría un	Fiesta	KD si ahora estuviera al precio de lanzamiento ,
7	seras ignorante , no es para gente con familia el	fiesta	KD es para gente q le gustan los AUTOS ,
8	cosa , la clase de consumidor que apunta a un	Fiesta	KD dudo mucho que tenga en cuenta un Fluence
9	En esta ciudad , **se lleva a cabo la**	fiesta	en honor a su patrona , con actos litúrgicos ,
10	folclórico Herencia Gaucha brilló **el domingo** en la	Fiesta	del Gaucho Carlos Andina , el profesor que los
11	comienza en el año 1997 **El domingo** en la	Fiesta	del Gaucho vimos en su plenitud al ballet folclórico
12	lo que saben hacer Pero no es la única	fiesta	, **este año** fueron a competir a Berazategui y a
13	un sábado muy especial en la XXXIX edición de la	Fiesta	Nacional del Gaucho , con la incorporación en la
14	más destacados de la **jornada** con que **se inicia la**	fiesta	El clima acompaño la primera jornada de la Fiesta
15	que propone la fiesta**dieron inicio a la** XXXIX	Fiesta	Nacional del Gaucho Con el usual espectáculo
16	registran distintos **momentos de la**	fiesta	con las nuevas tecnologías , haciendo un uso diferencial
17	encender la vela correspondiente **durante la**	fiesta	La torta Existen en plaza muchos diseños Los
18	bar mitzvá , con cintas de acuerdo al color de	fiesta	Se puede contratar con el servicio de catering y
19	en exposición **durante las horas que dure la**	fiesta	El candelabro y encendido de velas Se contratará el
20	de anticipación para que esté disponible **el día de la**	fiesta	Se lo puede pedir decorado con un arreglo de
21	una carpeta forrada en raso (del tono de la	fiesta	) que puede estar bordado con hilos plateados
22	se definen una o **dos semanas antes de la**	fiesta	en una reunión con el DJ y el grupo familiar
23	, y no al revés **Después que pasó la**	fiesta	Todo salió perfecto , como lo habían soñado El
24	camino de los Picos de Europa donde **tiene lugar la**	Fiesta	, muy cerca del lago Enol , siendo la única

Cuadro 3: Ejemplos de concordancias del sustantivos *fiesta* en el corpus. Las líneas 1 a 8 representan usos del nombre propio que refiere al automóvil *Ford Fiesta*. En negrita los elementos que consideramos predictores de la condición de nombre eventivo en el caso de *fiesta*.

conveniente, en el sentido de seguir el principio de parsimonia, atender primero a este grupo de predictores y no otros que, si bien pueden ser también confiables, van a ocurrir con menor frecuencia. Procedemos de esta manera también porque si es posible obtener el resultado esperado despejando algunas variables del problema, ello puede ser también beneficioso desde el punto de vista computacional, porque se traduce directamente en mayor capacidad para procesar más texto en menor tiempo.

La intuición general es entonces que los sustantivos eventivos tenderán a coaparecer más frecuentemente con elementos predictores eventivos en comparación con los no eventivos. En este punto es importante recalcar que los elementos predictores no son verdaderos diagnósticos de eventividad. Esa es precisamente la diferencia entre el pensamiento cuantitativo y el simbólico o basado en reglas: la presencia de un predictor en un determinado contexto no indica que esa sea una ocurrencia concreta de un nombre eventivo. La que cuestión es que si analizamos una gran cantidad de contextos de aparición del sustantivo podemos determinar si está asociado o no con esos predictores. Estos elementos predictores se pueden compilar en un listado que luego se coteja con el vocabulario de los contextos de ocurrencia del sustantivo analizado, lo que lo hace un procedimiento bien sencillo.

Clasificación de pistas contextuales

Mediante el análisis de concordancias de nombres eventivos y no eventivos, procedimos a analizar y clasificar los distintos elementos léxicos del contexto por frecuencia decreciente, reteniendo aquellos que consideramos predictores confiables de la categoría de nombre eventivo. Esto significa que utilizamos solo rasgos positivos, es decir, solo rasgos que son indicativos de la clase de eventivos y no tenemos en cuenta características que serían propias únicamente de los no eventivos. Disponemos de cuatro categorías de rasgos:

1. **Días de la semana y meses** (*lunes, martes, miércoles, jueves, viernes, sábado, domingo, enero, febrero, marzo, abril, mayo, junio, julio, agosto, septiembre, octubre, noviembre, diciembre*)
2. **Medidas temporales** (*semana, día, mes, año, hora, minuto*)
3. **Verbos aspectuales** (*ocurrir, comenzar, iniciar, efectuar, celebrar, (hubo, hubieron, habrán*))
4. **Otros ítems léxicos aspectuales** (*durante, antes, después, duración, constante, menudo, frecuente, rápido, lento*)

Como se puede apreciar, llevamos a cabo una considerable simplificación de los rasgos clasificadores con respecto al trabajo de Resnik (2010),

coincidiendo también con una simplificación importante del algoritmo de clasificación, al estar basado únicamente en frecuencias de coocurrencia y no en aprendizaje automático.

El algoritmo de clasificación

Implementamos un algoritmo que acepta como entrada un conjunto de sustantivos, que denotaremos como $X = \{X_1, ...X_n\}$ y el resultado es un conjunto tal que $(\forall x \in X)\ E(x)$, donde $E(x)$ resulta en un valor o ponderación que servirá para ordenar el conjunto X en un listado, en función de la probabilidad del candidato x de ser un nombre eventivo.

El algoritmo está basado en el análisis de los contextos de aparición de los sustantivos analizados en un corpus de gran tamaño. Por tanto, por cada sustantivo analizado $x \in X$, extrae una muestra aleatoria de 5.000 contextos de aparición de x, de extensión oracional, y recorre estos contextos para encontrar ocurrencias a derecha o izquierda de los elementos predictores descritos en el apartado 3.2.

Podemos representar cada contexto de aparición de x a su vez como un conjunto $C_j = \{t_1, ...t_{|C_j|}\}$, es decir ignorando el orden de aparición. Cada unidad léxica $t \in C_j$ es una palabra apareciendo a derecha o izquierda del elemento analizado x.

Una vez clasificados los predictores o pistas contextuales, que definimos aquí como un conjunto L, el algoritmo de clasificación puede detectar aquellos sustantivos que muestren una alta proporción de estos elementos predictores en sus contextos, y acusarlos como candidatos a nombres eventivos.

Definimos una función que asigna un valor $E(x)$ en (4), que medirá la cantidad de apariciones de cualquier elemento predictor de la condición de nombre eventivo en los contextos de x.

$$E(x) = \sum_{j=1}^{|C|} \begin{cases} 1 & \exists t \in C_j \wedge t \in L \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

La medida $E(x)$ pondera entonces al candidato x para indicar la frecuencia relativa de dichos elementos predictores en sus contextos. En función de esta estimación, podemos exponer al candidato a un ordenamiento más cercano a las primeras posiciones, o bien directamente tomar una decisión binaria de tipo eventivo/no eventivo. Esto último también puede ser llevado a cabo por medio de la aplicación de una ordenación de todos los candidatos y la aplicación posterior de un umbral de corte arbitrario, pero también se puede aplicar una regla de eliminación como $K(x)$, como se indica en la ecuación (5) y eliminar, así, a los nombres simples. Esto disminuye drásticamente el tamaño de los listados resultantes, lo cual facilita su posterior examen y procesamiento computacional. Reservamos un valor "I" para los casos indefinidos, tal que el elemento no se puede clasificar debido a que no hay suficientes contextos de aparición, ignorándolo como un elemento no analizable si se encuentra por debajo un umbral de frecuencia arbitrario u.

$$K(x) = \begin{cases} \text{I} & |C| \leq u \\ \text{E} & E(x) > p \\ \text{N} & \text{otherwise} \end{cases} \tag{5}$$

Añadimos además un criterio penalizador consistente en determinar si en alguno de los parámetros 1 a 4, descritos en la sección 3.2 se encuentra que en total esas unidades aparecen en menos del 2 % de la muestra. Si este es el caso, entonces el sustantivo se considera no eventivo.

4 Resultados

Para nuestros experimentos, procedimos en primer lugar a reproducir la clasificación del conjunto de datos elaborado por Resnik (2010), presentado en el Cuadro 2.

Los resultados obtenidos en la prueba con la muestra de 200 sustantivos eventivos y no eventivos tomada de Resnik, con los parámetros definidos en la sección anterior, revelan una precisión de 95 %, cobertura de 63 % y F1 de 75, por tanto un aumento significativo de la cobertura –casi 20 puntos porcentuales– con respecto a Bel et al. (2010).

Luego, y con el objeto de superar una de las limitaciones del trabajo de Resnik, que era el trabajar con una muestra pequeña y no aleatoria, intentamos reproducir el experimento con una muestra de sustantivos sensiblemente más grande. Como ya hemos indicado, aprovechamos por un lado la simplificación de nuestra metodología, que no requiere la utilización de software de aprendizaje automático, y por otro lado el avance en materia de hardware de los últimos siete años, más la disponibilidad actual de un corpus de gran tamaño como el EsTenTen.

Como listado de sustantivos a analizar, tomamos 65.000 sustantivos de la taxonomía *open source* ofrecida por Nazar & Renau (2016). A partir de ese listado, obtuvimos un reordenamiento de los sustantivos de ese listado en función de la ponderación que recibieron con nuestra medida $E(x)$ como probablemente eventivos.

Candidato	Evaluación
duración	0
pascuilla	1
bisemana	1
día	1
semanada	1
mesta	1
triduo	1
anaplastia	1
chaquetía	1
ramadán	1
nisán	0
prejornada	1
interescuadra	0
madrigada	0
novenario	1
conmemoración	1
crismal	1
vendimiario	1
...	...

Cuadro 4: Algunos ejemplos de los sustantivos obtenidos y su evaluación

A partir de estos nuevos resultados, examinamos manualmente una muestra y encontramos que, como era de esperar, no todos son eventivos y de los eventivos no todos son no deverbales. La precisión es variable en función de la ponderación que recibieron, pero incluso cuando esta es alta, la tasa de error en la clasificación es mayor que la obtenida en el primer experimento con la lista de 200 unidades tomada de Resnik (2010).

El Cuadro 4 muestra algunos ejemplos de sustantivos en la muestra analizada. En el examen de esta muestra nos limitamos a evaluar la condición de eventivo del sustantivo, suspendiendo por el momento la distinción con el sustantivo deverbal. Esto es porque lo que nos interesa evaluar de momento es el hecho de que el sustantivo ha sido acusado como eventivo por los elementos de su contexto, y no porque hayamos tenido en cuenta aspectos morfológicos. La morfología del castellano ofrece la posibilidad de detectar (y, si cabe, eliminar) los sustantivos deverbales, ya que los morfemas que indican tal condición pertenecen a una categoría cerrada.

Examinando los resultados encontramos casos de nombres eventivos deverbales, como *conmemoración*, casos indiscutibles de nombres eventivos como *ramadán, crimal* o *pascuilla* (cabe destacar la alta contribución de sustantivos eventivos de los diferentes ritos religiosos) y otros que, a pesar de su morfología, como *duración*, no pueden ser considerados sustantivos eventivos. En el caso de este error, la explicación es sencilla: el sustantivo *duración* también aparece acompañado, en gran medida, de aquellos elementos que consideramos predictores en el apartado 3.2. Otros casos son más discutibles. En el caso de *interesescuadra*, su presencia allí se debe a su uso como *torneo interesescuadra*, que es terminología perteneciente al campo del deporte, y que en ese sentido, sí puede ser leído como eventivo. Pero creemos que no podemos considerarlo, en forma aislada, como un sustantivo eventivo, al menos desde un criterio lexicográfico.

Para tener una medida más precisa de la calidad general de los resultados obtenidos en el segundo experimento, la Figura 1 muestra la precisión acumulada. Allí, el eje vertical presenta el porcentaje de precisión y en el horizontal están los candidatos ordenados según la ponderación dada por el algoritmo. Tal como cabía esperar, la precisión disminuye a medida que se consideran más candidatos. La pendiente de ese descenso no es excesivamente acusada, pero permite prever una tasa de disminución del desempeño bastante significativa.

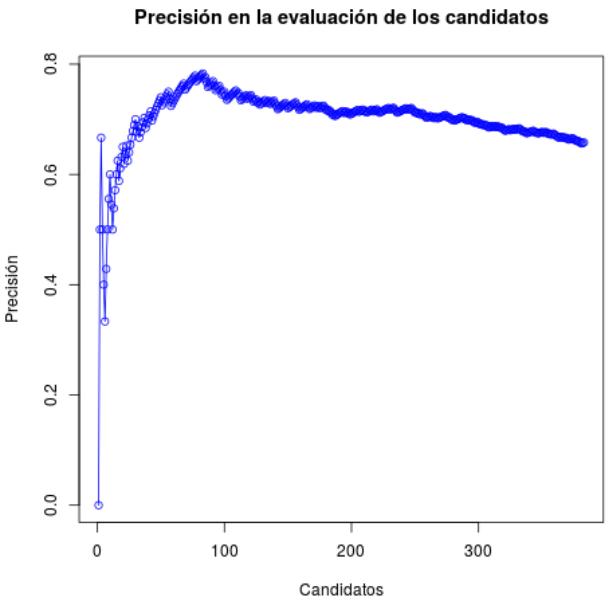

Figura 1: Precisión acumulada en una muestra de 400 candidatos seleccionados como nombres eventivos

Lo primero que salta a la vista en el examen del segundo experimento, y que apunta a explicar la diferencia en desempeño con el anterior, es que los sustantivos son en su mayoría extremadamente raros. Esto no resulta sorprendente ya que es lo que dicta la ley zipfeana de la distribución de frecuencias del vocabulario. La mayor parte del vocabulario estará constituido por

hapax legomena y *dislegomena*, y en muchos casos los autores nos encontramos examinando sustantivos que no conocíamos pero que, en general, se encuentran documentados en al menos un diccionario de la lengua. En el conjunto del Cuadro 2, en cambio, la frecuencia era una variable controlada. Pero justamente por esta diferencia entre el primer y el segundo resultado, creemos que es mejor estimación la del segundo experimento, por resultar más realista.

Otro aspecto relevante a tener en cuenta es que, al examinar la muestra, no solamente nos encontramos con palabras que no conocíamos: también encontramos casos en que no estábamos de acuerdo en si el sustantivo podía o no ofrecer una lectura eventiva. Con el objeto de cuantificar este desacuerdo, tomamos una submuestra de 100 resultados que fueron evaluados por los tres autores. En total, cada uno revisó 200 unidades, pero la mitad de estas unidades eran las mismas en los tres casos, por tanto se revisaron 400 unidades léxicas en total, y utilizamos la intersección de 100 unidades para el cálculo del acuerdo entre anotadores. En el 82 % de los casos existe acuerdo entre los tres anotadores, lo que resulta en un Kappa de 0.526, que se puede considerar un "acuerdo moderado" según Artstein & Poesio (2008). No es infrecuente que exista desacuerdo en materia de clasificaciones lingüísticas, pero al menos confirmamos que tenemos en común una intuición que nos permite reconocer un nombre eventivo en la mayoría de los casos.

5 Conclusiones

En este trabajo hemos propuesto un análisis cuantitativo de corpus –de tipo inductivo y deductivo– para la identificación de las palabras que suelen coocurrir con nombres eventivos no deverbales, con el fin de caracterizar e identificar de forma automática esta clase de sustantivos estudiada por Resnik (2010). Un argumento a favor del enfoque que proponemos en este trabajo es que la metodología es considerablemente más simple y menos costosa desde el punto de vista computacional en comparación con un enfoque basado en aprendizaje automático.

A partir del trabajo realizado, se advierte que la revisión en un corpus de los contextos de aparición de los nombres eventivos ofrece información empírica sobre las relaciones sintácticas que son frecuentes en esta clase de palabras y que, por ende, pueden constituir una fuente confiable para su caracterización e incluso de validación y/o confrontación de las propuestas que se han realizado sobre la base de la introspección.

Reconocemos, sin embargo, las limitaciones de nuestra investigación y encontramos que aún queda mucho trabajo por realizar. Como primera medida, es necesario revisar el algoritmo de clasificación identificando las causas de error en los resultados. Posiblemente no todos los sustantivos puedan ser tratados de la misma manera, y la variable frecuencia es de seguro un factor que debe ser controlado. Otras vías de acción están en la exploración de nuevos elementos predictores y ahondar en mayor complejidad, de ser necesario, en el tratamiento de la información contextual. Otra posibilidad sería incluir también rasgos predictores negativos, es decir rasgos que predicen la condición de no eventivo o nombre simple.

En cualquier caso, creemos que la presente investigación ofrece un precedente más en una línea de trabajo que merece seguir abierta, y de la que existen pocos referentes además de los citados, no ya en castellano sino también en el resto de las lenguas. Esperamos, además, que el presente trabajo resulte un aporte desde el punto de vista metodológico, ya que las herramientas que hemos desarrollado, que son abiertas y de muy sencilla aplicación y uso, pueden alentar a otros investigadores a probar con otros elementos predictores y así mejorar, posiblemente, las tasas de éxito que hemos conseguido hasta ahora. La simplicidad de la propuesta, a su vez, es suficiente motivación como para intentar reproducir los experimentos en otras lenguas, al menos las lenguas europeas, en las que cabría esperar resultados similares.

Por el momento, la utilidad práctica inmediata que tiene para nosotros este trabajo es el de poder enriquecer, por medio de este método, a una taxonomía de sustantivos en castellano que alberga en sí la categoría de "eventos", como es el caso de la ya mencionada taxonomía *open source* de Nazar & Renau (2016).

Agradecimientos

Este proyecto ha sido posible gracias a la financiación de Fondecyt Iniciación (Ref. 11140686), adjudicado por la agencia Conicyt del Gobierno de Chile al primer autor como Investigador Principal. Además, ha recibido también financiación por parte de Conicyt en forma de las becas CONICYT-PCHA/Doctorado Nacional/2016-21160915, concedida a la segunda autora, y CONICYT-PCHA/Doctorado Nacional/2016-21161057, concedida a la tercera autora.

Referencias

Adelstein, Andreína, Marina Berri & Victoria Boschiroli. 2012. Polisemia regular y representación lexicográfica: los nombres locativos en español. *Terminàlia* 5. 33–41.

Alexiadou, Artemis. 2001. *The functional structure in nominals: Nominalization and ergativity*. John Benjamins.

Alonso Ramos, Margarita. 2004. *Las construcciones con verbos de apoyo*. Visor Libros.

Apresjan, Juri. 1974. *Lexical semantics. user's guide to contemporary russian vocabulary*. Karoma Publishers.

Artstein, Ron & Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596.

Baeza-Yates, Ricardo & Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. Addison-Wesley.

Balvet, Antonio, Lucie Barque, Marie-Helene Condette, Pailne Haas, Richard Huyghe, Rafael Marín & Aurélie Merlo. 2011. Nomage: an electronic lexicon of French deverbal nouns based on a semantically annotated corpus. En *International Workshop on Lexical Resources (WoLeR'2011)*, 8–15.

Bel, Núria, Maria Coll & Gabriela Resnik. 2010. Automatic detection of non-deverbal event nouns for quick lexicon production. En *23rd International Conference on Computational Linguistics*, 23–27.

Bosque, Ignacio. 1999. El nombre común. En *Gramática descriptiva de la lengua española*, 3–75. Espasa.

Cabré, M. Teresa, Carme Bach & Jorge Vivaldi. 2006. 10 anys del corpus de l'IULA. barcelona. Informe técnico. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.

De Miguel, Elena. 2006. Tensión y equilibrio semántico entre nombres y verbos: El reparto de la tarea de predicar. En *XXXV Simposio Internacional de la Sociedad Española de Lingüística*, 1289–1313.

Fábregas, Antonio. 2010. Los nombres de evento: clasificación y propiedades en español. *Pragmalingüística* 18. 54–73.

Graña López, Benilde. 1993. La prominencia del argumento externo: el diagnóstico de los nombres eventivos. *Revista española de lingüística aplicada* 9. 85–96.

Grimshaw, Jane. 1990. *Argument structure*. The MIT Press.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann & Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11(1). 10–18.

Kilgarriff, Adam & Irene Renau. 2013. esTenTen, a vast web corpus of Peninsular and American Spanish. En *V International Conference on Corpus Linguistics (CILC2013)*, 12–19.

Nazar, Rogelio & Irene Renau. 2016. A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. En *10th International Conference on Language Resources and Evaluation (LREC'2016)*, .

Picallo, M. Carme. 1999. La estructura del sintagma nominal: las nominalizaciones y otros sustantivos con complementos argumentales. En *Gramática descriptiva de la lengua española*, 363–393. Espasa.

Pustejovsky, James. 1995. *The generative lexicon*. MIT Press.

Quinlan, Ross. 1993. *C45: Programs for machine learning*. Morgan Kaufmann.

Real Academia Española. 2010. *Diccionario de la lengua española*. Espasa-Calpe 22ª ed.

Resnik, Gabriela. 2010. *Los nombres eventivos no deverbales en español*: Universidad Pompeu Fabra. Tesis Doctoral.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Vendler, Zeno. 1967. *Linguistics in philosophy*. Cornell University Press.

Propuesta recibida el 1 de noviembre 2017 y aceptada para publicación el 31 de diciembre 2017.

Extracción automática de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con un sistema basado en patrones lingüísticos

Automatic extraction of analytical definitions and hyponymy-hypernymy relations with a pattern-based system

M. Alejandro Dorantes Cruz
Universidad Nacional Autónoma de México
mdorantescr@iingen.unam.mx

Alejandro Pimentel Alarcón
Universidad Nacional Autónoma de México
apimentala@iingen.unam.mx

Gerardo Sierra Martinez
Universidad Nacional Autónoma de México
gsierram@iingen.unam.mx

Gemma Bel-Enguix
Universidad Nacional Autónoma de México
gbele@iingen.unam.mx

Claudio Molina
Escuela Nacional de Antropología e Historia
claudio.molina.salinas@enah.edu.mx

Resumen

En el presente trabajo se muestra parte de un proyecto en curso centrada en el diseño de un autómata lexicográfico. El objetivo principal de la investigación es la extracción de definiciones analíticas y relaciones semánticas de términos con datos tomados directamente de internet. Presentamos dos de las capacidades del sistema: la extracción de definiciones analíticas y de hiperónimos. La metodología consiste principalmente en la búsqueda automática de esta información con patrones construidos manualmente basados en la estructura léxica de definiciones analíticas en lenguaje natural.

Con este desarrollo, ha sido posible mejorar la precisión reportada en el estado del arte. Se ha conseguido una precisión de 92.5 % para la tarea de extracción de definiciones analíticas y de las relaciones de hiperonimia.

Palabras clave

extracción automática de definiciones, hipónimia-hiperonimia, patrones lingüísticos

Abstract

This work is part of an ongoing project that is focused on the design of a lexicographic automaton. The main objective of the research is the extraction of analytical definitions and semantic relations of terms with data taken exclusively from internet. We present two of the abilities of the system: the extraction of a) analytical definitions and b) hypernyms. The methodology consists of the automatic search of that information with manually-built patterns based on the lexical structure of analytic definitions in natural language.

This method has improved the precision reported in the state of the art. Se have reached a precision of 92.5 % for the extraction of analytical definitions and for the hypernymy relations.

Keywords

automatic extraction of definitions, hypernymy-hyponymy, linguistic patterns

1 Introducción

La generación de definiciones lexicográficas es un área de poco o nulo desarrollo dentro del procesamiento del lenguaje natural (PLN). En cambio, el crecimiento exponencial de los nuevos conceptos en ciencia y la tecnología, junto con la especialización del conocimiento, hacen de los diccionarios elementos cruciales para todos aquellos profesionales que, en un momento dado, se alejan de su campo habitual de trabajo.

En algunos ámbitos, existe la idea de que internet hace innecesaria la existencia de diccionarios. En efecto, es de destacar los beneficios de tiene la web como corpus, ya que se encuentran gran número de datos provenientes de múltiples fuentes. Pero en realidad, aunque la web ofrece una gran cantidad de información, no es fácil encontrar herramientas que la estructuren, y la conviertan en conocimiento especializado.

DOI: 10.21814/lm.9.2.257

Al mismo tiempo, si bien existen otros sistemas como wordnet, o enciclopedias, en algunas ocasiones estos se encuentran desactualizados si se comparan con la información que día a día se registra en internet.

Por todo ello, encontrar formas automáticas para extraer y procesar desde la www los elementos constituyentes de una definición y conectarlos se ha convertido en un desafío para la lingüística computacional actual.

La lexicografía aún no ha conseguido un método para automatizar la creación de definiciones. Para lograr esta meta se requiere una estructura base modelada con patrones, nutrida con información suficiente y estadísticamente pertinente.

En este artículo se sostiene que es posible generar definiciones analíticas de forma automatizada partiendo de un conjunto de candidatos a definiciones, en su mayoría aceptables, y de la identificación de sus hiperónimos.

Existen distintas técnicas para extraer candidatos a definiciones (Pearson, 1998; Meyer, 2001; Alarcón, 2009) e hiperónimos (Hearst, 1992; Ortega, 2007). En ambos casos los sistemas existentes tienen un recall aceptable, pero su precisión es muy baja, o insuficiente para asegurar el éxito en la generación de definiciones. Por ello, esta propuesta se enfoca a mejorar la precisión, aunque el recall pudiera verse afectado.

El artículo tiene la siguiente estructura: en la Sección 2, se describe una tipología de las definiciones en lexicografía computacional, prestando particular atención al tipo de definición analítica, que es la que interesa en esta investigación; más tarde (Sección 3), se señalan los antecedentes en la literatura de la extracción de definiciones y la extracción de relaciones léxicas de hiperonimia-hiponimia, sea en trabajos teóricos previos o en investigaciones aplicadas puntuales; en la Sección 4 se describe la arquitectura y metodología seguida para la extracción de definiciones analíticas y relaciones semánticas de hiperonimia-hiponimia en estos mismos contextos; por último (Secciones 5 y 6), se ofrecen algunos resultados, se presenta una evaluación y se dan algunas conclusiones generales, así como algunas líneas futuras de investigación.

2 Definiciones en lexicología computacional

En este apartado se hace un repaso de cuáles son los elementos constitutivos de una definición y se introduce una tipología estas. Posteriormente, se presenta la noción de contexto definitorio (CD) y de patrón definitorio (PD).

Tipología de las definiciones

Aristóteles (según señala Smith (2007)) indica que una definición consta de dos partes: el genus, conocido también como *kind* o *family*, mismo que indica qué tipo de cosa es el definendum (elemento que está siendo definido), y la *differentia*, que especifica o hace único al *definendum*. De estos dos elementos (*genus* y *differentia*) se puede asumir que el genus sea también un término más general o hiperónimo, mientras que la *differentia* es mayormente una predicación en la que se enumeran las características diferenciadoras o propias del término.

Con base en las implicaciones de la propuesta aristotélica relacionada con esta particularidad, se ofrece en la literatura al respecto (Sierra et al., 2008; Aguilar, 2009) una clasificación en cuatro tipos de definiciones: analíticas, sinonímicas, funcionales y extensionales.

Las *definiciones* analíticas son el tipo más prototípico considerando el modelo aristotélico que se ha descrito hasta ahora. Según lo explicado por Aguilar (2009), una definición de tipo analítico aporta un conocimiento inherente al término definido, aunque también suelen aportar características no esenciales o características adquiridas accidentalmente.

La *definición sinonímica* no ofrece diferencia específica, sino únicamente género próximo (Dorantes, 2016). Por ejemplo, para la entrada "Estado financiero" se propone la definición "Estado de situación financiera"; por su parte, "Impuesto" se define como "Gravamen, arancel".

Una *definición funcional* ofrece, por medio de la diferencia específica, una eplicación que aclara la función, utilidad o fin de lo referido por la entrada (Aguilar, 2009). Por ejemplo: "Estado financiero" se explica como algo que "Sirve para calcular la utilidad o pérdida neta que generará el proyecto hecho a los estados de resultados" y "Servicio financiero" es algo que "Sirve para mejorar la calidad de vida y el desarrollo de los hogares".

Las *definiciones extensionales*, por su parte, enumeran las partes o componentes que forman al término definido, por ejemplo: "Estado financiero" / "Se compone de la cuenta de resultados y el balance"; "Impuesto" / "Se compone del objeto, el sujeto, la base y la tasa o la tarifa".

Nuestro sistema se basa en las definiciones analíticas porque son el tipo más arquetípico dentro de lexicografía. Como afirma Lara (1997), "la mayor parte de la definición lexicográfica y enciclopédica contemporánea (...) se rige en mayor o menor grado por la teoría aristotélica".

Es importante recordar que una definición analítica proporciona dos tipos de conocimiento:

- *Genus*: este conocimiento indica a qué clase o grupo pertenece el término de entrada por medio de un término más general o hiperónimo.
- *Differentia*: este conocimiento especifica qué hace que el término de entrada sea único del resto de los elementos en su clase de pertenencia.

Las definiciones analíticas resultan muy interesantes porque "género y diferencia se convierten en condiciones necesarias y suficientes para (el) reconocimiento de todo objeto" (Lara, 1997, pg. 208).

Contexto definitorio y patrón definitorio

Autores como Alarcón et al. (2008) señalan que un contexto definitorio (CD) es todo fragmento de tamaño indeterminado dentro de un documento en donde se describe clara y precisamente la definición de un término. Estos autores afirman que los CDs están formados por un término y una definición que se encuentran relacionados entre sí por sintagmas como "se define" o "se entiende como" entre otros, también conocidos como patrones definitorios (PDs).

Por un lado, el término es uno de los elementos constitutivos, no accesorio, del contexto definitorio y es el único elemento sobre el cual se introduce información relevante en el contexto (Alarcón et al., 2008); mientras que la definición es un elemento constitutivo del CD que contiene la información relevante que se aporta sobre el término. Esta definición constituye una explicación del término (Sager, 1993, pg. 67).

El sistema que presentamos incorpora el *nexus differentia* (ND) (Dorantes, 2016), una expansión de los patrones de CDs cuya ventaja es la división de una definición analítica en su término genérico (*genus*) y diferencia específica (*differentia*). Este elemento, según Dorantes (2016), es esencial para la estructura de las definiciones analíticas en español porque contiene una regularidad basada en la revisión de diccionarios, así como una gran cantidad de candidatos a definiciones. El autor señala que, aunque la *differentia* ya se había propuesto como un elemento de la definición analítica, no hay trabajos que muestren las características lingüísticas de la partícula o la forma en que dicha partícula introduce la diferencia específica.

Por último, conviene referir que un patrón definitorio (PD) es un elemento lingüístico que relaciona al término y a su definición, dándole a la predicación existente entre ellos una orientación de tipo sinonímica. En la literatura, se identifica que los PD en español podrían conformarse por verbos que, siguiendo a Rodríguez (1999), se denominan verbos metalingüísticos (definir, denominar, describir...), aunque autores como Alarcón (2006) señalan que es posible que verbos con una semántica distinta también funcionan como PD (ser, conocer o identificar) y cambien su comportamiento argumental a una predicación metalingüística.

Los PD, dependiendo del tipo de complemento que requieran, podrían determinar el tipo de orientación de la definición según la clasificación que ya se ha explicado anteriormente.

Un patrón verbal analítico establece una relación de predicación entre el término y su definición, en la que la segunda remite o describe características inherentes o adquiridas del primero. Según se desprende del estudio de Aguilar (2009), los verbos que orientan este tipo de definiciones son "referir", "representar", "significar" y "ser".

Todo lo anterior constituye un marco de referencia que contiene los elementos teóricos constituyentes de las definiciones analíticas. Sin embargo, es indispensable que inmediatamente se traten algunos aspectos complementarios relacionados con la extracción de definiciones y a la determinación de los *genus* de estas. Estos se desarrollan a continuación.

3 Trabajos previos

Desde la perspectiva que se ha planteado para la presente investigación, conviene referir algunos antecedentes tanto para la extracción de definiciones como para la extracción de relaciones de hiponimia-hiperonimia, mismas que a continuación se explican.

Trabajos sobre extracción de definiciones

En el estudio de la extracción automática de definiciones se ha trabajado desde distintas perspectivas, por ejemplo, uno de los primeros estudios en el área fue el de Pearson (1998) en el que presenta el comportamiento de los contextos en los que aparece un término que se describe. Pearson afirma que cuando un autor define un término, comúnmente se utilizan patrones que llaman la atención hacia la presencia de un elemento importante sobre el que se está trabajando y dando una definición. El autor identifica, además, la aparición de patrones léxicos que conectan las definiciones con los términos.

Por otra parte, Meyer (2001) refuerza estas ideas y encuentra que los patrones definitorios también proveen elementos clave para la identificación del tipo de definiciones aplicadas a los términos. Así, la principal motivación para trabajar con contextos definitorios surge de la necesidad de obtener conocimiento semántico de los términos que aparecen dentro de diferentes áreas de especialidad.

Uno de los proyectos mexicanos que trabaja en la extracción de contextos definitorios es el corpus CORCODE en español (Sierra et al., 2006), de uso público a través de internet.[1]

Este enfoque supone la creación de metodologías para la extracción automática de definiciones, por ejemplo, Klavans & Muresan (2001) se enfocan específicamente en métodos para la extracción de términos y definiciones en textos médicos con su sistema Definder. Este sistema basado en reglas trabaja de manera excepcional; al enfocarse en la búsqueda de terminología muy especializada, se reporta una precisión del 87 %.

Por otra parte, Espinosa-Anke et al. (2016) desarrollan DefExt, una herramienta capaz de extraer definiciones a partir de un corpus utilizando un enfoque semi-supervisado. En esta misma línea, el sistema utiliza etiquetado de partes de la oración y un ordenamiento de importancia de los documentos de un corpus; los autores reportan una precisión general de 50 %.

Adicionalmente, han surgido también buscadores que trabajan utilizando el internet en lugar de un corpus estático: GlossExtractor, desarrollado por Velardi et al. (2008) recupera información de la Web, este último se enfoca en glosarios y documentos especializados dentro de internet, a partir de los cuales extrae las definiciones de un listado de términos predefinidos. Este sistema está basado en inglés, utilizan un enfoque de aprendizaje por computadora sobre diccionarios y etiquetado automático de partes de la oración; reportan una precisión del 73.5 % sobre las definiciones extraídas de internet.

Un trabajo importante que se ha desarrollado en México es el buscador ECODE (Alarcón et al., 2008), sistema capaz de extraer contextos definitorios a partir de búsquedas en la red. El enfoque que se utiliza en ECODE consiste en la separación de la tarea en dos módulos: el primero se encarga de las búsquedas en línea, se utilizan 15 patrones rodeando a un término en una búsqueda textual exacta y el segundo módulo filtra resultados que no contienen definiciones mediante el uso de árboles de decisión y etiquetas sintácticas. En este trabajo se presentan de forma separada los estudios y resultados que se hicieron y obtuvieron para cada uno de los tipos de las definiciones. Para el caso de las definiciones analíticas, obtienen una precisión del 58 % y un recall del 83 %.

[1] En `http://www.corpus.unam.mx/corcode`.

Trabajos sobre extracción de hiperónimos

Las relaciones de hiponimia-hiperonimia son un tipo de relación léxico-semántica que es de vital importancia cuando se quiere estructurar construcciones lingüísticas como la definición analítica.

La extracción automática de relaciones semánticas es un tema clásico en lexicografía computacional, que se ha abordado principalmente utilizando diferentes enfoques basados en:

Diccionarios: las técnicas apoyadas por diccionarios (Calzolari, 1984; Alshawi, 1987; Richardson et al., 1993) son óptimas para descubrir relaciones hiponimia-hiperonimia y son capaces de alcanzar una gran precisión. Pero necesitan textos estructurados, que no siempre están disponibles. En algunos trabajos como el de Calzolari (1984), se llega a obtener una precisión de hasta el 90 %. En cambio, tienen la desventaja de que no se toman en cuenta términos específicos de un dominio, pues los diccionarios llegan a abarcar varios dominios.

En la actualidad existen otras herramientas que pueden presentar información de forma similar a un diccionario clásico. Por ejemplo Wikcionario, Wikipedia o Wordnet. Pero algunos de ellos no ofrecen la información como la que se quiere extraer. Wordnet sí contiene información estructurada, aunque el uso de la versión en español tiene muchos inconvenientes. Además, estos recursos y otros de parecidas características, no se actualizan con la asiduidad que permitiría un uso fiable para estos objetivos.

Agrupamiento (o co-ocurrencias): los métodos desarrollados para el enfoque de agrupamiento (Pereira et al., 1993; Riloff & Shepherd, 1997; Caraballo, 1999; Cimiano et al., 2004; Widdows, 2003; Mititelu Barbu, 2006) son capaces de encontrar hiperónimos incluso cuando no están explícitamente en el corpus de búsqueda. Sin embargo, necesitan textos muy amplios para obtener buenos resultados. Una de las ventajas que ofrece el método de agrupamiento radica en que es una alternativa que permite encontrar este tipo de relaciones semánticas, aun cuando no están explícitamente

en el corpus de búsqueda; dentro de sus desventajas está que no ofrece buenos resultados con textos pequeños (Ortega, 2007).

Patrones: las técnicas basadas en patrones léxicos simples manuales fueron desarrolladas por primera vez por Hearst (1992). Muchos autores han seguido este modelo (Ravichandran et al., 2004; Cederberg & Widdows, 2003; Pantel & Ravichandran, 2004; Oakes, 2005). Varios trabajos integran aprendizaje de máquina para optimizar los sistemas (Snow et al., 2004; Pasca, 2004; Pantel & Pennacchiotti, 2006; Pantel & Ravichandran, 2004; Bunescu & Mooney, 2007). Una de las principales desventajas de este método es la necesidad de corpus muy grandes para poder reportar resultados del 85 % en precisión.

Aunque la mayoría de la literatura está relacionada con el inglés, existen propuestas en varios idiomas que siguen métodos similares.

En español, Acosta et al. (2010) utiliza la teoría de prototipos y el aprendizaje de máquina para extraer las relaciones de un corpus, y luego calcula los pares hipónimo /hiperónimo. El trabajo de Ortega (2007) también se encuentra en el marco de los enfoques basados en patrones.

4 Metodología

En este apartado se explica cómo se llevan a cabo las tareas de extracción de definiciones y la obtención de las relaciones semánticas de hiponimia-hiperonimia. Creemos pertinente destacar que, con la intención de que el procesamiento fuera más ligero, el sistema no lleva a cabo ningún tipo de etiquetado gramatical, morfológico o sintáctico.

Arquitectura del sistema

El sistema presentado trabaja en tres etapas: en la primera se hace una serie de búsquedas con la finalidad de extraer de la web candidatos a contextos definitorios; en la segunda, y una vez que se tienen dichos contextos definitorios, se refina la búsqueda utilizando un elemento que forma parte constituyente de la definición analítica, el nexus differentia (mismo que se describe a continuación); por último, una vez que se han extraído candidatos a definiciones más precisas, se hace posible que de ellas se puedan obtener hiperónimos, pues estos quedan delimitados entre los patrones definitorios y los patrones diferenciales. A continuación se esquematizan a detalle las tres etapas anteriores.

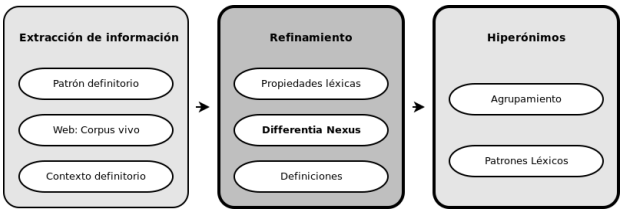

Figura 1: Etapas de la extracción de definiciones.

En la Figura 1 se describen gráficamente las etapas de este desarrollo. Los pormenores de cada una de ellas se describirán detalladamente en los apartados siguientes (4.2 Recuperación de información, 4.3 Extracción de la definición y *nexus differentia*, y 4.4 Extracción de hiperónimos).

Recuperación de información

La meta de esta etapa es la extracción de contextos definitorios de la web. En este estadio se genera un corpus de candidatos a definiciones, con base en algunos patrones definitorios del español (ser, definir y concebir). Este corpus tiene poca precisión, pero un alto recall, lo cual es perfecto para nuestro sistema, pues a mayor información mejores resultados.

Cabe destacar que los verbos utilizados tienen la propiedad de extraer el tipo de definiciones analíticas. A diferencia de otros métodos que únicamente hacen búsquedas en Google All, este sistema también hace búsquedas en Google Scholar y Google Books para que los datos obtenidos puedan servir como el respaldo de un saber especializado.

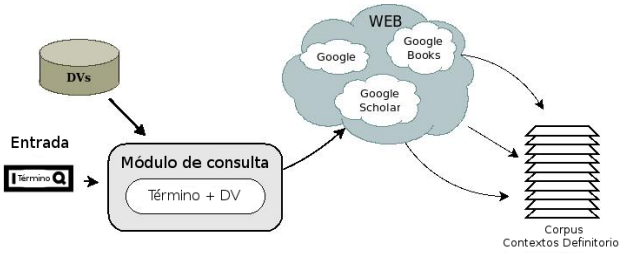

Figura 2: Extracción de Información.

En la Figura 2 pueden verse los pasos que sigue nuestro desarrollo para la extracción de contextos definitorios. Se ilustra al término siendo añadido a todos los verbos definitorios (DVs) considerados para ser buscados en cada uno de los módulos de Google. Los DVs se encuentran siempre en tercera persona del singular del presente de indicativo. La búsqueda genera un corpus de contextos definitorios que es utilizado en la siguiente etapa.

Extracción de la definición y nexus differentia

En esta etapa se procesa la salida obtenida en el paso anterior y se suma uno de los *nexus differentia*, así como algunas características léxicas. El término *nexus differentia*, introducido por Dorantes (2016), consiste en el "pronombre relativo simple", que se ha observado ser esencial en la estructura de definiciones analíticas en español, ya que en estas existe una regularidad comprobada con base en la revisión de diccionarios, así como de una gran cantidad de candidatos a definiciones. El único "pronombre relativo simple" que se ha considerado en este trabajo es "que", ya que es el más generalizado. En posteriores etapas del trabajo, se piensa incluir otros relativos, como "el/la cual", aunque su uso en las definiciones es mucho más bajo que el del genérico "que".

En la mecánica de construcción de los patrones se siguen las siguientes reglas:

- El término a buscar debe aparecer en conjunto con alguno de los verbos definitorios, así como con un artículo que lo identifique como un sustantivo.
- No puede aparecer una palabra funcional previo a la detección del patrón antes descrito.
- Se filtran aquellas protodefiniciones que comienzan o terminan con palabras funcionales.
- Si una protodefinición contiene el mismo término que se quiere definir se descarta como definición.
- Por último, el sistema se asegura de que contenga el *nexus differentia* en una posición adecuada, es decir, después del término que se busca definir.

Existen algunas particularidades en la forma como el español caracteriza un *genus* respecto de su *differentia*. Si bien la *differentia* se había planteado anteriormente como un elemento de la definición analítica, no se habían hecho trabajos que mostraran las características lingüísticas de dicha partícula ni tampoco la forma en que dicha partícula introduce la diferencia específica.

Por su parte, entre las características léxicas podemos encontrar aquellas que explican tanto la forma como la medida estándar de un término. Suponemos que los términos se encuentran también circunscritos por uno o más elementos de una lista cerrada de palabras que reducen la posibilidad de aportar un hiperónimo al aparecer al inicio y/o al final de un contexto. Dichos elementos se agregan a una búsqueda en la que se fuerza su aparición, con lo que se consigue una definición que contiene un genus con su *differentia*. Lo anterior logra un aumento de la precisión en la extracción de definiciones. El proceso que sigue el desarrollo se ilustra en la Figura 3. La entrada del sistema es un corpus de contextos definitorios y la salida son las definiciones analíticas y, en una etapa intermedia, la suma de las características léxicas y el nexus differentia que permiten hacer el filtrado y ofrecer definiciones analíticas precisas.

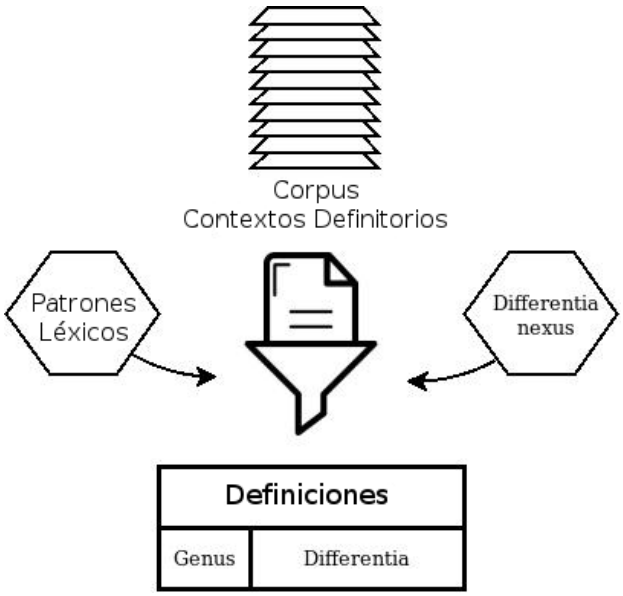

Figura 3: Refinamiento del corpus.

El Cuadro 1 puede servir para ilustrar qué es un contexto definitorio, la estructura de los patrones. Además, ejemplifica una definición analítica que sigue la formación del patrón.

Extracción de hiperónimos

El sistema de extracción de hiperónimos trabaja en dos etapas: en la primera se extraen todos los genus y en la segunda se agrupan por frecuencia, ofreciéndonos así los candidatos a hiperónimos con su respectiva frecuencia de aparición.

A continuación se clarifican ambas etapas: en la primera se extraen todos aquellos elementos que se encuentran circunscritos tanto por los DVs como por el ND; mientras que en la segunda se agrupan todos los hiperónimos que son completamente iguales con la intención de mostrar qué hiperónimo podría ser el más pertinente para cada término. Tal como se puede ver en la Figura 4.

Al final de cada uno de los procesos es posible obtener definiciones precisas y candidatos a hiperónimos. Dichos resultados son posibles gracias

Lexicografía	Macroestructura						Microestructura						
	Entrada, lema, definido, definendum o		Verbo definitorio				Definición, definiens, expresión explicativa, descripción, explicación, exposición, explanación o declaración						
	entidad léxica				GENUS		DIFFERENTIA						
Artículo lexicográfico	Cajero automático				Máquina		que permite retirar y depositar dinero en efectivo a los clientes de un banco en casi cualquier parte del mundo a cualquier hora del día						
Contexto definitorio	El	cajero automático	es	la	máquina	que	permite	retirar y depositar dinero en efectivo	a los clientes de un banco	en case cualquier parte del mundo	a cualquier hora del dia		
Sintaxis	Art. 1	Sintagma nominal	Verbo 1	Art. 2	Sintagma nominal	Pron. Rel. Nexus differentia	Verbo 2	Sintagma nominal	Sintagma prep.	Sintagma prep.	Sintagma prep.		
	M.D.	Núcleo Sujeto	Núcleo Pred		Atributo	M.D.	Núcleo Pred.	CD	CI	CCL	CCT		
	Oración principal					Oración subordinada relativa especificativa							
	Oración compuesta												

Cuadro 1: Estructura de una definición y del contexto definitorio de donde se extrae.

a la suma del *nexus differentia* que nos permite elevar la precisión en la tarea de la extracción de definiciones.

La extracción de hiperónimos, al igual que la de definiciones analíticas, es automática tomando en cuenta los resultados arrojados en el primer proceso (la extracción de definiciones). La tarea es sencilla, pues lo único que se hace es asociar el género del término definido (Figura 4).

A continuación se muestran los resultados y la evaluación de nuestro método con la intención de esclarecer la mejora propuesta.

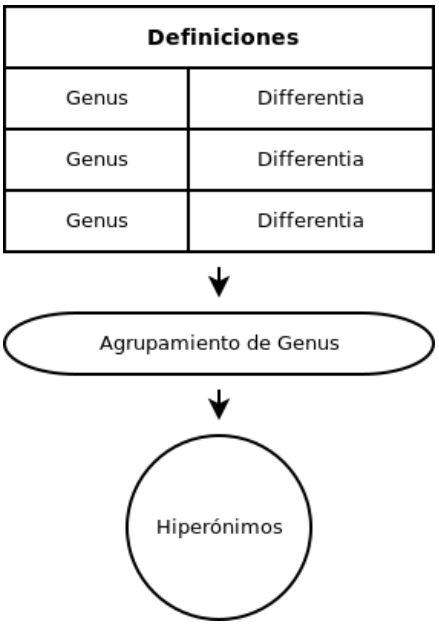

Figura 4: Extracción de hiperónimos.

5 Resultados

Los resultados de la extracción de las definiciones analíticas e hiperónimos se evalúan cuantitativamente y cualitativamente, respecto a otros métodos ya descritos anteriormente.

Resultados cuantitativos de la extracción de definiciones

Tras correr las pruebas de extracción de definiciones y compararlas con un conjunto de definiciones analíticas obtenidas del CORCODE, ésta herramienta nos ofreció 985 definiciones analíticas dentro de un total de 1426 contextos definitorios recopilados.

El Cuadro 2 muestra la matriz de confusión que se presenta en el sistema; por un lado, la columna de datos de la izquierda muestra cantidad de resultados que el sistema extrajo como definiciones; por otra parte, en la derecha de la tabla

se muestra la cantidad de contextos descartados por el sistema al buscar definiciones. Al final, la primera fila de datos corresponde a la cantidad de contextos que efectivamente tienen definiciones, mientras que la segunda muestra la cantidad de los contextos que no presentan una definición.

Real / Sistema	Positivo	Negativo
Verdadero	147	838
Falso	12	429

Cuadro 2: Matriz de confusión que se presenta en el sistema.

Como se puede ver, se extrajeron 159 definiciones, de las cuales 147 fueron correctamente catalogadas, es decir, hay una gran cantidad de contextos que se clasifican correctamente como definiciones dentro del total de contextos catalogados por el sistema; sin embargo, la relación entre las definiciones obtenidas, del total de definiciones que se encontraban presentes, es muy baja. Por lo que se logra una precisión del 92.5 % y una exactitud del 40.5 %. Dicho resultado se obtiene de dividir las definiciones analíticas correctas entre las definiciones analíticas obtenidas (frecuencia relativa).

En este caso los datos podrían parecer desalentadores si lo que se busca no fuera la precisión, pero como creemos que este es un elemento importante para la generación de definiciones, este dato es el que consideramos que tiene mayor peso para desarrollos futuros.

Se hace pertinente contrastarlos con los ya existentes en el método que se pretende mejorar. De esta manera, podemos observar que en ECODE la precisión es de 58 %, mientras que la de nuestro desarrollo es de 92.5 %. Lo anterior nos permite decir que nuestro desarrollo mejora en un 32.5 % la precisión de ECODE.

En general, estos datos son buenos con miras a la generación automática de definiciones, pues ofrecen resultados confiables y más precisos.

Resultados cualitativos de la extracción de definiciones

Con la intención de mostrar también los resultados cualitativos, es decir, algunas de las definiciones analíticas ofrecidas por el desarrollo, a continuación se presenta el Cuadro 3 donde se pueden ver cinco términos con las definiciones que ofrece el sistema propuesto.

En el Cuadro 3 se puede apreciar la pertinencia de las definiciones analíticas propuestas. Lo cual se debe a que todas cuentan con el ND del cual se ha hablado anteriormente.

Término	**Definiciones ofrecidas por nuestro desarrollo**
Factor de activación de la transcripción	Proteína celular que en principio fue identificada como un factor estimulador de la transcripción de la unidad de transcripción e4 de adenovirus, la cual se activa en la fase inicial de la infección.
Servidor	Programa que se ejecuta en un terminal remoto y trabaja conjuntamente con el cliente.
Turbocompresor	Dispositivo de sobrealimentación que produce el ingreso del aire a una presión por encima de la atmosférica.
Relevador	Dispositivo que provoca un cambio brusco en uno o más circuitos eléctricos de control, cuando la cantidad o cantidades medidas a las cuales responde cambian de una manera predeterminada.
Entorno del Servidor de Internet de SGI (ISE)	Solución muy efectiva que incluye herramientas avanzadas de administración, monitoreo y seguridad además de programas integrados de instalación y una interfaz basada en la Web.

Cuadro 3: Resultados cualitativos de nuestro sistema.

Resultados cualitativos de la extracción de hiperónimos

Para esta sección, se han buscado ocho términos y han sido divididos en dos grupos según si los términos pertenecían a las ciencias de la salud (Biología, Cerebro, Hepatitis y Oncología) o no (Física, Matemáticas, Ontología, Sintaxis). La elección de estos términos ha venido determinada por los recursos con los que se contaba para evaluar. Así, se han elegido campos en los que existían ontologías, y en último caso se ha recurrido a wordnet, versión 3.0.

Para el primer grupo se ha utilizado la base de datos de Descriptores en Ciencias de la Salud (DeCS), cuyos conceptos se organizan con una estructura jerárquica para permitirle usarlo como una taxonomía (Cuadro 4). Mientras que para la otra mitad de los términos se ha utilizado el tesauro de la UNESCO,[2] que es una lista controlada y estructurada de términos utilizados para el análisis de temas y la búsqueda de documentos en los campos de educación, cultura, ciencias naturales, ciencias sociales y comunicación (Cuadro 5).

[2] http://vocabularies.unesco.org/browser/thesaurus/es/

Término	**Nuestro sistema**	**WordNet**	**DeCS**	**Humano**
Biología	Ciencia Rama de las ciencias naturales	Ciencia de la vida Bio-ciencia Vida Colección Agregación Acumulación Montaje	Ciencia Disciplina de ciencia Biológica	Ciencia Ciencia exacta
Cerebro	Órgano	Estructura neuronal Inteligencia Cognición Conocimiento Noesis Intelecto Variedad de carne Órgano	Telencéfalo	Órgano Cuerpo Anatomía Aparato nervioso central Cabeza
Hepatitis	Inflamación de hígado Enfermedad inflamatoria Enfermedad	Enfermedad infecciosa Enfermedad del hígado	Hepatopatía Virosis	Enfermedad ETS
Oncología	Especialidad médica Especialidad Rama de la medicina Ciencia	Medicina Especialidad médica	Medicina interna	Medicina Cáncer Disciplina Estudio

Cuadro 4: Para los términos de ciencias de la salud, se muestran los hiperónimos de la base de datos de DeCS.

Término	**Nuestro Sistema**	**WordNet**	**UNESCO**	**Humano**
Física	Ciencia	Ciencia Natural	Ciencia Ciencias Físicas	Ciencia Ciencia exacta
Matemáticas	Ciencia	Ciencia Disciplina científica	Ciencia Matemáticas y estadística	Ciencia Ciencia exacta Números
Ontología	Rama de la metafísica de la filosofía Ciencia de las esencias Base de datos	Disposición Organización Sistema Metafísica	Metafísica	Filosofia Estudio Metafísica Rama de la filosofía
Sintaxis	Parte de la gramática Conjunto de reglas	Estructura Sistema Esquema Gramática	Gramática	Lingüística Nivel de lengua Orden Lengua

Cuadro 5: Tablas comparativas para los resultados dela evaluación del sistema. Ambas tablas incluyen los hiperónimos obtenidos de WordNet y de la anotación humana.

Como se puede ver, entre los términos ofrecidos por nuestro sistema y los ofrecidos por otros recursos, existe una semejanza y, en algunos casos como el de Hepatitis, una aportación que podría resultar pertinente a la hora de tratar de entender a la Hepatitis no sólo como enfermedad, sino también como una inflamación del hígado. Cabe destacar que ninguno de los conceptos cae fuera del área de especialidad del que proviene el término y la mayoría de ellos es el mismo o es similar al resultado consultado o al dado por (DeCS), WordNet, UNESCO o humanos. Adicionalmente, se ha preguntando a colegas universitarios por los hiperónimos relevantes de los términos dados, así como una comparación manual entre diferentes recursos para estimar la precisión de nuestro sistema.

La evaluación elaborada mediante una metodología cualitativa arroja los resultados de el Cuadro 6.

Precisión	Exactitud	Exhaustividad	F-Score
92.5 %	40.4 %	14.9 %	25.7 %

Cuadro 6: Resultados de la evaluación cualitativa.

La precisión nos indica un factor de pertinencia. Consiste en encontrar la proporción que existe entre los resultados que se extrajeron como definiciones y los que realmente son definiciones.

Por su parte, la exhaustividad se refiere a la relación de la cantidad de definiciones que fueron extraídas contra todas las definiciones que pudieron haber sido obtenidas del corpus.

La exactitud es una medida que obtiene la proporción de oraciones clasificadas correctamente, es decir la unión de verdaderos/positivos y falso/negativo en proporción con el total de elementos en la evaluación.

Por último, el F-score es un valor único ponderado de la precisión y la exhaustividad. Se trata de una media armónica que combina ambos valores.

Resultados cuantitativos de la extracción de definiciones

Como ya se dijo, se aplicó una encuesta a un grupo de 170 voluntarios para recuperar la aceptación que el estándar humano podría dar a cada uno de los candidatos a hiperónimos dados por el sistema. Ellos pertenecen a los últimos semestres de los estudios de Lengua y Literatura Hispánicas de la UNAM. Se recurrió a ellos tomando en cuenta todo su conocimiento lingüístico adquirido.

En la encuesta se presentó a los voluntarios el par de candidatos a término con la instrucción de calificar el candidato como un hiperónimo correcto o incorrecto para el término. Una tercera opción fue dada para el caso donde el voluntario no podría decir la categoría por sus propios medios.

Como se puede observar en el Cuadro 7, la aprobación de los humanos es muy satisfactoria: pasan el radio del 75 % llegando hasta el 96 %. Esto permite afirmar que, si bien aun se pueden mejorar los resultados, se va por buen camino hacia una mejora sustancial en esta tarea.

Término	Radio
Biologia	82.1 %
Cerebro	88.3 %
Física	96.3 %
Matemáticas	84.0 %
Sintaxis	75.1 %
Oncologia	84.3 %
Ontologia	63.6 %
Hepatitis	95.7 %
Evaluación del sistema: 83.67 %	

Cuadro 7: Aprobación desde la perspectiva humana de un set de candidatos a hiperónimos dados por el sistema.

6 Conclusiones y trabajo futuro

El método propuesto ha mejorado la precisión en la extracción de definiciones y ha mostrado candidatos a hiperónimos que, aunque no han sido evaluados, parecen pertinentes para los términos de entrada. A diferencia de los patrones utilizados en contextos definitorios, lo que fuerza la salida tanto de una definición como de un hiperónimo es el *nexus differentia* que nos permite recuperar definiciones como las ya mostradas anteriormente.

Los resultados arrojados por el sistema apuntan hacia que los *nexus differentia* son un factor decisivo para el mejoramiento de la precisión en la extracción de definiciones analíticas e hiperónimos, por ello, como trabajo futuro se plantea explorar el resto de *nexus differentia* del español, con la finalidad de aumentar el recall de nuestro sistema.

En cuanto a los hiperónimos, asumimos que los resultados son aceptables. Aunque los términos resultantes no siempre pertenecen al mismo nivel semántico, están relacionados. Además, los cambios se pueden atribuir a diferentes niveles de especialización entre las fuentes, es decir, cuanto más especializada sea una fuente, más "cercano" será el hiperónimo dado para una palabra. Por ejemplo, una fuente no especializada puede dar como hiperónimo de perro: "animal", una fuente más especializada puede recuperar "canino", y una fuente aún más especializada podría devolver "canis lupus", todos los cuales son hiperónimos correctos.

Dicho todo lo anterior, creemos que en futuros trabajos será posible comenzar a generar definiciones que sean adecuadas para cada término solicitado y que al mismo tiempo nos permitan estructurar el conocimiento.

Agradecimientos

Este artículo ha sido elaborado gracias a los proyectos PAPIIT IA400117 y Fronteras de la Ciencia 2016-01-2225.

Referencias

Acosta, Olga, César Aguilar & Gerardo Sierra. 2010. A method for extracting hyponymy-hypernymy relations from specialized corpora using genus terms. En *Workshop in Natural Language Processing and Web-based Technologies*, 1–10.

Aguilar, Cesar. 2009. *Análisis lingüístico de definiciones en contextos definitorios*: Universidad Nacional Autónoma de México. Tesis Doctoral.

Alarcón, Rodrigo. 2006. Extracción automática de contextos definitorios. propuesta para el desarrollo de un eccode (extractor de candidatos a contextos definitorios). Proyecto de tesis de doctorado. Universitat Pompeu Fabra.

Alarcón, Rodrigo. 2009. *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*: Universitat Pompeu Fabra. Tesis Doctoral.

Alarcón, Rodrigo, Carme Bach & Gerardo Sierra. 2008. Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica. *Revista Española de Lingüística* 37. 247–278.

Alarcón, Rodrigo, Gerardo Sierra & Carme Bach. 2008. ECODE: a pattern based approach for definitional knowledge extraction. En *XIII EURALEX International Congress*, 923–928.

Alshawi, Hiyan. 1987. Processing dictionary definitions with phrasal pattern hierarchies. *Computational Linguistics* 13(3–4). 195–202.

Bunescu, Razvan C. & Raymond J. Mooney. 2007. Learning to extract relations from the web using minimal supervision. En *45th Annual Meeting of the Association for Computational Linguistics (ACL'2007)*, 576–583.

Calzolari, Nicoletta. 1984. Detecting patterns in a lexical data base. En *22nd Annual Meeting of the Association for Computational Linguistics (ACL'1984)*, 170–173.

Caraballo, Sharon. 1999. Automatic constraction of a hypernym-labeled noun hierarchy from text. En *37th Annual Meeting of the Association for Computational Linguistics (ACL'1999)*, 120–126.

Cederberg, Scott & Dominic Widdows. 2003. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. En *The SIGNLL Conference on Computational Natural Language Learning (CoNLL'2003)*, 111–118.

Cimiano, Philipp, Andreas Hotho & Steffen Staab. 2004. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. En *16th European Conference on Artificial Intelligence (ECAI'2004)*, 435–439.

Dorantes, Miguel A. 2016. La estructura definitoria en lexicografía: sintaxis de la deinición analítica para sustantivos en un diccionario especializado. Tesis de Licenciatura. Universidad Nacional Autónoma de México.

Espinosa-Anke, Luis, Roberto Carlini, Horacio Saggion & Francesco Ronzano. 2016. DEFEXT: a semi supervised definition extraction tool. En *GLOBALEX Workshop, co-located with LREC'2016*, 24–28.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. En *14th Conference on Computational Linguistics*, 539–545.

Klavans, Judith L. & Smaranda Muresan. 2001. Evaluation of the DEFINDER system for fully automatic glossary construction. En *American Medical Informatics Association Symposium (AMIA'2001)*, 324–328.

Lara, Luis Fernando. 1997. *Teoría del diccionario monolingüe*. COLMEX.

Meyer, Ingrid. 2001. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. En *Recent advances in Computational Terminology*, 279–302. John Benjamins.

Mititelu Barbu, Virginica. 2006. Automatic extraction of patterns displaying hyponym-hypernym co-occurrence from corpora. En *First Central European Student Conference in Linguistics*, s.pp.

Oakes, Michael. 2005. Using Hearst's rules for the automatic acquisition of hyponyms for mining a pharmaceutical corpus. En *Recent Advances in Natural Language Processing (RANLP'2005)*, 63–67.

Ortega, Rosa M. 2007. *Descubrimiento automático de hipónimos a partir de texto no estructurado*: Instituto Nacional de Astrofísica, Óptica y Electrónica. Trabajo de Fin de Máster.

Pantel, Patrick & Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. En *Conference on Computational Linguistics (COLING'2006)*, 113–120.

Pantel, Patrick & Deepak Ravichandran. 2004. Automatically labeling semantic classes. En *North American Chapter of the Association for Computational Linguistics Conference (NAACL'2004)*, 321–328.

Pasca, Marius. 2004. Acquisition of categorized named entities for web search. En *13th ACM international conference on Information and knowledge management*, 137–145.

Pearson, Jennifer. 1998. *Terms in context*. John Benjamins.

Pereira, Fernando, Naftali Tishby & Lillian Lee. 1993. Distributional clustering of English words. En *31st Annual Meeting of the Association for Computational Lingistics (ACL'1993)*, 183–190.

Ravichandran, Deepak, Patrick Pantel & Eduard Hovy. 2004. The terascale challenge. En *KDD Workshop on Mining for and from the Semantic Web*, 1–11.

Richardson, Stephen D., Lucy Vanderwende & William Dolan. 1993. Automatically deriving structured knowledge bases from on-line dictionaries. En *The Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, 69–79.

Riloff, Ellen & Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. En *2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'1997)*, 117–124.

Rodríguez, C. 1999. Operaciones metalingüísticas explícitas en textos de especialidad. Trabajo de investigación. IULA, Universitat Pompeu Fabra.

Sager, Juan C. 1993. *Curso práctico sobre el procesamiento de la terminología*. Pirámide.

Sierra, Gerardo, Rodrigo Alarcón, Cesar Aguilar & Carme Bach. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication* 14(1). 74–98.

Sierra, Gerardo, Rodrigo Alarcón, César Aguilar, Alberto Barrón, Valeria Benítez & Itzia Baca. 2006. Corpus de contextos definitorios: una herramienta para la lexicografía y la terminología. En *X Simposio Iberoamericano de Terminología*, .

Smith, Robin. 2007. Aristotle's logic. Consultado el 3 de julio de 2016, de http://plato.stanford.edu/archives/win2007/entries/aristotle-logic/.

Snow, Rion, Daniel Jurafsky & Andrew Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. En *17th International Conference on Neural Information Processing Systems (NIPS'2004)*, 1297–1304.

Velardi, Paola, Roberto Navigli & Pierluigi D'Amadio. 2008. Mining the web to create specialized glossaries. *IEEE Intelligent Systems* 23(5). 18–25.

Widdows, Dominic. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. En *North American Chapter of the Association for Computational Linguistics Conference (NAACL'2003)*, 276–283.

Proposta rebuda el 15 de d'octubre 2017 i acceptada per a publicació el 31 de de desembre 2017.

Creació d'un motor de TAE especialitzat en farmàcia i medicina per a la combinació romanés–castellà

Creating an SMT engine for pharmaceutical and medical texts in the Romanian–Spanish language pair

Adrià Martín-Mor*
Universitat Autònoma de Barcelona
adria.martin@uab.cat

Víctor Peña-Irles
Universitat Autònoma de Barcelona
victorpenairles@gmail.com

Resum

Aquest article[1] descriu el procés de creació d'un motor de traducció automàtica estadística especialitzat en medicina per a la combinació lingüística romanés–castellà a partir de corpus lliures disponibles a internet. S'utilitza la plataforma MTradumàtica, creada en el marc d'un projecte de recerca del grup Tradumàtica per a fomentar l'ús de la TA entre els traductors. L'article es pot interpretar com una mostra que aquest propòsit s'ha assolit en el cas d'ús que presentem, la qual cosa suggereix que el perfil dels traductors és vàlid per dur a terme processos de personalització de TA.

Paraules clau

traducció automàtica, traducció automàtica estadística, personalització de motors de traducció

Abstract

This article[2] describes the process of creation of a statistical machine translation engine specialised in medicine for the Romanian–Spanish language pair. The engine was based on free corpora available in internet. The article describes the use of the platform MTradumàtica developed in the context of a research project by the Tradumàtica research group, aimed at promoting the use of MT among translators. The article can be interpreted as the evidence that the aim of promoting MT among translators has been attained in this particular case, and it suggests that the profile of the translators is valid to carry out processes of customisation of MT engines.

Keywords

machine translation, statistical machine translation, statistical machine translation customisation

1 Introducció

En el marc de ProjecTA,[3] el grup de recerca Tradumàtica es va proposar analitzar l'estat de la traducció automàtica (TA) en el teixit empresarial de Catalunya i de l'Estat espanyol (Torres-Hostench et al., 2016). Els resultats de l'anàlisi van conduir a la creació d'una plataforma per a la personalització de motors de traducció automàtica estadística (TAE) per tal d'acostar la TA als professionals de la traducció. Aquest article descriu el procés de creació d'un motor de TA especialitzat en medicina per a la combinació lingüística romanés–castellà. L'article està dividit en 6 apartats. L'apartat 2 (Personalització de motors de TAE) descriu què és la personalització de motors i quines plataformes existeixen actualment per a aquestes tasques. L'apartat 3 (Recursos per a la creació del motor) avalua els recursos disponibles per a la creació del motor de TAE en la combinació lingüística esmentada, amb referències a altres traductors automàtics existents i una avaluació crítica dels punts forts i les febleses. Finalment, l'apartat 4 (Descripció del procés de creació) descriu els recursos utilitzats i la seua preparació, abans que els resultats i les conclusions (5 i 6 respectivament) tanquen l'article.

*ORCID: 0000-0003-0842-3190

[1] 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 exili d'activistes polítics i membres del govern i en solidaritat amb els ciutadans que van patir la repressió de l'Estat espanyol arran del referèndum d'autodeterminació de l'1 d'octubre del 2017.

[2] This article is signed, as citizens of the Catalan Republic proclaimed by the legitimate government of Catalonia, in protest against the imprisonment and exile of political activists and members of the Catalan government and in solidarity with all the citizens who suffered reprisals by the Spanish state following the Catalan self-determination referendum held on the 1st October 2017.

[3] http://www.projecta.tradumatica.net. Referència FFI2013-46041-R, finançat pel Ministerio de Economía y Competitividad del Gobierno de España. Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad.

DOI: 10.21814/lm.9.2.254

2 Personalització de motors de TAE

Actualment, diverses plataformes permeten la personalització de motors de TAE. En l'àmbit del programari privatiu, existeixen programes com ara KantanMT,[4] LetsMT,[5] Microsoft Translator Hub[6] o Slate Desktop (anteriorment, DoMosesYourself).[7] Moses, programari lliure amb llicència GNU Lesser General Public License,[8] és un dels programes més utilitzats per a la creació de motors de TAE. Segons LT-Innovate (2013, p. 71), Moses és "widely used within the industry to build customized MT engines" i, justament, es destaca que, com que es tracta d'una plataforma lliure, "people wishing to develop a custom engine can focus on obtaining the training corpora rather than writing their own statistical machine translation engine (a difficult task that is beyond the abilities of most developers)." Malgrat tot, tal com continua LT-Innovate (2013, p. 72), Moses és "difficult to administer", començant pel fet que no té interfície gràfica d'usuari (GUI) i, per tant, requereix un cert coneixement de sistemes UNIX i del terminal, la qual cosa sol suposar una barrera d'entrada per a una gran part dels usuaris potencials. Probablement per aquest motiu, hi ha hagut en els últims anys intents de desenvolupar sistemes per a un públic menys expert en tecnologia. Per exemple, Machado & Fontes (2014) presenten un conjunt d'eines de programari lliure (desenvolupades "by a translator for translators", p. 2) per a la creació de motors de TAE, com ara eines de conversió de formats o materials de suport. Més recentment, han aparegut sistemes basats en Moses amb interfície gràfica, com ara ModernMT,[9] Machine Translation Training Tool (MTTT)[10] o MTradumàtica, tots tres amb llicències lliures.

MTradumàtica[11], actualment en versió experimental, és una plataforma web basada en Moses per a la creació de motors de TAE personalitzats (Martín-Mor, 2017). La llicència LGPL de Moses permet la modificació del codi font i la redistribució de programari, la qual cosa comporta que qualsevol usuari pot complementar o adaptar el programa original per als seus objectius, en el cas de ProjecTA, acostar la TA als traductors. A tal efecte, la plataforma desenvolupada es proposava:

1. Desenvolupar una interfície gràfica prenent en consideració una dimensió educativa envers l'usuari final.
2. Permetre l'ús via web, per tal d'evitar instal·lacions en local, la qual cosa converteix, *de facto*, el programa en multiplataforma.
3. Permetre la instal·lació en servidors propis, per tal d'assegurar una major confidencialitat en l'àmbit professional.

Des del punt de vista de la política de la recerca, el fet de contribuir al desenvolupament de programari lliure garanteix alhora que el producte de projectes d'investigació finançats amb fons públics esdevé també públic i disponible per a tota la societat.

Així, MTradumàtica segueix el següent esquema per a la creació de motors.

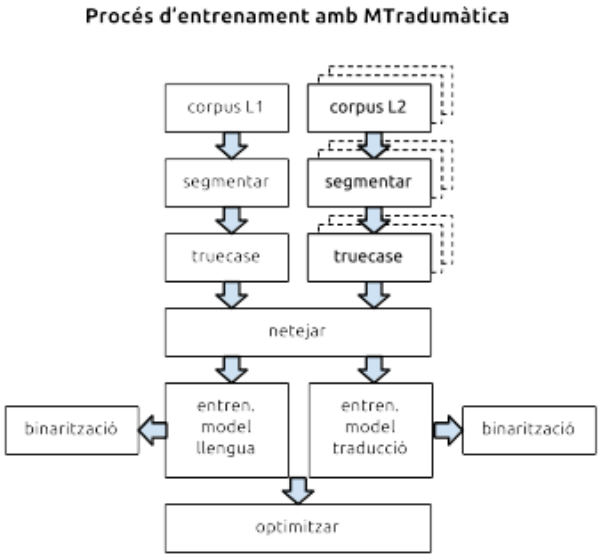

Figura 1: Esquema de processos en MTradumàtica.

A partir d'un corpus paral·lel bilingüe (per al model de traducció, en endavant, MT) i d'un o més corpus monolingües (per al model de llengua, en endavant, ML), MTradumàtica, com Moses (Koehn, 2016, p. 36), duu a terme els processos de segmentació (*tokenisation*), *truecasing* i neteja dels corpus. *Segmentar* vol dir separar amb espais les paraules dels signes de puntuació. En altres paraules, aïllar la puntuació permet incrementar les probabilitats d'obtenir coincidències amb els futurs textos que es traduiran automàticament. El procés de *truecasing*, en canvi, consisteix a determinar la caixa més probable de cada paraula, majúscules o minúscules. Tal com afirma Koehn al seu glossari de termes de Moses (Koehn, 2016, p. 361), "[t]his process typically leaves all words unchanged except for the

[4] https://www.kantanmt.com/.
[5] https://www.letsmt.eu/.
[6] https://hub.microsofttranslator.com/.
[7] https://slate.rocks/.
[8] Vegeu https://www.gnu.org/copyleft/lesser.html
[9] http://www.modernmt.eu.
[10] https://github.com/roxana-lafuente/MTTT.
[11] http://m.tradumatica.net

first word in the sentence, which may be lowercased." S'evita així que els vocabularis continguin entrades diferents per a la mateixa paraula en majúscules i en minúscules, i per tant les dades són menys esparses i es facilita l'entrenament. La neteja consisteix en la supressió de les frases llargues i mal alineades dels corpus amb l'objectiu de minimitzar els problemes en la fase d'entrenament.

Una vegada duts a terme aquests tres processos, el sistema processa les dades lingüístiques proporcionades en la fase d'entrenament, en la qual, a partir de l'anàlisi de coocurrències de paraules i segments en les dues llengües, s'infereixen de manera automàtica correspondències de traducció. El resultat de l'entrenament és el model de traducció, format per una taula de frases, un model de llengua i, ocasionalment, una taula de reordenament. Atés que la consulta de les taules pot ser lenta, els models es binaritzen per tal que es carreguen més ràpidament.

Finalment, l'optimització (o *tuning*) és un procés que determina automàticament els valors òptims d'una sèrie de paràmetres per tal que el motor generi "the best possible translations" (Koehn, 2016, p. 12). L'optimització consisteix en la traducció automàtica de milers de frases d'un subconjunt dels models (anomenat *development* o tuning set), la comparació amb les traduccions humanes de referència i l'ajustament automàtic dels valors de cada paràmetre per tal de millorar la qualitat del motor, mesurada mitjançant mètriques automàtiques com ara BLEU (Papineni et al., 2002). MTradumàtica es basa en els paràmetres per defecte de Moses per a fer l'optimització (no permet personalitzar-los ni tampoc té paràmetres diferents per a cada combinació lingüística). Un cop acabada l'optimització, el motor de TA estarà a punt. Actualment, MTradumàtica no permet dur a terme processos de postedició en la mateixa plataforma.

Pel fet que MTradumàtica ha estat dissenyat amb l'objectiu de facilitar l'acostament dels traductors a la TA, la interfície del programa conté referències als processos esmentats anteriorment. Tal com es pot observar a la Figura 2, malgrat que no és imprescindible tenir coneixements avançats sobre aquests processos per a la creació d'un motor de TAE amb MTradumàtica, l'eina també es proposa *formar* l'usuari en les nocions bàsiques de l'àmbit.

A tal efecte, la interfície inicial del programa presenta un procés lineal de sis passos (set, si es té en compte la funció *Inspect*, actualment en desenvolupament, v. més avall):

1. Càrrega de fitxers
2. Generació de monotextos
3. Generació de models de llengua
4. Generació de bitextos
5. Generació de traductors automàtics
6. Traducció

Els sis passos són visibles des de la pàgina inicial amb una breu explicació i indicacions addicionals. A més, al llarg de tot el procés, la barra superior mostra a l'usuari en quin pas es troba.

El procés comença amb la càrrega dels fitxers que posteriorment es faran servir per a la generació dels models de llengua i de traducció. De fet, la pàgina inicial, tal com es pot observar a la Figura 2, conté un enllaç al projecte Opus, el repositori de corpus lliures (Tiedemann, 2009). La pestanya *Files* mostra els textos carregats amb informació quantitativa (nombre de línies, paraules i caràcters) i la llengua del fitxer, detectada automàticament pel programa (l'usuari té la possibilitat de corregir la detecció automàtica en els casos en què falla). Davall dels textos carregats, hi ha un camp per a la càrrega de fitxers. En el moment de donar per tancat aquest article (desembre 2017), i tal com s'informa davall del camp esmentat, només es poden carregar fitxers de text amb una sola frase per línia.[12]

Al pas següent, *Monotexts*, l'usuari ha de generar monotextos amb l'objectiu de generar un model de llengua posteriorment, a la pestanya *LMs*. Es poden combinar diversos fitxers monolingües per tal d'obtenir un model de llengua més gran.

Un cop generat el model de llengua, la pestanya *Bitexts* permet —de manera paral·lela a com s'ha fet a la pestanya *Monotexts*— crear corpus bilingües mitjançant la càrrega de parelles de textos monolingües. Com en el cas dels monotextos, l'usuari pot combinar fitxers (sempre que siguen paral·lels) per tal d'obtenir un model de traducció més gran. El pas següent, *Translators*, permet crear traductors automàtics, amb model de llengua o sense. L'últim pas, *Translate*, permet utilitzar el motor creat, siga mitjançant la interfície web o mitjançant la càrrega de fitxers. Tal com expliquen Martín-Mor & i Huerta (2017,

[12] Es preveu que properament aquest pas permeti la càrrega de fitxers TMX, atés que és un format àmpliament utilitzat pels traductors. Mentrestant, es pot recórrer a programes com ara Okapi Rainbow per a la conversió de TMX a format *Parallel Corpus Files*: `http://okapiframework.org/wiki/index.php?title=Format_Conversion_Step` [última visita setembre 2017].

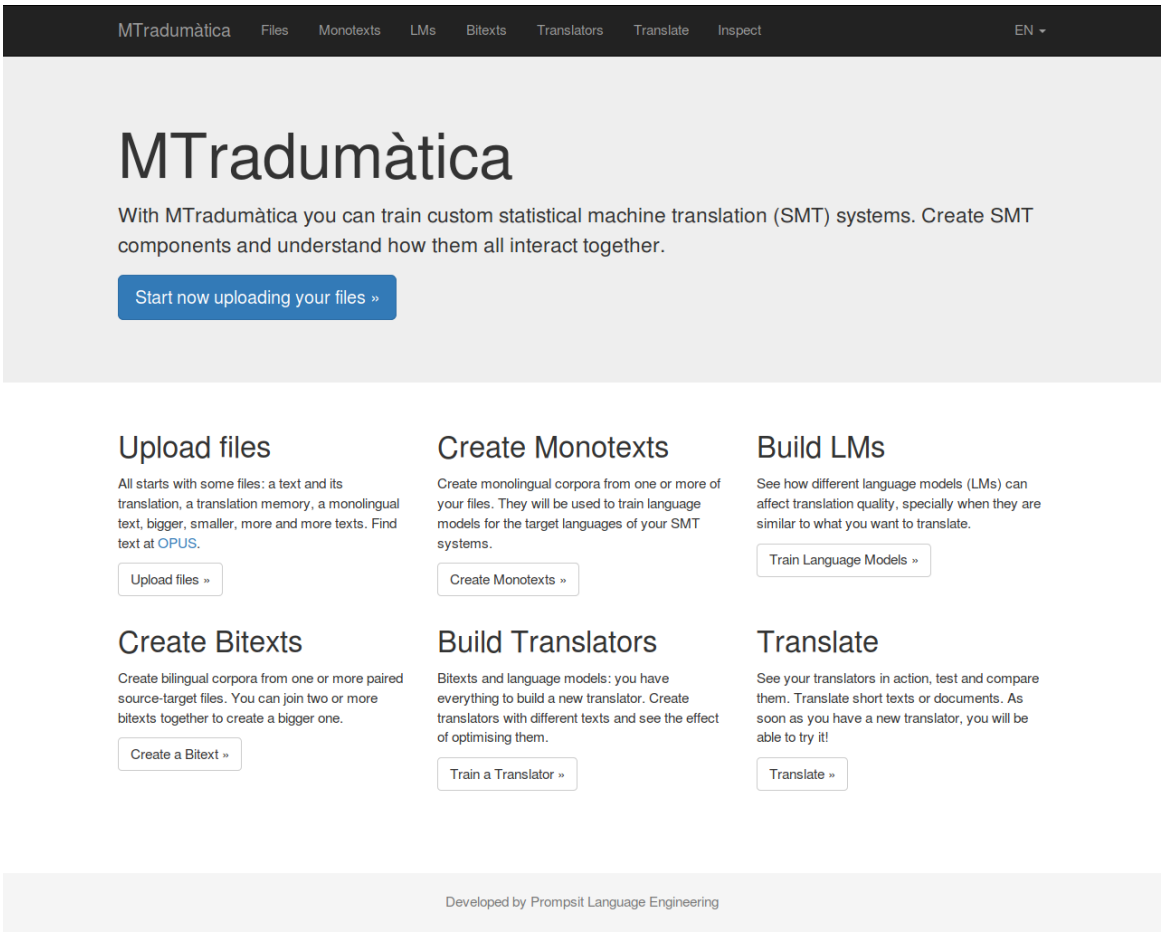

Figura 2: Interfície gràfica de MTradumàtica.

p. 112), està previst que MTradumàtica permeti a eines de traducció assistida accedir als motors mitjançant API.

Tal com s'ha esmentat anteriorment, la funció Inspect —visible en la versió actual de MTradumàtica però encara en desenvolupament— permetrà la consulta de les taules i els models de cadascun dels motors amb l'objectiu d'identificar possibles accions de millora.[13]

3 Recursos per a la creació del motor

En aquest cas pràctic d'entrenament de motors a la plataforma MTradumàtica l'objectiu ha estat crear dos motors de TAE: un del castellà al romanés i un altre del romanés al castellà. Cal precisar que, per a crear aquests motors de TAE, no és necessari emprar una plataforma com MTradumàtica, tot i que sí que facilita la tasca per la interfície gràfica i l'ús intuïtiu.

Hui en dia hi ha diversos motors de TA disponibles per a aquesta combinació de llengües, entre els quals, el traductor de Google,[14] el de Yandex,[15] Bing de Microsoft —el qual especifica de manera explícita que utilitza l'anglés com a llengua pont—,[16] i el motor de TA basat en regles d'Apertium, tan sols en la direcció romanés–castellà.[17]

Els motors esmentats són genèrics i no estan especialitzats en cap camp del coneixement. Els motors que es presenten en aquest article han estat entrenats amb corpus especialitzats en l'àmbit de la medicina i la farmàcia, per la qual cosa han estat necessaris:

- Corpus bilingües sobre medicina i farmàcia ro↔es per a tots dos MT i tots dos ML.
- Corpus monolingüe en castellà per a millorar el ML (es).
- Corpus monolingüe en romanés per a millorar el ML (ro).

[13] Per a més informació sobre els futurs desenvolupaments, vegeu Martín-Mor (2017).

[14] https://translate.google.com/.
[15] https://translate.yandex.com/.
[16] https://www.bing.com/translator/.
[17] https://www.apertium.org/.

Finalment, s'han seleccionat els corpus bilingües següents:

- Corpus ro↔es de l'Agència Europea del Medicament (EMEA): conté 12,9 milions de paraules en castellà i 11,9 milions en romanés, i un milió de segments alineats. El corpus naix a partir de l'alineació de fitxers en PDF d'aquest organisme i es pot descarregar de manera lliure del web del projecte Per-Fide (Almeida et al., 2014).[18] D'aquest corpus s'ha reservat el contingut des de la línia 961 fins a la 3461 (39 559 paraules en castellà i 36 828 en romanés) per tal de no emprar-los per a l'entrenament del motor i mantenir-los com a text de referència per a futures tasques d'optimització o per a l'avaluació automàtica de la TA (v. apartat 5). Així, s'ha entrenat el motor amb un corpus final amb 12 629 507 paraules en castellà i 11 690 520 en romanés.

- Corpus ro↔es del Centre Europeu per a la Prevenció i el Control de les Malalties (ECDC): conté 40 392 paraules en castellà i 37 105 paraules en romanés i 2 285 segments alineats. El corpus s'ha descarregat de manera lliure del portal EU Science Hub,[19] del servei de ciència i coneixement de la Comissió Europea.

Aquests dos corpus també s'han emprat per a entrenar els models de llengua. Tot i això, per tal de millorar aquests models, s'ha decidit entrenar-los amb continguts monolingües addicionals. Així, s'han creat dos corpus, un per a cada llengua de destí, a partir de continguts de la Viquipèdia del domini de la medicina i la farmàcia, com es veurà amb més detall a l'apartat 4. En total, els corpus monolingües tenen 378 000 paraules en castellà i 216 000 paraules en romanés.

Una de les particularitats de la combinació lingüística és la codificació dels caràcters en romanés. Les lletres diacrítiques del romanés *Ș* i *Ț* (i les minúscules *ș* i *ț*) es van incloure a Unicode per primer cop al setembre del 1999, en la versió 3.0.0 (Consortium, 2000) i ISO les publica a la ISO/IEC 8859-16 un any més tard. D'altra banda, als sistemes operatius i programes no s'han incorporat els caràcters correctes de manera homogènia, fet que ha provocat problemes de compatibilitat.

Corpus	**ro**	**es**
Viquipèdia (ro)	216 000	
Viquipèdia (es)		378 000
ECDC ro↔es	37 105	40 392
EMEA ro↔es (total)	11 901 523	12 939 973
EMEA ro↔es (per a l'entrenament)	11 690 520	12 629 507
EMEA ro↔es (com a referència)	36 828	39 554

Taula 1: Nombre de paraules dels corpus.

Això ha provocat que molts textos informatitzats en romanés no continguen diacrítics o s'hi hagen emprat durant les darreres èpoques els caràcters turcs anàlegs (*ş* i *ţ*), com descriu Kaplan (2011) amb més detall. Actualment, hi ha diversitat pel que fa a l'ús d'aquests diacrítics, el qual encara no és homogeni. Aquest és un aspecte que s'ha de tenir en compte tant durant el procés de creació del motor com durant a l'ús mateix del motor entrenat, com es veurà a l'apartat 4.

4 Descripció del procés de creació

En aquest apartat s'explica el procés de creació dels motors de TAE ro↔es a MTradumàtica. En primer lloc, s'hi descriuen les tasques prèvies per al processament dels corpus i, a continuació, la creació mateixa dels motors a MTradumàtica.

Processament previ dels corpus

Com s'explica a l'apartat 2, MTradumàtica només accepta fitxers de text pla amb una sola frase per línia. Per conseqüent, per a cada corpus ha sigut necessari aconseguir fitxers de text pla per a cada llengua amb una frase per línia i que conservessen l'alineació, en el cas dels corpus bilingües.

Quant al corpus de l'EMEA, descarregat del web de l'OPUS, només ha calgut descarregar els fitxers en format Moses, els quals ja compleixen aquestes característiques necessàries per a la creació del motor.

Pel que fa al corpus de l'ECDC, el format per defecte és un TMX multilingüe, tot i que al paquet s'inclou un programa Java que n'extreu els parells de llengües en un TMX bilingüe. Posteriorment, tal com s'ha esmentat més amunt, s'ha convertit en fitxers en format Moses amb el programa lliure Okapi.

[18] Vegeu http://per-fide.di.uminho.pt/ [última visita setembre 2017].

[19] Vegeu https://ec.europa.eu/jrc/en/language-technologies/ecdc-translation-memory [última visita setembre 2017]. Els drets d'autor pertanyen a la EU/ECDC i se'n permet l'ús tant comercial com no comercial. Vegeu: http://optima.jrc.it/Resources/ECDC-TM/2012_10_Terms-of-Use_ECDC-TM.pdf [última visita setembre 2017].

D'altra banda, pel que fa a la recopilació de corpus de la Viquipèdia per als ML, s'han extret articles per categories per mitjà de la funció Exporta.[20] Dins dels fitxers d'exportació, en format XML, ha calgut netejar el codi i extraure'n tan sols el text aprofitable per a entrenar el motor. Per a fer-ho s'ha emprat l'editor de textos de codi lliure Notepad++[21] mitjançant l'ús de macros. Aquest procés, inclosa la programació de les macros amb expressions regulars, s'explica amb detall a Peña-Irles (2017).[22]

Paga la pena afegir que, pels motius que s'expliquen a l'apartat 3, hi ha molts textos romanesos que no tenen diacrítics o que els tenen amb la codificació incorrecta. En el cas d'estudi de l'article no s'ha emprat cap text sense diacrítics, tot i que sí que s'ha observat heterogeneïtat dels caràcters per a l'escriptura de diacrítics romanesos. Per aquest motiu, ha calgut unificar-los. Notepad++ ha facilitat la cerca i substitució d'aquests caràcters pels caràcters de la norma Unicode 3.0.0 de l'any 2000 (Consortium, 2000).

Creació dels motors a MTradumàtica

Com s'explica a l'apartat 2, el primer pas és la pujada de tots els fitxers preprocessats a la plataforma MTradumàtica, tant en romanés com en castellà. Ha estat convenient modificar prèviament l'extensió dels fitxers per ".es" i ".ro" en funció de la llengua, per tal de facilitar el reconeixement de la llengua per part del sistema (v. apartat 2).

A continuació, s'ha creat un únic corpus per a cada llengua (a la pestanya *Monotexts*) a partir dels corpus previstos per als ML. Dit altrament, s'han creat dos corpus generals, un per a cada llengua, amb els fitxers d'EMEA, ECDC i els continguts de la Viquipèdia. Aquest pas és necessari per a poder completar el procés següent, és a dir, l'entrenament dels ML. A la pestanya LM, s'han entrenat dos ML de destinació, un per a cada motor de TAE, a partir dels monotextos acabats de crear. Aquest procés té una durada variable en funció de la quantitat de paraules i de la capacitat del servidor en què s'allotja MTradumàtica. En el cas del servidor de Tradumàtica, el ML en castellà s'ha entrenat en 5 minuts i 33 segons, mentre que el ML en romanès ha tardat 4 minuts i 44 segons a fer-ho.

Per als bitexts, en canvi, s'han ajuntat els corpus bilingües en un de general per tal d'entrenar els MT. En aquest cas, s'han seleccionat els corpus EMEA i ECDC, ja alineats. A diferència del procés descrit per als monotextos en el paràgraf anterior, atés que els bitextos són bidireccionals, no cal crear un corpus per a cadascun dels motors, sinó que el mateix permet entrenar tant el motor romanés–castellà com el traductor castellà–romanés.

Finalment, a la pestanya *Translators* s'han creat els traductors automàtics, un per a cada direcció, amb tot el que s'ha preparat prèviament: un ML entrenat i un bitext. Una vegada fet això, s'inicia el procés d'entrenament estadístic del motor. La durada també varia en funció de la longitud dels corpus i les especificitats del servidor en què s'allotja MTradumàtica. A tall indicatiu, en el cas del motor ro→es s'ha tardat 4 hores, 47 minuts i 43 segons, mentre que el motor es→ro ha tardat 4 hores, 46 minuts i 34 segons.

Després de l'entrenament, el pas següent per a la construcció dels motors és l'optimització. Aquest pas és optatiu i permet millorar-ne la qualitat. En el nostre cas es va ometre l'optimització dels motors pel fet que, en el moment de crear-los, aquesta funcionalitat estava en desenvolupament en MTradumàtica (vegeu l'apartat 6).

Un cop completat l'entrenament, a la pestanya Translate els traductors automàtics creats ja es poden utilitzar, tant introduint-hi un text a la interfície web, com mitjançant la càrrega de fitxers. Per a la combinació romanés–castellà cal tenir en compte que és necessari que els textos tinguen els caràcters de l'Unicode 3.0.0 abans d'introduir un text per a traduir-lo. Altrament, el traductor no és capaç de reconéixer els caràcters no estandarditzats, ja que no apareixen als corpus amb què s'ha entrenat el motor.

5 Resultats

L'objectiu d'aquest apartat és analitzar el rendiment dels motors mitjançant mètriques automàtiques per tal de demostrar la viabilitat d'MTradumàtica per a l'entrenament dels motors i comparar aquestes mètriques amb altres motors de TAE. S'han avaluat els resultats dels motors de traducció mitjançant tres mètriques d'avaluació automàtica de la TA: BLEU (Papineni et al., 2002; KantanMT), METEOR-ex (Banerjee & Lavie, 2005) i TER (Snover et al., 2006). Els dos primers mètodes es mesuren mitjançant valors de l'1 al 0, i n'és l'1 el valor òptim segons

[20] La funció Exporta permet la descàrrega d'articles per categories. Disponible a l'adreça `https://ro.wikipedia.org/wiki/Special:Export%C4%83` [última visita setembre 2017].

[21] Aquest editor de textos es pot descarregar des del web `https://notepad-plus-plus.org/` [última visita setembre 2017].

[22] El fitxer de macros ha estat publicat amb llicència lliure a `http://www.github.com/tradumatica`.

aquest mètode. D'altra banda, el mètode TER és un indicador que mesura l'esforç de postedició, de manera que, com més baix és el valor, menor és l'esforç de postedició (Peña-Irles, 2017). Aquests mètodes són avaluadors automàtics que indiquen el rendiment del motor, tot i que no expressen necessàriament la qualitat del resultat de la TA.

Per a dur a terme les avaluacions automàtiques és necessari disposar d'un text original, una o diverses traduccions automàtiques i una o diverses traduccions amb qualitat humana. S'ha fet servir el conjunt d'eines d'avaluació automàtica de la TA anomenat Asiya Online,[23] desenvolupat per la Universitat Politècnica de Catalunya, un "open toolkit aimed at covering the evaluation needs of system and metric developers along the development cycle" (Giménez & Màrquez, 2010). És accessible de manera lliure pel web i permet valorar els resultats de la TA mitjançant més de quinze mètodes d'avaluació (Peña-Irles, 2017). Cal precisar que Asiya Online mostra l'indicador TER en valors negatius (–TER), per la qual cosa n'ha estat necessària la conversió a valors positius.

Pel que fa a l'avaluació dels motors entrenats, se n'han dut a terme dues per a cada combinació de llengües, la primera a partir d'un text de referència extret del corpus de l'EMEA,[24] i la segona a partir d'un prospecte mèdic posteditat per a cada llengua.[25] D'altra banda, s'han dut a terme les mateixes avaluacions amb tres altres motors de TA genèrics disponibles en aquestes combinacions d'idiomes: Google, Yandex i Apertium — només per a la combinació romanés–castellà— (v. l'apartat 3), per tal de comparar-ne els resultats. Recordem també (vegeu l'apartat 4, Descripció del procés de creació) que els motors que s'analitzen a continuació no han estat optimitzats. La metodologia i els resultats d'aquesta anàlisi s'analitzen amb més detall a Peña-Irles (2017).

[23] L'URL de l'eina és `http://asiya.lsi.upc.edu/demo/asiya_online.php` [última visita setembre 2017].

[24] El text de referència s'ha pres d'una part d'un corpus que s'ha extret prèviament a l'entrenament del motor i que s'empra amb la finalitat d'avaluar el rendiment del motor i per a l'ajustament dels paràmetres de la TAE o *tuning*. Els textos emprats per a l'avaluació tenen 2 203 paraules en castellà i 2 083 en romanés.

[25] Per al castellà s'ha descarregat un prospecte del web del Ministeri de Sanitat espanyol (de 185 paraules): `https://www.aemps.gob.es/cima/dochtml/p/69429/Prospecto_69429.html` [última visita setembre 2017]. Per al romanés, s'ha extret un prospecte del web Ce se întâmplă doctore (de 269 paraules): `http://www.csid.ro/medicamente/omeprazol-terapia-20-mg-capsule-gastrorezistente-11474561/` [última visita setembre 2017].

Motor romanés–castellà

Els resultats del motor entrenat a MTradumàtica del romanés al castellà amb les mètriques esmentades es mostren a la taula següent. També s'hi comparen els resultats amb els dels traductors d'Apertium, Google i Yandex:

	BLEU	METEOR-ex	TER
Text de referència EMEA (ro→es)			
MTradumàtica	0,60	0,73	0,35
Apertium	0,19	0,35	0,68
Google	0,43	0,58	0,47
Yandex	0,35	0,54	0,54
Prospecte posteditat (ro→es)			
MTradumàtica	0,54	0,69	0,32
Apertium	0,33	0,51	0,45
Google	0,40	0,59	0,35
Yandex	0,52	0,66	0,29

Taula 2: Avaluació de la TA ro→es (Mtradumàtica, Apertium, Google i Yandex).

La taula anterior mostra que els resultats de les mètriques d'avaluació amb MTradumàtica són similars, i fins i tot, en la majoria dels casos, superiors, als de productes existents, la qual cosa indica que el motor descrit en aquest article podria ser viable en aplicacions de disseminació (Forcada, 2009). Els resultats del motor romanés–castellà presenten unes mètriques molt positives i elevades, que podrien suposar la viabilitat del motor en aplicacions de disseminació. D'altra banda, en comparar-lo amb la resta de motors de TA analitzats, s'obtenen uns resultats superiors. Hi destaca el resultat de Google, a la taula 3, amb diferències d'entre 0,15 i 0,2 punts al paràmetre BLEU, i el baix rendiment d'Apertium. A més, el resultat de Yandex en el text posteditat és semblant i, fins i tot, superior en el cas de TER.

Motor castellà–romanés

Els resultats del motor entrenat a MTradumàtica del castellà al romanés amb les mètriques esmentades es mostren a la taula següent. També s'hi comparen els resultats amb els dels traductors de Google i Yandex:

En analitzar els resultats per a la combinació castellà–romanés s'observa que les mètriques són molt positives i que, a més a més, s'obtenen els millors resultats en comparació amb els altres dos motors analitzats. Els resultats tant de Google com de Yandex obtenen unes mètriques inferiors.

	BLEU	METEOR-ex	TER
Text de referència (es→ro)			
MTradumàtica	0,73	0,51	0,24
Google	0,33	0,30	0,54
Yandex	0,29	0,28	0,58
Prospecte posteditat (es→ro)			
MTradumàtica	0,54	0,47	0,34
Google	0,35	0,30	0,44
Yandex	0,30	0,29	0,49

Taula 3: Avaluació de la TA es→ro (Mtradumàtica, Apertium, Google i Yandex).

6 Conclusions

Aquest article ha presentat un estudi de cas d'aplicació d'un producte de recerca a un projecte real. La plataforma MTradumàtica, desenvolupada en el marc d'un projecte públic per a l'acostament de la traducció automàtica als traductors, i amb un èmfasi en l'aspecte formatiu, ha estat utilitzada per part de traductors per crear un traductor automàtic especialitzat en farmàcia i medicina per a la combinació lingüística romanés–castellà. D'una banda, això ens permet constatar que l'objectiu per al qual naixia el producte en certa manera es compleix: s'ha creat un motor de TAE especialitzat a partir de recursos lliures de la xarxa utilitzant la interfície gràfica de la plataforma. El motor de TAE, a més, no sols és funcional, sinó que dóna bons resultats pel que fa al rendiment amb indicadors aproximats com BLEU, com es veu a l'apartat 5. Cal tenir en compte, com hem dit a l'apartat 4, que les mètriques automàtiques presentades en aquest article s'han generat a partir de motors no optimitzats. És bastant raonable pensar que els motors optimitzats obtindrien millors resultats, per la qual cosa podem considerar que els valors obtinguts són un bon punt de partida que només podria millorar. Malgrat tot, creiem que l'interés de l'experiència es troba no tant en el rendiment del resultat, sinó principalment en la constatació que és possible dur a terme processos de creació de motors de qualitat per part de traductors. En aquest sentit, com a línia de recerca en un futur, seria interessant analitzar la qualitat real dels resultats d'aquests motors optimitzats amb indicadors de l'esforç de postedició, com ara HTER, mitjançant experiments reals de posteditors. D'altra banda, l'experiència descrita suggereix, tal com plantejava el projecte de recerca de Tradumàtica, que el perfil dels traductors és un perfil vàlid per dur a terme processos relacionats amb la personalització de motors de TA. Tot això ens fa entreveure l'impacte que pot tenir per als programes de formació en traducció el desenvolupament de sistemes de personalització de TA amb interfície gràfica.

L'article ha descrit detalladament cadascun dels passos que s'han seguit per al desenvolupament d'un motor amb l'objectiu que l'experiència siga replicable per part d'altres traductors amb necessitats similars, per a la mateixa o per a altres combinacions lingüístiques i camps d'especialitat. És per aquest motiu que s'ha utilitzat no sols programari lliure sinó també recursos disponibles amb llicències lliures. També els recursos generats, com ara el paquet de macros per a la neteja dels corpus descarregats de la Viquipèdia han estat posats a disposició de la comunitat amb llicència lliure.

L'experiència descrita apunta a la necessitat que les plataformes per a la personalització de motors permeten el preprocessament dels corpus mitjançant regles senzilles. En casos com l'esmentat a l'apartat 3, seria útil configurar una sèrie de regles, com ara per mitjà de la utilitat d'Unix *Stream EDitor* (`sed`),[26] amb llicència GPLv3, útil per a aplicar transformacions a un text, amb l'objectiu que qualsevol codificació incorrecta en el text original no genere una traducció errònia o desconeguda, sinó que es convertisca a la codificació correcta abans de ser traduïda.

Referències

Almeida, José João, Sílvia Araújo, Nuno Carvalho, Idalete Dias, Ana Oliveira, André Santos & Alberto Simões. 2014. The Per-Fide corpus: A new resource for corpus-based terminology, contrastive linguistics and translation studies. En *Working with Portuguese Corpora*, 177–200. Bloomsbury Publishing.

Banerjee, Satanjeev & Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. En *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Consortium, Unicode. 2000. *The unicode standard, version 3.0.* Addison-Wesley.

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

Giménez, Jesús & Lluís Màrquez. 2010. Asiya: An open toolkit for automatic machine trans-

[26] Vegeu `http://sed.sourceforge.net/sedfaq2.html#s2.1` [última visita setembre 2017].

lation (meta-)evaluation. *The Prague Bulletin of Mathematical Linguistics* 94(1). 77–86.

KantanMT. 2017. What is BLEU score? Recuperat de https://www.kantanmt.com/whatisbleuscore.php.

Kaplan, Michael S. 2011. The history of messing up Romanian on computers. MSDN blogs. 24 d'agost. Recuperat de http://archives.miloush.net/michkap/archive/2011/08/24/10199324.html.

Koehn, Philipp. 2016. *Moses user manual and code guide*.

LT-Innovate. 2013. Status and potential of the European language technology markets. http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=4267.

Machado, Maria José & Hilário Leal Fontes. 2014. *Moses for mere mortals. tutorial. a machine translation chain for the real world*. https://github.com/jladcr/Moses-for-Mere-Mortals/blob/master/Tutorial.pdf.

Martín-Mor, Adrià & Ramon Piqué i Huerta. 2017. MTradumàtica i la formació de traductors en traducció automàtica estadística. *Tradumàtica* 15. 97–115.

Martín-Mor, Adrià. 2017. MTradumàtica: Statistical machine translation customisation for translators. *Skase Journal of Translation and Interpretation* 11(1). 25–40.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. En *40th annual meeting on Association for Computational Linguistics (ACL'2002)*, 311–318.

Peña-Irles, Víctor. 2017. Entrenament de motors de traducció automàtica estadística especialitzats en farmàcia i medicina entre el castellà i el romanés. Treball de recerca de màster. Universitat Autònoma de Barcelona.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciula & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. En *Association for Machine Translation in the Americas Conference*, s.pp.

Tiedemann, Jörg. 2009. News from OPUS: A collection of multilingual parallel corpora with tools and interfaces. En *Recent Advances in Natural Language Processing*, 237–248. John Benjamins.

Torres-Hostench, Olga, Marisa Presas & Pilar Cid-Leal. 2016. L'ús de traducció automàtica i postedició a les empreses de serveis lingüístics de l'estat espanyol. informe de recerca Project-TA 2015. Universitat Autònoma de Barcelona.

Projetos, Apresentam-se!

Propuesta recibida el 1 de noviembre 2017 y aceptada para publicación el 26 de diciembre 2017.

GECO, un Gestor de Corpus colaborativo basado en web

GECO, A Web-based Collaborative Corpus Manager

Gerardo Sierra Martínez
Universidad Nacional Autónoma de México
gsierram@iingen.unam.mx

Julián Solórzano Soto
Universidad Nacional Autónoma de México
jsolorzanos@iingen.unam.mx

Arturo Curiel Díaz
Universidad del Bío-Bío, Chile
me@arturocuriel.com

Resumen

Este artículo presenta GEstor de COrpus (GECO), un software de gestión de corpus en línea que permite a los usuarios subir colecciones de documentos y volverlos corpus digitales. En el sistema, los corpus pueden ser procesados por otras aplicaciones, las cuales están implementadas como módulos integrados a la infraestructura de GECO. En este documento se describen a detalle sus características, así como la funcionalidad del generador de concordancias desarrollado en torno a él.

Palabras clave

gestor de corpus, generador de concordancias, software de anotación

Abstract

This paper presents GEstor de COrpus (GECO), an online corpus management software that lets users upload document collections and turn them into digital corpora. Inside the system, corpora can be further processed by other applications, which are implemented as modules over the GECO framework. In this document, GECO's features are described in detail, as well as the functionality of a concordance generator module developed on top of it.

Keywords

corpus manager, concordance generator, annotation software

1 Introducción

Los gestores de corpus son aplicaciones especializadas que permiten a los usuarios cargar archivos de texto y ejecutar consultas (Ntoulas et al., 2001). Están diseñados para manejar grandes cantidades de información y vienen normalmente con funcionalidades adicionales, tales como el cálculo de estadísticas del corpus.

A grandes rasgos, lo que constituye la parte de gestión de corpus del software es aquella que administra los documentos: permite a los usuarios añadir o eliminar textos de una colección y permite anotar los documentos con diversos metadatos (tal como autor, género, época, tema, etc.). Sin embargo, el aspecto de creación de corpus no suele ser el enfoque principal de los gestores de corpus sino las aplicaciones que proveen, como los generadores de concordancias, que son sistemáticamente incluidos en los sistemas más populares (Kouklakis et al., 2007).

En este artículo se presenta a detalle un sistema de gestión de corpus llamado GECO[1] (GEstor de COrpus). Se pone énfasis en la descripción de los principios de diseño en los que está basado GECO, y como eficientiza el proceso de creación de nuevos corpus.

La sección 2 hace una breve comparación con software existente similar. La sección 3 describe los objetivos de diseño y la filosofía de GECO. En la sección 4 se explican a detalle las funcionalidades del software. La sección 5 presenta un ejemplo detallado de cómo se integra al sistema un módulo aplicativo, un generador de concordancias. En la sección 6 se presenta una descripción más técnica del funcionamiento del sistema. Finalmente, en la sección 7 se presentan algunas conclusiones y trabajo futuro.

2 Software relacionado

Hoy en día existe una gran variedad de software de gestión de corpus en el mercado, cada uno con diferentes capacidades para análisis cuanti-

[1] Está disponible en la página: http://www.corpus.unam.mx/geco/.

DOI: 10.21814/lm.9.2.256

tativos del texto (Manning & Schutze, 1999), como lo son anotación de metadatos, generación de concordancias y cálculo de colocaciones (Kouklakis et al., 2007). Los gestores de corpus pueden proporcionar desde listas de palabras simples (Anthony, 2005), hasta robustos marcos de trabajo de desarrollo capaces de usar el procesamiento del lenguaje natural (NLP) para aplicaciones concretas (Kilgarriff et al., 2015). Por ejemplo, el proyecto Corpógrafo (Sarmento et al., 2006) es un gestor de corpus que utiliza técnicas de procesamiento de lenguaje para extracción de términos y extracción de relaciones léxicas. Asimismo, LinguaKit es una herramient multilingüe para el análisis, la extracción, antotación y corrección lingüística (Gamallo & Garcia, 2017).

En los siguientes párrafos se describen las características de dos gestores de corpus bien conocidos y sobre los cuales están construidos muchos otros sistemas: el IMS Open Corpus Workbench (CWB) y Manatee. Ambas herramientas ofrecen poderosos lenguajes de consulta e implementan una arquitectura similar a la que otros gestores de corpus han usado antes (Christ, 1994) y proporcionan al usuario una interfaz gráfica basada en web. Las siguientes descripciones se dan con fines comparativos.

El IMS Open COrpus Workbench (CWB)

El CWB es una colección de herramientas de código abierto para la gestión de corpus y anotación lingüística (Evert & Hardie, 2011). Está diseñado para manejar grandes cantidades de información eficientemente. El CWB puede codificar e indexar corpus de cualquier tamaño. Para poder ser procesados adecuadamente, los archivos de entrada deben ya estar segmentados y anotados. Para esto, el CWB proporciona dos tipos de anotaciones: los atributos posicionales (atributos-p) y los atributos estructurales (atributos-s), ambos expresados como etiquetas XML. A grandes rasgos, un atributo-p es un atributo a nivel palabra: ligan una posición del corpus a un valor. Por ejemplo, las palabras en sí son atributos-p que corresponden a valores que aparecen en una posición única dentro del corpus. Por el otro lado, los atributos-s son atributos ligados a rangos de posiciones: permiten asociar etiquetas a secuencias de palabras en cualquier parte del documento. Por ejemplo, las colocaciones son anotadas como atributos-s.

La función más importante de CWB es el procesador de consultas de corpus (CQP, por sus siglas en inglés), un sistema de generación de concordancias con un lenguaje de consulta muy flexible que permite ingresar complejos patrones de búsqueda de palabras o frases. Soporta expresiones regulares y es capaz de encontrar atributos de palabras (por ejemplo, etiquetas de parte de la oración) así como elementos entre etiquetas estructurales.

Existe un paquete separado llamado CQPWeb (Hardie, 2012) que proporciona una interfaz web para el software, permitiendo al usuario instalar nuevos corpus y utilizar el CQP desde el navegador web. La Figura 1 muestra un ejemplo de la interfaz usando el Brown Corpus (Francis, 1965).

CQPWeb ofrece otras funcionalidades adicionales no incluidas en CWB, tales como colocaciones y ordenamiento de los resultados de búsqueda, metadatos y otros. Puede ser instalado tanto localmente (en computadoras individuales) como en un servidor público donde los usuarios pueden registrarse y acceder a los corpus instalados usando una cuenta. Para ligar la interfaz con los corpus, el administrador puede apuntar el sistema a un recurso ya existente o bien subir los archivos directamente desde la interfaz web, lo cual creará inmediatamente el índice. La interfaz solicita la descripción de la estructura del corpus, dada a partir de los atributos-p y atributos-s. Una vez que los corpus están cargados en el sistema, el administrador puede dar permisos específicos a cada usuario dependiendo de a qué corpus éste tendrá acceso. Los desarrolladores planean en el futuro permitir que los usuarios puedan subir sus propios documentos, aunque al momento de este escrito esta funcionalidad aún no está implementada.

Como ejemplo de aplicaciones construidas sobre el CWB tenemos el Bwananet de la Universidad Pompeu Fabra (Vivaldi, 2009) y el proyecto Gramateca de la Linguateca (Simões & Santos, 2014).

Manatee + Bonito / SketchEngine

Manatee (Rychlý, 2007) es un software de gestión de corpus de propósito general. Ofrece las mismas funcionalidades que CWB: procesamiento del texto (codificación, etc), administración de metadatos, segmentación del texto, generación de concordancias, anotación (tanto atributos-p como atributos-s) y el cálculo de estadísticas del corpus.

También proporciona un lenguaje de consulta propio, el cual es una extensión de CQP.

El sistema fue diseñado modularmente, proporcionando módulos para la compresión, el indexado y la evaluación de consultas, entre otros.

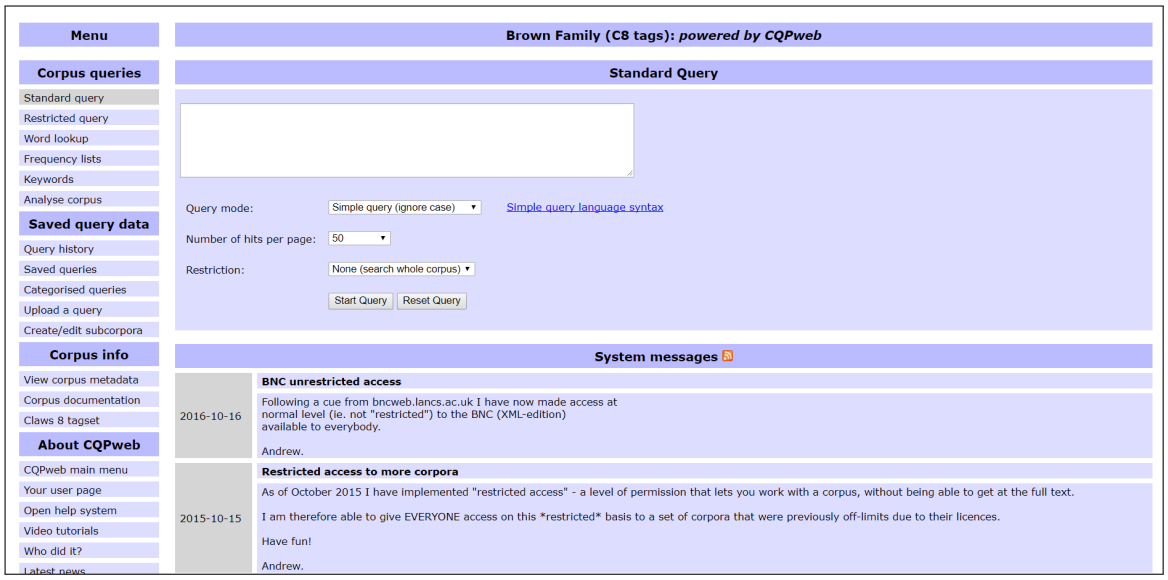

Figura 1: La página de bienvenida del CQPWeb para el Brown Corpus.

Incluye una interfaz de línea de comandos con la cual los corpus son creados y mantenidos. También tiene dos interfaces gráficas: Bonito y Bonito2.

Bonito es un módulo de Manatee que proporciona una interfaz amigable al usuario. Funciona con una arquitectura cliente-servidor, siendo Bonito un cliente de Manatee. Esto significa que Bonito puede ser instalado independientemente de una instancia de Manatee, -incluso en diferentes computadoras- en tanto el primero pueda acceder localmente o vía Internet al segundo. Una versión web de Bonito, llamada Bonito2, también es distribuida junto con Manatee. La mayor diferencia entre los dos es que Bonito2 es accesible vía el navegador web, por lo cual los usuarios pueden evitar instalar el cliente y simplemente dirigir el navegador a la dirección en donde Manatee está funcionando. La Figura 2 muestra un ejemplo de la interfaz web (Rychlý, 2007).

Gracias a su diseño modular, otros sistemas de gestión de corpus pueden ser construidos con base en Manatee + Bonito para agregar funcionalidad adicional. Un ejemplo sería la aplicación comercial SketchEngine. Este software añade nuevas funcionalidades a Manatee + Bonito, ofreciendo bosquejos de palabras (word sketch) además de las herramientas de análisis de corpus tradicionales. A grandes rasgos, los bosquejos de palabras son resúmenes derivados del corpus del comportamiento gramatical y colocacional de una cierta palabra. Se da un ejemplo en la Figura 3.

SketchEngine además tiene el módulo "Arquitecto de COrpus", el cual es una extensión enfocada a la construcción de corpus. Este módulo permite a los usuarios cargar archivos en diversos formatos para crear corpus. Los archivos pueden convertirse en lo que se conoce como un archivo vertical, en el cual cada línea es una palabra junto con sus respectivas anotaciones (por ejemplo, etiquetas POS). Cuando los archivos son cargados, se puede compilar e indexar el corpus vía la interfaz. La funcionalidad del "Arquitecto de COrpus" es además complementada por el módulo "WebBootCaT" (Baroni et al., 2006), el cual está diseñado para construir corpus a partir de la web. Funciona por medio de consultas a la web vía un motor de búsqueda existente (por ejemplo Google) y los documentos que se recuperan se integran en un corpus. También, el usuario puede proporcionar URLs específicos para recuperar sus contenidos.

3 Objetivos de diseño de GECO

A diferencia de los gestores presentados anteriormente, GECO fue concebido en su totalidad como una aplicación web, enfocada en la construcción colaborativa de corpus. En ese sentido, tiene la finalidad de ser un repositorio central de documentos para una variedad de aplicaciones orientadas al PLN, que pueden ser integradas a GECO por desarrolladores con herramientas de código abierto. Además de esto, tiene un enfoque muy marcado hacia la construcción y publicación de corpus como fin en sí mismo, por lo cual también provee herramientas para dar a conocer los corpus creados a través de un portal web personalizado, el cual puede desplegar el nombre de los participantes del proyecto, agradecimientos, referencias, etc.

Figura 2: Generador de concordancias de Bonito2.

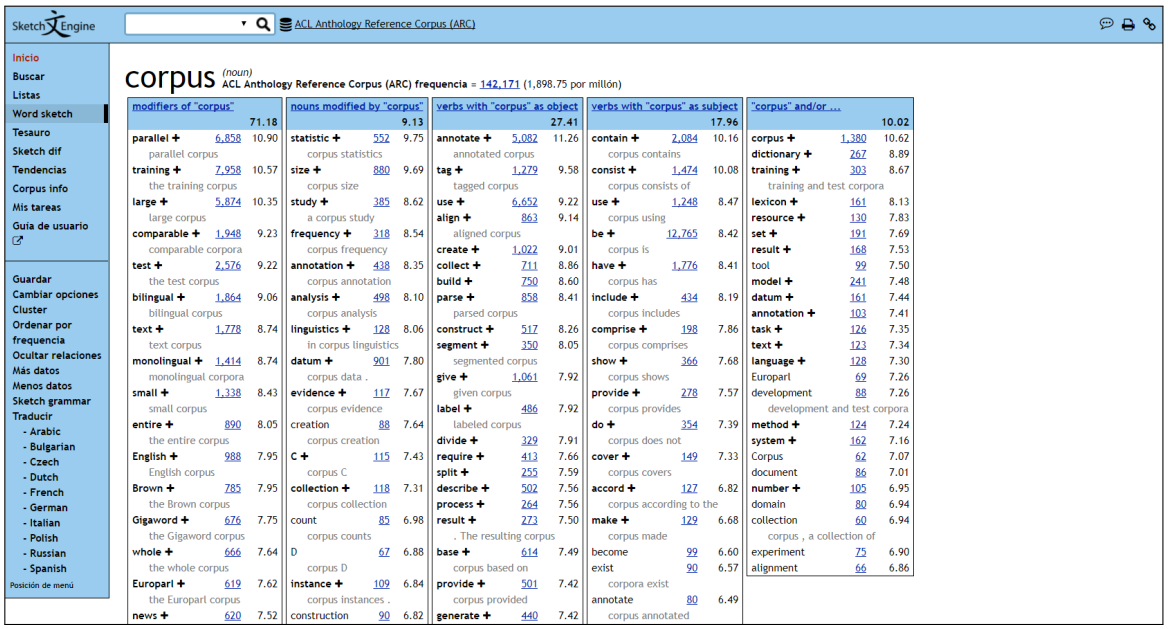

Figura 3: Word Sketch de la palabra corpus, obtenida del ACL Antology Reference COrpus, (Birda et al., 2008), generado con SketchEngine.

Colaboración

Para GECO el proceso de administración de documentos es una tarea colaborativa: varios usuarios pueden participar en la creación de un corpus. Para ello el software organiza archivos similar a como lo haría un sistema de archivos tradicional (Arpaci-Dusseau & Arpaci-Dusseau, 2016, p. 478). Esto permite a los usuarios agrupar documentos en carpetas y compartirlas en línea si así lo desean, dejando que ellos selectivamente decidan quién puede y quién no puede acceder a sus archivos. De manera similar, el administrador del sistema puede controlar el nivel

de acceso que los usuarios tienen de las carpetas. Por ejemplo, solo los usuarios con el permiso de escritura pueden subir documentos a la carpeta y modificar sus metadatos.

Diseño modular

Uno de los aspectos más importantes de GECO es su diseño modular. Fue desarrollado para integrarse con otras aplicaciones de software. Los corpus creados con GECO son visibles para módulos externos vía una Interfaz de Programación de Aplicación (API) (Reddy, 2011). Esto permite a dichas aplicaciones consultar la información acerca de los documentos, recuperando tanto su contenido como sus metadatos. De esta manera, la API facilita la implementación de aplicaciones de PLN, haciendo transparente para los desarrolladores las tareas de preprocesamiento de los textos.

Los módulos de GECO pueden comunicarse unos con los otros, ya que comparten la misma base de datos. Esta comunicación interna permite a los usuarios crear flujos de procesamiento, direccionando los resultados de cualquiera de las aplicaciones existentes para fungir como entrada de otra. Una base de datos compartida también permite a los administradores controlar el acceso individual a cada uno de los módulos, decidiendo quien puede ejecutar qué funcionalidad, dependiendo de las necesidades de cada usuario. Por defecto, los módulos respetan los permisos de archivos: los usuarios pueden ejecutar la funcionalidad de un módulo sólo sobre los datos a los que tienen acceso. De esta manera, los usuarios que tienen una sesión iniciada en GECO pueden visualizar únicamente los módulos y recursos compartidos sobre los cuales tienen permiso, y no los datos privados creados por otros usuarios.

4 Funcionalidades básicas

Desde una perspectiva más técnica, GECO implementa una serie de módulos destinados a cumplir las metas descritas en la sección anterior: un manejador de archivos, un motor PLN de anotación de textos, un sistema de metadatos y varias herramientas de gestión de proyectos. Además, algunos módulos aplicativos basados en la API de GECO son distribuidos junto con el sistema, proporcionando la misma funcionalidad mínima que otros gestores de corpus proporcionan (por ejemplo, generación de concordancias).

La Figura 4 muestra un diagrama de bloques de la organización interna de GECO. De manera breve, El "Núcleo GECO" expone una API web, referido en el diagrama como el "Webservice GECO". Éste se comunica directamente con el módulo "Interfaz Web GECO", construido con web2py (Di Pierro, 2011), el cual funciona como mascarilla del sistema. Las aplicaciones externas comunican la información de sus vistas a la interfaz, mostrado en el diagrama como "Vistas de las Aplicaciones". Los usuarios solo pueden interactuar con el sistema a través de la interfaz web. Sin embargo, la operación y los permisos de acceso son controlados en su totalidad por el núcleo del sistema, a través de la API Web. El "núcleo GECO" se comunica con un servidor de conversión de documentos, el cual transforma documentos en varios formatos a texto plano. También tiene acceso a un servidor FreeLing (Padró & Stanilovsky, 2012), una suite de PLN de código abierto. Finalmente, los datos de configuración del sistema y los archivos del corpus son almacenados en una base de datos relacional y en un sistema de archivos independiente, respectivamente. A continuación se presenta una explicación más a fondo de cómo funcionan estos componentes en conjunto.

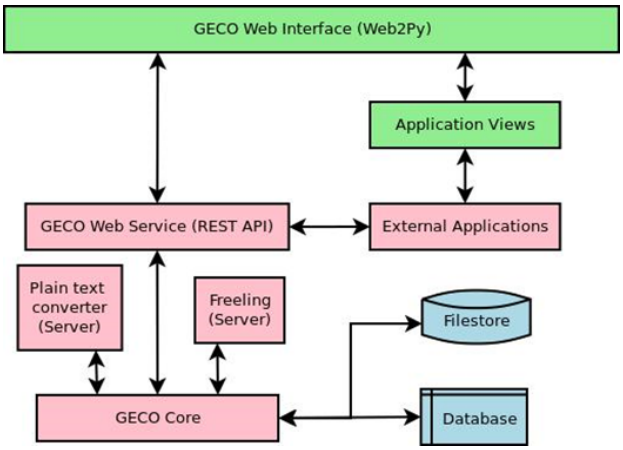

Figura 4: Diagrama de la arquitectura de GECO. Solo la Interfaz web y las Vistas de las Aplicaciones son accesibles a los usuarios. Los documentos son almacenados en un repositorio de archivos central, mientras que la información administrativa del sistema es almacenada en una base de datos relacional.

Manejador de archivos

El "núcleo GECO" es esencialmente un manejador de archivos. Controla la organización de los documentos, incluyendo adición, remoción y modificación. Permite a los usuarios navegar a través de los documentos, organizados en carpetas, como lo harían en cualquier sistema operativo. De esta forma, los usuarios tienen la capacidad de crear nuevas carpetas y llenarlas con

archivos, donde cada carpeta servirá como una posible fuente de documentos para un nuevo proyecto de corpus. La creación de archivos corresponde a una llamada iniciada por el módulo "Interfaz Web GECO".

Como se mencionó anteriormente, los corpus leíbles por máquina están compuestos típicamente de archivos de texto plano, y GECO no es diferente en este aspecto. Sin embargo, a veces los archivos del usuario no son texto plano. El módulo núcleo tiene la responsabilidad de organizar y transformar los diferentes tipos de archivo soportados por GECO como sea necesario. Al recibir un archivo, interpreta y convierte los diferentes tipos de formatos de archivos (por ejemplo, .doc, .pdf, etc) a texto plano UTF-8 (Yergeau, 2003), por medio del módulo "Convertidor de texto plano". Los archivos originales pueden ser conservados por GECO como metadatos multimedia, para no perder información. Para los usuarios, el proceso completo es transparente, ya que no tienen que preocuparse por convertir los archivos antes ellos mismos. La idea detrás de esta conversión es adaptarse a las necesidades del usuario y no viceversa. GECO también puede puede procesar archivos comprimidos en formato ZIP, para cargas masivas.

Cabe mencionar que, por lo menos hasta este momento, GECO no es capaz de importar textos previamente anotados, sea cual fuere su esquema de anotación. Por el momento GECO solo recibe textos sin anotaciones, y es el sistema quien encarga de anotarlo. El archivo anotado resultante es un XML sencillo que puede ser utilizado con otros indexadores de corpus, por ejemplo Manatee. Esta es la razón principal por la cual en esta fase del proyecto no se haya optado por manejar otros esquemas de anotación, como podrían ser los estándares XML-TEI(Areta et al., 2007), aunque bien en el futuro podría beneficiarse de ello.

Preprocesamiento automático del texto

Al insertar documentos de texto plano en el sistema de archivos, el "Núcleo GECO" envía los archivos a un servidor FreeLing, para realizar un procesamiento básico de PLN. Este servicio permite que GECO pueda ejecutar funciones de análisis de lenguaje (segmentación, análisis morfológico, detección de entidades nombradas, etiquetado POS, etc.). El análisis proporcionado por FreeLing permite a GECO segmentar los archivos en tokens, obtener el lema de cada token, y anotarlos con sus respectivas etiquetas POS. Como se verá adelante, este análisis de FreeLing es ampliamente usado en las aplicaciones de GECO.

Por cada archivo, GECO almacena dos versiones: el texto plano UTF-8 (tal cual como se subió), y el texto vertical (con sus respectivas anotaciones tras ser procesado por FreeLing). El sistema permite a los usuarios descargar ambos en cualquier momento. Juntos, el manejo de archivos y el módulo de preprocesamiento actúan como una caja negra, etiquetando el texto plano con información lingüística básica normalmente requerida en los primeros pasos de cualquier tarea de PLN. Adicionalmente, el archivo vertical puede ser enriquecido con metadatos definidos por el usuario. Para esto, el "Núcleo GECO" provee a los usuarios con herramientas para manejar metadatos, como se explica en la siguiente sección.

Manejo de metadatos

Un aspecto importante de la construcción de un corpus es la captura de metadatos. Dependiendo de la aplicación, los usuarios pueden requerir de cualquier número de anotaciones adicionales a nivel documento. En general, el tipo de información a ser añadida depende del uso que se le pretenda dar al corpus y a qué datos puedan ser útiles para obtener conocimiento adicional de esa colección de documentos en específico (Biber et al., 1998). Para este fin, GECO provee funcionalidad para indicar el tipo de campos de anotación que una colección pueda tener. Los campos son definidos como pares (nombre, valor). La herramienta de metadatos funciona por carpeta. Los archivos dentro de la carpeta obtienen el mismo conjunto de campos. Para asignar los valores de cada campo a cada archivo, GECO ofrece una interfaz parecida a una hoja de cálculo. La Figura 5 muestra un ejemplo de edición de metadatos.

En la tabla, cada campo aparece como una columna, con su nombre como encabezado, mientras que cada fila corresponde a un archivo de la colección. El sistema permite a los usuarios agregar o eliminar campos directamente en esta pantalla, es decir, agregar o eliminar columnas. Una vez que los metadatos están capturados, la vista de documentos de GECO permite ordenar y filtrar con base en estos valores. La Figura 6 muestra cómo un campo definido por el usuario puede ser usado para ordenar una colección.

Proyectos

Los documentos que se cargan a una carpeta a GECO no constituyen un corpus automáticamente. Uno de los aspectos flexibles del sistema dentro de un marco de colaboración, es que los usuarios pueden crear un corpus usando archivos

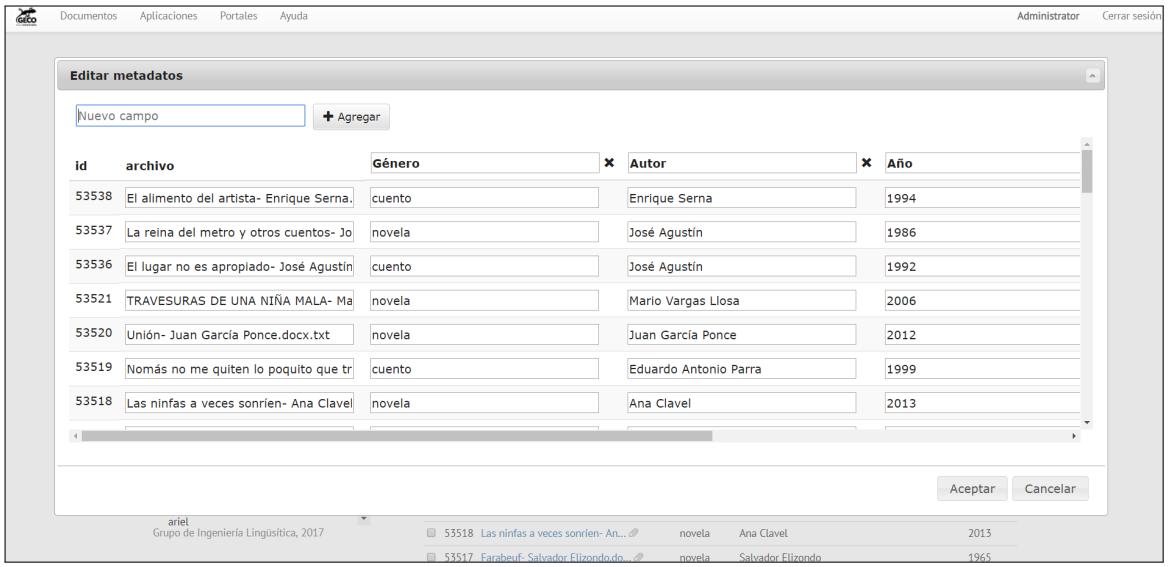

Figura 5: Pantalla de edición de metadatos de GECO.

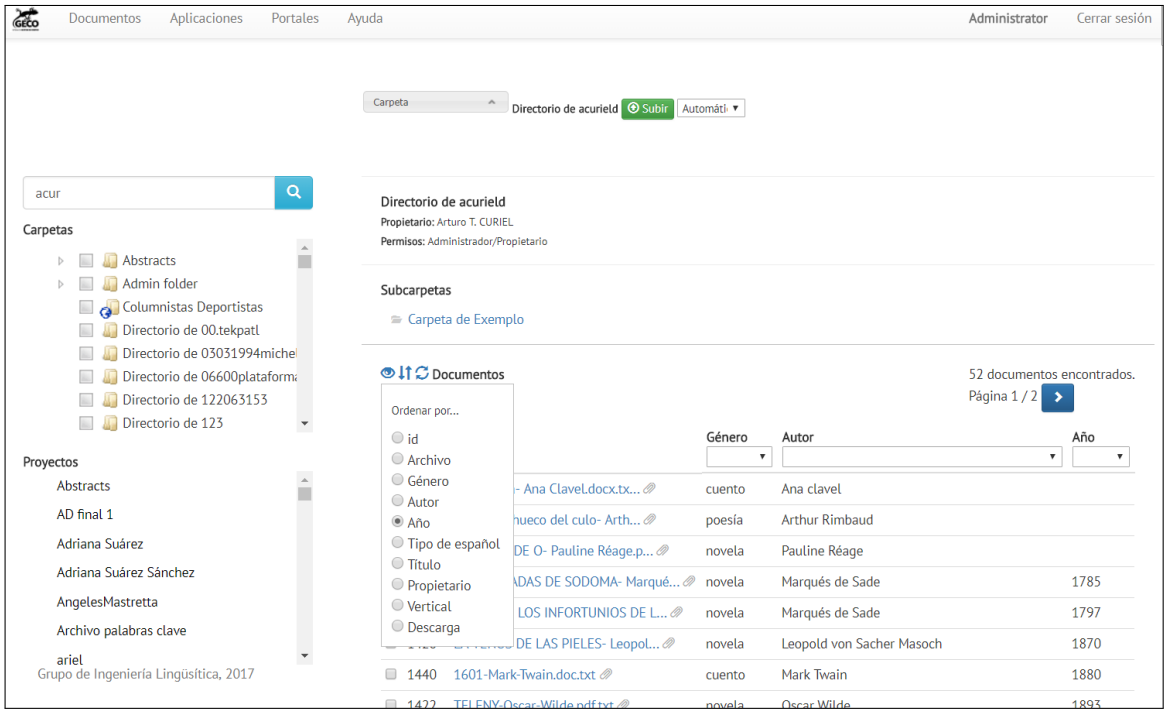

Figura 6: Ordenando una colección por año en GECO.

de varias carpetas. En un caso de uso común, varios usuarios colaborando en la construcción de un solo corpus, pueden tener acceso a diferentes carpetas individuales, donde recolectan sólo un subconjunto del material. No obstante, el corpus requiere ser utilizado como una unidad, aún cuando está contenido en más de una ubicación.

Geco soluciona este problema usando el concepto de "Proyectos", que permite a los usuarios manejar sus corpus como "proyectos". Para el sistema, un proyecto es una colección de documentos, incluso contenidos en diferentes carpetas, unificados bajo un mismo nombre y descripción. Para construir un corpus, el sistema permite a los usuarios escoger archivos individualmente de varias carpetas y encapsularlos en un proyecto, y asignar a un nombre a dicha selección: el nombre del corpus.

Cuando se crea un proyecto, el sistema permite al propietario crear un portal web para el mismo, en el cual los internautas pueden consultar información acerca del corpus, tal como los participantes del proyecto, agradecimientos, publicaciones relacionadas, entre otros. Este portal básico ofrece una estructura genérica para todos los proyectos, que consiste en una serie de pes-

tañas de menú para navegar. Los usuarios pueden elegir la combinación de colores para el portal, la imagen de fondo y el logo que aparecerá en el encabezado de la página. Un ejemplo de la interfaz de configuración del portal se muestra en la Figura 7.

Los proyectos quedan catalogados en el sistema y pueden ser listados mediante el "Web Service GECO". Esto permite a aplicaciones de terceros acceder a colecciones existentes en GECO, siempre y cuando el usuario tenga las credenciales adecuadas para acceder los datos solicitados.

Políticas de seguridad

El manejo de permisos es una de las características más importantes de GECO. Se refiere a qué información los usuarios pueden ver y escribir. Los permisos de lectura y escritura son asignados con base en los tres objetos principales que son: documentos, carpetas y proyectos. Entre estos objetos y cada usuario existirá un "rol", el cual determinará cómo cada usuario puede interactuar con cada objeto.

Los dos principales roles que existen en el sistema son Propietario y Usuario. Los usuarios que tienen el rol de propietario sobre un objeto, pueden otorgar a otros usuarios permisos de acceso sobre ese objeto. Esto es, los Propietarios pueden cambiar las relaciones que existen entre los objetos de los cuales son propietarios, y los demás usuarios. A continuación se describe cómo operan los diferentes permisos más específicamente sobre cada objeto.

Sobre documentos: Los documentos heredan los permisos del contenedor en el que fueron creados: si la carpeta que lo contiene es leíble por un conjunto de usuarios U_f, entonces todo usuario $u \in U_f$ será capaz de leer el documento. Lo mismo pasa con los permisos de un proyecto: si el documento es incluído en el proyecto P, y un grupo de usuarios U_p pueden leer P, entonces todo usuario $u \in U_p$ podrá leer el documento. Lo mismo ocurre con los permisos de creación y escritura. El usuario que crea el archivo (lo carga a la carpeta), obtiene el rol de Propietario. Los propietarios tienen control total sobre sus documentos sin importar la carpeta o proyecto en el que se encuentren. Un documento solo puede ser eliminado por su propietario. Finalmente, un usuario con permisos de lectura sobre el documento lo puede descargar. Es importante que los usuarios verifiquen quién tiene acceso a una carpeta o proyecto antes de incluir un documento en él, para evitar problemas de derechos de autor.

Sobre carpetas: En el caso de las carpetas existe el rol de Colaborador. Adicionalmente, las carpetas tienen dos tipos de nivel de acceso: público y privado. Las carpetas públicas proporcionan a todos los usuarios permiso de lectura sobre las carpetas. Por otro lado, las carpetas privadas solo son visibles a los Propietarios y Colaboradores. Por defecto, toda nueva carpeta es creada con el nivel Privado. El rol de Colaborador proporciona a los usuarios permiso de lectura y permiso de escritura limitado: los colaboradores pueden subir archivos a la carpeta, crear subcarpetas y editar los metadatos de los documentos. Al igual que en el caso de los documentos, solo los Propietarios pueden eliminar la carpeta.

Sobre proyectos: Tienen un comportamiento similar a las carpetas. Como éstas, también pueden ser privados y públicos. Globalmente, los proyectos públicos pueden ser leídos por cualquier usuario, incluyendo usuarios anónimos (no tienen sesión iniciada). Esto significa que cualquier persona puede recuperar los contenidos de un proyecto público, ya sea a través de la interfaz o haciendo una llamada a la API. De manera similar, los proyectos privados solo son visibles para los usuarios con rol Propietario o aquellos que el propietario haya otorgado acceso. Los usuarios con el rol de colaborador pueden agregar o quitar documentos del proyecto (solo quitar del proyecto, no borrarlos completamente). Al igual que en los otros dos tipos de objetos, solo los propietarios pueden eliminar los proyectos. Finalmente, aunque los portales no son un tipo de objeto como tal, son tratados como carpetas en términos de su visibilidad. Hay dos niveles de acceso a los portales: públicos y privados. Los públicos pueden ser visitados por cualquier persona que tenga la URL mientras que las páginas privadas solo pueden ser visitadas por aquellos que tengan permiso explícito, que en el caso de los portales puede ser un usuario en particular o "cualquier usuario con sesión iniciada". Estas políticas pueden ser implantadas a nivel página, de tal manera que la página principal del portal sea accesible a todo el público mientras que otras secciones requieren que el usuario inicie sesión para verlas.

Aplicaciones

El paso final de GECO es el más versátil: implementar aplicaciones sobre el sistema. A grandes rasgos, las aplicaciones son herramientas de software que proveen funcionalidades a los módulos explicados anteriormente. En general, hacen uso de los proyectos GECO expuestos a través de la API, cuyos documentos ya tienen un cierto grado de preprocesamiento, y están listos para ser

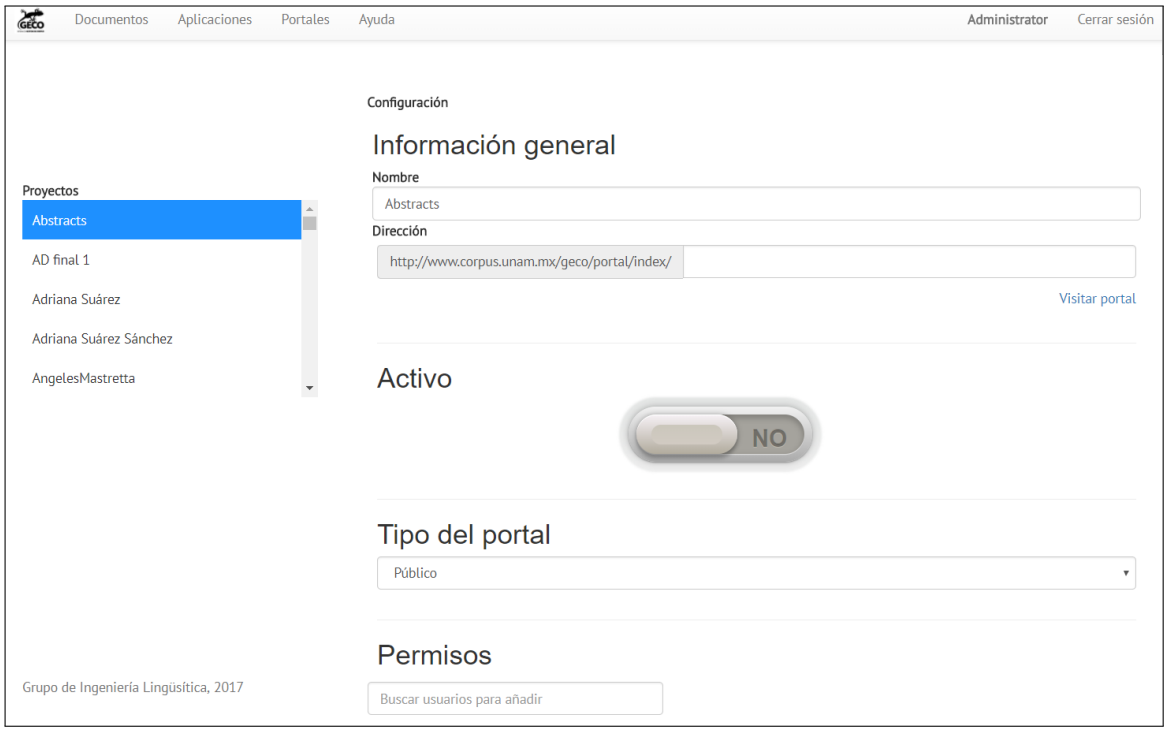

Figura 7: Configuración de portal.

utilizados en otras tareas. Estrictamente hablando, cualquier funcionalidad que estas aplicaciones proveen, no son una funcionalidad de GECO: las aplicaciones son herramientas de software a nivel usuario, diseñadas para aprovechar el contenido de los proyectos GECO.

Como en los proyectos, la API registra y lleva un catálogo de las aplicaciones externas disponibles en GECO y las despliega en la interfaz en el navegador web. La interfaz permite a los usuarios seleccionar cualquier proyecto como entrada a cualquiera de las aplicaciones listadas. Lo contrario también es posible: las aplicaciones pueden ser embebidas en el portal de un proyecto, de tal manera que los internautas pueden hacer uso de las herramientas directamente desde la página del corpus. La Figura 8 muestra un ejemplo de cuatro aplicaciones registradas en GECO, que pueden ser lanzadas con cualquier proyecto de los listados a la izquierda.

Todas las aplicaciones integradas a GECO son expuestas por la API como parte de los recursos disponibles. Esto permite a nuevas aplicaciones llamar programáticamente a las existentes para crear un flujo de trabajo: los nuevos módulos puedes recuperar resultados de otras aplicaciones para usarlos en su propio ciclo de procesamiento. GECO incluye cuatro aplicaciones preregistradas en el catálogo de recursos. Sus funcionalidades se describen brevemente en las siguientes subsecciones.

SAUTEE

El Sistema Automático para Estudios Estilométricos (SAUTEE) es una herramienta que lleva a cabo análisis estilométricos de corpus. Le da al usuario control sobre cuáles marcadores estilométricos utilizar y cómo combinarlos.

A grandes rasgos, los documentos son vectorizados de acuerdo con los marcadores seleccionados y se calcula una distancia entre los vectores resultantes. La salida de esta herramienta es un gráfico de dispersión creado por medio de un escalamiento multidimensional en la matriz de distancias resultante. La imagen producida puede ser usada para inspeccionar cómo los documentos se aglomeran. Para enriquecer el análisis y hacer la visualización más clara al usuario, cada punto del gráfico puede ser coloreado de acuerdo con los metadatos de los documentos. Se presenta un ejemplo en la Figura 9.

TermExt

TermExt es una herramienta de extracción de términos basada en el algoritmo del valor-C (C-value) (Frantzi et al., 2000). La aplicación está diseñada para recibir el contenido de un proyecto registrado en GECO, y extraer de éste una lista de términos, ordenados por puntaje. La Figura 10 muestra un ejemplo de los resultados producidos.

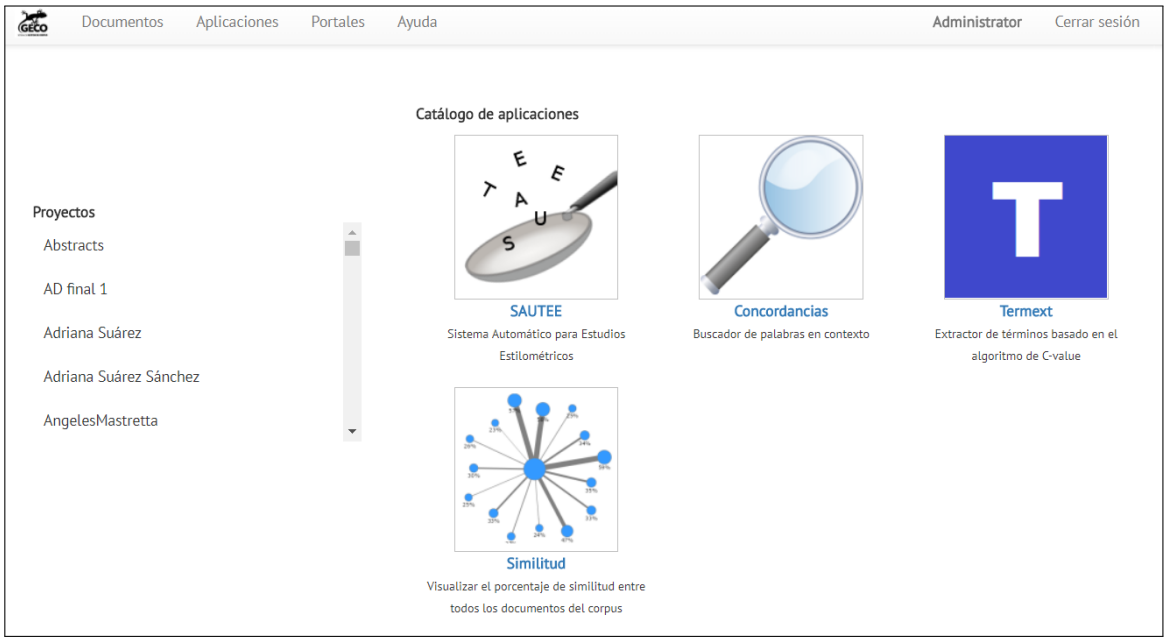

Figura 8: Aplicaciones registradas en GECO, accesibles vía el navegador web.

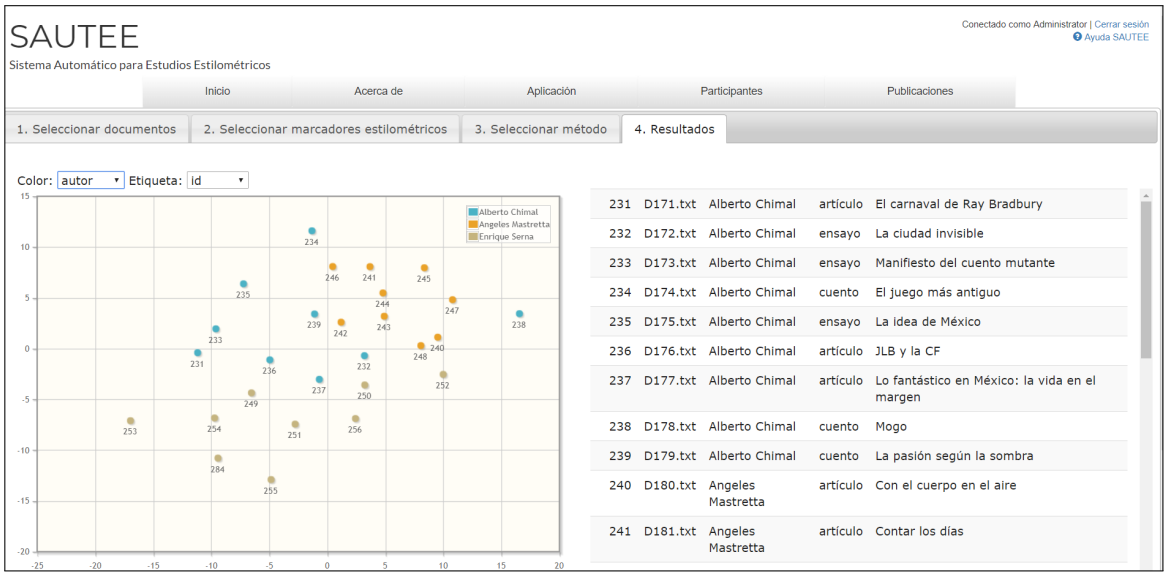

Figura 9: Ejemplo de gráfica generada por SAUTEE.

Similitud

Similitud es una aplicación web que, de manera similar a SAUTEE, calcula una medida de similitud entre todos los documentos de un corpus, solo que la medida está basada en el contenido textual de los documentos, y no en características estilísticas. Su salida es un diagrama que muestra qué tan similar es un documento con respecto a todos los demás. Los valores de similitud son expresados como porcentajes y son desplegados en los nodos de un gráfico de similitud generado por la aplicación. La Figura 11 muestra un ejemplo de los gráficos resultantes.

Concordancias

Concordancias es una aplicación clásica de generación de concordancias que soporta la recuperación tanto de palabras como frases en contexto. Ya que los documentos de GECO están segmentados, lematizados y etiquetados con el etiquetado-POS de Freeling, esta herramienta es capaz de recuperar concordancias en cualquiera de estas tres formas. Adicionalmente, la aplicación puede dividir el corpus en subcorpus más pequeños, filtrando los documentos por sus metadatos. La siguiente sección presenta un análisis más a fondo de está aplicación, así como todos los detalles de su integración con el gestor de corpus.

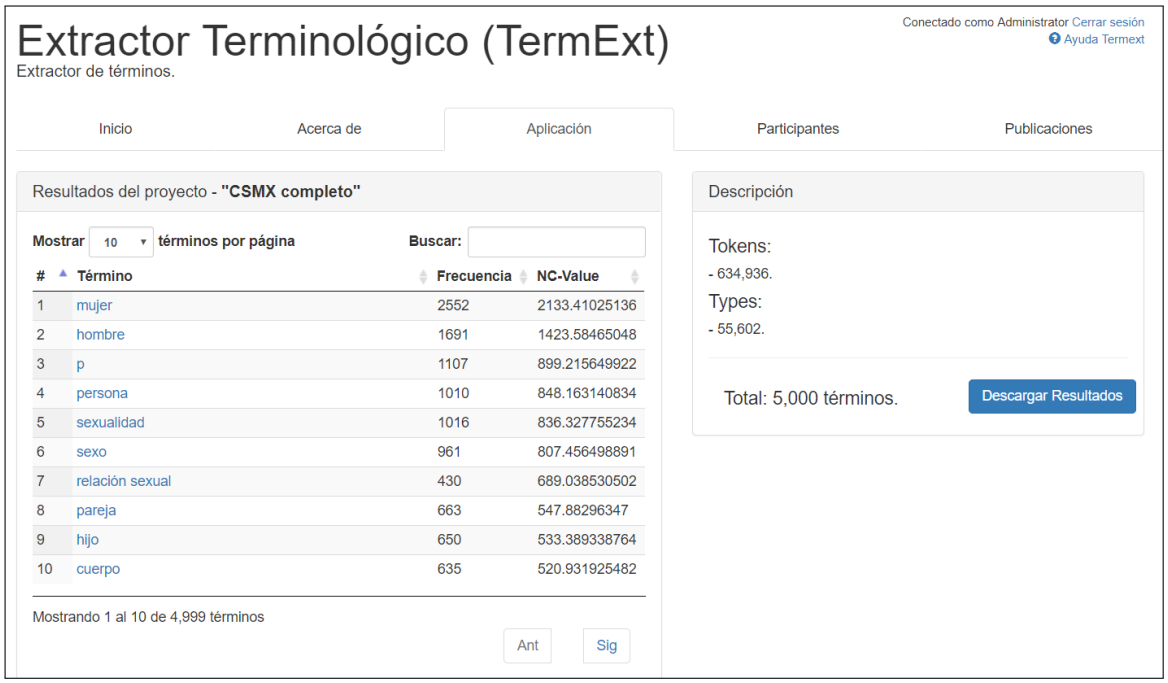

Figura 10: Ejemplo de resultados obtenidos con TermExt.

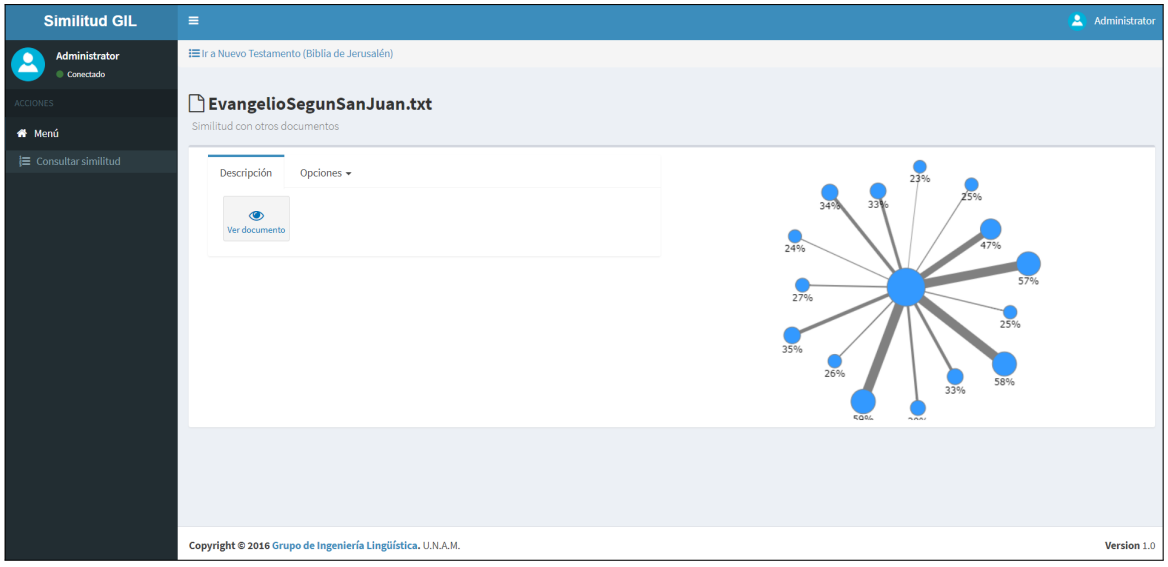

Figura 11: Ejemplo de la gráfica obtenido con Similitud.

5 Aplicación de ejemplo: concordancias

Las concordancias son fragmentos de texto extraídos de un conjunto de documentos, resultado de consultas hechas por un usuario (Manning & Schutze, 1999). Tradicionalmente se muestran en el formato de tres columnas de "Palabra clave en contexto" (Keyword in context o KWIC): la columna de en medio muestra el término buscado (la palabra clave), y las columnas de los lados muestran las palabras que aparecen a la izquierda y a la derecha (el contexto). Los contextos tienen una longitud determinada (tamaño de la ventana), la cual en GECO es un parámetro definido por el usuario. La Figura 12 muestra la pantalla principal de la aplicación.

Selección del corpus

Las concordancias son calculadas a partir de los proyectos registrados en el catálogo de GECO. La aplicación hace una solicitud a la API de GECO, quien retorna una lista de proyectos disponibles y a los que el usuario tenga permiso de accesar según las políticas de seguridad manejadas por GECO, a fin de que pueda seleccionar qué corpus utilizar.

Figura 12: Pantalla de Concordancias.

Por cada proyecto, la aplicación debe hacer un procesamiento propio. Específicamente, se debe crear un índice de los documentos (Zobel et al., 1998). Para esto se recurre a un motor ya mencionado anteriormente en este trabajo: Manatee. Los documentos verticales creados al cargar los documentos a GECO, combinados con los metadatos que hayan sido capturados en los documentos, son enviados como entrada a Manatee para su indexado y compresión. La aplicación de Concordancias guarda este índice localmente, así como la fecha de último indexado, en el caso de que se requiera actualizar el índice si se añadieran nuevos documentos o nuevos metadatos.

Consultas y filtrados

Una vez que los documentos se han indexado, se hace posible ejecutar consultas eficientes. Se propone un lenguaje de consulta especializado que permite buscar por palabra, lema o etiqueta POS. Por ejemplo, uno puede buscar "`perro`" o bien encerrar el término entre corchetes para buscar el lema "`[niño]`" (lo cual traerá resultados para niño, niña, niños, niñas). Se pueden usar diples para buscar una etiqueta POS; por ejemplo, "`<V>`" busca todos los verbos. Los comodines `*` y `?` buscarán cualquier subcadena y cualquier caracter, respectivamente. Estos elementos se pueden combinar; por ejemplo, "`[niño] <V>`" buscará el lema niño seguido de cualquier verbo.

Finalmente, el lenguaje permite búsquedas de proximidad (Goldman et al., 1998), especificando la distancia de una cadena con respecto a otra. Para ello se escribe un número entre llaves, correspondiente a la distancia (en palabras) deseada. Por ejemplo, la consulta "`el niño {3} ayer`" traerá todos los resultados que correspondan a "el niño" seguido entre cero y tres palabras cualquiera más la palabra "ayer".

Además, los metadatos capturados en GECO quedan registrados por manatee como atributos de una etiqueta que envuelve a todo el documento, por lo cual es posible filtrar documentos basados en sus valores. Esto permite efectivamente crear subcorpus en el momento. Para esta clase de filtrado la interfaz presenta selectores de pares campo-valor para restringir el dominio de búsqueda.

Esta sección ha descrito a un alto nivel, cómo es que aplicaciones externas interactúan con los módulos de GECO, mostrando así cómo los principios de diseño de éste promueven la integración. En la siguiente sección se presenta una explicación más técnica del sistema para dar una idea más concreta de su funcionamiento interno.

6 Resumen técnico de la arquitectura de GECO

Las secciones anteriores han presentado el funcionamiento en general de GECO y los principios bajo los que fue diseñado. En esta sección, damos un recorrido más técnico de la funcionalidad del software, dando algunos detalles de implementación de bajo nivel. En particular, los siguientes párrafos describen cómo se guardan los archivos internamente y qué tecnologías son usadas para transportar la información.

Almacenamiento de archivos

Todas las operaciones relacionadas con el almacenamiento de archivos son ejecutadas por Odoo (Reis, 2015), el backend administrativo de GECO y motor del módulo "Núcleo GECO" de la Figura 4.

Odoo es un sistema de gestión empresarial que ofrece varias funcionalidades pre-hechas que permitieron la aceleración del desarrollo. Ejem-

plos de estas funcionalidades por las cuales se incluyó en la arquitectura son: manejo de usuarios (registro y autenticación), visualización de gráficos a partir de cubos de datos (para análisis de estadísticas de uso del sistema), creación de páginas web, y manejo de archivos por medio de una base de datos, lo que se detalla a continuación.

Para insertar nuevos archivos, Odoo asigna un Identificador Único Universal (UUID) a cada archivo cargado, el cual se vuelve su nombre interno. Esto ayuda al sistema a evitar conflictos por nombre, ya que tendrán un nombre único sin importar el nombre del archivo original. Los metadatos de los documentos son almacenados directamente en la base de datos. Con este esquema, solo se leen archivos del sistema de archivos cuando se requieren para procesamiento, mientras que otras operaciones pueden trabajar más rápidamente únicamente manipulando registros de la base de datos.

Base de datos

GECO usa PostgreSQL (Stonebraker & Rowe, 1986), un conocido software de bases de datos relacional. Su función principal es almacenar la información acerca de los usuarios, aplicaciones y permisos de acceso que tienen los objetos en el sistema. También contiene información de los documentos, aunque, como se explicó anteriormente, no los documentos en sí.

Odoo simula un sistema de archivos por medio de los registros de la base de datos. La estructura de las carpetas que se muestra en realidad no es la estructura físicamente en el disco. Odoo construye una representación externa valiéndose únicamente de relaciones entre registros, para evitar operaciones constantes de entrada/salida (lo cual es más lento).

El sistema crea entradas en la base de datos para documentos, carpetas y proyectos, y los relaciona entre sí usando referencias en sus registros. También se guardan referencias que apuntan a la ubicación real del archivo dentro del sistema de archivos.

API

GECO se comunica con aplicaciones externas por medio de una API HTTP mediante la cual provee acceso a todos los recursos y funciones del sistema. La funcionalidad de la API y el web service son provistas por Odoo, el cual recibe peticiones en formato JSON (Bray, 2014). La API permite a aplicaciones externas enviar peticiones JSON al sistema, indicando las operaciones que se desean realizar. Asimismo, permite a las aplicaciones cargar documentos, crear carpetas, integrar proyectos, descargar corpus y modificar metadatos. Está enfocada a los desarrolladores que deseen extender GECO con su propia funcionalidad, en forma de módulos de aplicaciones. La interfaz gráfica de GECO, por ejemplo, está implementada por medio de esta API.

Interfaz gráfica de usuario (GUI)

La GUI de GECO es la manera en que los usuarios finales se comunican con el sistema. Está diseñada para mostrar solo los elementos a los cuales el usuario tiene permiso de accesar. Es accesible a través de cualquier navegador web moderno. La Figura 13 muestra un ejemplo del GUI, en la pantalla de administración de proyectos.

En la figura se puede apreciar que la interfaz muestra una lista de todas las carpetas disponibles en el lado izquierdo de la pantalla. Al hacer click sobre un elemento, se despliegan al usuario sus archivos y subcarpetas. Las carpetas solo se muestran si el usuario tiene permiso de por lo menos lectura. Dentro de las carpetas, los documentos se pueden seleccionar para formar un proyecto o agregarlos a uno existente. Además del control de recursos básico, los menús de navegación de la GUI dan acceso al catálogo de aplicaciones (módulo externos registrados en GECO) y al listado de portales, como se muestra en la Figura 8.

Catálogo de recursos

El catálogo de recursos es donde GECO lista todos los corpus y aplicaciones disponibles, aquellos que pueden ser llamados por medio de la API. Sirve como un índice global por medio del cual las aplicaciones externas que se conecten pueden obtener un listado de todos los corpus y aplicaciones disponibles. Las aplicaciones deben autenticarse para poder acceder a los catálogos, de tal modo que el sistema sabrá qué recursos mostrarle y cuáles no, dependiendo de los permisos de acceso.

Finalmente, el catálogo permite que las aplicaciones se ejecuten de tres modos diferentes:

Normal: Usado cuando el usuario visita el url de la aplicación directamente. En este caso la aplicación debe mostrar un catálogo de corpus para que el usuario pueda seleccionar con cuál proyecto quiere trabajar.

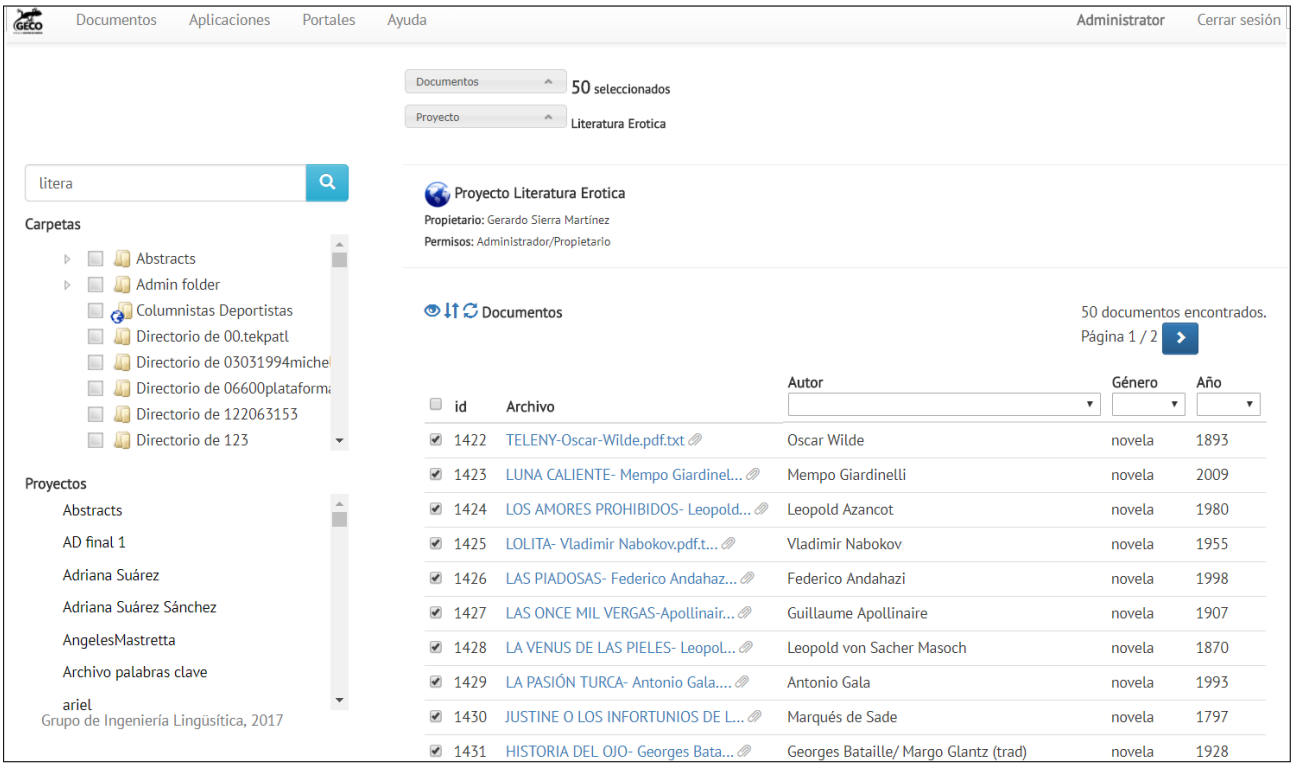

Figura 13: Pantalla de gestión de proyectos en la interfaz de GECO.

Implícito: Usado cuando el usuario lanza una aplicación desde la interfaz de GECO, seleccionando un proyecto de la lista. En este caso la aplicación se abre directamente en el corpus seleccionado.

Embebido: Usado en los portales. Como el implícito, se abre directamente en un corpus preseleccionado. En este caso dicho corpus será aquel que corresponda al proyecto del portal en cual está embebido.

7 Conclusiones y trabajo futuro

En este trabajo se describieron las funcionalidades, principios de diseño y características salientes de un gestor de corpus recientemente desarrollado, GECO, el cual se enfoca en la creación de corpus, y delega la explotación del mismo (por ejemplo, generación de concordancias) por medio de un enfoque modular. También se presentó un módulo para generación de concordancias implementado con Manatee e integrado a GECO por medio de su API.

Las mayores ventajas de GECO sobre otros gestores de corpus es que puede ser usado para construcción de corpus colaborativamente. También tiene la ventaja de que produce corpus anotados, leíbles por máquina, directamente de las fuentes documentales independientemente de su formato, de tal modo que el usuario no tenga que preparar los textos de antemano. Otra funcionalidad que diferencia a GECO de otros softwares existentes es que permite publicar un portal web del corpus, así como embeber aplicaciones como el generador de concordancias en las páginas del portal.

El trabajo futuro incluye agregar soporte para una gran gama de esquemas de anotación. Idealmente GECO debería ser capaz de interpretar cualquier número de atributos-p y atributos-s que pudieran resultar de usar otras herramientas o esquemas de anotación además del análisis básico que actualmente provee Freeling. Esto es un problema abierto en el área, a la vez que propuestas como el modelo de datos Ziggurat (Evert & Hardie, 2015) aún no están estandarizados.

Finalmente, una meta importante para GECO a largo plazo es expandir el catálogo de módulos oficiales. La idea de GECO es serle útil a la comunidad ofreciendo una suite de herramientas de análisis, que de otra manera para un usuario que no sea experto en computadoras le resultarían difíciles de acceder y usar.

Agradecimientos

Este artículo ha sido elaborado gracias a los proyectos PAPIIT IA400117 y Fronteras de la Ciencia 2016-01-2225.

Referencias

Anthony, Laurence. 2005. AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. En *An Interactive Workshop on Language e-Learning (IWLeL'2004)*, 7–13.

Areta, Nerea, Antton Gurrutxaga, Igor Leturia, Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza, Nerea Ezeiza & Aitor Sologaistoa. 2007. ZT Corpus: Annotation and tools for Basque corpora. En *Corpus Linguistis Conference*, s.pp.

Arpaci-Dusseau, Remzi & Andrea Arpaci-Dusseau. 2016. Operating systems: Three easy pieces. Electronic Version 0.91.

Baroni, Marco, Adam Kilgarriff, Jan Pomikálek & Pavel Rychlý. 2006. WebBootCat: a web tool for instant corpora. En *EuraLex Conference*, 123–132.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Birda, Steven, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Josepha, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev & Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. En *Language Resources and Evaluation Conference (LREC'2008)*, 1755–1759.

Bray, Tim. 2014. The JavaScript object notation (json) data interchange format. Internet Engineering Task Force. RFC 7159.

Christ, Oliver. 1994. A modular and flexible architecture for an integrated corpus query system. En *3rd Conference on Computational Lexicography and Text Research (COMPLEX'1994)*, 7–10.

Di Pierro, Massimo. 2011. web2py for scientific applications. *Computing in Science & Engineering* 13(2). 64–69.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. En *Corpus Linguistics Conference*, s.pp.

Evert, Stefan & Andrew Hardie. 2015. Ziggurat: A new data model and indexing format for large annotated text corpora. En *3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, 21–27.

Francis, W. Nelson. 1965. A standard corpus of edited present-day American English. *College English* 26(4). 267–273.

Frantzi, Katerina, Sophia Ananiadou & Hideki Mima. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries* 3(2). 115–130.

Gamallo, Pablo & Marcos Garcia. 2017. LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática* 9(1). 19–28.

Goldman, Roy, Narayanan Shivakumar, Surech Venkatasubramanian & Hector Garcia-Molina. 1998. Proximity search in databases. En *24rd International Conference on Very Large Data Bases*, 26–37.

Hardie, Andrew. 2012. CQPweb: combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3). 380–409.

Kilgarriff, Adam, Fredrik Marcowitz, Simon Smith & James Thomas. 2015. Corpora and language learning with the Sketch Engine and SKELL. *Revue française de linguistique appliquée* XX(1). 61–80.

Kouklakis, George, George Mikros, George Markopoulos & Ilias Koutsis. 2007. Corpus manager: A tool for multilingual corpus analysis. En *Corpus Linguistics Conference*, s.pp.

Manning, Christopher & Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT Press.

Ntoulas, Alexandros, Sofia Stamou, Manolis Tzagarakis, Ioanna Tsakou & Dimitris Christodoulakis. 2001. Viewing web search engines as corpus query systems. En *6th Conference on Computational Lexicography and Corpus Research*, s.pp.

Padró, Lluís & Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. En *Language Resources and Evaluation Conference (LREC'2012)*, 2473–2479.

Reddy, Martin. 2011. *API design for C++*. Morgan Kaufmann.

Reis, Daniel. 2015. *Odoo development essentials*. Packt Publishing.

Rychlý, Pavel. 2007. Manatee/bonito: a modular corpus manager. En *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 65–70.

Sarmento, Luís, Belinda Maia, Diana Santos, Ana Pinto & Luís Cabral. 2006. Corpógrafo V3: From simple word-concordance to semi-automatic knowledge engineering. En *Language Resources and Evaluation Conference (LREC'2006)*, 1502–1505.

Simões, Alberto & Diana Santos. 2014. Nos bastidores da Gramateca: uma série de serviços. En *1st Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish*, 97–104.

Stonebraker, Michael & Lawrence Rowe. 1986. The design of postgres. En *ACM SIGMOD international conference on Management of data*, 340–355.

Vivaldi, Jordi. 2009. Catálogo de herramientas informáticas relacionadas con la creación, gestión y explotación de corpus textuales. *Tradumática* 7. s.pp.

Yergeau, François. 2003. UTF-8, a transformation format of ISO 10646. RFC 3629.

Zobel, Justin, Alistair Moffat & Kotagiri Ramamohanarao. 1998. Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems* 23(4). 453–490.

http://www.linguamatica.com/

linguaMATICA