Volume 6, Número 2- Dezembro 2014

linguaMATICA

ISSN: 1647-0818

UNIVERSIDADE DE VIGO

Universidade do Minho

Volume 6, Número 2 – Dezembro 2014

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Editorial

Chegamos ao noso sexto ano de vida e aos 13 números da revista co nada desprezábel caudal acumulado de 82 artigos publicados, dos cales 50 en portugués, 19 en español, 6 en galego, 4 en euskara, 2 en catalán e 1 en inglés; e cunha media por revista de 6,3 artigos. Nestes seis anos de existencia, logramos mancomunar a máis de 150 autoras e autores de universidades e centros de investigación das dúas beiras do atlántico e de diversos continentes no noso propósito de consolidar unha publicación científica relevante no ámbito do procesamento das linguas da península ibérica e redactada nestas mesmas linguas.

Neste sexto aniversario da Linguamática, queremos agradecer a todas as autoras e a todos os autores a súa elección da nosa revista como plataforma para a divulgación científica do seu traballo. Así mesmo, queremos expresar de novo o noso agradecemento persoal máis sincero ao medio cento de membros que conforman o Comité Científico internacional da revista e que participan semestre a semestre na revisión e selección das propostas recibidas.

Os seis anos da revista talvez puidesen ter ido mellor, mais non se pode dicir de ningún xeito que non fosen bos e proveitosos, nin que faltasen os eforzos para obter os mellores resultados. Esperamos poder seguir levando adiante a Linguamática outros seis anos, cando menos cos mesmos azos, esperando que o vento e a fortuna nos sexan favorábeis.

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,
Universidade Beira Interior

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 Carlos Medeiros,
Porto Editora

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

Lluís Padró,
Universitat Politècnica de Catalunya

Marcos Garcia,
Universidade de Santiago de Compostela

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

Salvador Climent Roca,
Universitat Oberta de Catalunya

Susana Afonso Cavadas,
University of Sheffield

Tony Berber Sardinha,
Pontifícia Univ. Católica de São Paulo

Xavier Gómez Guinovart,
Universidade de Vigo

Revisora convidada: **Itziar Gonzalez-Dios**, Euskal Herriko Unibertsitatea.

Artigos de Investigação

Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua

Basque time structures. Analysis and annotating experiment

Begoña Altuna
Ixa Taldea. UPV/EHU
begona.altuna@ehu.es

María Jesús Aranzabe
Ixa Taldea. UPV/EHU
maxux.aranzabe@ehu.es

Arantza Díaz de Ilarraza
Ixa Taldea. UPV/EHU
adiazdeilarraza@ehu.es

Laburpena

Denbora-informazioa erauztea oso erabilgarria da hizkuntzaren prozesamenduan (HP), besteak beste, testuen sinplifikazioan, informazio-erauzketako eta itzulpen automatikoko sistemetan balia baitaiteke. Lan honetan, euskaraz informazio hori baliagarri bihurtzeko egin diren lehen urratsak azaltzen dira: batetik, euskaraz denbora adierazteko erabiltzen diren egiturak zein eratakoak diren aztertu da gramatiketan oinarrituta, eta bestetik, egitura horiek testuetan etiketatzeko lehen erabakiak hartu dira. Halaber, ekonomiari buruzko corpus bat osatuta egin den etiketatze-lanaren esperimentua azaltzen da.

Gako-hitzak

Denbora-informazioa, denbora-adierazpenak, denborazko erlazio-hitzak, abiarazle lexikoak

Abstract

Time information extraction is very useful in natural language processing (NLP), as it can be used in text simplification, information extraction and machine translation systems. In this paper we present the first steps of making that information accessible for Basque language: on one hand, Basque structures that convey time have been analysed based on grammars and, on the other hand, first decisions on tagging those on real texts have been taken. Also, we give account of an annotating experiment we have carried out on a financial news corpus.

Keywords

Time information, time expressions, time function words, lexical triggers

1 Sarrera

Gizakiok denbora nozioa hitzez adierazteko gaitasuna dugu. Ez dugu mekanismo bakarraren bidez egiten, ordea. Ikus dezagun adibide bat:

> [*Duela 13.800 milioi urte*] [*Big-Banga gertatu zenean*] [*berehala*] inflazio kosmikoa izan zela uste da. Hau da, [*ikaragarrizko denbora tarte txikian*] sekulako hedapena izan zuela unibertsoak. [*Une horretan*] gertatutakoa ez dago [*gaur egungo*] ezerekin alderatzerik. (Berria, 2014-03-18)[1]

Denborako uneak (*Big-Banga gertatu zenean*, *berehala*, *gaur egungo*) edo iraupenak (*ikaragarrizko denbora tarte txikian*) adierazten dituzten denbora-egiturak identifikatu ditugu goiko testuan. *Duela 13.800 milioi urte* irakurriz gero ere, *Big-Bang*aren eta *gaur egun*en artean 13.800 milioi urteko tartea igaro dela inferi dezakegu eta bigarren perpauseko *Une horretan* egiturak zein uneri egiten dion erreferentzia (Big-Bangari) ere erraz atzeman dezakegu. Gertatzen diren ekintza edo egoerek ere denbora-egiturak dira: denek kronologiako une edo tarte batean jazotako gertaera bat adierazten dute eta aditz jokatuek denbora gramatikala adierazten dute (*izan zela*, *izan zuela*, *dago*, etab.).

Hizkuntzaren prozesamenduan (HP), testuaren ulermen automatiko sendoa helburu bada, hizkuntzaren azterketa sakona egin behar da eta azterketa horretan testuko gertaerak noiz jazotzen diren ere aztertu behar da. Denbora-informazioa automatikoki prozesatzea konplexua da, makinek ez baitute gizakiok dugun hizkuntza prozesatzeko berezko gaitasunik. HPrako denbora-egiturak identifikatu eta egitura bakoitzetik ahalik eta informazio gehien esplizitu egin behar da informazio hori makinarentzat ulergarria izateko. Informazio hori HPko hainbat sistematan erabili ahal izango da: informazio-erauzketan, galdera-erantzunetan, testuen sinplifikazioan eta itzulpen automatikoan.

Prozesu horretan, lehenik eta behin, hizkuntzaren azterketa egin behar da, hizkuntza bakoitzean denbora nola adierazten den eta erabiltzen diren egiturak zein diren jakiteko. Infor-

[1] http://paperekoa.berria.info/plaza/2014-03-18/036/002/Uhin_egiaztatzaileak.htm

mazio hori guztia aztertu eta modu sistematikoan adierazi behar da. Horretarako, markaketa-lengoaia baten bidez testuko hitzei edo egiturei euren informazioa azaleratuko duen etiketa bana emango zaie. Hala, denbora-egitura bakoitzak etiketa bat hartuko du eta atributuen bidez denbora-informazioa azaleratuko da. Euskarazko denbora-informazioa makina bidez eskuragarri izateko, TimeML (Pustejovsky et al., 2003a) markaketa-lengoaia erabili dugu, hainbat hizkuntzatan (frantsesa (Bittar, 2010), koreera (Im et al., 2009), italiera (Caselli et al., 2011), etab.) ere erabili eta denbora-informazioa adierazteko estandar bihurtu baita.

Hizkuntza bakoitzak denbora adierazteko berezko erak ditu eta, horregatik, hizkuntza guztietara molda daitekeen etiketatze-lengoaia estandarra sortu bada ere, hizkuntza bakoitzarentzako denbora-egituren analisia beharrezkoa da. Euskarazko denbora-adierazpenei buruzko informazioa hainbat gramatikatan (*Euskal Gramatika Lehen Urratsak (EGLU) I eta II* (Altuna et al., 1985; Altuna et al., 1987), *Euskal Gramatika Osoa* (Zubiri & Zubiri, 1995)) aurki daiteke, baina ez da HPra bideratutako denbora-egituren azterketarik egin. Hutsune hori betetzera dator lan hau.

Lan honetan, euskarazko gramatika horietan oinarrituta, denbora-adierazpenak, denborako une edo tarte bat adierazten dutenak, zerrendatuko ditugu 2. atalean. Jarraian, 3. atalean denboraren tratamendu konputazionala azalduko dugu, denbora-informazioaren prozesamendua azalduz eta TimeML markaketa-lengoaia deskribatuz. 4. atalean, euskarazko denbora-egitura batzuk etiketatzeko hartu ditugun erabakiak eta etiketatze horretan laguntzeko hartu ditugun irizpideak deskribatuko ditugu. 5 atalean, hiru anotatzailek aurrera eramandako anotazio esperimentua azalduko dugu. Amaitzeko, 6. atalean, denboraren prozesamendu automatikoan aurrera egiteko jarraituko ditugun urratsak azalduko ditugu.

2 Denbora-adierazpenen sailkapena

Sarreran aipatu bezala, denbora-informazioa duten formak dira denbora-egitura. Egitura horietako batzuk gertaerak denboran kokatzeko baliatzen direnak dira: denbora-adierazpenak. Atal honetan, euskaraz denborako une bat edo iraupen bat adierazteko erabiltzen diren denbora-adierazpenak deskribatuko ditugu. Denbora-adierazpenek perpausaren denbora-kokapena adierazten dute; hau da, eurek seinalatzen dute perpauseko gertaerari dagokion denbora-unea edo luzapena. Adierazpide horiek guztiek, formaz ugariak badira ere, *noiz?*, *noiztik?*, *noiz arte?*, *noizko?* eta *zenbat denbora?* galderei erantzuten diete eta denbora une edo tarte mugatu bat (iraupen bat) irudikatzen dute.

Egitura horien azterketa beharrezkoa da horiek etiketatzeko jarraituko dugun etiketatze-eskema aukeratzeko eta IXA ikerketa-taldean[2] garatu ditugun tresnen berrerabilera aztertzeko. Hori kontuan hartuta eta gramatiketan oinarrituta (*Euskal Gramatika, Lehen Urratsak (EGLU) I eta II* (Altuna et al., 1985; Altuna et al., 1987), *Euskal Gramatika Osoa* (Zubiri & Zubiri, 1995) eta Sareko Euskal Gramatika[3]), denbora-adierazpenen sailkapena egingo dugu jarraian.

2.1 azpiatalean denbora adierazten duten aditzondo eta adizlagunak ikusiko ditugu. 2.2 eta 2.3 azpiataletan egitura zabalagoak, postoposizio-lokuzioak eta hitz anitzeko unitate lexikalak (HAUL), izango ditugu hizpide eta 2.4 azpiatalean orduei eta datei arreta berezia eskainiko diegu.

2.1 Aditzondoak eta adizlagunak

Denborazko aditzondoek zirkunstantzia bat adierazten dute. EGLUn jasotzen den moduan, adberbioen artean bi sail nagusi bereizten dira irizpide morfologikoak erabiliz: i) aditzondoak, berez eta postposizio-atzizkirik gabe aditzari laguntzen dioten aditz-sintagmako elementuak, eta ii) adizlagunak, izen-sintagmak adizlagun bilakatzeko atzizki edo postposizio jakin batzuen beharra dutenak.

Aditzondoak formaz anitzak dira; Euskaltzaindiak (Altuna et al., 1987) hainbat motatakoak batzen ditu: aditzondo bakunak (*atzo* (1), *maiz* (2) edo *luze* (3)), zehaztugabe konposatuak ((4) adibideko *noizbait* eta (5) adibideko *inoiz*) eta bestelakoak ((6) adibideko *dagoeneko*).

(1) *Atzo* programa berria estreinatu zuten telebistan.

(2) *Maiz* joaten gara Frantziara ostrak jatera.

(3) Ebakuntzak *luze* joko du erizainaren arabera.

(4) *Noizbait* entzuna nuen Mikelek gaur kontatu duen istorioa.

[2] http://ixa.si.ehu.es

[3] http://www.ehu.es/seg/aurkezpena

(5) *Inoiz* horrelako soineko politik dendan ikusten baduzu, eros iezadazu.

(6) *Dagoeneko* euritakoa eta neguko arropa atera behar izan dugu.

Gainera, 2.2. atalean ikusiko dugun moduan, aditzondoak egitura handiagoetan, postposizio-lokuzioetan, ager daitezke eta *gaurtik aurrera* (7) edo *noizean behin* moduko esamoldeak (8) sortu.

(7) *Gaurtik aurrera* ez duela gehiago erreko zin egin du.

(8) *Noizean behin* jatetxe garestietan bazkaltzea gustatzen zaio Maiteri.

Horiez gain, adizlagun itxura duten *askotan* (9), *sasoiz* (10) eta gisakoak aipatu behar ditugu, adizlagun itxura izanagatik ere, aditzondoak baitira.

(9) *Askotan* ahazten ditut eguzkitarako betaurrekoak hondartzara noanean.

(10) Lanera *sasoiz* heltzeko ordubete lehenago esnatu ohi naiz.

Adizlagunak, aditzondoak ez bezala, postposizio-atzizkidun sintagmak dira. Orokorrean leku-denborazko atzizkiak (inesiboa nagusiki (11), adlatiboa (12), ablatiboa (13), leku genitiboa (14)) hartzen dituzte sintagmok, baina batzuetan soziatiboa (15) eta instrumentala (16) ere hartzen dute. Postposizio askeak ere har ditzakete sintagma horiek 2.2. azpiatalean aztertuko dugun bezala.

(11) *Goizean* irten zen etxetik.

(12) Iluntzetik *egunsentira* eten behar izan dituzte erreskate operazioak.

(13) *Gabonetatik* egon zen Mikelen eskutitzaren zain.

(14) Eskatutakoa *sanferminetarako* prest izango duzu.

(15) Andra Mari eguna *ostegunarekin* jausten da aurten.[4]

(16) *Bi orduz* egon naiz zure zain.

Aditzondoek bezala, denborako une (11–15) edo iraupen bat (16) adieraziko dute. Bereizketa hori oso erabilgarria izango da prozesamendu automatikoan ikusiko dugunez, une edo iraupen izan ezberdin tratatuko direlako.

2.2 Postposizio-lokuzioak

Postposizio-lokuzioak "adposizio-sintagma baten burua izateko gauza diren unitate fraseologikoak" (Lorente, 2001) direla esan izan da tradizioan. Postposizio atzizkidun forma osagarri eta postposizio aske batez osatzen dira (17–22).

(17) Kontzertua *hamaikak aldera* hasi zen, iragarritakoa baino ia ordubete beranduago.

(18) *Bi minutu barru* ez bada agertzen, joan egingo gara.

(19) *Bostak irian* gertatu zen istripua.

(20) Eskabidea *epez kanpo* aurkeztu zuen eta ez zioten diru-laguntza eman.

(21) *Berandura arte* esna egoteak osasunari kalte egiten diola adierazi du neurozientzialariak.

(22) *Gaur eta bihar bitartean* amaituko dut.

Ez dira, ordea, egitura guztiz zurrunak. Postposizio-lokuzioetan kontuan izan behar da forma bat baino gehiago izan ditzaketela, nahiz eta esanahiari eusten dioten. Horren adierazle ditugu, esaterako, (23–25) adibideetako *Ostegun arte*, *osteguna arte* eta *ostegunera arte* postposizio-lokuzioak:

(23) *Ostegun arte* egongo gara Bartzelonan.

[4] *Andra Mari eguna* ere denbora-adierazpena da, baina adibide honekin soziatiboaren denbora balioa nabarmendu nahi dugu.

(24) *Osteguna arte* egongo gara Bartzelonan.

(25) *Ostegunera arte* egongo gara Bartzelonan.

Denboraren adierazpenak berezkoak ditu zenbait postposizio-lokuzio: oraintsu aipatutako *alde*, *arte* eta *barru*, adibidez, baina horiek ez dira bakarrak. Usuenak eta euren formak (Aduriz et al., 2008) 1. taulan aurki ditzakegu. Taulan, postposizio askearekin edo beregainarekin batera, forma osagarriaren kasua eta elementu beregainarena berarena ere ageri dira, hala nola, denborazko erabilera adibideak.

Taulako denborazko postposizio-lokuzio arruntenen zerrenda honakoek osatzen dute: *-en aitzin*, *-tik aitzina*, *-en barren*, *-tik goiti*, *-0 irian*, *-z kanpo*, *-0 ondoren*, *-en oste*, *-0 parte* eta *-en pe*. Horien erabilera urriagoa da.

Postposizio-lokuzio horiei aparteko sail bat eskaini nahi izan diegu, elementu osagarriaren burua denbora-adierazpen izango delako eta elementu beregaina "seinale". Seinaleok erlaziohitzak izango dira eta denbora-adierazpena gertaera edo beste denbora-adierazpen batekin lotzeko balio izango dute:

(26) Liburutegia *seiak arte* egongo da zabalik.

(26) adibidean ikus dezakegunez, *arte* elementu beregainak denbora-muga adierazten du, kasu honetan liburutegia zabalik egoteko tartea. *Seiak*ek, bere aldetik, kronologiako une bat adierazten du, liburutegia itxiko den ordua, hain zuzen ere.

2.3 Hitz anitzeko unitate lexikalak (HAUL)

Hitz anitzeko unitate lexikalek arreta berezia hartzen dute, aurretik azaldutako egituren forma badute ere, esanahi konposizionala duten egitura ihartuak baitira. Gure eguneroko jardunean ere oso arruntak dira eta prozesamendu automatikoari begira oso garrantzitsua izango da horiek identifikatu eta unitatetzat hartzea analisi zuzena lortu nahi izanez gero. (27) eta (28) adibideetako denborazko HAULak eta beste hainbat (Urizar, 2012)-an aurki daitezke.

(27) Patata-arrautzopila egiteko *lehenik eta behin* patatak zuritu behar dira.

(28) Gaurko pilota partida *luze gabe* hasiko da, pilotariak berotze-ariketak amaitu baitituzte.

2.4 Orduak eta datak

Esan bezala, orduei eta datei atal berezia eskaintzea erabaki dugu. Egitura horiek erraz identifikatzen diren forma zurrunak hartzen dituzte euskaraz. Prozesamendu automatikoari begira, beste denbora-adierazpenetatik aparte aztertzea komenigarria da erregelen bidez aise erazagut baitaitezke. Normalean adizlagun funtzioa betetzen dute ondoko adibideetan ikus daitekeenez:

(29) *Bostetan* etorri zen menditik.

(30) Autobusa *ordu bata eta hamarrean* pasatzen da nire etxe azpitik.

(31) *15:00etan* hasiko da emanaldia.

(32) Gernikako bonbardaketa *1937ko apirilaren 26an* gertatu zen.

(33) *Bilbon, 2013ko ekainaren 19an.*

(34) *Gaur zortzi* izango da liburuaren aurkezpena.

Baina ezaugarri sintaktikoei dagokienez, perpauseko beste funtzio batzuk ere har ditzakete, subjektua (35) eta objektuarena (36), baita izenlagunarena ere (37)an ageri denez:

(35) *Ordu biak jota* ziren heldu zirenerako.

(36) Gaur *hamalau* ditu hilak.

(37) *Martxoaren 25eko* greba eguna oso jendetsua izatea espero da.

Denbora-egiturak aurkeztu ondoren, 3. eta 4. ataletan azalduko dugu egitura horien analisia nola egin dugun.

3 Ikuspegi konputazionala

Lan honen sarreran aipatu bezala, euskarazko denbora-adierazpenen azterketa hizkuntzaren prozesamenduari begira egin dugu, denbora-informazioaren prozesamenduari begira bereziki. Ataza horretan zer urrats eta baliabide erabili edo erabiliko diren azalduko dugu jarraian.

Postposizio beregaina	**Forma osagarria**	**Elementu beregaina**	**Adibidea**
alde (IZE)	-ABS	-ra	*Hirurak aldera* bazkalduko dugu
arte (IZE)	-ABS	-0/-ko	Ikastaroa *zortziak artekoa* da
	-0	-0	*Bihar arte* ez dago autobusik
aurre (IZE)	-tik	-ra	*Gaurtik aurrera* ez du gehiago erreko
barru (IZE)	-0	-0	*Bi egun barru* entregatuko dut lana
bitarte (IZE)	-ra	-0/-an/-ko	*Etxera bitartean* kontatuko dizut hori
	-ABS	-an	*Bostak bitartean* hemen egongo zarete
buru (IZE)	-en	-an	*Bi egunen buruan* jakin zuen emaitza
gero/geroztik (ADB)	-z	-0/-ko	*Istripuaz gero* ez du ezer gogoratzen
inguru (IZE)	-ABS	-0/-an	*Zortziak inguruan* esnatu gara

1. taula: Denborazko postposizio-lokuzioak.

3.1 Denbora-informazioaren prozesamendua

Denbora-informazioa esplizitu egiteko, lehenik eta behin denbora adierazten duten egiturak identifikatu behar dira. Lan hau denbora-adierazpenei eta zenbait erlazio-hitzi baino ez diegu eskaini, baina gertaerak (gertatzen diren egoera edo ekintzak) ere denbora-egiturak dira. Azken horiek denbora-informazioaren ikuspegitik lantzeke ditugu oraindik.

Denbora-adierazpenak azpimultzotan sailkatuko ditugu: data, ordua, iraupena edo errepikapena, eta definituko ditugun atributu batzuen bidez, euren informazioa azaleratuko dugu. Denbora-erlazioak adierazten dituzten erlazio-hitzak ere identifikatuko ditugu. Denbora-adierazpen eta gertaeren artean denbora-erlazioak sortzen dira; gertaera bat zein unetan gertatu den edo zenbat iraun duen, adibidez, eta erlazio horietako batzuk testuan esplizituki adierazten dira erlazio-hitzen bidez. Denbora-adierazpenak eta erlazio-hitzak etiketatzeko jarraian azalduko dugun TimeML markaketa-lengoaia (Pustejovsky et al., 2003a) erabili dugu.

TimeML baliatuz etiketatu dugu azterketarako lagina. Eskuz etiketatutako lagina abiapuntu hartuta, denbora-informazioa etiketatzeko eskema definitzea eta ikerketa taldeko tresnen berrerabilgarritasuna aztertzea lortuko dugu. Ondoren, erdi-automatikoki etiketatuko litzateke corpusa eta, behin eskuzko gainbegiratzea eginda, denbora-informazioa *gold standard*a izango litzatekeen corpusa sortuko genuke.

3.2 TimeML

TimeML lengoaia naturaleko testuetako denbora-informazioa etiketatzeko markaketa-lengoaia da. Setzerren lanari (Setzer, 2001) eta TIDES proiektuaren barruko TIMEX2 etiketa multzoari (Ferro et al., 2003) jarraipena emanez sortu zen TimeML (Pustejovsky et al., 2003a); horien proposamenak hobetuz eta osatuz denbora-egitura guztiei etiketa bat esleitzeko etiketatze-eskema sendoa proposatu zen. TimeML XML (eXtensible Markup Language) markaketa-lengoaiaren gainean garatu zen eta hizkuntza guztietan aplikagarri egin zen. Egun ISO-TimeML markaketa-lengoaia (ISO-TimeML working group, 2008), International Organization for Standarization-ek (ISO) estandartzat hartutakoa, hainbat hizkuntzatan erabili izan da. TimeMLn oinarrituta zenbait hizkuntzatan denboraren arabera etiketatutako corpusak sortu dira, TimeBank 1.1 (Pustejovsky et al., 2003b) eta TimeBank 1.2 (Pustejovsky et al., 2006) besteak beste.

TimeMLk bai uneak, bai iraupenak etiketatzeko balio du eta etiketatutako denbora-adierazpenek `<TIMEX3>` etiketa hartzen dute, baita euren informazioa esplizitu egingo duen hainbat atributu ere, hala nola, mota (`type`), denborako unea edo denbora-tartearen luzera adierazten duen zenbakizko balioa (`value`) eta hasiera (`beginPoint`) eta amaiera puntuak (`endPoint`) iraupenen kasuan.

Lengoaia horretan, erlazio-hitzak etiketatzeko `<SIGNAL>` etiketa erabiltzen da. Gertaerek, "gertatzen diren egoerak" (Pustejovsky et al., 2003a), `<EVENT>` etiketa hartuko dute eta denbora-erlazioek, denbora-adierazpen eta gertaeren artean sortzen diren erlazioek, `<TLINK>`.

Denbora adierazpen mota	Adibideak
DATE	sanferminak, 2014ko martxoaren 20an, etzi
TIME	arratsaldeko bostetan, 17:26, gaur goizean
DURATION	50 urte, bost minutuz, bi asteko
SET	egunero, astean birritan

2. taula: Denbora-adierazpenen motaren araberako sailkapena.

Euskararako ere estandar hori baliatzea erabaki dugu, egun denbora markaketa-lengoaia osoen eta aurreratuena baita.

TimeML euskarara moldatzeko lehen urratsa deskribatzen dugu; alegia, markaketa-lengoaia hori erabilita euskarazko denbora-egiturak identifikatzeko lehen urratsak. Hala, 2. atalean, gramatiketan oinarrituta sailkatu ditugun denbora-adierazpenak TimeMLk proposatzen duen sailkapenaren arabera antolatzen saiatuko gara orain. TimeMLk denbora-adierazpenak lau motatan sailkatzen ditu: DATE (eguna bezain luze edo luzeagoak diren datak), TIME (eguna baino laburragoak diren denborak), DURATION (iraupena) eta SET (errepikapena) (2. taula). Seinaleak 4.1.1 azpiatalean azalduko ditugu.

4 Euskarazko denbora-adierazpenen etiketatzea

Lan honen sarreran esan bezala, testuko denbora-adierazpenak etiketatzeko, TimeML markaketa-lengoaian oinarritutako etiketak baliatuko ditugu. Jarraian, etiketa horiek erabilita sarrerako testuko denbora-egituren informazioa agerian utzi dela ikus dezakegu:

```
<TIMEX3 tid="t1" type="DURATION"
        value="P13800000000Y"
        beginPoint="-13799997986"
        endPoint="2014">
Duela 13.800 milioi urte </TIMEX3>.
Big-Banga gertatu zenean berehala inflazio
kosmikoa izan zela uste da.
Hau da, <TIMEX3 tid="t2" type="DURATION"
         value="PTXS">
ikaragarrizko denbora tarte txikian </TIMEX3>
sekulako hedapena izan zuela unibertsoak.
<TIMEX3 tid="t3" type="TIME" value="PAST_REF">.
Une horretan </TIMEX3> gertatutakoa ez dago
<TIMEX3 tid="t4" type="DATE" value="PRESENT_REF">
gaur egungo </TIMEX3> ezerekin alderatzerik.
(Berria, 2014-03-18)
```

Denbora-egiturek etiketak hartu dituzte eta etiketa horien barruko atributuek egituren ezaugarriak (mota, balioa, hasiera-puntua, etab.) adierazten dituzte. Identifikatzailea (tid) esleitu diegu lehenik eta behin, eta gero deskribapenezko atributu dei ditzakegunak: type, value, beginPoint eta endPoint. Informazio hori guztia <TIMEX3> XML etiketa baten barruan batu dugu, ondoren automatikoki prozesagarria izateko.

Ikus daitekeenez, ez diegu denbora-adierazpen guztiei etiketa eman (*Big-Banga gertatu zenean* eta *berehala*). Egitura horiek denborazkoak badira ere, lehen urrats honetan ez ditugu etiketatu, denborazko menderagailuak nola tratatu eta zein balio esleitu erabakitzeko azterketa-lan zehatzagoa egin behar delako. *Berehala* moduko aditzondoei balioa esleitzea zaila da eta horiek ere azterketa sakonagoaren beharra izango dute.

Jarraian, aurreko testutik erauzitako denbora-adierazpen baten etiketatze adibide bati helduko diogu. (38) adibidean denbora-adierazpen baten denboraren araberako etiketatzea ikus daiteke; zein atributu hartzen duen eta atributuok zein balio hartzen duten.

(38)
```
<TIMEX3 tid="t1" type="DURATION"
value="P13800000000Y"
beginPoint=-13799997986"
endPoint="2014» Duela 13.800 milioi urte
</TIMEX3>
```

(38) adibideko egitura denbora-adierazpena dela adierazteko, <TIMEX3> etiketa esleitu zaio, baita "tid"identifikatzailea ere. Iraupen bat adierazten du egitura horrek eta hala adierazi da "DURATION"mota esleitzean eta, iraupena izaki, hasiera puntua eta amaiera puntua ere esleitu zaizkio. Denbora-adierazpenaren balio kronologikoa valueren bidez islatzen da; kasu honetan, iraupen bat denez, balio horrek denbora tartea adieraziko du. Halaber, beginPoint eta endPoint atributuek iraupen horren hasiera eta amaiera puntua adieraziko dituzte.

4.1 Seinaleak eta abiarazle lexikoak

Denbora-informazioaren azterketan, beste elementu batzuk ere izan behar dira kontuan. Jarraian denborazko erlazio-hitzak (seinaleak) eta denbora-adierazpenak identifikatzeko baliagarri

izango diren denborazko abiarazle-lexikoak azalduko ditugu. Lehenek denborazko erlazio baten berri ematen dute eta bigarrenek denbora-adierazpenak antzematen laguntzen dute, berez denbora-adierazten duen elementu bat edo gehiagoz osatuta baitaude.

4.1.1 Seinaleak

Seinale deituko diegu denbora erlazio-hitz edo egiturei, ingelesez TimeML lengoaian (Pustejovsky et al., 2003a) "signal"erabiltzen baita horiek izendatzeko eta hurbileko kalkoa erabiltzea egokia dela uste dugulako. Horiek perpauseko gertaeren artean edo denbora-adierazpenen artean sortzen diren erlazioak adieraziko dituzte. Hainbat kategoriatakoak izan daitezke seinaleak:

- denborazko postposizio-lokuzioak:[5] *-ABS/-0 arte*, *-ra/-ABS bitarte*, *-z gero*, *-tik aurre*.
- denborazko lokailuak: *bitartean*.
- denborazko adberbioak eta adberbio lokuzioak: *ondoren*, *eta gero*, *baino lehenago*.
- karaktere bereziak: - eta /

(39) Sanferminak astebete *barru* amaituko dira, baina ordura *arte* ez da aspertzeko betarik izango.

(40) Patatak frijituko ditugu. *Bitartean* arrautzak irabiatuko ditugu.

(39) adibidean *-0 barru* postposizio-lokuzioa dugu *astebete* denbora-adierazpena eta *amaituko dira* gertaeraren arteko erlazioa azalduz. Kasu honetan amaiera unea inferi dezakegu: iterazio mementoa baino astebete beranduago. *Arte*k, ordea, *ordura* denbora-adierazpena eta *ez da izango* egoera lotzen ditu baldintza berria noiz hasiko den adieraziz. (40) adibidean, *frijitu* eta *irabiatu* gertaeren arteko denbora-erlazioa nolakoa den adierazten du *bitartean*ek.

Funtzio bera betetzen dute denborazko mendeko perpausetako erlazio-atzizkiek, *-(e)n*, *-(e)la*, baina TimeMLn denbora-informazioaren etiketatzea hitzaren mailan egiten denez (ez morfema mailan), horiek nabarmentzea oso zaila da eta alde batera uztea erabaki dugu. Ondorioz, erraztasunaren izenean, denborazko erlazio-atzizkiak ez markatzea erabaki dugu.

[5]Etiketatzean postposizioen elementu beregaina baino ez da etiketatzen.

4.1.2 Abiarazle lexikoak

Abiarazle lexikoak (Ferro et al., 2003)-en arabera, "*a word or numeric expression whose meaning conveys a temporal unit or concept*"(denborazko unitate edo kontzeptu bat adierazten duen hitzezko edo zenbakizko adierazpenak) dira. Euskaraz ere, berez denbora adierazten duten esapideak aurki ditzakegu: *ordu*, *gaur*, *ekain*, eta abar. 2.1 ataleko adibideetan ikusi ahal izan denez, denbora-adierazpenek berez denbora adierazten duen unitatea izango dute buru askotan ((11) adibideko *goizean* eta *gauera*, (14) adibideko *sanferminetarako*, (15) adibideko *ostegunarekin* eta (16) adibideko *orduz*). Abiarazle lexiko horiek identifikatzea garrantzitsua izango da hizkuntzaren tratamendu automatikoa egitean, egiturok berez denbora adierazten baitute.

Forma horiek orokorrean denbora adierazten duten esapideetan agertzen dira. Ondorioz, abiarazle lexiko horiek aurretik deskribatutako egituraren batean agertuz gero, denbora adierazten dutela, izan denbora-une edo iraupen bat, suposa dezakegu. Hartara, euskarazko hainbat abiarazle lexiko zerrendatu ditugu (ikusi eranskina). Denbora adierazten duten hitzon zerrenda zabaldu eta osatzeko, euskarazko WordNet (Pociello, 2008) kontsultatu dugu, ingelesezko *time* eta euskarazko *denbora* hitzen euskarazko hiponimoak bilatuz. Besteak (Altuna et al., 1987) eta (Ferro et al., 2003)-ko adibideetatik itzulita edo moldatuta lortu ditugu. 3. taulan abiarazleotako lagin bat kategoria gramatikalaren arabera sailkatu dugu.

Hala ere, denbora-adierazpen guztiek ez dute abiarazle lexiko bat hartzen (*askotan* (9), *bostetan* (29)) eta hori ere kontuan hartzekoa da egiturok automatikoki nola erauzi proposatzean. Egiturok denborazkoak direla erabakitzeko testuinguruaren beharra izango dugu. (41) adibidean *bostetan* denborazko adizlaguna da eta ordua adierazten du; (42) adibideko *bostetan* sintagmak, ordea, lekua:

(41) *Bostetan* geratu ginen kafea hartzeko.

(42) Bost etxe zituen eta *bostetan* jarri zuen jacuzzia.

Bestalde, 3. taulan adierazitako abiarazle lexikoak denbora adierazle moduan hartu izan diren arren, gerta daiteke testuinguru batzuetan hitzok denborarik ez adieraztea:

(43) Arratsalde on!

Kategoria	**Abiarazle lexikoak**
Izen arruntak	aro, arratsalde, asteburu, astelehen, eguerdi, etorkizun, mende, minutu, ordu, otsail, sanfermin, sasoi, seiurteko, solstizio, udazken
Izen bereziak	Gabon, San Joan
Adjektiboak	eguneroko, hilabetekari, bienal
Denbora-patroiak	13:03, 2014/02/12, 1992ko
Adberbioak	gaur, berandu, orain, lehen, berehala, egundo
Zenbakiak	5etan, 6an

3. taula: Kategoriaren araberako abiarazle lexikoen sailkapena.

(44) Lan hau lau pertsonaren artean egitekoa da, *orduan* (= ondorioz) Mikeli deitu behar izango diogu lana amaitu nahi badugu.

(43) adibideko *arratsalde* denborazkotzat ez hartzea erraza izan daiteke unitate fraseologikoen azterketa eginez gero, esaldiko bi tokenak egitura fosilizatu batean agertzen baitira. (44) adibideko *orduan*, ordea, askoz ere zailagoa izango da hitz anbiguoa baita.

5 Esperimentazioa

Arestian aipatu bezala, denbora-adierazpen eta seinaleen deskribapenarekin batera, etiketatze saiakera bat egin dugu. NewsReader proiektuaren[6] metodologia jarraituz, hiru anotatzailek euskarazko 4 kazetaritza-testu etiketatu dituzte denbora-adierazpenak eta denbora erlaziohitzak (seinaleak) markatuz. Testuak euskarazko egunkari bateko albisteak dira eta enpresa baten itxiera eta horren erosketari buruzkoak dira. Erabili dugun tresna CAT (CELCT Annotation Tool) (Bartalesi Lenzi, Moretti, & Sprugnoli, 2012) izan da, denbora-informazioa etiketatzeko aproposa baita.

Etiketatzaileek guztira 56 esaldi aztertu dituzte euskararako denbora-egiturak etiketatzeko osatu diren etiketatze-gidalerroak jarraituz (Altuna, Aranzabe, & Díaz de Ilarraza, 2014). Gidalerro horiek TimeML markaketa-lengoaian eta NewsReader proiektuko eskuzko etiketatze-irizpideetan (Tonelli, Sprugnoli, & Speranza, 2014) oinarritutakoak izan dira. Etiketatze tresna, halaber, proiektu horretan erabiltzen ari direna da.

[6]http://www.newsreader-project.eu/

5.1 Anotatzaileen arteko adostasuna

Anotatzaileen arteko adostasuna neurtu dugu ondoren, testuan ezarritako etiketa kopurua eta etiketa horien luzera (token berak hartzen dituzten) kontuan hartuta. Horretarako, Diceren koefizientea (Dice, 1945) erabili dugu denbora-adierazpen eta seinaleen gaineko adostasuna neurtzeko (ikusi 4. eta 5. taulak). Diceren koefizienteak etiketentzako token berak aukeratu diren neurtzen du adostasun osoa islatuz.

4. eta 5. tauletan etiketaren gaineko adostasuna ikus daiteke. *Markable*k[7] etiketaren luzera zehatzaren gaineko adostasuna adierazten du eta *token*ek token batzuetan adostasuna izan dela. Denbora-adierazpenetan oso hurbileko emaitzak lortu izanak egiturak identifikatzean anotatzaileek adostasun handia izan dutela adierazten du. Seinaleen kasuan, token bakarrekoak izanik orokorrean, *markable* eta *token*entzat emaitza berak lortu dira.

Testuan ezarritako `<TIMEX3>` etiketekin batera, bost `<TIMEX3>` etiketa huts ere etiketatu dituzte. Etiketa hutsak informazio inplizitua azaleratzeko erabiltzen diren testuaz kanpoko etiketak dira. Horien gaineko adostasuna oso eskasa izan da A anotatzaileak 5, Bk 1 eta Ck 0 etiketatu baititu. Hiru anotatzaileek jarri dituzten etiketak, testukoak zein hutsak, guztira 33 izan dira eta adostasuna % 66,7koa izan da. Seinaleen aldetik, 16 izan dira markatu direnak eta % 31,2ko adostasuna lortu da hiru anotatzaileen artean.

Denbora-adierazpenen balio eta motaren gaineko adostasuna ere neurtu da. 6. taulan `type` eta `value` atributuen gaineko adostasuna ikus dezakegu, bai bikotekakoa bai orokorra.

Emaitza orokorrak emateaz aparte, binakako ebaluazioak egin ditugu, adostasuna anotatzaile guztien artean maila berekoa den edo ez aztertzeko.

[7]Markatu nahi den edozein entitate.

Anotatzaile bikoteak	**Micro-average (Markable)**	**Micro-average (Token)**	**Macro-average (Markable)**	**Macro-average (Token)**
A – B	0.96	0.976	0.969	0.977
A – C	0.943	0.965	0.923	0.965
B – C	0.902	0.94	0.892	0.942
Guztira	0.935	0.96	0.928	0.961

4. taula: Denbora-adierazpenen (`TIMEX3`) anotatzaileen arteko adostasuna.

Anotatzaile bikoteak	**Micro-average (Markable)**	**Micro-average (Token)**	**Macro-average (Markable)**	**Macro-average (Token)**
A – B	0.75	0.75	0.79	0.79
A – C	0.556	0.556	0.479	0.479
B – C	0.444	0.444	0.393	0.393
Guztira	0.583	0.583	0.554	0.554

5. taula: Denbora erlazio hitzen (`SIGNAL`) anotatzaileen arteko adostasuna.

5.2 Emaitzen ebaluazioa

Etiketatutako denbora-adierazpen eta seinaleak alderatuz gero, adostasuna non lortu den edo ez ikusi ahal izan dugu. Alde batetik, *markable* bati etiketa bera jarri zaion neurtu dugu eta beste aldetik, etiketa horiek hedapen bera (token berak) hartu duten. Denbora-adierazpenen kasuan emaitza onak lortu dira, adostasun osoa beti izan da 0,89tik gorakoa testuan jarritako etiketetan, baina etiketa hutsak kontuan hartuz gero, emaitzek behera egin dute. Seinaleen kasuan emaitzak nabarmen baxuagoak dira. Anotatzaileak bat ez etortzeko arrazoi nagusiak ondokoak izan dira:

- Denbora-adierazpen edo seinale bat etiketatu ez izana edo etiketa huts osagarriak sortu ez izana.
- Etiketatu behar ez zen tokena etiketatu izana. Hau da, denbora-egituratzat hartu da denbora-egitura ez dena.
- Tokenei etiketa ezberdina eman izana. Denbora-egitura zuzen identifikatu, baina etiketa okerra eman izana.
- Denbora-adierazpenen atributuei balio ezberdinak eman izana.

Emaitzen ebaluaziorako, anotatzaileak batu eta adostasunaren edo ez-adostasunaren arrazoiak aztertu ditugu. Gidalerroen interpretazio okerra izan da arazorik handiena; eta, ondorioz, anbiguotasuna edo argitasun falta agertzen zuten atalak berrikusi eta zuzendu ditugu. Halaber, gidalerroak idaztean kontuan hartu ez ziren denbora-egiturak, lagintzat hartutako corpusean agertu ez zirelako, gehitu ditugu, egituren zerrenda osatzeko. Horrekin batera, denbora-egitura batzuk, "jadanik"eta parekoak, esaterako, alde batera uztea erabaki dugu, ez baitute denbora-informazio berririk gehitzen eta beste hizkuntzetan ere albo uzten direlako. Aldaketa horien bidez espero dugu seinaleetan nagusiki hobekuntza nabaritzea, horien etiketatzea mugatu baita gehien aldaketak egitean.

Etiketa hutsen sorrera eta erabilera ere arazo iturri izan da. Horien bidez, testuan esplizitu agertzen ez den, baina deduzi daitekeen denbora-informazioa adierazten da. Gidalerroetan horien erabilera argiago azaldu dugu eta adibide praktikoak ere egin ditugu anotatzaileen lana hobetzeko.

6 Ondorioak eta etorkizunerako lanak

Lan honen helburua euskarazko testuen ulermen automatikoan urratsa egitea izan da. Horretarako, denbora-informazioa adierazten duten euskarazko denbora-adierazpenak eta erlazio-hitzak aztertu ditugu. Behin azterketa hori eginda, elementu horiek etiketatzeko erabiliko den TimeML denbora markaketa-lengoaia (Pustejovsky et al., 2003a) aukeratu dugu beste hizkuntzatan erabili delako eta euskararako ere egokia delako.

Eredu horri jarraituz, eta zenbait egokitzapen egin ondoren, euskarazko denbora-egiturak kodetzeko baliatuko den etiketatze-eskema definitu dugu. Etiketatze-eskema horren egokitasuna egiaztatzeko, hiru etiketatzailek 56 esaldiko lagin bat etiketatu dute NewsReader proiektuaren metodologia jarraituz eta CAT tresna (Bartalesi Lenzi, Moretti, & Sprugnoli, 2012) erabi-

Anotatzaile bikoteak	A–B	A–C	B–C	Orokorra
`type`	0.76	0.64	0.75	0.55
`value`	0.76	0.68	0.46	0.45

6. taula: `type` eta `value` atributuen gaineko adostasuna.

lita. Emaitzak ebaluatu ondoren, definitutako etiketatze-eskeman zenbait zuzenketa egin ditugu argi ez zeuden edota anbiguotasuna sortzen zuten atalak zehazteko. Lan horren ondorioz, gidalerroen lehen bertsioa (Altuna, Aranzabe, & Díaz de Ilarraza, 2014) osatu dugu.

Etorkizuneko lanen artean, gidalerro horiek zabaltzea aurreikusi dugu; izan ere, tamaina handiagoko corpus bat osatzea dugu helburu eta corpus horren azterketak denbora-egituren kasuistika gehituko du.

Corpus horretan, denbora-egitura guztiak (denbora-adierazpenak, erlazio-hitzak eta gertaerak) eta horien artean sortzen diren denbora-erlazioak markatuko dira denboraren arabera finkatutako irizpideak jarraituz eta garatutako etiketatze-tresnak baliatuz. Corpus hori *gold standard* modura erabiltzea eta eskuragarri jartzea da gure asmoa.

Amaitzeko, etiketatze-lanetan laguntzeko, berez denbora adierazten duten hitzen zerrenda egin dugu, baina abiarazle-lexikoen zerrenda hori osatzeko asmoa ere badugu. Lan hau idatzi bitartean TempoWordNet (Dias et al., 2014) garatu da. Datu-base horretan *synset* bakoitzari denbora-informazioa esleitu zaio. Abiarazle lexikoak bezala, TempoWordNet-eko informazioa denbora-egiturak identifkatzeko baliagarria izango dela uste dugu. Denboraren araberako analisi ona lortzeko, baliabide hori nola integratu ere erabaki behar izango dugu.

Eskerrak

Ikerketa lan hau Eusko Jaurlaritzak emandako ikertzaileak prestatzeko doktoretza-aurreko A modalitateko bekari esker (Erref. zk.: PRE_2013_1_959) egin da.

Eskerrak Itziar Aldaberi corpusa osatzeko bere laguntzagatik eta Arantxa Otegiri formatu arazoetan hain baliagarri izateagatik.

Eskerrak errebisatzaileei lan honen hobekuntzan hartutako lanagatik.

Eranskina - ABIARAZLE LEXIKOAK

Jarraian abiarazle lexiko izendatu ditugun formen adibideak batu eta sailkatuko dira.

- IZEN ARRUNTAK

  Abendu, abuztu, aldi, apiril, aro, arrats, arratsalde, aste, astearte, asteazken, astelehen, asti, azaro, betikotasun, bider, continuum, denbora, eguerdi, egun, ekain, etorkizun, garai, gau, gaur-bihar, gero, geroaldi, goiz, goizalde, hamarkada, hilabete, igande, iluntze, infinitu, iragan, iraganaldi, irail, iraunaldi, iraupen, larunbat, lehen, lehenaldi, maiatz, martxo, memento, mende, milurteko, minutu, momentu, negua, orain, orainaldi, ordu, oren, ostegun, ostiral, otsail, sasoi, segundo, uda, udaberri, udazken, une, urri, urtaro, urtarril, urte, uztail.

- IZEN BEREZIAK

  Gabon, San Joan.

- ADJEKTIBOAK

  Berantiar, egunkari, goiztiar, hilabetekari.

- ADBERBIOAK

  Arestian, aspaldi, atzo, aurki, aurrenik, aurtemein, aurten, bart, behiala, behin, behin edo behin, behinik behin, berandu, berriki, berritan, beti, bihar, birritan, edonoiz, egun, etzi, etzidamu, gaur, gaurgero, gero, geurtz, goiz, goizik, harrezkero, herenegun, hirutan, honezkero, horrezkero, iaz, inoiz, lehen, lehen baino lehen, lehenbailehen, lehenengo eta behin, lehenik, luzaz, maiz, noizbait, noizbehinka, noizean behin, noiznahi, ondoan, orain, oraino, orduan, ostean, ostera, sarri, sasoiz, sekula, usu.

- DENBORA-ESAPIDE BEREZIAK

  2014ko apirilaren 15ean, 2015/09/19, 12:40.

Bibliografia

Aduriz, Itziar, Izaskun Aldezabal, María Jesús Aranzabe, Jose Mari Arriola, Klara Ceberio, Ainara Estarrona, Mikel Iruskieta, Mikel Lersundi, Elisabete Pociello, Larraitz Uria, Ruben Urizar, & Edurne Aldasoro. 2008. Euskarazko postposizio-lokuzioen tratamendu konputazionala. Txosten teknikoa, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV / EHU LSI / TR 07-2008. `http://ixa.si.ehu.es/Ixa/Argitalpenak/`

Barne_txostenak/1220881618/publikoak/Postposizioen%20barne-txostena.

Altuna, Begoña, María Jesús Aranzabe, & Arantza Díaz de Ilarraza. 2014. Euskarazko denbora-egiturak etiketatzeko gidalerroak. Txosten teknikoa, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV / EHU LSI / TR 01-2014. http://ixa.si.ehu.es/Ixa/Argitalpenak/Barne_txostenak/1414871293/publikoak/Denbora-egiturak%20etikeatzeko%20gidalerroak.

Altuna, Patxi, Pello Salaburu, Patxi Goenaga, María Pilar Lasarte, Lino Akesolo, Miren Azkarate, Piarres Charriton, Andolin Eguskitza, Jean Haritschelhar, Alan King, Jose Mari Larrarte, Jose Antonio Mujika, Beñat Oyharçabal, & Karmele Rotaetxe. 1985. *Euskal Gramatika Lehen urratsak (EGLU) I.* Euskaltzaindiko Gramatika batzordea, Euskaltzaindia, Bilbo.

Altuna, Patxi, Pello Salaburu, Patxi Goenaga, María Pilar Lasarte, Lino Akesolo, Miren Azkarate, Piarres Charriton, Andolin Eguskitza, Jean Haritschelhar, Alan King, Jose Mari Larrarte, Jose Antonio Mujika, Beñat Oyharçabal, & Karmele Rotaetxe. 1987. *Euskal Gramatika Lehen Urratsak (EGLU) II.* Euskaltzaindiko Gramatika Batzordea, Euskaltzaindia, Bilbao.

Bartalesi Lenzi, Valentina, Giovanni Moretti, & Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 333–338, Istanbul, Turkey. European Language Resources Association (ELRA).

Bittar, André. 2010. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard.* Ph.D. thesis, Université Paris Diderot, Paris. http://www.linguist.jussieu.fr/~abittar/docs/Bittar-PhD.pdf.

Caselli, Tomasso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, & Irina Prodanof. 2011. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151, Association for Computational Linguistics, Portland, Oregon, USA.

Dias, Gael, Mohammed Hasanuzzaman, Stéphane Ferrari, & Yann Mathet. 2014. TempoWordNet for Sentence Time Tagging. In *Proceedings of the 4th ACM Temporal Web Analytics Workshop (TEMPWEB) associated to the 23rd International World Wide Web Conference*, pages 833–838, Seoul, South Korea. International World Wide Web Conferences Steering Committee.

Dice, Lee Raymond. 1945. Measures of the Amount of Ecologic Association between Species. *Ecology*, 26:297–302.

Ferro, Lisa, Laurie Gerber, Inderjeet Mani, Beth Sundheim, & George Wilson. 2003. TIDES 2003 Standard for the Annotation of Temporal Expressions. Txosten teknikoa, MITRE, McLean, USA, September. http://www.mitre.org/sites/default/files/pdf/ferro_tides.pdf.

Im, Seohyun, Hyunjo You, Hayun Jang, Seungho Nam, & Hyopil Shin. 2009. KTimeML: Specification of Temporal and Event Expressions in Korean Text. In *Proceedings of the 7th workshop on Asian Language Resources in conjunction with ACL-IJCNLP 2009*, pages 115–122, Suntec City, Singapore. Association for Computational Linguistics.

ISO-TimeML working group. 2008. Language resource management — Semantic Annotation Framework (SemAF) — Part 1: Time and events. International Standard ISO/CD 24617-1(E), ISO. http://lirics.loria.fr/doc_pub/SemAFCD24617-1Rev12.pdf.

Lorente, Mercé. 2001. Altres elements lèxics. In *Gramàtica del Català Contemporani.* Empúries, Barcelona, pages 831–888.

Pociello, Elisabete. 2008. *Euskararen ezagutzabase lexikala: Euskal WordNet.* Ph.D. thesis, Euskal Filologia Saila, Euskal Herriko Unibertsitatea, Leioa. http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1204622545/publikoak/2008Tesi-txostena-eranskinak-aurkezpena.rar.

Pustejovsky, James, José M Castaño, Robert Ingria, Roser Saurí, Robert J Gaizauskas, Andrea Setzer, Graham Katz, & Dragomir R Radev. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

Pustejovsky, James, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David

Day, Lisa Ferro, & Marcia Lazo. 2003b. The TimeBank Corpus. In D. Archer, P. Rayson, A. Wilson, & T. McEnery, editors, *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster, UK. UCREL, Lancaster University.

Pustejovsky, James, Marc Verhagen, Roser Saurí, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, & Andrea Setzer. 2006. TimeBank 1.2. Txosten teknikoa, Linguistic Data Consortium. `https://catalog.ldc.upenn.edu/LDC2006T08`.

Setzer, Andrea. 2001. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study.* Ph.D. thesis, University of Sheffield, Sheffield, UK. `ftp://ftp.dcs.shef.ac.uk/home/robertg/theses/setzer_thesis.pdf`.

Tonelli, Sara, Rachele Sprugnoli, & Manuela Speranza. 2014. NewsReader Guidelines for Annotation at Document Level. version 4.1. Txosten teknikoa, Fondazione Bruno Kessler. `http://www.newsreader-project.eu/files/2013/01/NWR-2014-2.pdf`.

Urizar, Ruben. 2012. *Euskal lokuzioen tratamendu konputazionala.* Ph.D. thesis, Euskal Filologia Saila, Euskal Herriko Unibertsitatea, Donostia. `http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1342621075/publikoak/TESIA`.

Zubiri, Ilari & Entzi Zubiri. 1995. *Euskal Gramatika Osoa.* Didaktiker, Bilbao.

Avaliação de métodos de desofuscação de palavrões

Evaluation of profanity deobfuscation methods

Gustavo Laboreiro
Faculdade de Engenharia da Universidade do Porto - DEI - LIACC
gustavo.laboreiro@fe.up.pt

Eugénio Oliveira
Faculdade de Engenharia da Universidade do Porto - DEI - LIACC
eco@fe.up.pt

Resumo

Os palavrões são uma forma de expressão notável pela sua intensidade. Quando uma pessoa recorre a este tipo de expressão emite uma opinião ou avaliação espontânea e crua, que muitas vezes é suprimida em nome dos "bons costumes" e sensibilidades. Acontece que esta forma de expressão tem também valor aquando de certas análises de opinião e de sentimentos, que são operações já feitas de forma rotineira nas redes sociais. Daí que neste trabalho procuramos avaliar os métodos que permitem recuperar e reconhecer estas formas de expressão que foram disfarçadas através de ofuscação, muitas vezes como forma de evasão à censura automática.

Keywords

palavrões, obscenidades, conteúdo gerado pelo utilizador.

Abstract

Cursing is a form of expression that is noted by its intensity. When someone uses this form of expression they are emitting a spontaneous and raw form of opinion, usually suppressed for the "mild ways" and sensitive people. As it happens, this sort of expression is also valuable when doing some sort of opinion mining and sentiment analysis, now a routine task across the social networks. Therefore in this work we try to evaluate the methods that allow the recovery of this forms of expression, disguised through obfuscation methods, often as a way to escape automatic censorship.

Keywords

profanity, cursing, user-generated content.

1 Introdução

A *Web 2.0* trouxe um novo paradigma à Internet, ao encorajar os utilizadores a partilhar conteúdo de sua própria autoria, a deixar os seus juízos e avaliações, e até a interagir com os autores e outros comentadores através de conversas e trocas de impressões. Mas os responsáveis pela plataforma têm de zelar belo seu bom funcionamento, e garantir que alguns utilizadores mal-comportados não afastam o público-alvo. Esta situação é mais significativa quando o conteúdo primário é da responsabilidade dos donos da plataforma (como é o caso de jornais *on-line*).

Um dos abusos que se pretende evitar é o uso de discurso insultuoso, e em particular o recurso aos palavrões. Tem sido este o objetivo dos estudos sobre o tema: procurar palavras presentes num léxico "proibido". Mas há muito mais que os palavrões podem dizer-nos a diversos níveis, e que tem sido ignorado.

O nosso objetivo é procurar uma forma de permitir que os palavrões sejam usados como um recurso adicional na análise automática de utilizadores ou mensagens. Mas para tal é necessário ser capaz de os identificar e reconhecer no meio do texto ruidoso, mesmo que estejam "disfarçados". Neste trabalho iremos analisar a principais formas de abordar este desafio, os seus problemas e os seus resultados.

Iremos explorar primeiro os palavrões e os seus usos, assim como os trabalhos científicos que os abordaram, a fim de estabelecer o contexto em que o presente trabalho se insere. De seguida será apresentada a coleção anotada que serviu de base à análise que é feita, a coleção "SAPO Desporto", que se encontra disponível on-line. A metodologia da experiência é apresentada na Secção 4, onde detalhamos os processos que foram considerados relevantes para a correta identificação dos palavrões, e como podem afetar os resultados. Na Secção 5 detalhamos os nossos testes, cujos re-

sultados acompanhamos dos nossos comentários e análise. Concluímos o nosso trabalho sumariando as nossas conclusões e apresentamos as nossas propostas para trabalho futuro.

1.1 O que são os palavrões, e qual a sua utilidade?

Neste trabalho definimos palavrões como palavras socialmente convencionadas indecentes que são usadas com intenções ofensivas ou vulgares. A origem, a disseminação, a interpretação e o uso de palavrões podem ser estudados em várias áreas, como a psicologia, a linguística, a antropologia e a sociologia.

Sabemos que alguns estados emocionais podem ser expressos de forma adequada apenas através do uso de palavrões (Jay e Janschewitz, 2007), em particular a frustração, a fúria e a surpresa (Jay, 2009) ou a tristeza e a fúria (Wang et al., 2014). Daí que será expectável que os palavrões sejam significantes na análise automática de sentimentos ou opiniões (Constant et al., 2009).

O uso de palavrões tende também a variar em função de diversos fatores associados ao orador, nomeadamente com a sua idade, sexo, identidade de grupo, fatores de personalidade e classe social (Thelwall, 2008; Jay, 2009), e como tal pode ajudar na elaboração de perfis de utilizadores.

1.2 Quão frequente é o uso de palavrões?

É claro que a resposta a esta pergunta depende de muitos fatores — quem, onde, em que contexto, etc.. Mas a título inicial podemos começar pelo estudo de Mehl e Pennebaker, que analisaram 4 dias de gravações contendo as conversas e interações de 52 estudantes universitários (Mehl e Pennebaker, 2003), resultando na estimativa de que 0,5% das palavras proferidas eram palavrões. Este estudo estabelece uma base de comparação que podemos usar na análise dos três estudos seguintes, que abordam o uso de palavrões em três comunidades on-line diferentes.

1.2.1 No MySpace

Em 2006, Thelwall fez um estudo relativo à rede social MySpace, onde comparou o uso de palavrões entre 9376 perfis, divididos entre utilizadores dos Estados Unidos da América e do Reino Unido (Thelwall, 2008). O autor encontrou palavrões nas *homepages* da maioria de jovens (considerando como jovens todos os utilizadores com idade igual ou inferior aos 19 anos), sendo que

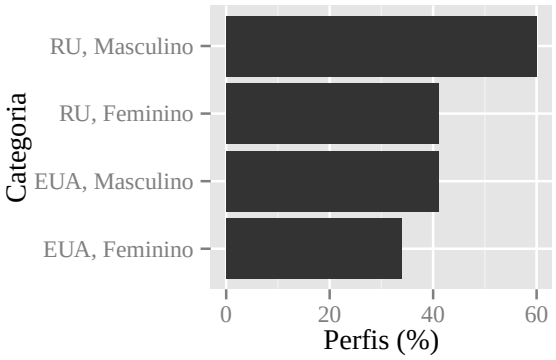


Figura 1: Proporção dos 9376 perfis do MySpace analisados que continham profanidade (média, grave ou muito grave), agrupados por sexo e nacionalidade.

a nível geral quase 40% dos perfis apresentavam este vocabulário.

A Figura 1 ilustra a prevalência de palavrões entre os dois sexos e as duas nacionalidades em estudo, mostrando que esta linguagem é mais prevalecente entre os utilizadores do sexo masculino e na comunidade do Reino Unido. A grande maioria dos palavrões encontrados foram classificados como sendo "fortes", como mostra a Figura 2. Os palavrões moderados tinham maior presença no Reino Unido (onde é 2 a 3 vezes mais comum), enquanto que o recurso a linguagem "muito forte", apesar de comparativamente rara, se encontrava também muito mais presente na comunidade britânica.

O estudo de Thelwall focou-se mais nos palavrões fortes e muito fortes aquando do estudo da frequência das palavras. Com base neste vocabulário mais restrito aponta como rácio máximo de palavrões por palavra empregue de 5% de palavrões para um utilizador do Reino Unido, ou seja, observaram um britânico empregava um palavrão a cada 20 palavras na sua *homepage*. Para um norte-americano o rácio máximo que foi visto foi de 11%, o que é significativamente superior. No entanto, estes valores extremos não são muito representativos, como se pode perceber pela Tabela 1, que compila os valores médios.

1.2.2 No Twitter

Mais recentemente, em 2014, Wang et al. estudaram 14 milhões de utilizadores do Twitter, que entre si englobavam 51 milhões de mensagens em inglês (Wang et al., 2014). Nelas, 7,73% continham pelo menos um palavrão (1 em 13 *tweets*), sendo que estes eram observados com uma frequência de 1,15% referente ao total de pala-

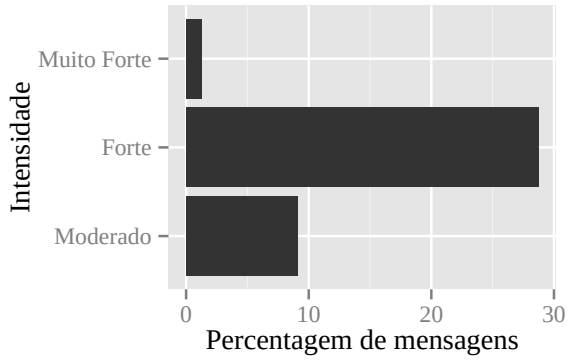


Figura 2: Classificação do nível de intensidade dos palavrões encontrados nas mensagens analisadas do MySpace.

País	Classe	Percentagem
RU	Geral	0,2
	Masculino	0,23
	Feminino	0,15
EUA	Geral	0,3
	Masculino	0,3
	Feminino	0,2

Tabela 1: Percentagem média de palavras consideradas como palavrões fortes ou muito fortes, nos perfis amostrados do MySpace.

vras escritas — o que faz uma média de um palavrão por cada 87 palavras observadas. Dado que 1% das palavras tipicamente usadas em conversas orais nessa língua são pronomes pessoais na primeira pessoa (e.g. "we", "us", "our") (Mehl e Pennebaker, 2003), e os linguistas não os consideram como termos raros, podemos afirmar que este tipo de linguagem tem presença significativa no Twitter.

Os autores referem também que os palavrões são associados principalmente com emoções negativas. As mensagens com palavrões exprimiam mais emoções como tristeza ou fúria, e muito poucas abordavam o amor, como transcrito na Tabela 2. Por outro lado, como apresentado na Tabela 3, mais de um em cada cinco *tweets* zangados contém palavrões, estando comparativamente muito ausentes das mensagens positivas.

	Tristeza	Fúria	Amor
Com palavrões	21,83	16,79	6,59
Sem palavrões	11,31	4,50	nd

Tabela 2: Percentagem de *tweets* contendo ou não palavrões que exprimem as seguintes emoções.

Emoção	Percentagem
Fúria	23,82
Tristeza	13,93
Amor	4,16
Agradecimento	3,26
Alegria	2,50

Tabela 3: Percentagem de *tweets* exprimindo diversas emoções que contêm palavrões.

1.2.3 No Yahoo! Buzz

Sood, Antin e Churchill publicaram vários trabalhos relacionados com o estudo da linguagem e comunidades on-line. Baseado na análise de 6500 mensagens extraídas do Yahoo! Buzz (Sood, Churchill e Antin, 2011), uma comunidade social centrada na submissão e comentário de notícias on-line, derivada do Digg. Este estudo tem a particularidade de ter recorrido à plataforma de *croudsourcing* da Amazon, o Mechanical Turk, para analisar as mensagens.

Os restantes estudos aqui citados limitaram-se a procurar por palavras presentes num léxico, assim como algumas variantes, o que lhes permitiu abordar grandes quantidades de dados rapidamente. Contudo, os palavrões ou suas variantes que não foram considerados no léxico (e há sempre vários — abordaremos também a questão da ofuscação mais à frente) não foram contabilizados. Logo, há um problema de *cobertura*, ou seja, há a forte possibilidade de haver *falsos negativos* (por vezes chamados de "erros do Tipo II"). Os números referidos tendem por isso a ser mais comedidos e a pecar por diferença.

Sood e seus colaboradores recorreram a anotadores que classificam todas as mensagens, revelando o número de palavrões existentes na coleção (salvo a discordância de interpretação entre os anotadores). No entanto, a consequência é que as coleções tendem a ser de dimensão menor, e consequentemente, menos representativas da população.

Os resultados desta análise (Sood, Antin e Churchill, 2012a) indicam que das 6354 mensagens (146 não obtiveram consenso entre os anotadores), 9,28% continham 1 ou mais palavrões. Infelizmente não foram apresentados valores para a percentagem de palavras que eram palavrões — possivelmente porque a anotação foi feita ao nível da mensagem. A Figura 3 apresenta os resultados referentes a 10 secções de notícias, mostrando que este tipo de vocabulário era mais prevalente na secção de política e menos usado nos comentários desportivos.

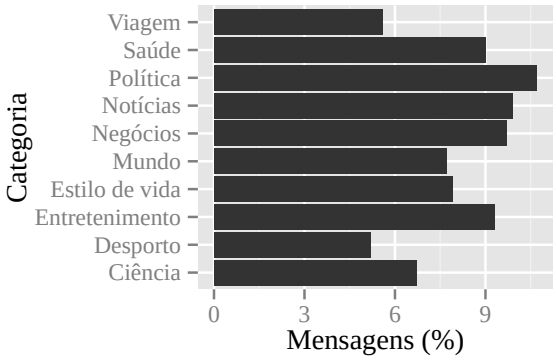


Figura 3: Proporção das mensagens nas diversas secções no Yahoo! Buzz que continham profanidade.

Por esta altura podemos concluir duas coisas: os palavrões têm uma presença significativa nas interações on-line, e que eles variam de acordo com uma série de fatores — como o país, o sexo, a idade, os sentimentos e os assuntos em discussão. Esta variação não se reflete apenas na frequência, mas também nas palavras que são empregues.

1.3 Quantos palavrões existem?

Pode afirmar-se à partida que é impossível compilar uma coleção que contenha todos os palavrões, e por vários motivos.

Como tudo o que é de cultura tradicionalmente oral, os palavrões evoluíram de forma divergente e apresentam múltiplas variações. Basta consultar uma compilação de expressões carroceiras (Almeida, 2014) para perceber que o contexto e a interpretação são muito importantes na identificação e compreensão de muitas palavras e expressões. Muitas palavras inócuas podem ser combinadas numa expressão insultuosa — e, por vezes, via este processo uma palavra pode começar a adquirir uma conotação mais "mal vista".

Focando-nos mais em palavras individuais, os palavrões são encontrados raramente na forma escrita em publicações de referência (como dicionários, livros e imprensa); e a pouca exposição à sua grafia correta favorece a confusão sobre como se escrevem. Assim, torna-se mais provável a uma pessoa escrever mal um palavrão de forma acidental. Por exemplo, escreve-se "murcão" ou "morcão"? A consequência desta dúvida comum é a facilidade de se ir encontrando, ao longo do tempo, o mesmo palavrão escrito de várias formas diferentes, tornando indistinta a forma correta. A este problema devemos ainda acrescentar as variações regionais e culturais que são transversais a todo o português. Desta forma cada palavrão acumula um número variado de formas de ser escrito.

Aliadas também às questões etnográficas há ainda a questão que temos ignorado até aqui de que a nossa definição de palavrão está sujeita a *interpretação*. Por exemplo, no Brasil "veado" é um termo insultuoso associado à homossexualidade ou efeminação, e o mesmo acontece com muitos termos que lhe são associados, como "Bambi" ou o número 24 — o número do veado no Jogo do Bicho[1]. Em oposição, a cultura Portuguesa tem o veado como um símbolo antigo de masculinidade. Outro exemplo de interpretação diferente de é a palavra "cu", que pode ser considerada indelicada ou mal vista por algumas pessoas, enquanto que outras a usam sem qualquer hesitação, independentemente do contexto social em que se encontram.

Sood, Antin e Churchill abordam de certa forma esta questão, referindo que o contexto dita qual o sentido do uso de uma palavra, e como pode ditar quando uma palavra é ou não aceitável (Sood, Antin e Churchill, 2012a). Por exemplo, "cornudo" é um termo aceitável quando o tema são animais, como discussões de zoologia, biologia ou pecuária; mas noutros pode visto como insultuoso. Nem sempre é fácil deduzir o contexto em que um termo é empregue, e até lá não podemos assumir que um sistema automático de deteção de palavrões baseados em listas consiga ter sucesso pleno.

Devemos também recordar que os palavrões estão sujeitos às mesmas variações que afetam todas as outras palavras, tais como género, número, diminutivos, aumentativos, conjunções, etc.. Existe também a possibilidade de criar novas palavrões através de aglutinação (Thelwall, 2008), embora seja uma ocorrência mais comum no inglês. Contabilizando todos estes fatores, mesmo começando com uma lista de palavrões modesta, facilmente se expande até às muitas centenas de elementos se tentarmos contabilizar as variantes possíveis.

Se não é fácil reconhecer de forma automática um palavrão dadas todas as formas como ele pode ser escrito quando o autor *quer* que ele seja percebido, o problema fica muito mais complicado quando o autor *não quer* ser óbvio, e ergue propositadamente uma barreira à sua compreensão. Iremos de seguida abordar a questão da ofuscação de palavrões, que traz complicações acrescidas a todos os processos que lidam com os palavrões.

[1] Uma espécie de lotaria popular no Brasil, que é ilegal mas tolerada pelas autoridades.

1.4 O que é a ofuscação?

Chamamos ofuscação ao desvio ou alteração propositada da grafia de uma palavra da sua versão canónica. Podemos enunciar vários motivos que podem levar o autor a recorrer à ofuscação na elaboração do seu texto.

Estilístico Escrita que invoca uma determinada pronúncia ou forma de falar (por exemplo: "Bibó Puorto!");

Diferenciador Escrever de forma diferente a fim de mostrar uma certa identidade de grupo. Por exemplo, o chamado "leet speak" (1337);

Crítica A alteração mais ou menos grave, normalmente de um nome, de forma a acarretar algum juízo de valor sobre o mesmo. Por exemplo, "Micro$oft" para indicar que considera que a empresa liga apenas a questões financeiras, ou "Pinócrates" para sublinhar que José Sócrates um mentiroso na sua opinião;

Auto-censura Para quando uma pessoa não quer ser perfeitamente explícita, como por exemplo "m*rda", ou quando se pretende contornar mecanismos de controlo baseados no conteúdo das mensagens, como filtros anti-spam (e.g. "V!@gra");

Abreviação Quando não quer escrever todos os carateres, como em "FdP".

Dado que as intenções dos autores nem sempre são aparentes, não temos sempre conhecimento se dada palavra foi escrita de forma diferente propositadamente ou de forma acidental. Como do ponto de vista do nosso trabalho essa distinção não é significativa, iremos considerar todos os desvios gráficos na escrita de palavrões como tentativas de ofuscação.

1.5 Porquê apostar na desofuscação?

O principal foco dos trabalhos que lidam com palavrões é saber se determinada mensagem tem ou não palavrões, ou quanto muito se determinada palavra é um palavrão. Isto porque a tarefa em causa é *filtrar* mensagens — daí o grande foco no léxico. O nosso objetivo é *reconhecer* os palavrões, isto é, se este aglomerado de símbolos for um palavrão, que palavrão é?

A diferença que este passo pode fazer no âmbito da análise automática de comunidades, sentimentos, opiniões ou outra forma de estudo assente sobre conteúdos textuais gerados por utilizadores, é que se conserva (recupera, até se poderá dizer) informação adicional emitida pelo autor. Se alguém se refere ou dirige a uma pessoa usando um palavrão — vamos imaginar, a título de tecer um cenário, que estudamos a imagem de uma figura pública nas redes sociais, e sabemos apenas que alguém usou um palavrão em referência a essa figura pública — então há muito pouca informação que se pode inferir. No entanto, se reconhecermos o palavrão podemos perceber que pode estar a referenciá-la como homossexual, como traída na sua relação amorosa, como desprovida de valor ou como sendo sexualmente desejada, para referir apenas alguns exemplos. O que procuramos fazer é aumentar a cobertura de situações desse género com que conseguimos lidar. Observando os números que dispomos sobre a ofuscação de mensagens na Internet (Laboreiro e Oliveira, 2014), ilustrados na Figura 5, estes números podem, em alguns casos, ser significativos.

Existem outras tarefas semelhantes à desofuscação, na medida em que visam "regularizar" textos, como é o caso da correção de erros ortográficos, que trata mudanças de grafia acidentais sobre um léxico muito mais vasto. A desofuscação e a correção de erros são, por natureza, tarefas de *normalização*, ou seja, tornar a grafia das palavras previsível, constante e com o mínimo de variação possível.

2 Estado da arte

Os palavrões, do ponto de vista da informática, têm sido abordado poucas vezes e com poucos avanços. Numa análise dos principais problemas dos sistemas de deteção de palavrões (Sood, Antin e Churchill, 2012a), os autores avançam que a aparente facilidade desta tarefa, que é muito difícil em contextos reais, pode ter contribuído para que esta área tenha sido negligenciada em termos de investigação.

Iremos abordar três vertentes associadas a este problema, que cobrem essencialmente os trabalhos científicos que tratam a questão dos palavrões: trabalhos simples baseada em listas, o tratamento de mensagens insultuosas ou ofensivas, e estudos linguísticos que abordam o uso de palavrões. No entanto nenhum destes trabalhos se propôs a tratar o mesmo problema que abordamos neste trabalho.

2.0.1 Trabalhos baseados em listas

Estes são os trabalhos mais simples. O problema que se tentava resolver pode ser descrito de forma muito sintética: há uma lista de palavras que deviam *nunca* surgir em mensagens lidas ou escritas pelos utilizadores. O objetivo era pois filtrar (ou no mínimo identificar) estas mensagens corretamente. Na verdade, a responsabilidade residia toda na componente lexical, e como tal o funcionamento destes sistemas era muito elementar.

Inicialmente estas aplicações destinavam-se a proteger as crianças da exposição a linguagem rude na Internet (Jacob et al., 1999), e a manter a devida compostura nas empresas (Cohen, 1998) — os anos 90 trouxeram novas ferramentas eletrónicas que, aparentemente, inspiravam informalidades indevidas nos locais de trabalho.

A patente norte-americana identificada pelo número 5 796 948 (Cohen, 1998) documenta a sua invenção como "um método para a interseção de comunicações ou correspondência em rede, que contenha palavras ofensivas ou profanas, fragmentos de palavras, frases, parágrafos ou outra unidade de linguagem que possa ser formulada em qualquer linguagem natural (...) ou artificial." O documento é bastante omisso na descrição do método usado para reconhecer esse tipo de ofensas (não era este o foco principal da patente), mas tudo indica que consiste unicamente em encontrar substrings na mensagem que estejam enunciadas numa lista pré-determinada.

No ano seguinte, Jacob et al. abordaram um sistema de classificação automática de websites (Jacob et al., 1999), que passava por classificá-los, referindo a sua adequação à exploração por crianças e jovens em várias faixas etárias. A avaliação dos websites era feita por humanos, o que os autores admitem poder vir a criar a um problema de escalabilidade. No entanto o artigo mencionava que as abordagens baseadas em listas de palavras-chave têm uma aplicação demasiado "cega" das regras, só funcionam com texto, e mesmo o texto pode não ser bem validado. Com isto queremos salientar que já na altura se procurava uma alternativa para estes métodos, mas passaram já 15 anos sem que melhores soluções se tenham implantado.

Com a massificação do acesso à Internet, que aconteceu mais tarde, olhou-se de novo e com mais atenção para a questão de proteger os utilizadores do conteúdos indesejados. O foco expandiu-se, e passou de palavrões para todo o tipo de insultos escritos.

2.0.2 Lidar com mensagens insultuosas

A procura de mensagens insultuosas tem alguma afinidade com a identificação de palavrões. Aliás, todos estes sistemas integram um léxico de palavrões a que recorrem durante o seu funcionamento, e mostram uma aplicação direta do reconhecimento deste vocabulário.

Já em 1997 Spertus descrevia *Smokey*, um sistema que detetava "mensagens hostis" (Spertus, 1997), também chamadas de "flames" em inglês. Como refere o autor, estas mensagens não são o mesmo que "expressões obscenas", já que apenas 12% continham vulgaridades, e mais de um terço das mensagens contendo vulgaridade não eram consideradas abusivas — os palavrões podem ser usados como expletivo, por exemplo.

O Smokey fazia uso de uma série de regras fixas e pré-definidas, algumas das quais recorrendo a uma lista de palavrões. Por exemplo, as regras que lidam com palavrões são ativadas quando um dos palavrões considerados é encontrado, mas opera de forma diferente caso um "vilão" seja mencionado na mesma frase (um "vilão" é uma entidade — como um político ou personagem de uma série de TV — da qual é usual dizer mal na comunidade em questão, e como tal o uso de palavrões é *esperado*).

Spertus parece ter evitado o texto ruidoso no seu trabalho, visto referir que removeu "mensagens sem sentido (alguém pressionando teclas ao calhas) das coleções", e posteriormente, quando discute as limitações do sistema, refere uma mensagem que "não foi detetada devido à sua tipografia pouco usual", que era "G E T O V E R I T" (em português seria algo como "S U P E R A I S S O").

Mais tarde, em 2008, Mahmud, Ahmed e Khan apresentaram a sua abordagem à deteção automática de mensagens hostis e insultos recorrendo a informação semântica (Mahmud, Ahmed e Khan, 2008). Os autores admitem as limitações do seu sistema, afirmando que "não lidava ainda com texto erróneo, tal como a má colocação da vírgula, sinais de pontuação não correspondidos [à falta de esclarecimento dos autores, assumimos que se referiam a símbolos usados normalmente aos pares, como aspas ou parêntesis, dos quais só se encontrou um no texto], etc.".

Em 2010, Razavi et al. desenvolveram um sistema de classificação multi-nível para a deteção de textos abusivos (Razavi et al., 2010), mas no estudo os autores descartaram das mensagens todos os símbolos que não fossem alfabéticos ou de "pontuação expressiva".

Também nesse ano, Xu e Zhu propuseram uma abordagem para filtragem de mensagens ofensivas que operava ao nível das frases (Xu e Zhu, 2010). O seu objetivo era remover o conteúdo desagradável mantendo a integridade global da frase. O leitor, idealmente, não daria pela operação sobre a mensagem original. Referem que a ofuscação traz um problema difícil que deve ser tratado como uma questão específica.

Dois anos depois, Xiang et al. procuram detetar *tweets* ofensivos recorrendo a uma abordagem de *bootstrapping* (Xiang et al., 2012). Consideram apenas palavras compostas por letras e os símbolos - e '.

A título de resumo, enquanto que os trabalhos da subseção anterior descreviam trabalhos que consistiam em verificar se uma palavra constava numa lista de palavrões (Cohen, 1998; Jacob et al., 1999), aqui falámos de vários trabalhos que usam esse mecanismo para procurar ou filtrar insultos. Foram diversos os métodos adotados para este fim: regras pré-definidas (Spertus, 1997), análise semântica (Mahmud, Ahmed e Khan, 2008; Xu e Zhu, 2010) ou classificação automática (Razavi et al., 2010; Xiang et al., 2012). No entanto, tenham sido usados métodos mais simples ou mais sofisticados, nenhum destes trabalhos procuram lidar com palavras que não estejam bem escritas, sejam palavrões ou não. Isto compreende-se, já que o foco do trabalho situava-se mais ao nível do reconhecimento semântico das mensagens. De seguida iremos abordar estudos que tratam o tema dos palavrões em grandes coleções de mensagens, e têm em atenção a questão da grafia variável.

2.0.3 Estudos em coleções de mensagens

Alguns trabalhos focaram-se mais no estudo do uso ou no reconhecimento de palavrões. No entanto, a fim de podermos comparar os seus resultados (e saber o que podemos esperar da aplicação dos seus métodos) necessitamos de considerar algumas decisões que foram tomadas no âmbito do trabalho que foi desenvolvido. Nomeadamente é relevante especificar como o léxico foi obtido e como foi feito o reconhecimento dos palavrões em cada um dos casos estudados.

MySpace Na análise da sua coleção de *homepages* do MySpace (Thelwall, 2008), foram compiladas duas listas de palavrões. Para a lista britânica, Thelwall começou com palavrões que constavam no guia oficial da BBC, acrescentando variantes comuns, como sufixos. De seguida encontrou palavras formadas por aglutinação, procurando nos textos pelo tronco das palavras já consideradas. Devido ao grande número de resultados obtidos desta forma, excluiu as que surgiam em menos de 0,1% dos perfis. De seguida procurou no léxico por variantes gráficas das palavras (acidentais ou propositadas). Por fim, acrescentou um pequeno número de palavrões bem conhecidos.

Palavrões...
Muito fortes: *cunt, motherfuckin, motherfucking, muthafucker, muthafuckin, mutherfucker*
Fortes: *fuck, fucked, fucken, fucker, fuckin, fucking, fuckstick*
Médios: *aresehole, asshole, bastard, follock, piss, pissin, pissing, shagged, shagging, twat, wank, wanker, wanking*

Tabela 4: Palavrões considerados por Mike Thelwall na sua análise principal dos perfis do MySpace.

Para a lista criada só para os utilizadores dos EUA o autor baseou-se nas "sete palavras sujas" ("seven dirty words"), uma lista de palavras conhecidas como estando proibidas de serem transmitidas em sinal aberto pela Comissão Federal de Comunicações desse país[2], sendo elas *shit, piss, fuck, cunt, cocksucker, motherfucker* e *tits*.

O autor não usou todas as palavras nas experiências, fazendo uso de apenas 6 palavrões muito fortes, 7 palavrões fortes e 13 palavrões de intensidade média. Não é perfeitamente claro no artigo quais os palavrões que co-existiam nas duas listas, listando a Tabela 4 as palavras numa forma agregada.

Yahoo! Buzz Sood, Antin e Churchill reutilizaram a coleção que tinham anotado para a deteção de insultos pessoais (Sood, Churchill e Antin, 2011) para a deteção de palavrões. Primeiro abordaram os problemas associados ao uso de sistemas de correspondência direta com listas de palavrões (Sood, Antin e Churchill, 2012a), onde afirmam que as listas são fáceis de contornar (recorrendo à ofuscação), são de adaptação difícil (não conseguem lidar com abreviaturas e erros), e que dificilmente conseguem acomodar mais que um domínio.

Recorrendo à suas 6354 mensagens anotadas e a duas listas de palavrões disponíveis on-line, os autores demonstram a inadequação da corres-

[2]Apesar desta lista ser bem conhecida, é discutido se alguma vez foi criada uma enunciação formal de palavras inapropriadas para transmissão em televisão.

pondência direta em tarefas de identificação de palavrões. Os autores procuraram identificar de forma automática quais as mensagens que continham palavrões, tendo por base duas listas de palavrões disponíveis on-line. O melhor resultado que obtiveram com este método (usando a medida F1 como principal critério de avaliação) foi precisão 0,53, cobertura 0,40 e F1 0,46, recorrendo à radicalização (*stemming*).

Avaliando as palavras que aparentavam ser as mais discriminatórias na distinção entre uma mensagem ter ou não ter palavrões (recordamos que a granularidade desta anotação é ao nível da mensagem), apenas 8 em 33 palavras estavam bem escritas e metade das palavras na lista usavam símbolos não alfabéticos para fins de ofuscação. Para ilustrar como o uso destes métodos é difícil de resolver de forma automática, 40% dos usos do símbolo "@" na sua coleção destinavam-se a ofuscar palavrões, enquanto os restantes 60% se distribuíam entre endereços de email, direcionar mensagens a utilizadores, e outros fins não insultuosos.

Um trabalho subsequente dos mesmos autores (Sood, Antin e Churchill, 2012b) propõe duas soluções para este problema. A primeira recorre à distância de Levenshtein (Levenshtein, 1966) para comparar as palavras no texto com os palavrões na lista. Esta comparação é feita apenas no caso da palavra não ser reconhecida diretamente como palavrão ou como uma palavra no dicionário de inglês. A palavra é considerada um palavrão se o número de edições (o número total de operações de inserção, remoção e/ou alteração) igualar o número de "sinais de pontuação[3]" no termo ou for inferior a um certo valor que varia com o seu comprimento.

A segunda solução apresentada faz uso de SVMs (Support Vector Machines), com uma abordagem de "bag of words" baseada em bigramas, com vetores binários representando a presença ou ausência destes bigramas, e empregando um *kernel* linear. Este classificador, após ter sido treinado com 1/10 das mensagens numa sequência de testes de validação cruzada, prevê se cada uma das restantes mensagens contém ou não algum palavrão.

Individualmente as SVMs obtiveram precisão 0,84, cobertura 0,42 e F1 0,56, enquanto que a combinação (por disjunção) das SVMs, Levenshtein e uma das listas empregues no trabalho anterior resultou em precisão 0,62, cobertura 0,64 e F1 0,63. Ou seja, reconheceu mais palavrões, mas à custa dos falsos positivos. Estes valores são, relembra-se, relativos à classificações binária da presença de palavrões em *mensagens*, e não à identificação de palavras que são palavrões.

Twitter Wang et al. elaboraram um léxico mais elaborado que o de Sood, Antin e Churchill na sua análise da plataforma de *microblogging* (Wang et al., 2014). Começaram por juntar várias listas de palavrões existente na Internet, que foram compiladas com o propósito de servirem sistemas de filtragem de palavrões. Retiveram apenas as palavras em inglês que são usadas principalmente de forma ofensiva. Esta filtragem foi feita manualmente, resultando em 788 palavras, contendo também variações gráficas usadas com o intuito de ofuscar as palavras.

Os autores procuraram inovar também no reconhecimento de palavrões ofuscados. Para todas as palavras que não são constam no léxico de palavrões, procederam primeiro à normalização das palavras removendo letras repetidas. Depois substituíram algarismos e símbolos pelas letras mais semelhantes graficamente. Por fim fizeram uso da distância de Levenshtein para efetuar a correspondência com a lista de palavrões. Consideraram que havia correspondência se a distância de edição for igual ao número de símbolos de máscara (símbolos usados para disfarçar palavras, normalmente em substituição das letras), mais concretamente, os seguintes sete símbolos: _ % - . # \ ' . Fica por esclarecer se, por exemplo, "####" seria interpretável como um palavrão, ou porque os asteriscos não foram considerados (assim como, possivelmente, no trabalho de Sood, Antin e Churchill, caso tenham mesmo considerado apenas sinais de pontuação).

Na sua experiência, os autores anotaram manualmente uma amostra de 1000 tweets escolhidos aleatoriamente da sua coleção, como contendo ou não palavrões. Usando o método descrito acima obtiveram uma precisão de 0,99, cobertura de 0,72 e medida F1 de 0,83 na classificação automática das mensagens. Ou seja, muito poucas mensagens sem palavrões foram reconhecidas erradamente como tendo um, e a grande maioria das mensagens que efetivamente tinham palavrões foi reconhecida como tal.

SAPO Desporto A análise de 2500 mensagens extraídas do SAPO Desporto, um website português de notícias desportivas, que foram anotadas manualmente (Laboreiro e Oliveira, 2014), revelou que uma em cada cinco mensagens continha um ou mais palavrões.

Apesar do tamanho ser relativamente reduzido, há alguns aspetos a salientar deste traba-

[3]É possível que os autores se estejam a referir a símbolos não alfanuméricos e não espaços em branco.

lho que o tornam, de uma forma ou de outra, relevante.

Primeiro, todas as mensagens encontram-se anotadas com granularidade mais fina que nos corpus anotados que mencionámos até aqui, ou seja, é anotado ao nível da palavra invés de o ser ao nível da mensagem, que é o mais comum. Consequentemente é possível calcular a medida de cobertura assim como trabalhar com um léxico perfeito — isto é, todos os falsos negativos (palavrões que não são encontrados) correspondem a falhas no processo de reconhecimento e nunca a um palavrão que, por infeliz acaso, não foi considerado durante elaboração do léxico. Na Secção 1.3 vimos como é fácil isso acontecer.

Em segundo lugar, esta coleção está disponível para uso livre em licença aberta[4]. Isto possibilita que várias abordagens e algoritmos sejam comparados e medidos contra o mesmo padrão, o que é essencial na ciência. Tanto quanto é do conhecimento dos autores, este é o único *corpus* dedicado ao estudo dos palavrões disponível desta forma.

Finalmente é importante consagrar a língua portuguesa que, apesar de ser uma das línguas mais populares na Internet, tem sido menosprezada neste tema. Todas as mensagens estão escritas em português.

Todos estes motivos levaram a que esta coleção fosse escolhida como base para o nosso estudo, que passamos a descrever, começando até por apresentar alguns aspetos relevantes do corpus SAPO Desporto.

3 A coleção de mensagens utilizada

O objetivo deste trabalho consiste na identificação e reconhecimento de palavrões que foram escritos de uma forma que dificulta a sua compreensão. Deste ponto em diante será apresentado o trabalho que foi desenvolvido nesse sentido. Mais concretamente, será avaliada a adequabilidade dos métodos geralmente utilizados para a identificação de palavrões (ou seja, dizer se é palavrão ou não), e a sua capacidade no reconhecimento dos mesmos (dizer qual foi o palavrão encontrado).

Iremos aproveitar esta secção para apresentar os dados que foram utilizados na avaliação, já que tal permitirá ao leitor perceber melhor o problema da ofuscação e as dificuldades da sua remoção.

Como foi dito anteriormente, optou-se por usar uma versão revista da coleção SAPO Desporto, uma coleção destinada ao reconhecimento de palavrões. A revisão dos dados permitiu resolver algumas inconsistências pontuais na anotação. Dado que a coleção SAPO Desporto encontra-se já devidamente descrita (Laboreiro e Oliveira, 2014), apenas alguns pontos mais relevantes serão aqui destacados.

A fonte original destas mensagens foi o website de notícias SAPO Desporto, que foi escolhida exatamente devido à profusão de palavrões existente nos comentários. É interessante observar esta oposição (por certo cultural) com o Yahoo! Buzz, onde o desporto foi a categoria com menos palavrões observados (Figura 3).

Nas 2500 mensagens foram assinalados 776 usos de palavrões no total (palavrões repetidos na mesma mensagem só contam uma vez se forem escritos da mesma forma). E apesar dos palavrões estarem presentes em mais de 20% das mensagens, quase 30% das mensagens com palavrões têm mais do que um. A forma como estes se acumulam está ilustrada na Figura 4.

Como o SAPO empregava um sistema (muito simples) de filtragem de palavrões, contorná-lo tornou-se uma prática comum. Na verdade os 5 palavrões mais frequentes só se encontram em forma ofuscada, e o filtro do SAPO bloqueava 23 dos 30 palavrões mais vistos (ou seja, continuam a ser muito utilizados apesar da filtragem). Esta filtragem proporcionou numa amostra rica em variedade de palavrões, através da procura de palavras ausentes do léxico do SAPO. Foram identificados 40 palavrões-base, expandidos para 111 quando considerando variações (indicados na Tabela 7, em apêndice). Mas o registo apresenta-se rico também em métodos de ofuscação via grafias alternativas, como ilustrado na Figura 5. Encontrou-se uma média de 3 grafias para cada um dos 111 palavrões. Estes métodos de ofuscação foram compilados em 17 categorias por Laboreiro e Oliveira, que destacaram as seguintes como sendo as mais comuns:

- A substituição de uma letra por outra letra (normalmente mantendo a pronúncia da palavra inalterada), por exemplo: “ku”;
- A repetição de letras, como em “puuta”;
- O uso de sinais de pontuação ou espaços como separadores de letras, e.g. “m.e.r.d.a”;
- O recurso a algarismos para tomar o lugar de letras graficamente semelhantes, como visto em “m3rda”;
- A substituição de letras por símbolos, graficamente semelhantes ou não, como nas palavras “put@” ou “p*ta”; e

[4] `http://labs.sapo.pt/2014/05/obfuscation-dataset/` visto em 2014-12-20

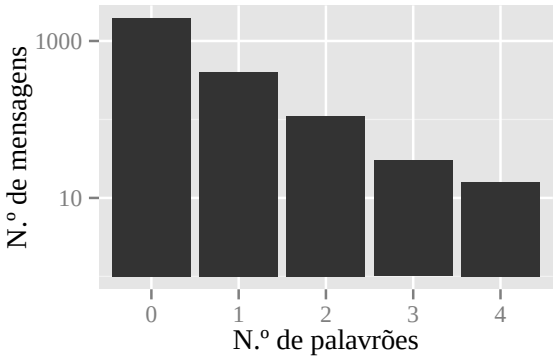


Figura 4: Número de mensagens contendo determinado número de palavrões assinalados. Grafias repetidas na mesma mensagem não foram contabilizadas.

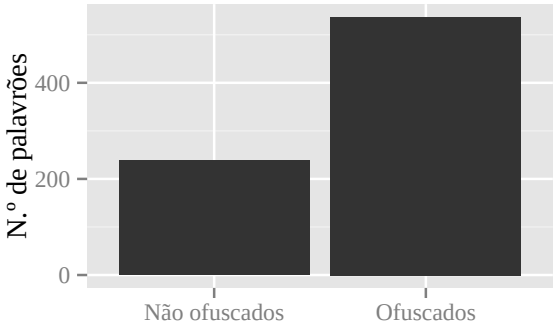


Figura 5: Comparação do número palavrões não ofuscados com o número de palavrões ofuscados. Grafias repetidas na mesma mensagem não foram contabilizadas.

- A supressão ou modificação de acentos, como em "cabrao".

Ao ofuscarem palavrões os utilizadores parecem preferir a substituição de caracteres mais do que a inserção ou remoção de símbolos, possivelmente porque a substituição preserva a dimensão da palavra, o que funciona como uma pista para a sua descodificação.

Da anotação da coleção resultam três derivados importantes. Em primeiro lugar, estão assinaladas quais as palavras no texto que são palavrões, o que é necessário para a tarefa de identificação dos palavrões. Resulta também a correspondência entre as palavras ofuscadas e os palavrões que representam, que é informação relevante para a tarefa de reconhecimento. É também daqui que se recolhe o léxico, que é simplesmente a lista de todos os palavrões encontrados e ao mesmo tempo todos os palavrões que nos interessam. Por fim, desta correspondência é possível extrair informação sobre *como* é feito o processo de ofuscação. Esta informação foi usada apenas para o estudo geral da coleção (Laboreiro e Oliveira, 2014), mas a sua importância não se esgota aí. Na secção 7 são apresentadas algumas propostas como trabalho futuro para o seu melhor aproveitamento.

De seguida tratamos dos métodos empregues para a desofuscação automática de palavrões.

4 Metodologia

Descrevemos já algumas estratégias empregues na identificação de palavrões, como é o caso da procura das palavras vistas num léxico ou o cálculo da distância de Levenshtein. Para avaliar as abordagens mais comuns e verificar se conseguem identificar os palavrões que não estão escritos na sua forma canónica, procedemos à divisão da tarefa em duas componentes: pré-processamento, onde manipulamos o texto original de forma a reduzir o seu ruído e o reconhecimento propriamente dito do palavrão, que se espera que fique mais facilitado devido ao tratamento prévio do texto.

A tarefa de pré-processamento é composta pela atomização e normalização, descritas de seguida, enquanto os métodos usados para fazer a correspondência entre as palavras vistas e os palavrões no léxico são descritos mais à frente, na Secção 4.3.

4.1 A atomização

A atomização (em inglês, *tokenization*) é o processo de dividir o texto em átomos lógicos para processamento individual, tradicionalmente palavras, pontuação e números. Hoje em dia há mais tipos de átomos que necessitam de ser tomados em consideração, como URLs, endereços de email e smileys. A atomização deficiente é suscetível de causar a perda de informação em situações em que o texto é ruidoso (Laboreiro et al., 2010).

Dois métodos de atomização foram usados no nosso trabalho. Um simples, baseado em poucas regras, que delimita os átomos na fronteira entre carateres alfanuméricos e carateres não alfanuméricos (como espaço em branco ou pontuação). Mais concretamente fez-se uso da expressão regular `\b[\w\d_-]+\b` para identificar os átomos relevantes.

Os delimitadores usados por esta expressão regular são semelhantes aos empregues nos atomizadores mais antigos, como é o caso do Penn Tre-

ebank Tokenizer,[5] e apesar de ser fácil expandir esta expressão para abarcar situações mais complexas, o objetivo era mesmo a simplicidade.

O outro atomizador utilizado foi desenvolvido tendo em consideração texto mais ruidoso, como o do Twitter, e está disponível on-line para uso livre (Laboreiro et al., 2010), na coleção "Sylvester"[6]. Este atomizador usa uma SVM treinada com mensagens do Twitter anotadas manualmente e visa encontrar os pontos de descontinuidade (entre carateres alfanuméricos e não alfanuméricos) ideais para separar o texto. Espera-se que esta ferramenta consiga lidar melhor com palavras ofuscadas, como "m.rda", sem que sejam separadas em três átomos, dada a abordagem mais sofisticada e os testes com textos ruidosos.

Dado que as etapas de processamento subsequentes lidam com os átomos a título individual, serão necessárias condições muito favoráveis para recuperar um palavrão cortado durante o processo de atomização. No entanto, algum símbolo extra que permaneça agregado à palavra pode comprometer a correspondência, ao diminuir a similaridade com o palavrão. Este não reconhecimento é mais provável se múltiplas palavras estiverem "agarradas", como por exemplo, "se-o-texto-for-escrito-desta-forma". O equilíbrio entre a decisão de cortar ou manter um símbolo ao formar átomos é algo ténue, e por esse motivo testamos dois métodos.

Sabendo que a nossa tarefa de atomização se foca unicamente em *dividir* o texto em átomos, e nunca trata de *agregar* sequências de letras separadas pelo caráter espaço (que é o separador de átomos *de facto* nos nossos textos), pode considerar-se à partida que as palavras ofuscadas com recurso a este símbolo estão irremediavelmente mal atomizadas. Olhando para os nossos dados encontrámos 37 instâncias em que isso acontece, muitas das vezes usando mais do que um espaço para esconder a palavra. Será muito difícil reconhecer algum destes palavrões fazendo uso dos métodos descritos neste trabalho.

Uma situação também problemática (embora menos) é o uso de pontuação no processo de ofuscação, visto que é um excelente candidato a definir fronteiras de átomos. Na nossa coleção encontrámos 48 instâncias em que isso acontece.

É importante realçar que os objetivos deste trabalho passam por fazer uma correspondência perfeita entre o palavrão ofuscado e o palavrão desofuscado, o que exige o reconhecimento do átomo correto na mensagem. Ou seja, pretende-se que ao fazer a operação de clarificação da mensagem não se deixe "ruído" adicional na mesma, nem, por outro lado, se suprimam carateres necessários. Por exemplo, imagine-se que a mensagem contém o palavrão "m.rda" e o atomizador age de forma incorreta, produzindo "m . rda". Se o reconhecedor conseguir fazer a correspondência entre "rda" e "merda", consideramos como estando errado, visto que, ao corrigir a mensagem fica o texto "m . rda". Portanto, encontrar o palavrão correto não é suficiente.

4.2 A normalização

A normalização procura reduzir o número de formas diferentes como se escrevem as palavras. Ao restringir o alfabeto com que se trabalha, reduz-se também a complexidade do mapeamento entre o texto e o léxico. Por exemplo, um erro encontrado frequentemente consiste na omissão do acento em algumas palavras, duplicando assim o número de grafias a considerar para reconhecer essas palavras. Se abolirmos os acentos, ambas as representações (com e sem acento) convergem numa só forma (sem acento). Este processo é de certa forma análogo a converter *strings* para minúsculas antes de as comparar.

É certo que este processo também pode originar efeitos colaterais, promovendo a confusão entre palavras diferentes. Contudo, o número de casos que devem afetar o nosso léxico será insignificante no máximo (é muito improvável que alguém invoque *cágados* num website dedicado ao desporto), e como tal os benefícios deverão compensar largamente os problemas que possam surgir. As experiências deverão verificar se isto é verdade.

Definiram-se quatro níveis de normalização:

Nenhum em que nada é alterado, servindo apenas para efeitos de comparação;

Mínimo onde se efetuam modificações simples:

- As letras são todas convertidas para minúsculas;
- Os acentos e cedilhas são removidos;
- Os espaços em branco repetidos são condensados;
- Referências a utilizadores, como "@xpto", são trocadas por um símbolo;
- URLs são substituídos por um símbolo;
- As hashtags perdem o caráter "#" e passam a ser palavras normais;

[5] `http://www.cis.upenn.edu/~treebank/tokenizer.sed` visto em 2014-12-20

[6] `http://labs.sapo.pt/2011/11/sylvester-ugc-tokenizer/` visto em 2014-12-20

Símbolos	Letra	Símbolos	Letra
0	o	5	s
1	i	6	g
2	z	7	t
3 €	e	8	b
4 @	a	9	g

Tabela 5: Tabela de substituição de símbolos por letras através de equivalência gráfica.

Básico onde se efetuam as tarefas do nível "Mínimo", mais:

- Os números são transformados nas letras graficamente mais semelhantes, como enunciado na Tabela 5;

Máximo onde se efetuam as tarefas do nível mínimo, acrescido do seguinte:

- Vários símbolos mais incomuns são convertidos para os símbolos da tabela ASCII, usando como base uma tabela de carateres confundíveis, obtida do Consórcio Unicode[7].

Estas transformações são aplicadas também ao conteúdo do léxico, a fim de possibilitar a devida correspondência na fase do reconhecimento.

4.3 O reconhecimento

O processo de reconhecimento é responsável por fazer corresponder os átomos encontrados no texto (como palavras) com os elementos constantes no léxico de palavrões. Usámos três métodos de correspondência amplamente conhecidos:

Igual que indica apenas se ambas as sequências de carateres são idênticas;

Substring que indica se a palavra no léxico está contida no átomo avaliado; e

Levenshtein que mede o número de operações de edição do algoritmo de Levenshtein necessárias para transformar o átomo em questão no palavrão com que está a ser comparado (Levenshtein, 1966).

Os dois últimos métodos aceitam um parâmetro, compreendido entre 0 e 1, que define o grau mínimo de semelhança que é exigido para reconhecer a correspondência. O valor 0 é o mais permissivo enquanto o 1 é mais exigente.

No caso do método "Substring", o parâmetro define a proporção da palavra que deve ser idêntica entre os dois valores comparados, a fim de idealmente permitir encontrar "cu" em "cus" mas não em "curvilíneo", por exemplo. No caso do método "Levenshtein" o parâmetro corresponde ao nível de semelhança, mas de acordo com a fórmula $1 - E/C$, em que E é o número de edições e C é o comprimento do átomo analisado. Em ambos os métodos o comportamento converge para o do "Igual" à medida que o valor do parâmetro se aproxima de 1.

[7] `http://www.unicode.org/review/pri273/` visto em 2014-12-20

5 Descrição da experiência

Iremos agora elaborar sobre a estrutura do sistema de avaliação que foi concebido para medir o contributo dos diversos métodos atrás descritos na tarefa de desofuscação de palavrões. Esta experiência prevê a avaliação de todas as combinações de atomizadores, normalizadores e reconhecedores. Desta forma é possível identificar aqueles que são mais ou menos adequados, tal como calcular as diferenças de desempenho que cada um potencia.

Usou-se a coleção de 2500 mensagens anotadas do SAPO Desporto que possui todos os palavrões já assinalados. No entanto os testes não foram adaptados às condições particulares deste corpus, ou seja, não se tomou partido do conhecimento prévio dos métodos de ofuscação mais comuns.

Para cada método de reconhecimento configurável testaram-se os valores de semelhança em intervalos de 0,1, entre 0,1 e 1,0, que corresponde ao nível máximo de exigência. O algoritmo de Levenshtein foi configurado para atribuir o mesmo peso às suas três operações de edição.

O desempenho foi quantificado recorrendo às medidas de precisão, cobertura e a média harmónica de ambas, conhecida como F1. Considera-se um sucesso apenas um mapeamento correto entre o átomo que é observado no texto e o palavrão que ele representa.

Deve ser salientado desde já que o teste é bastante exigente, pois esta avaliação exige que num caso como "seus grandes cabr.." consiga resolver o palavrão correto assim como o seu género e número. Caso contrário é declarado como um erro (falso positivo), visto que a identificação de palavrões não é contabilizada só por si.

Parte da dificuldade inerente à tarefa de desofuscação prende-se com os casos em que pode existir mais do que uma solução possível. Por exemplo, "m3rda" pode ser desofuscado como "merda" ou "morda", e "p*ta" pode resultar em "pata" ou "puta". Para um computador é efetivamente igual, já que se trata apenas de trocar um símbolo por outro.

Uma possível solução possível para este problema seria optar pela solução mais frequente nos textos, dado que tanto "morda" como "pata" são palavras algo improváveis. Acontece que "merda" e "puta" são inexistentes nos textos da coleção, visto que são palavras detetadas pelo sistema de filtros do SAPO. Optou-se por assumir que no presente contexto não há motivo para os utilizadores ofuscarem palavras que não sejam ofensivas, e como tal, as soluções que são palavrões são mais valorizadas.

6 Resultados e análise

Apresentar todos os resultados que foram obtidos no decorrer da nossa experiência seria muito confuso e ineficiente. Por isso optámos por focar a nossa atenção, primeiro no tema do pré-processamento e depois no processamento em si, convergindo na melhor solução que foi alcançada.

Assim sendo, comecemos por observar como as medidas de precisão e de cobertura são afetados pela atomização e normalização, como representado na Figura 6, que engloba seis gráficos. A precisão é representada nos eixos das abcissas enquanto que a cobertura ocupa o eixo das ordenadas. Cada coluna de gráficos apresenta os resultados fixando cada um dos métodos de atomização, enquanto que as linhas representam cada um dos métodos de normalização. Cada método de reconhecimento surge representado por um símbolo diferente, como indicado na legenda, e apresentam os resultados referentes à variação dos seus parâmetros. O método igual surge menos vezes no gráfico por dois motivos: primeiro, não tem parâmetro, e segundo, porque é menos sensível ao pré-processamento, e como tal os seus pontos tendem a estar sobrepostos.

6.1 A Atomização

Começando pela atomização, podemos notar como a escolha do método utilizado não tem muita influência nos valores obtidos, visto que o método "Sylvester" parece trazer nenhum ganho sobre o "Simples". Na verdade, se observado com atenção, nota-se que os resultados são até ligeiramente piores, quer em precisão quer em cobertura.

Dos 776 palavrões assinalados, 134 falham devido à atomização mal feita com o método "Simples" e 148 com o método "Sylvester". Estes casos correspondem a falsos negativos provocados quer por situações complexas que a atomização não resolve (e.g. ofuscação usando espaços), quer por problemas que a atomização

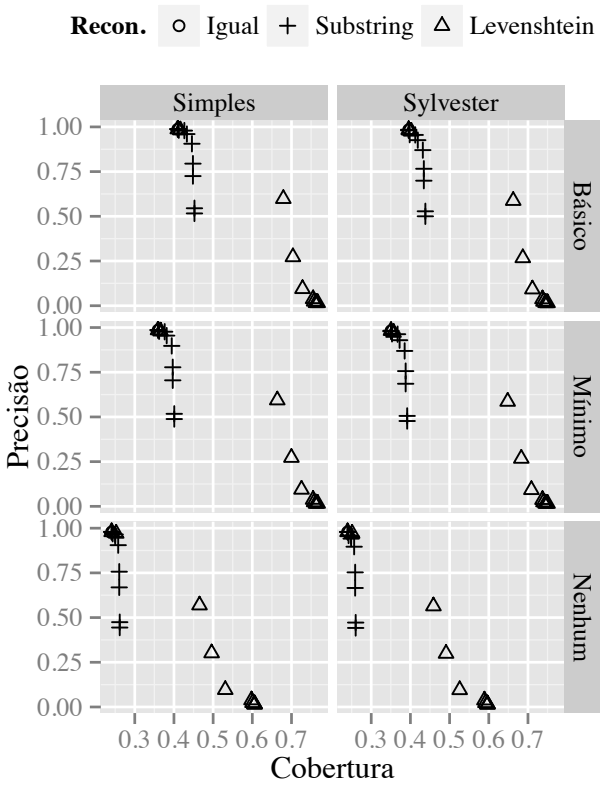


Figura 6: Comparação do desempenho dos dois atomizadores (as colunas) e três métodos de normalização do texto (as linhas) com todos os métodos de reconhecimento.

cria (e.g. ofuscação usando pontuação). De qualquer forma, nestas situações, que correspondem a 17–19% de cobertura, o átomo certo nunca chegará a ser devidamente avaliado, e estão desde já perdidas.

6.2 A Normalização

Por seu lado a normalização consegue fazer muito pelo reconhecimento de vários palavrões que passariam despercebidos de outra forma (os falsos negativos), e cada nível de operação que se adiciona parece contribuir neste sentido.

Ainda assim, a normalização mais completa (a que chamámos o nível "Máximo") não difere do nível "Mínimo", devido à ausência de substituições mais elaboradas nas mensagens usadas. Por esse motivo não surge ilustrada na Figura 6. Possivelmente o filtro do SAPO, que era bastante simples, permitia aos utilizadores o mesmo resultado trocando um "e" por "3" ou por "€", que estão facilmente acessíveis no teclado, o que não acontece com o "ε". Por esse motivo não destacamos os resultados do nível de normalização "Máximo" durante a nossa análise.

Os principais ganhos de cobertura foram obtidos no nível "Mínimo" de normalização, com

benefícios mais modestos fornecidos pelo nível "Básico". O método de reconhecimento "Levenshtein" foi o reconhecedor que beneficiou mais deste pré-processamento.

Focando-nos agora na precisão, não foram notadas diferenças significativas entre os níveis de normalização "Mínimo" e "Básico". O método de reconhecimento "Igual" permaneceu quase inafetado pela mudança de normalização. Já com o "Substring" nota-se que os parâmetros mais permissivos melhoram com a normalização, permitindo-lhe ascender acima dos 0.50 quando esta é mais ativa. Por fim, o algoritmo de Levenshtein apresenta ganhos muito tímidos de precisão, possivelmente devido à sua capacidade de tolerância superior.

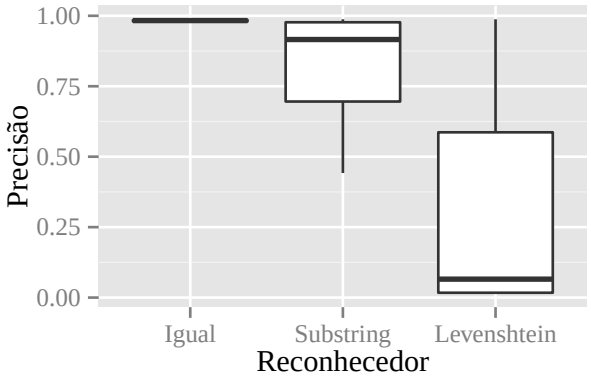


Figura 7: Comparação do desempenho dos três reconhecedores em termos de precisão.

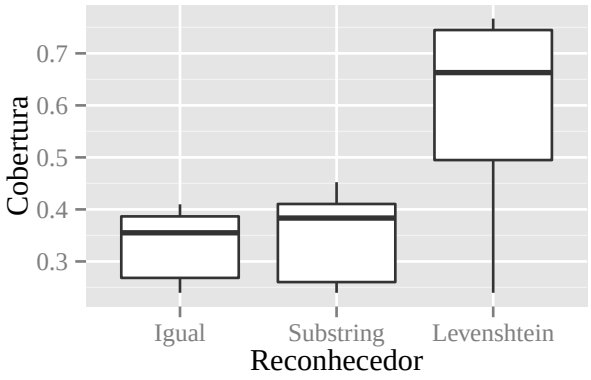


Figura 8: Comparação do desempenho dos três reconhecedores em termos de cobertura.

6.3 O Reconhecimento

Continuamos a nossa análise focando-nos apenas na atomização com o método "Simples" e normalização "Básica", cuja combinação foi a que proporcionou os melhores resultados. Iremos primeiro comparar o desempenho dos três reconhecedores, e de seguida analisar como se comportam perante os diversos valores de tolerância.

6.3.1 Os Três Reconhecedores

A Figura 7 mostra as medidas de precisão obtidas por todos os reconhecedores, enquanto a Figura 8 representa os valores de cobertura alcançados com os vários valores de parâmetro. Os gráficos de caixa e bigodes representam o primeiro, segundo e terceiro quartis com linhas horizontais, sendo a mediana destacada. As linhas verticais representam os valores máximos e mínimos, não havendo qualquer valor além de 1,5 vezes a distância inter-quartil.

Podemos confirmar como os métodos "Igual" e "Substring" são precisos, pouco dados a falsos positivos, mas rígidos e incapazes de lidar com ofuscações mais elaboradas. O método "Igual" é bastante fiável no que toca à precisão, mas a sua cobertura é bastante baixa. É curioso também comparar o potencial ganho de cobertura que uma comparação por substring pode trazer sobre uma igualdade, com o potencial risco que isso pode trazer ao nível da precisão.

Por seu lado, o método de Levenshtein permite encontrar mais palavrões, mas à custa de muitos falsos positivos que penalizam a sua medida de precisão muitas vezes. No entanto, o método de Levenshtein mostra que consegue ser quase tão preciso quanto o método "Substring", ao passo que nenhum outro método consegue ascender aos níveis de cobertura elevados atingidos pelo reconhecedor "Levenshtein".

6.3.2 O Parâmetro de Tolerância

Iremos agora focar-nos nos métodos de reconhecimento "Substring" e "Levenshtein", que apresentam os resultados mais interessantes. Na Figura 9 encontra-se representado o impacto que os diferentes níveis de tolerância têm nos resultados, desta vez baseando a comparação na medida F1, onde se observa como estes dois métodos de reconhecimento se comportam de forma diferente. O eixo das ordenadas representa o valor do parâmetro de tolerância, que vai de muito tolerante (à esquerda) até completamente intolerante (à direita). Para ajudar a formar uma imagem mais completa, apresentam-se também a Figura 10, que mostra a contagem de falsos positivos, e a Figura 11, que apresenta a contagem de falsos negativos aquando do uso dos mesmos parâmetros.

O método "Substring" apresenta uma redução continuada dos falsos positivos, ou seja, uma melhoria da precisão, que acompanha o aumentar do grau de exigência. Porém, os falsos negativos, que se mantinham quase constantes com parâmetros inferiores a 0,5, aumentam quando se usam valores superiores. Ainda assim, o valor de F1 permanece algo estável, oscilando entre os 0,48 e os 0,60.

Por seu lado, o método "Levenshtein" revela resultados muito maus com valores de parâmetro muito permissivos. Produz 36 811 falsos positivos quando usado com o valor de parâmetro 0,1. No entanto estes valores decrescem de forma muito acelerada com o aumentar do valor do parâmetro, sem alguma vez ficarem abaixo do número de falsos positivos produzidos pelo método "Substring". A cobertura vai também reduzindo, mas a um ritmo muito inferior até ao valor de parâmetro 0,8, altura em que a medida F1 atinge o seu máximo de 0,64. Quando o método opera com valores de parâmetro 0,9 ou 1,0, difere pouco nos resultados do método "Substring".

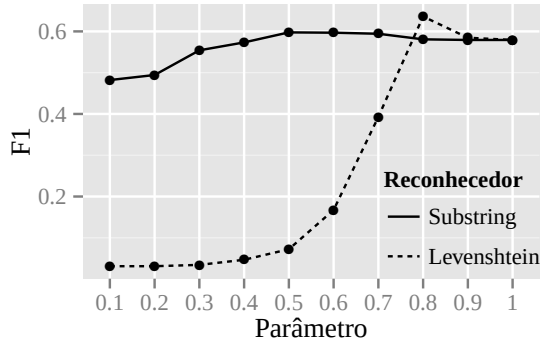


Figura 9: Efeito do nível de similaridade imposto aos métodos de reconhecimento na medida F1.

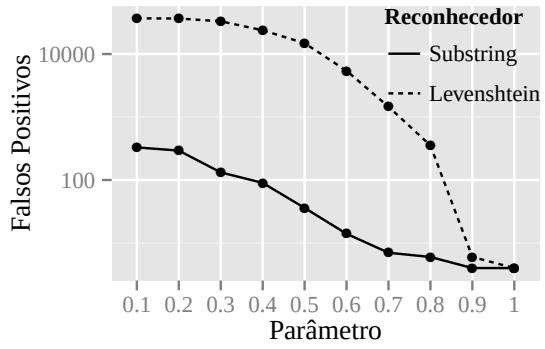


Figura 10: Contagem do número de falsos positivos obtidos com os métodos de reconhecimento "Substring" e "Levenshtein", quando usando diversos valores como parâmetro.

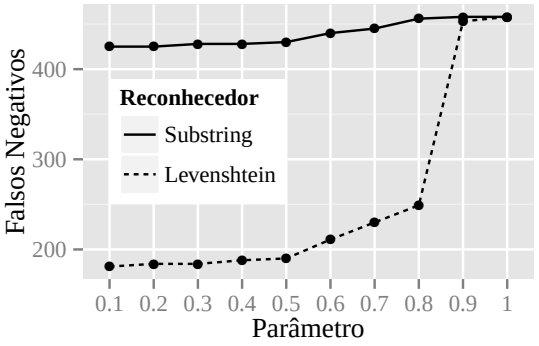


Figura 11: Número de falsos negativos gerados pelos métodos de reconhecimento "Substring" e "Levenshtein" em função do valor de parâmetro.

As medidas de 0,60 na precisão, 0,68 de cobertura e 0,64 de F1 deixam ainda uma margem significativa para melhoria, mesmo sendo este um teste muito exigente. Os objetivos de cobertura revelam-se muito difíceis de cumprir sem gerar uma torrente de falsos positivos, daí que se consideraram duas soluções possíveis. A primeira hipótese consiste no afrouxamento da avaliação de uma forma que permita ainda satisfazer as necessidades que propomos colmatar. A segunda recorre a uma fonte de dados externa, mais concretamente um dicionário, como forma de evitar muitos falsos positivos.

6.3.3 Reconhecimento mais tolerante

Se fôssemos menos rígidos na avaliação e em vez de exigir a variação exata do palavrão nos contentássemos em reconhecer qual dos 40 palavrões-base está correto, como é que os resultados melhorariam?

Na verdade, ao correr a experiência nesses parâmetros não se obtiveram ganhos significativos. Em vez dos 40 palavrões-base agruparam-se os palavrões em 22 *conceitos*, que abrangendo cada um 1 ou mais palavrões-base e suas variantes, mesmo que graficamente distintos.

Nestes moldes conseguiu-se apenas um ganho de 0,01 na medida F1 máxima, o que não é significativo. Isto poderá significar que os autores tendem a deixar indícios suficientes para compreender a variação correta do palavrão, e preferem ofuscar a parte mais comum do palavrão — que se compreende que seja a secção mais facilmente reconhecida por uma pessoa e consequentemente mais tolerante a alterações.

Visto que esta reformulação do problema não trouxe ganhos significativos, não foi mais perseguida.

6.3.4 Empregar um dicionário

Um dos problemas de aumentar a permissividade no processo de reconhecimento é que existem palavras graficamente semelhantes a palavrões mas que estão bem escritas. Por exemplo, "poder", "conta", "pilar" ou "nisso". O número de palavras cresce exponencialmente com o aumento do nível de tolerância que é permitida.

Por forma a ignorar todas as palavras bem escritas e que não são palavrões, optou-se por usar um dicionário, o que é uma decisão óbvia e não é nova (Sood, Antin e Churchill, 2012b).

Extraímos as 994 921 palavras do dicionário base de português do GNU/Aspell[8], do qual foram removidas as palavras do nosso léxico de palavrões. Repetiu-se depois a experiência, mas descartando imediatamente todos os átomos encontrados que constavam no dicionário, não as submetendo sequer ao reconhecimento de palavrões. Apresentamos apenas os valores com atomização "Simples" e normalização "Básica" que continuam a revelar-se a melhor aposta, analisando só o reconhecedor "Levenshtein".

Na Figura 12 pode ser visto o impacto que o dicionário teve na cobertura e na precisão, comparando os valores obtidos anteriormente sem o dicionário com os valores obtidos com o dicionário. Os pontos no extremo esquerdo (menor cobertura, maior precisão) representam os resultados obtidos com o mínimo de tolerância (parâmetro com valor 1.0). Observa-se que ambos os métodos convergem quando se exige maior similaridade das palavras, aproximando-se do comportamento do reconhecedor "Igual". A ligeira diferença nas medidas de cobertura pode ser explicada pela ofuscação de palavrões escrevendo-os como palavras no dicionário, como por exemplo a expressão "filho da pata", que já não chega a ser processada.

No outro extremo das linhas encontram-se os resultados gerados pelos parâmetros mais permissivos, onde é mais significativa a diferença de precisão conseguida, mas os níveis de cobertura são semelhantes. Isto acontece porque o efeito principal da filtragem de palavras é a eliminação de possíveis falsos positivos, enquanto o número de falsos negativos quase não é alterado.

O reflexo destes novos resultados na medida F1 está representado na Figura 13. Aqui comparam-se os resultados do reconhecedor "Levenshtein" quando usa e não usa o dicionário. É notória a melhoria dos valores de F1 para quase todos os níveis de exigência de similaridade, graças à presença deste filtro; mas 0,8 continua a ser o melhor valor de parâmetro. A partir desse valor, quando se dá pouca margem de manobra ao algoritmo de Levenshtein, usar ou não o dicionário faz pouca diferença.

[8] `http://aspell.net` visto em 2014-12-20

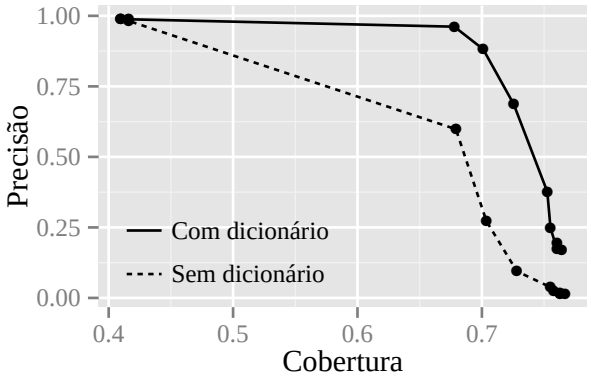


Figura 12: Precisão vs. Cobertura do método "Levenshtein" usando ou não um dicionário de português como filtro de palavras avaliadas.

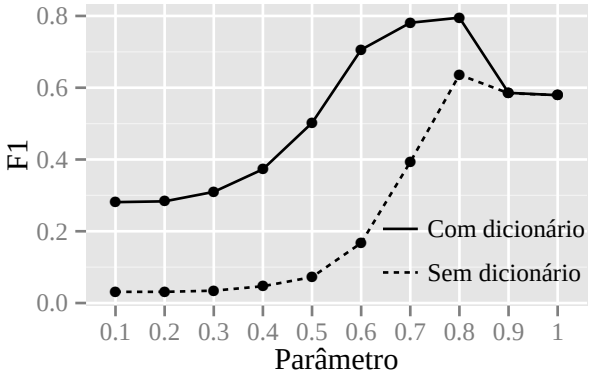


Figura 13: Impacto do dicionário na medida F1 para cada valor de parâmetro do reconhecedor "Levenshtein".

A conclusão que se pode tirar é que a exclusão das palavras presentes no dicionário permite que o reconhecedor opere com maior liberdade, sem no entanto se comprometer com falsos positivos. Assim é possível aumentar o nível de cobertura sem decréscimo significativo na precisão, como acontecia anteriormente. Os melhores resultados que foram obtidos são agora 0.96 de precisão, 0.68 de cobertura e 0.80 de F1.

7 Conclusão e trabalho futuro

O nosso estudo referente à desofuscação de palavrões dedicou-se principalmente a observar, medir e analisar. Observámos a prevalência do uso de palavrões no website SAPO Desporto, bem como os métodos usados para os disfarçar. Me-

abadalhocado	encornada	merdoso
badalhoca	encornadinhos	mijo
badalhoco	encornado	morcona
bardamerda	encornador	morconas
bastardos	encornados	morcão
bico	encornar	morcãozada
bicos	encornei	morcões
bosta	enraba	pachacha
bostas	enrabada	pachachinha
broche	enrabado	panasca
broches	enrabados	panascos
brochista	enrabar	paneleirice
cabrão	enrabá-lo	paneleiro
cabrões	enrrabar	paneleiros
cagadeira	esporra	paneleirote
cagado	esporrada	panisgas
cagalhão	foda	peida
cagalhões	foda-se	picha
caganeira	fodam	pila
cagarolas	fode	pilas
cago	fodei-vos	piroca
caguem	fodem	pirocas
caralhinhos	fodendo	pisso
caralho	foder	pizelo
caralhos	fodesses	pizelos
chulecos	fodeu	piça
cocó	fodida	piças
colhões	fodido	porcalhota
cona	fodidos	punhetas
conas	fodo	puta
corno	mamadas	putas
cornos	mamões	putinha
cornudas	maricas	putéfia
cornudo	mariconço	rabeta
cornudos	masturbar-se	rabetas
cu	merda	rameira
cuzinho	merditas	tomates

Tabela 6: Lista dos 111 palavrões anotados na coleção SAPO Desporto.

dimos o desempenho de vários métodos de pré-processamento de texto e reconhecimento de palavrões ofuscados, a fim de compreender qual é o contributo de cada um deles para o resultado final. Por fim, analisámos os valores obtidos e conseguimos melhorá-los significativamente.

Este estudo procurou essencialmente identificar os pontos fortes que convém manter na análise de palavrões, e os pontos fracos que necessitam de ser melhorados ou repensados por serem inadequados.

Começando pelo positivo, mostrámos que é possível usar métodos e técnicas comuns e frequentes para identificar e reconhecer palavrões, obtendo níveis de desempenho minimamente satisfatórios (0.96 de precisão, 0.68 de cobertura e 0.80 de F1). Foi também possível mostrar a importância do pré-processamento neste processo, designadamente a atomização e normalização, assim como comparar três formas de estabelecer correspondência entre os átomos no texto e os palavrões no léxico. Por fim, conseguiu-se medir o impacto resultante de uma filtragem dos átomos baseada num dicionário no processo de identificação e reconhecimento de palavrões.

A título de aspetos negativos, há a salientar todo o trabalho que resta ainda fazer para se poder atingir níveis de desempenho verdadeiramente bons nesta tarefa. Destacamos também o atomizador "Sylvester", por exemplo, que não materializou os resultados que se esperava obter do uso de uma ferramenta mais especializada, mesmo tendo em atenção que o estávamos a usar numa situação bastante específica e difícil para qualquer atomizador. Talvez exemplos de treino mais focados com texto ofuscado ajudassem a resolver parte do problema. A questão da ofuscação com espaços persistirá até que sejam desenvolvidos atomizadores com capacidade de aglutinação e que funcionem bem em ambientes ruidosos.

A normalização revelou-se muito útil, apesar de ter ainda aspetos passíveis de melhorar. Com uma língua como o inglês, por exemplo, a remoção dos acentos não teriam um impacto idêntico, e a tradução dos números para letras teria de ser bem pensada, visto que podem ser usados de forma mais fonética (e.g. "4ever" [forever], "l8er" [later], "2b or not 2b" [to be]). Como trabalho futuro propõe-se tomar em consideração o som das palavras, já que as substituições de letras tendem a manter a sua pronúncia (Laboreiro e Oliveira, 2014).

A distribuição estatística das letras pode também ser relevante para a normalização. Por exemplo, poucas palavras têm "k" em português, e por isso esta letra é propensa a ser removida do nosso alfabeto em prol de um "c" ou "q". Também é provável que na fase de normalização se consiga tratar da remoção de letras repetidas, permitindo que o processo de reconhecimento seja ainda mais simples por não ter de lidar com esta situação.

Ao nível dos reconhecedores, o algoritmo de Levenshtein revelou-se o mais versátil, conseguindo uma maior flexibilidade ao regular o equilíbrio entre cobertura (limite mais permissivo) e precisão (limite mais rígido) num espetro mais alargado que os outros dois. Ainda assim acreditamos que este algoritmo pode ser adaptado de forma a adequar-se melhor a esta tarefa. Uma possibilidade é modificar os pesos

das operações em função do caráter da palavra observado. Por exemplo, o custo de substituir um símbolo não alfabético por uma letra seria próximo de zero. Outra possibilidade é usar a coleção SAPO Desporto para treinar um mecanismo de desofuscação baseado em aprendizagem automática.

Outra linha de trabalho a estabelecer passa pela análise do contexto, que nós humanos usamos para nos ajudar a resolver as formas de ofuscação mais agressivas. Esta poderia basear-se na análise da frequência de n-gramas.

Agradecimentos

Este projeto teve o financiamento do Co-Laboratório Internacional para Tecnologias Emergentes UT Austin | Portugal, projeto UTA-Est/MAI/0006/2009, e do SAPO Labs UP.

Referências

Almeida, José João. 2014. Dicionário aberto de calão e expressões idiomáticas, Outubro, 2014. `http://natura.di.uminho.pt/jjbin/dac`.

Cohen, Elliot D. 1998. Offensive message interceptor for computers, Agosto, 1998. Patent 5796948 A.

Constant, Noah, Christopher Davis, Christopher Potts, e Florian Schwarz. 2009. The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung: International Journal for Language Data Processing*, 33:5–21.

Jacob, Varghese, Ramayya Krishnan, Young U. Ryu, R. Chandrasekaran, e Sungchul Hong. 1999. Filtering objectionable internet content. Em *Proceedings of the 20th international conference on Information Systems*, ICIS '99, pp. 274–278, Atlanta, GA, USA. Association for Information Systems.

Jay, Timothy. 2009. The utility and ubiquity of taboo words. *Perspectives on Psychological Science*, 4(2):153–161.

Jay, Timothy e Kristin Janschewitz. 2007. Filling the emotional gap in linguistic theory: Commentary on Pot's expressive dimension. *Theoretical Linguistics*, 33:215–221.

Laboreiro, Gustavo e Eugénio Oliveira. 2014. What we can learn from looking at profanity. Em Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A. S. Pardo, e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language*, volume 8775 of *Lecture Notes in Computer Science*. Springer International Publishing, pp. 108–113, Setembro, 2014. http://labs.sapo.pt/2014/05/obfuscation-dataset/.

Laboreiro, Gustavo, Luís Sarmento, Jorge Teixeira, e Eugénio Oliveira. 2010. Tokenizing micro-blogging messages using a text classification approach. Em *Proceedings of the fourth workshop on analytics for noisy unstructured text data*, AND '10, pp. 81–88, New York, NY, USA, Outubro, 2010. ACM. http://labs.sapo.pt/2011/11/sylvester-ugc-tokenizer/.

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

Mahmud, Altaf, Kazi Zubair Ahmed, e Mumit Khan. 2008. Detecting flames and insults in text. Em *6th International Conference on Natural Language Processing (ICON-2008)*. Center for research on Bangla language processing (CRBLP), BRAC University.

Mehl, Matthias R. e James W. Pennebaker. 2003. The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, 84(4):857–870.

Razavi, Amir Hossein, Diana Inkpen, Sasha Uritsky, e Stan Matwin. 2010. Offensive language detection using multi-level classification. Em Atefeh Farzindar e Vlado Keselj, editores, *Advances in Artificial Intelligence*, volume 6085 of *Lecture Notes in Computer Science*, pp. 16–27. Canadian Conference on Artificial Intelligence, Springer.

Sood, Sara Owsley, Judd Antin, e Elizabeth F. Churchill. 2012a. Profanity use in online communities. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pp. 1481–1490, New York, NY, USA. ACM.

Sood, Sara Owsley, Judd Antin, e Elizabeth F. Churchill. 2012b. Using crowdsourcing to improve profanity detection. Em *AAAI Spring Symposium: Wisdom of the Crowd*, volume SS-12-06 of *AAAI Technical Report*. AAAI.

Sood, Sara Owsley, Elizabeth F. Churchill, e Judd Antin. 2011. Automatic identification

of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, Outubro, 2011.

Spertus, Ellen. 1997. Smokey: Automatic recognition of hostile messages. Em *Proceedings of Innovative Applications of Artificial Intelligence (IAAI)*, pp. 1058–1065.

Thelwall, Mike. 2008. Fk yea I swear: cursing and gender in MySpace. *Corpora*, 3(1):83–107.

Wang, Wenbo, Lu Chen, Krishnaprasad Thirunarayan, e Amit P. Sheth. 2014. Cursing in English on Twitter. Em *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, Fevereiro, 2014.

Xiang, Guang, Bin Fan, Ling Wang, Jason Hong, e Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. Em *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1980–1984. ACM.

Xu, Zhi e Sencun Zhu. 2010. Filtering offensive language in online communities using grammatical relations. Em *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*.

Izen+aditz konbinazioen azterketa elebiduna, hizkuntza-aplikazio aurreratuei begira

Bilingual analysis of noun+verb combinations and their functionality in advanced language applications

Uxoa Iñurrieta, Itziar Aduriz[1], Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola
IXA taldea, Euskal Herriko Unibertsitatea
usoa.inurrieta@ehu.es,a.diazdeilarraza|gorka.labaka|kepa.sarasola@ehu.es
(1) Department of Linguistics, University of Barcelona
itziar.aduriz@ub.edu

Laburpena

Hiztegi elebidunak oinarritzat hartuta, euskarazko eta gaztelaniazko izen+aditz konbinazioak izan ditugu aztergai lan honetan. Konbinazioen eta euren ordainen ezaugarri morfosintaktiko zein semantikoei begiratu diegu, eta bi hizkuntzak parez pare jarri ditugu, zer alde eta antzekotasun duten aztertzeko. Artikulu honek agerian uzten du zeinen konplexuak diren era horretako egiturak eta, ondorioz, zeinen garrantzitsua den Hizkuntzaren Prozesamenduko aplikazioetan tratamendu egoki bat ematea, itzulpen automatikoan adibidez. Horrez gain, azterketatik lortutako emaitza guztiak interfaze publiko batean jarri ditugu, edonork bilaketak egin ahal izan ditzan guk landutako konbinazioen gainean.

Gako-hitzak

Hitz-konbinazioak, Hizkuntzaren Prozesamendua, Elhuyar hiztegiak

Abstract

This article deals with noun+verb combinations in bilingual Basque-Spanish and Spanish-Basque dictionaries. We take a look at morphosyntactic and semantic features of word combinations in both language directions, and compare them to identify differences and similarities. Our work reveals the high complexity of those constructions and, hence, the need to address them specifically in Natural Language Processing tools, for example in Machine Translation. All of our results are publicly available online, where users can query the combinations we have analysed.

Keywords

Word combinations, Natural Language Processing, Elhuyar dictionaries

1 Sarrera

Hizkuntza batean hitz-konbinazioak zuzen erabiltzeak agerian uzten du hiztunak hizkuntza hori zenbateraino ezagutzen duen, konbinazioek testua aberasten eta edertzen baitute gehienetan, eta hori egiteko gai denak derrigor ezagutu behar baitu ondo hizkuntza, gramatika-arauetatik eta oinarrizko hiztegitik haratago. Hizkuntza hori itzultzen edo tratamendu automatikorako gaitzelanetan hasten garenean, ordea, aberastasun eta edertasun horiek arazo-iturri bihurtzen dira sarri.

Izan ere, era askotakoak izan badaitezke ere, hitz-konbinazio guztiak datoz bat zerbaitetan: normalean, ez dituzte hizkuntzaren ohiko arau morfosintaktiko zein semantikoak jarraitzen. Horregatik, hain zuzen ere, era horretako egiturek –izan kolokazio[1], lokuzio[2], edo enuntziatu fraseologiko[3]– buruhauste galantak sortzen dituzte hainbat alorretan: itzulpengintzatik hasi eta Hizkuntzaren Prozesamendura (HP), atzerriko hizkuntzen ikaskuntza ere tartean dela.

Lan honetan, izen+aditz konbinazio mota guztiak aztertu ditugu, momentuz tipologia kontuak alde batera utzita. Gure helburua euskararen eta beste hizkuntza batzuen artean zein alde eta antzekotasun dagoen ikustea da, gerora, hitz-konbinazioek HPko aplikazioetan sortzen dituzten arazoak konpontzen laguntzeko. Azterketa hau EHUko IXA ikerketa-taldean egiten ari garenez, taldean bertan garatutako aplikazioak hobetzera bideratuko dugu gure ikerketa-lana, eta garrantzi berezia emango diogu gaztelaniatik euskararako Matxin[4] itzultzaile automatikoari (ikus 3. atala).

Azterketa linguistikoa corpus paralelo eta alderagarrietan oinarritzeko asmoa dugu, baina,

[1] *haizea ibili; gogotik barre egin; euskaldun peto*
[2] *ahalak eta leherrak egin; adarra jo; aita ponteko*
[3] *nolako zura, halako ezpala; ez horregatik*
[4] http://www.opentrad.com/eu/?

horren aurretik, hiztegien gainean egin nahi izan dugu lehen urratsa. Horretarako, Elhuyarren gaztelania-euskara eta euskara-gaztelania hiztegietatik[5] izen+aditz konbinazioak erauzi ditugu euren ordainekin batera, eta, bi hizkuntzak parez pare jarrita, hainbat ezaugarriri begiratu diogu: morfologikoei, sintaktikoei eta semantikoei.

Hortaz, artikulu honetan, izen+aditz konbinazioak ardatz hartu eta euskararen eta gaztelaniaren bat-etortze mailari buruz arituko gara, tratamendu automatikoan zenbateraino lagun dezaketen ikusteko. Azterketaren emaitzen bidez, erakutsiko dugu ez dagoela kidetasun handirik bi hizkuntzen artean, egitura horiei dagokienez.

Gainera, landu ditugun konbinazio guztiak datu-base batean bildu ditugu informazio linguistikoarekin batera, eta eskuragarri jarri ditugu interfaze publiko baten bidez (ikus 5. atala).

2 Aurrekariak

Hitz-konbinazioak, oro har, fraseologiaren alorrean kokatu ohi dira. Erabateko adostasunik ez dagoen arren, unitate fraseologikoak ohi baino sarriago elkarrekin agertzeko joera duten hitz-konbinaziotzat definitzen dira, eta hiru ezaugarri nagusi dituztela esaten da (Sanz Villar, 2011; Baldwin & Kim, 2010; Corpas Pastor, 2004): hitz batez baino gehiagoz osatuta daude, sintaktikoki finkoak dira, eta konbinazio osoaren esanahia ezin da osagaien esanahi indibidualetatik inferitu.

Gaiaren inguruan egindako lanak askotarikoak dira: azterketa teorikoek konbinazioak deskribatzen eta sailkatzen dituzte, idiomatikotasunaren zein ezaugarri morfosintaktikoen arabera (Rafel, 2004; Zabala, 2004; Corpas Pastor, 2001); beste lan batzuk, berriz, aplikazio jakinetara bideratuta egon ohi dira. Bigarren multzoan, azken hamarkadetan indar berezia hartu dute HPren alorrean egindako lanek (Seretan, 2013; Alonso Ramos, 1995), eta euskarak ere izan du bere lekua prozesu horretan.

Bi ikerketa-lan nagusi egin dira euskaraz: Urizarrena (2012) bata, eta Gurrutxagarena (2014) bestea. Lehenak euskarazko lokuzioen tratamendurako oinarri linguistikoak ematen ditu, eta bigarrenak, berriz, euskarazko aditz+izen konbinazioen detekzio eta karakterizazio automatikoa du aztergai. Ikuspegi elebidunetik, ordea, oraindik ez da ikerketarik egin, eta hutsune hori betetzera dator lan hau.

Izan ere, beste hizkuntzetan ere gehiago aztertu izan da konbinazioen detekzio eta erauzketa elebakarra, baina ikuspegi elebidunetik ere egin dira hainbat hurbilpen eta esperimentu. Lük eta Zhouk (2004), esaterako, corpus elebakarretatik erauzten dituzte kolokazioak, eta, gero, lexikoi elebidunak sortzeko, hizkuntza batean eta bestean lortutako emaitzak parekatzen dituzte. Champollion sistemak (Smadja, McKeown, & Hatzivassiloglou, 1996), berriz, corpus paraleloetatik erauzten ditu zuzenean ingelesezko eta frantsesezko kolokazioak eta ordainak.

Bestalde, Mel'cuk-en (1998) lana izan da, ziur asko, alor elebidunean egin diren hurbilpenen arteko ezagunena, non kolokazioak funtzio lexikalen bidez errepresentatzeko metodo bat proposatzen baita, kolokazioko osagaien artean dagoen erlazio semantikoaren arabera. Gerora, hainbat esperimentu egin dira hurbilpen horretan oinarrituta, kolokazioen inguruko informazioa itzultzaile automatikoetan gehitzeko (Heylen & Maxwell, 1994). Dena den, funtzio lexikaletan oinarritutakoez gain, badago itzulpen-prozesua beste transferentzia-metodo batzuen bidez egiten duen esperimentrik ere (Wehrli et al., 2009).

3 Itzulpen automatikora begira

Sarreran esan bezala, ikerketa-lan honetatik ateratako ondorioak HPko tresnak hobetzeko baliatzea da gure helburua. Horietako bat IXA taldean garatutako Matxin itzultzaile automatikoa da, gaztelaniatik euskarara itzultzen duena.

Sistemak erregela linguistikoak ditu oinarrian, eta itzulpen-prozesua hiru fasetan egiten du: analisia, transferentzia eta sorkuntza (Iñurrieta, 2013; Mayor et al., 2011). Lehen fasean, jatorrizko hizkuntzako itzulgaiaren analisi morfosintaktikoa egiten da. Bigarrenean, jatorrizko hizkuntzaren egitura xede-hizkuntzara ekartzen da, eta, sistemaren lexikoi elebidunetik abiatuta, gaztelaniazko hitz bakoitzari euskarazko baliokide bat ematen zaio. Azkenik, hirugarren fasean, euskarazko arau sintaktiko eta morfologikoak aplikatu, eta itzulpena bera sortzen da.

Mayorren (2011) arabera, itzulpenen % 69 ulertzeko modukoak dira; beraz, esan daiteke nahiko emaitza onargarriak lortzen direla oro har. Hitz-konbinazioei dagokienez, ordea, sistema oraindik ez da itzulpen txukunak sortzeko gai, eta baliteke itzulpen okerrek osatzen duten % 31 horretan konbinazioek garrantzia izatea. Izan ere, lexikoi elebidunean badaude hitz anitzeko sarrera batzuk –Elhuyar hiztegiko konbinazioak tarteko–, baina, batetik, ez dira asko; bestetik, ez dira beti ondo hautematen; eta, gainera, Matxinek hitz bakartzat hartzen ditu, ia informazio gehigarririk gabe.

[5] `http://hiztegiak.elhuyar.org/`

Konbinazio bat hitz anitzeko unitatetzat hauteman ahal izateko, osagai guztiek segidan eta ordena berean agertu behar dute esaldian, hala izan ezean, detekzioak huts egiten baitu (ikus 1. adibidea). Gainera, konbinazio osoari kategoria bakarra ematen zaio, eta ez konbinazioko osagai bakoitzari berea (ikus 2. adibidea). Izen+aditz konbinazioen kasuan, aditza lematizatu egiten da, baina izena eta izenarekin batera doazen determinatzaile eta preposizioak ez; bere horretan agertu behar dute, aldaketarik txikiena ere egin gabe.

Hori gutxi balitz, euskarazko sorkuntzarako ere ez dago hitz anitzekoentzako arau berezirik, eta, oro har, segida bat beste segida baten bidez ordezten da, bestelako informaziorik kontuan izan gabe (ikus 3. eta 4. adibideak).

Hori guztia aintzat hartuta, ez da harritzekoa emaitza traketsak lortzea nahiz eta itzulgaiak sinple samarrak izan. Hona hemen adibide batzuk:

(1) Ayer le **gastaron una broma**.
Atzo **broma egin** zioten.

La **broma** que le **gastaron** fue cruel.
***Gastatu** zioten **broma** krudela izan zen.

Lehen esaldia ondo itzuli da, *gastar una broma* konbinazioa jarraian agertzen delako osorik; bigarrenean, ordea, *broma* eta *gastar* ez daude bata bestearen ondoan, ez eta ordena berean ere, eta sistemak ez du konbinazioa hitz anitzeko unitatetzat hartu. Hori dela eta, euskaraz okerreko aditza aukeratu da: *gastatu*, *egin*en ordez (*Egin zioten broma krudela izan zen*).

(2) El colesterol **hace mal** al corazón.
Kolesterolak bihotzari **kalte egiten** dio.

Ha **hecho mal** el examen.
Azterketa ***kalte egin** du.

Bigarren adibide honetan, lehenengoaren kontrakoa gertatu da. *Hacer* eta *mal* jarraian ikusita, Matxinek unitate bakartzat hartu ditu bi hitzak, analisi sintaktikoari jaramonik egin gabe. Lehen esaldian ez dago arazorik, baina bai bigarrenean, beharrezkoa baitzen sistemak *mal* izentzat ez hartzea, baizik eta aditzondotzat, *kalte*ren ordez *gaizki* jarri ahal izateko (*Azterketa gaizki egin du*).

(3) ¿**Tienes frío**?
Hotz zara?

Miren **tiene frío** el plato.
Mirenek platera **hotz *da**.

Hirugarren adibidean ikus daitekeenez, hitz anitzeko bat tartean dagoenean, sistema ez da gai subjektuaren eta aditzaren arteko konmuztadurarik egiteko. Horregatik, lehen esaldiaren itzulpena egokia bada ere, bigarrenarena ez: aditzak *ukan* motakoa behar luke *izan* motakoa beharrean (*Mirenek platera hotz du*).

(4) Le **hizo daño**.
Min eman zion.

Se **hizo daño**.
Min *eman zen.

Hirugarrenean bezala, azken adibidean ere sorkuntzarako arau faltan datza arazoa. Matxinen lexikoi elebidunean, *hacer daño*ren baliokidetzat *min eman* bakarrik ageri da, eta, hori dela eta, sistemak ez du bereizten *min eman* eta *min hartu*ren artean. Informazio gehigarria behar du, *hacerse daño* bihurkaria denean *hartu* aditza erabili behar dela jakiteko eta dagokion aditza jartzeko (*Min hartu zuen*).

4 Hiztegien gaineko azterketa

Aipatu bezala, euskarazko eta gaztelaniazko hitz-konbinazioak aztertzeko eta alderatzeko ikerketalan honetan, lehen urratsa hiztegien gainean egin dugu. Gure azterketa corpusetan oinarritu nahi badugu ere, egokia iruditu zaigu abiapuntutzat hiztegi elebidunak hartzea, konbinazio usuenak behintzat jasotzen baitituzte normalean.

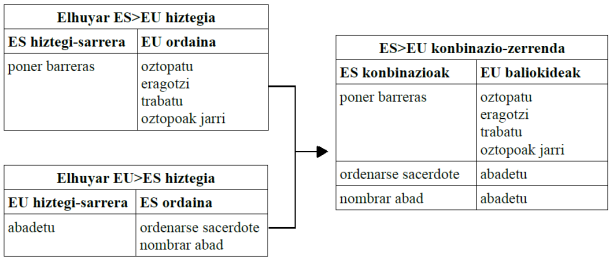
Jakin badakigu hiztegietan biltzen diren konbinazioak tipologia bakarrekoak izan ohi direla –lokuzioak[6], oro har (Urizar, 2012)– eta beste mota batzuetako konbinazioak ez direla hainbeste landu izan hiztegigintzan –kolokazioak[7], kasu–. Dena den, esan bezala, pauso honetan hasitako lana osatzeko, corpusetara joko dugu gerora.

Oinarritzat hartu ditugun konbinazio-zerrendak Elhuyarren gaztelania-euskara eta euskara-gaztelania hiztegietatik erauzi ditugu. Sarrera guztiak hartu, eta Eustagger analizatzaile morfologikoaz (Aduriz et al., 1996) baliatu

[6] *adarra jo, lan egin* edo *min eman* bezalakoak
[7] *legeak urratu* edo *konpromisoa berretsi*ren gisakoak

gara izen+aditz konbinazioak izan litezkeenak hautemateko. Prozesu automatiko gehienetan bezala, emaitza oker batzuk ere lortu ditugu benetako izen+aditz konbinazioekin batera, eta zerrenda eskuz gainbegiratu behar izan dugu.

Ahalik eta kasu gehien landu nahi genituenez, hiztegi-sarrera diren izen+aditz konbinazioez gain, beste egitura batzuetako sarreren ordaintzat ageri diren izen+aditz konbinazioak ere hartu ditugu. Hau da: gaztelaniazko *poner barreras* konbinazioa eta haren *oztopatu* ordainaz gainera, euskarazko *abadetu* sarrera eta haren ordaintzat ageri den *ordenarse sacerdote* ere erauzi ditugu, ordaina izen+aditz motakoa delako (ikus 1. irudia).



1. irudia: Konbinazio-zerrendak sortzeko prozesua.

Gaztelaniazko 2.650 konbinazio landu ditugu: gaztelania-euskara hiztegitik erauzitako 1.390 eta euskara-gaztelania hiztegitik erauzitako 1.260. Konbinazio gehienek ordain bat baino gehiago dutenez, euskarazko baliokideak askoz ere gehiago izan dira: 6.587.

Euskarazko konbinazioei dagokienez, berriz, 2.954ko zerrenda osatu dugu, euskara-gaztelania hiztegiko 1.354 eta gaztelania-euskara hiztegiko 1.600 batuta. Horien gaztelaniazko baliokideak 6.392 dira guztira.

Datozen azpiataletan bildu ditugu horien guztien azterketatik ateratako emaitzak. Lehenik, 4.1 atalean, euskarazko eta gaztelaniazko konbinazioetako aditzak eta izenak nolakoak diren azalduko dugu, bi hizkuntzetako emaitzak parez pare jarrita. Bigarrenik, 4.2 atalean, bestelako ezaugarri morfosintaktiko batzuk izango ditugu hizpide: zein kasu eta postposizio-marka hartzen dituzten euskarazko izenek, zein kategoriatako hitzez osatzen diren gaztelaniazko konbinazioak, zer gertatzen den egitura horiekin hizkuntza batetik bestera, eta zer portaera duten mugatasunak eta numeroak horrelako egituretan. Azkenik, 4.3 atalean, ezaugarri semantikoei begiratuko diegu, eta hizkuntza bateko eta besteko aditzak eta izenak elkarren artean baliokide ote diren ikusiko dugu.

4.1 Konbinazioetako aditzak eta izenak

Azterketa sakonagorik egin aurretik, gure lanaren oinarri diren zerrendak hartu, eta osagai nagusiei begiratu diegu: aditzei eta izenei. Hizkuntza bateko eta besteko konbinazioak hartuta, zein aditz/izen eta zenbateko maiztasunez erabili diren jakin nahi izan dugu, gero bi hizkuntzak parez pare jarri eta euren artean antzekotasunik ba ote dagoen ikusteko.

Fraseologiari buruzko lanetan maiz aipatu izan denez, hitz-konbinazioetan erabiltzen diren aditzak oso ohikoak izaten dira gehienetan, aditz "arinak", esanahi osorik gabeak. Ez da harritzekoa, beraz, zer aurkitu dugun: hizkuntza bateko eta besteko aditzik erabilienen artean, gehienak oso arruntak dira, eta, gainera, asko baliokideak dira euren artean. Hain zuzen ere, euskaraz gehien errepikatzen diren sei aditzen ordainak gaztelaniazko aditzetan bilatuta, zortzi aditzik erabilienen artean aurkitu ditugu: *egin – hacer, izan – ser/estar/tener, eman – dar, hartu – tomar, egon – estar, jarri – poner* (ikus 1. eta 2. taulak).

Aditza	**Agerr.**	**Ehunekoa**
egin	614	% 20,79
izan	296	% 10,02
eman	217	% 7,35
hartu	147	% 4,98
egon	133	% 4,50
jarri	92	% 3,11
Guztira	**1.499**	**% 50,75**

1. taula: Euskarazko konbinazioetan agerraldi gehien dituzten aditzak.

Aditza	**Agerr.**	**Ehunekoa**
hacer	233	% 8,79
dar	197	% 7,43
estar	122	% 4,60
poner	112	% 4,23
tener	103	% 3,89
echar	73	% 2,75
tomar	64	% 2,42
ser	56	% 2,11
Guztira	**960**	**% 36,22**

2. taula: Gaztelaniazko konbinazioetan agerraldi gehien dituzten aditzak.

Hala ere, agerraldiei begiratuz gero, bat aise konturatzen da euskaraz askoz ere alde handiagoa dagoela gehien errepikatzen diren aditzen eta gainontzekoen artean. Izan ere, euskarazko konbinazioetan 310 aditz desberdin agertu dira guztira, baina konbinazio guztien % 50,75 sortzen dira 1. taulan bildutako seiekin bakarrik.

Gaztelaniaz ere badago aldea, baina ez da inondik inora ere euskarazkoa bezain handia; zor-

tzi aditzik erabilienek konbinazioen % 36,23 osatzen dituzte. Kontuan hartzekoa da, dena den, agertu diren aditzak ere gehiago direla: 453.

Bestalde, izenen agerraldiak ere kontatu ditugu, antzeko ondoriorik atera ote genezakeen ikusteko. Bagenekien izen desberdinen kopurua aditzena baino askoz ere altuagoa izango zela, baina, era berean, susmoa genuen izen batzuek beste batzuek baino joera handiagoa dutela konbinazioetan agertzeko. Eta ez genbiltzan oker, 3. eta 4. taulek agerian uzten dutenez.

Izena	**Agerr.**	**Ehunekoa**
buru	65	% 2,20
begi	54	% 1,83
aurre	43	% 1,46
kontu	33	% 1,12
atze	24	% 0,81
bide	23	% 0,78
esku	21	% 0,71
Guztira	**263**	**% 8,91**

3. taula: Euskarazko konbinazioetako izenik erabilienak.

Izena	**Agerr.**	**Ehunekoa**
mano	28	% 1,06
cabeza	27	% 1,02
ojo	21	% 0,79
vista	21	% 0,79
oído	18	% 0,68
cuenta	18	% 0,68
Guztira	**133**	**% 5,02**

4. taula: Gaztelaniazko konbinazioetako izenik erabilienak.

Hortaz, izenetan ere badago antzekotasunik bi hizkuntzen artean, gehien-gehien errepikatzen direnetako batzuk bat baitatoz euskaraz eta gaztelaniaz: *buru – cabeza, begi – ojo, esku – mano, kontu – cuenta*. Horietako asko, gainera, gorputz-atalen izenak dira, multzo horretako hitzak oso ohikoak baitira hitz-konbinazioetan, lokuzioetan batez ere.

4.2 Morfosintaxia

Behin konbinazioetan zein aditz eta izen agertzen diren ikusita, konbinazioen beste ezaugarri morfosintaktiko batzuk aztertzeari ekin diogu. Hasteko, hizkuntza bateko eta besteko konbinazioak bereiz landu ditugu: euskarazko izenen kasu- eta postposizio-markak, eta gaztelaniazko osagaien kategoriak. Bigarrenik, bi hizkuntzetako egiturak parez pare jarri ditugu, kasu- eta postposizio-marken eta gaztelaniazko egituren artean loturarik ba ote dagoen argitu nahian. Eta, azkenik, mugatasunari eta numeroari ere begiratu diegu, ezaugarri horiek hizkuntza batetik bestera zenbateraino gordetzen diren jakiteko.

4.2.1 Euskarazko konbinazioen ezaugarri morfologikoak: kasu- eta postposizio-markak

Lehenik, euskarazko konbinazioetako izenak hartu, eta etiketa bana jarri diegu eskuz, kasu- edo postposizio-markaren arabera. Kontaketak egitean, berehala konturatu gara alde nabarmena dagoela etiketa bateko edo besteko multzoen artean.

Urizarren (2012) eta Zabalaren (2004) lanetan aipatzen denez, izen batez eta aditz batez osatzen diren konbinazio gehien-gehienek absolutibo-marka izan ohi dute. Horixe bera ondorioztatu dugu guk ere, konbinazio guztien hiru laurden baino gehiago baitira, hain zuzen ere, multzo horretakoak: *denbora galdu, dei egin, gobada jo, itzal egin, hitza bete...*

Inesiboa daramaten izenak ere nahiko sarri ageri dira besteen aldean (*sutan egon, jokoan jarri*), eta adlatibodunek (*aurrera egin, eskura ordaindu*), instrumentaldunek (*eskuz jo, aurrez prestatu*) eta ablatibodunek (*burutik egon, hutsetik hasi*) jarraitzen diete.

Gehien errepikatu diren bost markak 5. taulan ikus daitezke, agerpen-kopuruaren eta ehunekoaren arabera.

Marka	**Agerr.**	**Ehunekoa**
absolutiboa (abs)	2250	% 76,18
inesiboa (ine)	364	% 12,32
adlatiboa (ala)	101	% 3,42
instrumentala (ins)	87	% 2,94
ablatiboa (abl)	85	% 2,88
beste batzuk	62	% 2,27

5. taula: Euskarazko konbinazioen markak.

Gainontzeko markak nahiko bakanak dira, eta ez dira konbinazio guztien % 1era ere heltzen: ergatiboa (*deabruak hartu*), datiboa (*amuari lotu*), adlatiboa + abutiboa (*leporaino egon*), soziatiboa (*suarekin jolastu*), genitiboa + absolutiboa (*gorrarena egin*), lekuzko genitiboa + absolutiboa (*hitzekoa izan*), adlatiboa + lekuzko genitiboa (*bururako gorde*), eta lekuzko genitiboa (*zorioneko egon*).

4.2.2 Gaztelaniazko konbinazioen ezaugarri morfosintaktikoak: osagaien kategoriak

Gaztelaniari dagokionez, lau egituratako konbinazioak bildu ditugu. Batzuk izen eta aditzez bakarrik osatuak dira, eta beste batzuek determinatzaile eta/edo preposizioak dituzte tartean.

Azpiko taulan ikusten denez, gehien errepikatzen den egitura *aditza + determinatzailea + izena* da (*dar un toque, ser una pena, hacer un favor*), eta beste hiru multzoen artean ez dago alde handirik: *aditza + izena* (*meter baza, tener afecto*), *aditza + preposizioa + izena* (*saber de memoria, tener a favor*), *aditza + preposizioa + determinatzailea + izena* (*dejar a un lado, caerse por su peso*).

Egitura	Agerr.	Ehunekoa
adi + det + ize	999	% 37,70
adi + ize	629	% 23,74
adi + prep + ize	577	% 21,77
adi + prep + det + ize	445	% 16,79

6. taula: Gaztelaniazko konbinazioen egiturak.

4.2.3 Gaztelaniazko ordainak euskarazko konbinazioaren arabera

Bi hizkuntzen ezaugarri morfosintaktikoak alderatzen hasteko, euskarazko konbinazioen gaztelaniazko ordain guztiak analizatu, eta hainbat multzotan sailkatu ditugu. Honako sailkapen hau sortu dugu:

- Aditz soila (**adi**): *descentrar, lanzar, descansar.*
- Aditza, determinatzailea eta izena (**adi + det + ize**): *matar el tiempo, hacer la colada, echar la llave.*
- Aditza eta izena (**adi + ize**): *ganar tiempo, tener noticia, hacer caso.*
- Aditza, preposizioa eta izena (**adi + prep + ize**): *arder en deseos, llevar por acompañante, comer con apetito.*
- Aditza eta adberbioa edo adberbio-sintagma (**adi + AdbS**): *salir adelante, sentar bien, llegar lejos.*
- Aditza eta adjektiboa edo adjektibo-sintagma (**adi + AdjS**): *estar claro, volverse loco/a, estar arrepentido/a.*
- Aditza, eta izen soila ez den izen-sintagma (**adi + IS**): *poner mala cara, levantar la tapa de los sesos, hacer buen tiempo.*
- Aditza, preposizioa, eta izen soila ez den izen-sintagma (**adi + prep + IS**): *dejar a medio camino, entrar en uso de razón, pagar en dinero contante.*
- Aditza eta partizipioa (**adi + part**): *pillar inadvertido, dejar frito, estar preocupado.*
- Aditza eta gerundioa (**adi + ger**): *salir corriendo, estar ardiendo, andar endredando.*
- Aurreko multzoetan sartzen ez direnak (**beste**): *estar por suceder, dar mucho que decir, entablar amistad con alguien.*

Kasu- eta postposizio-marka guztiak kontuan hartuta, gehien ageri den ordain mota aditz soila da nabarmen: ordain guztien % 58,07. Ez da harritzekoa, bai baitakigu euskaraz beste hizkuntza batzuetan baino ohikoagoak direla hitz batez baino gehiagoz osatutako aditz konplexu edo elkartuak (Zabala, 2004; Azkarate Villar, 1990), eta askotan itzultzen direla gaztelaniazko aditz soilen bidez.

Interesgarria da, hala ere, beste ordainen egiturei ere begiratutxo bat ematea. Aditz soilak alde batera utzita, lau egitura nabarmentzen dira besteen gainetik: *adi + det + ize, adi + ize, adi + prep + ize* eta *adi + prep + det + ize*, hurrenez hurren. Egitura horiexek dira, hain zuzen ere, gaztelaniazko konbinazioetan bildu ditugunak (ikus 4.2.2 atala), eta, agerraldien arabera antolatuta, ordena ere mantendu egiten da.

Egitura	Agerr.	Ehunekoa
adi	3711	% 58,07
adi + det + ize	790	% 12,36
adi + ize	544	% 8,51
adi + prep + ize	363	% 5,68
adi + prep + det + ize	275	% 4,30
beste	230	% 3,60
adi + AdjS	155	% 2,43
adi + IS	96	% 1,50
adi + AdbS	76	% 1,19
adi + prep + IS	67	% 1,05
adi + part	63	% 0,99
adi + ger	21	% 0,33

7. taula: Gaztelaniazko ordainen agerraldiak.

Bestalde, markaz markako azterketa eginez gero, absolutiboan dauden konbinazioen ordainek atentzioa ematen dute, denen ia bi heren hartzen baitituzte aditz soilek bakarrik: *uzta bildu – cosechar; oreka galdu – desequilibrar...* Gainontzeko ordainetan, berriz, agerpen gehien dituztenak *adi + det + ize* eta *adi + ize* egiturak dira: *itxura egin – hacer el paripé; hotz izan – hacer frío.*

Horrez gain, izenek postposizio-markaren bat izateak ere badu eraginik ordainen egituretan (ikus 8. taula). Izan ere, sarri erabiltzen diren beste markei –ablatiboari, adlatiboari, inesiboari eta instrumentalari– begiratuz gero, aditz soilen ondoren preposiziodun egiturak nagusitzen direla ikusten da, absolutibodun konbinazioen ordainetan ez bezala: *armairutik atera – salir del armario; eskura ordaindu – pagar en mano; bistan egon – saltar a la vista; barrez ito – morirse de risa.*

EU	ES egitura	Ehunekoa
	adi	% 63,86
abs	adi + det + ize	% 14,52
	adi + ize	% 9,81
	adi	% 29,50
abl	adi + prep + det + ize	% 22,30
	adi + prep + ize	% 13,67
	adi	% 37,28
ala	adi + prep + det + ize	% 19,74
	adi + prep + ize	% 13,60
	adi	% 34,64
ine	adi + prep + ize	% 19,35
	adi + prep + det + ize	% 14,73
	adi	% 60,32
ins	adi + prep + ize	% 18,25
	adi + prep + det + ize	% 8,73

8. taula: Gaztelaniazko ordainik ohikoenak markaren arabera.

Hori horrela, eta euskarazko postposizio-markak gehienetan preposizioen bidez itzultzen direla jakinik, batek pentsa lezake konbinazioen itzulpena ez dela hain irregularra zentzu horretan, eta, hein batean, zuzen legoke. Izan ere, kasuan kasuko azterketa eginda, ikusi dugu postposizio-marka batzuen baliokidetzat ageri diren preposizioak nahiko sarri direla edozein hiztunek espero litzakeenak. Ablatiboaren pare, esaterako, *de* eta *por* ageri dira ordainen % 86,89tan (*ahotik kendu – quitar de la boca; zeharretik irten – salirse por la tangente*); eta adlatiboaren ordez, berriz, *a* erabili da % 76,47tan (*belarrira esan – decir al oído*).

Dena den, baliokidetza horiek hein batean baino ez dira erregularrak, postposizio-marka guztiekin ez baita gauza bera gertatzen. Inesiboa, adibidez, % 57,7tan baino ez da *en*, *por* edo *sobre* preposizioen bidez itzuli: *baxoerditan ibili – ir de poteo, txantxetan hartu – tomar a broma.*

4.2.4 Euskarazko ordainak gaztelaniazko konbinazioaren arabera

Gaztelaniazko ordainekin egin dugun bezala, euskarazkoak ere egituraren arabera multzokatu ditugu:

- Aditz soila (**adi**): *geldotu, neskazahartu, txunditu.*
- Absolutibodun izena eta aditza (**ize (abs) + adi**): *bizia arriskatu, ahotsa goratu, eskua jaso.*
- Postposizio-markadun izena eta aditza (**ize (pos) + adi**): *buruan sartu, berriketan ibili, negarrari eman.*
- Adberbioa edo adberbio-sintagma eta aditza (**AdbS + adi**): *azkarrago ibili, alferrik galdu.*
- Adjektiboa edo adjektibo-sintagma eta aditza (**AdjS + adi**): *nabaria izan, izugarria izan, argal mantendu.*
- Izen soila ez den izen-sintagma (absolutiboan) eta aditza (**IS (abs) + adi**): *kontu ezaguna izan, hitzaren jabe egon.*
- Izen soila ez den izen-sintagma (postposizio-markaren batekin) eta aditza (**IS (pos) + adi**): *bere onetik atera, bere kabuz utzi.*
- Aurreko multzoetan sartzen ez direnak (**beste**): *amaitutzat jo, tentuz ibiltzeko esan, hortzen artean hitz egin.*

Multzo bakoitzaren agerpen-kopuruak eta ehunekoak 9. taulan bildu ditugu.

Egitura	Agerr.	Ehunekoa
ize (abs) + adi	2321	% 35,24
adi	1550	% 23,53
ize (pos) + adi	876	% 13,30
beste	576	% 8,74
AdbS + adi	424	% 6,44
AdjS + adi	367	% 5,57
IS (abs) + adi	269	% 4,08
IS (pos) + adi	204	% 3,10

9. taula: Euskarazko ordainen agerraldiak.

Euskarazko ordain guztiak kontuan hartuz gero, argi eta garbi ikusten da bi egitura beste guztiak baino askoz ere gehiago errepikatzen direla: *ize (abs) + adi* motako konbinazioak lehenik, eta aditz soilak ondoren. Nolanahi ere, gaztelaniazko konbinazioak egituraren arabera bereizita, ikusi dugu determinatzaileen eta preposizioen agerpenak baduela zerikusia ordain motarekin.

Har ditzagun, esaterako, aditzez eta izenez osaturiko konbinazioak batetik, eta aditzez, determinatzailez eta izenez osaturikoak bestetik. Agerpen gehien dituzten ordainak *ize (abs) + adi* multzokoak eta aditz soilak dira bi kasuetan, eta gainontzeko ordain motak oso gutxi agertzen dira bi horien aldean. Hala ere, gaztelaniaz determinatzailerik ez duten konbinazioetan, *ize (abs) + adi* multzoko ordainen eta aditz soilen arteko aldea ez da hain nabarmena agerraldiak kontuan hartuz gero; determinatzailea duten konbinazioetan, aldiz, *ize (abs) + adi* egiturako ordainak aditz soilak baino ia hiru bider gehiago dira (ikus aldea 10. taulan).

ES	EU egitura	Ehunekoa
adi+det+ize	ize (abs) + adi	% 50,34
	adi	% 18,56
adi+ize	ize (abs) + adi	% 46,34
	adi	% 31,38

10. taula: Preposiziorik gabeko konbinazioen ordainik ohikoenak.

Horrez gain, hemen ere badago loturarik preposizioen eta postposizio-marken artean. Izan ere, aurreko bi kasuetan absolutibo-markadun izenak nagusi baziren ere, preposiziodun konbinazioen ordainetan ohikoagoak dira postposizio-markaren bat daramaten izenak. Hain zuzen ere, gaztelaniazko konbinazioak *adi + prep + ize* edo *adi + prep + det + ize* multzoetakoak direnean, sarrien agertzen den ordain mota *ize (pos) + adi* da.

ES	EU egitura	Ehunekoa
adi+prep+det+ize	ize (pos) + adi	% 29,78
	adi	% 25,51
	AdbS + adi	% 11,25
	ize (abs) + adi	% 9,31
adi+prep+ize	ize (pos) + adi	% 26,16
	adi	% 23,00
	ize (abs) + adi	% 14,42
	AdbS + adi	% 12,10

11. taula: Preposiziodun konbinazioen ordainik ohikoenak.

Aurreko atalean bezala, hemen ere preposizioen baliokidetzat agertzen diren postposizio-markak, batzuetan, aurreikusteko modukoak dira: *en* preposizioa duten konbinazioen ordainetatik % 87,68k inesibodun izen bat daramate (*estar en la inopia – ametsetan egon*); *con* preposizioaren ordez instrumentala eta soziatiboa agertzen da ordainen % 87,5etan (*andar con cuidado – kontuz ibili, ir con el cuento – koplarekin etorri*); eta *por* preposizioa ere ablatiboaren eta inesiboaren bidez ordezkatu da % 88,56etan (*pasar por el tamiz – galbahetik pasatu, pasar por las armas – armetan iragan*).

Hemen ere, ordea, hein batean baino ez da erregulartasun hori gordetzen. *A* preposizioa, esaterako, % 56 kasutan bakarrik itzuli da adlatiboaren edo datiboaren bidez (*traer a la memoria – gogora ekarri*), eta gainontzeko ordain guztiek beste marka batzuk dituzte: *andar a la greña – istilutan ibili, saltar a la vista – begi-bistakoa izan.*

Azkenik, interesgarria da nabarmentzea adberbioak ere nahiko maiz agertzen direla preposiziodun konbinazioen ordainetan; 11. taulan ikus daitekeenez, *adi + prep + det + ize* multzoan, absolutibodun izenekin sortzen diren egiturak baino are sarriago agertzen dira adberbioekin sortzen direnak: *caer de su peso – argi egon, dar en el clavo – bete-betean asmatu,* etab.

4.2.5 Mugatasuna eta numeroa: euskarazko eta gaztelaniazko konbinazioen alderaketa

Ezaugarri morfosintaktikoak aztertzen eta alderatzen jarraitzeko, mugatasunari eta numeroari begiratu diegu, hizkuntza batetik bestera mantentzen ote diren jakiteko. Konbinazio-zerrenda osotik zati bat bakarrik hartu dugu, hain zuzen ere, ordaintzat ere konbinazioak dituzten konbinazioez osatua.

Euskaraz hiztegi-sarrera diren konbinazioetan, mugatu singularrean (*umea izan, ahoa garbitu*) daude izenen ia erdia, eta mugagabearen (*ohar egin, zerraldo utzi*) eta mugatu pluralaren (*goraintziak eman, erroak bota*) agerraldien artean ez dago desberdintasun handirik. Bestalde, konbinazioen hamarren bat baino gehiago zalantzazko kasuak dira (*denbora egin, lotsa izan*), ezin baita zehatz jakin mugatu singularrean ala mugagabean dauden.

Euskarazko konbinazioen mugatasunari eta numeroari buruzko informazioa 12. taulan bildu dugu, ehunekotan, bai eta ezaugarri hori gaztelaniaz zenbatetan gordetzen den ere.

	EU konbinazioak	ES gordea
s	% 48,48	% 54,23
mg	% 21,99	% 80,72
pl	% 18,08	% 23,44
*	% 11,46	–

12. taula: Mugatasuna eta numeroa euskaratik gaztelaniara.

Eskuineko zutabean, euskarazko ezaugarria gaztelaniaz zenbatetan gorde den ikus daiteke. Deigarriena mugatu pluralaren kasua da beharbada, baliokideen laurdena baino ez baitago mugatu pluralean gaztelaniaz. Bestalde, mugagabearen kasuan kontrakoa gertatzen da, ordain gehienak mugagabeak baitira hizkuntzaz aldatuta ere.

Gaztelaniazko hiztegi-sarrerak abiapuntutzat hartuz gero, ordea, emaitzak aldatu egiten dira.

	ES konbinazioak	EU gordea
mg	% 50,25	% 27,09
s	% 40,56	% 61,94
pl	% 9,19	% 50,60

13. taula: Mugatasuna eta numeroa gaztelaniatik euskarara.

Batetik, hiztegi-sarrera gehien-gehienak mugagabean daude (*dar alas, estar en auge*), eta mugatu singularrean ere asko (*perder el juicio, quitar de la cabeza*); mugatu pluralean, berriz, oso konbinazio gutxi daude (*parar los pies, subirse por las paredes*).

Bestetik, baliokideek mugatasuna eta numeroa zenbatetan gordetzen duten begiratuz gero, mugagabeak atentzioa ematen du. Euskaratik gaztelaniara oso sarri gordetzen bazen ere, gaztelaniatik euskarara ordainen % 27,09tan baino ez da gorde. Mugatu pluralari dagokionez, ordea, baliokideen erdiak daude mugatu pluralean, eta ez, euskaratik gaztelaniara bezala, laurdena baino gutxiago.

4.3 Aditzak eta izenak hizkuntza batetik bestera

Gure azterketaren lehen atalean (ikus 4.1. atala), hizkuntza bateko eta besteko konbinazioen osagai nagusiak aztertu ditugu: aditzak eta izenak. Oraingoan, berriz, konbinazioen osagaiak euren ordainenekin alderatu ditugu. Mugatasuna eta numeroa aztertzeko egin dugun bezala (ikus 4.2.5. atala), izenez eta aditzez osatutako ordainak bakarrik hartu ditugu, dagozkien hiztegi-sarrerekin batera. Gero, bi hizkuntzetako aditzak eta izenak parekatu, eta hiztegian bilatu dugu ea bata bestearen ordaintzat ageri diren.

Adibidez, *deabruetara bidali – mandar al infierno* bikotea hartuta, *deabru – infierno* eta *bidali – mandar* hiztegian baliokidetzat jasota ote dauden begiratu dugu. Kasu horretan, aditza bakarrik mantendu da, *deabru* sarreraren ordainetan ez baita *infierno* agertzen.

Hasi aurretik, gure irudipena zen nahiko gutxitan agertuko zirela ordaintzat bai izena eta bai aditza, eta, oro har, gure ustea bete da, nahiz eta proportzioa desberdina izan euskaratik gaztelaniarako eta gaztelaniatik euskararako bikoteetan.

Gure kontaketetatik atera ditugun portzentaiak 14. taulan jaso ditugu.

	ize	**adi**	**biak**	**bat ere ez**
eu–es	% 23,84	% 23,71	% 28,01	% 24,44
es–eu	% 19,14	% 22,84	% 21,79	% 36,23

14. taula: Izenen eta aditzen baliokidetza hiztegietan.

Taulako azken zutabeak atentzioa ematen du, izan ere, gaztelaniatik euskararako konbinazio-pareetan, herena baino gehiago dira ez izenik eta ez aditzik baliokide ez dutenak. Euskaraz, berriz, zifrarik altuena kontrako kasuena da, bai izena eta bai aditza ordaintzat dituzten pareena, alegia.

Nolanahi ere, zentzu batean zein bestean, argi ikusten da izen+aditz konbinazioen itzulpena oso gutxitan egin daitekeela hitzez hitz, osagaietako bat behintzat ezin baita normalean hartzen duen ordainez ordezkatu.

Horrez gain, kontuan izan behar da guk hitzak hiztegian agertzen diren ala ez bakarrik begiratu dugula baina, baiezko kasuetan, ez dugula zehaztu ordaina ohikoa den ala ez. Esaterako, *bidea urratu – abrir camino* parean, aditzak baliokidetzat hartu ditugu, *urratu* sarreraren barruan *abrir* agertzen delako ordainen artean; euskaraz eta gaztelaniaz egiten dugunok, ordea, nekez erabiliko genituzke ordaintzat, *bide* eta *camino*ren alboan ez bada.

Hortaz, 14. taulak ere agerian uzten du hitzkonbinazioak zenbateraino diren irregularrak eta, hori kontuan hartuta, zenbateraino den beharrezkoa beste itzulpen-estrategia batzuk bilatzea, xede-hizkuntzan testu txukun bat sortu nahi bada.

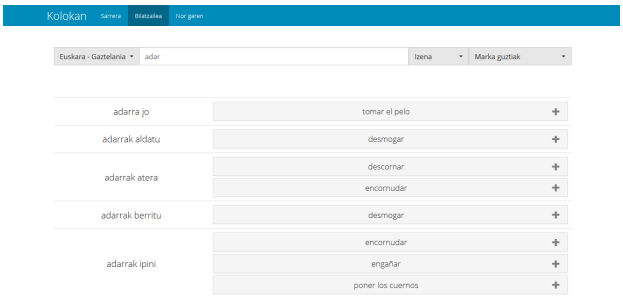
5 Koloka: izen+aditz konbinazioen informazio linguistikoa biltzen duen datubasea

Hiztegien gainean egin dugun azterketan, izen+aditz konbinazioei buruzko informazio ugari bildu dugu, eta datu-base bat sortu dugu informazio horrekin guztiarekin. *Koloka*[8] interfaze publikoak aukera ematen du datu-base horretan bilaketak egiteko, eta erabiltzaileen eskura jartzen du euskarazko eta gaztelaniazko 5.604 konbinazioren eta 12.979 ordainen inguruko informazio linguistikoa.

Bilatzaileak hainbat aukera eskaintzen ditu. Batetik, bilaketa zein hizkuntzatatik abiatuta egin nahi den aukera daiteke: euskaratik gaztelaniara, ala gaztelaniatik euskarara. Bestetik, interfazeak aukera ematen du bilaketa konbinazio osoaren, aditzaren edo izenaren arabera egiteko. Eta, horrez gain, euskara-gaztelania zentzuan, euskarazko izenaren kasu- eta postposiziomarka zehaz daiteke, eta gaztelania-euskara zentzuan, berriz, konbinazioaren egitura.

Bigarren irudian ikusten denez, *adar* izena bilatuta, landu ditugun konbinazioen artean hitz horrekin sortu direnak erakusten dira, ordainekin batera. Euskarazko konbinazio bat gaztelaniara nola itzultzen den jakin nahi duten erabiltzaileek, ziur asko, nahikoa izango dute emaitza horrekin, baina informazio linguistiko gehiago lortu nahi dutenek ere badute horretarako aukera, ordain bakoitzaren ondoan dagoen [+] ikonotxoa sakatuta.

[8] `http://ixa2.si.ehu.es/koloka`



2. irudia: *Koloka* interfazearen itxura, *adar* izena bilatuta.

Zubiak eraiki konbinazioa bilatuta, esaterako, bi ordain agertzen dira: *construir puentes* eta *tender puentes* (ikus 3. irudia). Bigarrenaren informazioa zabalduz gero, erabiltzaileak hainbat datu jasoko ditu: *zubiak eraiki* absolutiboan dagoela, eta *tender puentes*, berriz, aditz batez eta izen batez osatu dela; euskarazko konbinazioa mugatu pluralean dagoen bitartean, gaztelaniazkoa mugagabean dagoela; izenak hiztegian ordaintzat jasota daudela, baina ez aditzak; eta konbinazio hori Elhuyar euskara-gaztelania hiztegitik jaso dugula.

3. irudia: *Zubiak eraiki* bilaketaren emaitza.

6 Ondorioak eta etorkizuneko lanak

Azterketa linguistiko honen helburua hitz-konbinazioen konplexutasuna agerian uztea izan da, argi gera dadin zeinen garrantzitsua den Hizkuntzaren Prozesamenduko aplikazioetan horrelako egiturei tratamendu berezia ematea.

Gure ikerketa-lan nagusia corpusetan oinarritzeko asmoa badugu ere, hiztegi elebidunak erabili ditugu lehenik, konbinaziorik usuenak biltzen dituztelakoan.

Horretarako, hizkuntza bateko eta besteko izen+aditz konbinazioak eta euren ordainak aztertu ditugu hiru ataletan. Lehenik, euskarazko eta gaztelaniazko konbinazioetan ageri diren aditzei eta izenei begiratu diegu, eta ikusi dugu badagoela antzekotasunik euren artean –aditzen kasuan, bereziki–, nahiz eta gaztelaniazko aditzzerrenda euskarazkoa baino dezente zabalagoa izan. Bigarrenik, euskarazko zein gaztelaniazko konbinazioen eta ordainen egiturak parekatu ditugu, eta, loturaren bat aurkitu badugu ere, gure emaitzek argi uzten dute ez dagoela konbinazioen itzulpenetarako arau morfosintaktiko orokorrik sortzerik. Azkenik, ezaugarri semantikoei ere eman diegu begiratu bat, eta, ordaintzat konbinazio bat duten konbinazioetan, izenak izenekin eta aditzak aditzekin baliokide ote diren jakin nahi izan dugu. Atal horretan ere agerian geratu da hitzez hitz itzul daitezkeen konbinazioak oso bakanak direla.

Oro har, beraz, gure ustea bete da: antzekotasunak antzekotasun, izen+aditz konbinazioak oso aldakorrak dira euskaratik gaztelaniara eta gaztelaniatik euskarara, eta behar-beharrezkoa da egitura horiei tratamendu berezi bat ematea, HPko tresnen bidez emaitza egokiak lortu nahi badira.

Azkenik, gure azterketaren emaitza guztiak sarean jarri ditugu eskuragarri, `http://ixa2.si.ehu.es/koloka` helbidean, edozein erabiltzailek bilaketak egiteko moduan.

Aurrera begira, hiztegien gainean hasitakoa corpusetara eramatea da gure helburua, orain arte aztertutako konbinazioak testu errealetan zenbat erabiltzen diren ikusteko, eta gure konbinazio-zerrendak zabaltzeko. Hartara, gure beharretara hobeto egokituko den datu-base bat osatu ahal izango dugu.

Horrez gain, azterketa hau alderdi praktikora eraman nahi dugu, eta ikuspegi linguistikotik landu ditugun konbinazioetako bakoitzak itzulpen automatikoan zer motatako arazoak sortzen dituen aztertu.

Eskerrak

Lan hau Ekonomia eta Lehiakortasun Ministerioko doktoretza aurreko diru-laguntza bati esker egin ahal izan dugu (BES-2013-066372), SKATeR proiektuaren barruan (TIN2012-38584-C06-02).

Eskerrak eman nahi dizkiogu Ruben Urizarri gurekin aholkulari-lanetan aritu izanagatik, eta baita Mikel Artetxeri ere, bilaketak egiteko interfazea prestatu izanagatik.

Horrez gain, Elhuyarren laguntza ere ezinbestekoa izan da azterketa honetarako, haiek eman baitigute gure lanaren oinarri izan den materiala.

Bibliografía

Aduriz, Itziar, Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Nerea Ezeizai, & Ruben Urizar. 1996. Euslem: A lemmatiser/tagger for basque. *Proc. Of EURALEX'96*, pages 17–26.

Alonso Ramos, Margarita. 1995. Hacia una definición del concepto de colocación: de jr firth a ia mel'cuk. *Revista de Lexicografía*, 1:9–28.

Azkarate Villar, Miren. 1990. *Hitz elkartuak euskaraz*. Filosofi-Letren Fakultatea, Deustuko Unibertsitatea.

Baldwin, Timothy & SuÑam Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing, second edition. Morgan and Claypool*.

Corpas Pastor, Gloria. 2001. La traducción de unidades fraseológicas: técnicas y estrategias. *La lingüística aplicada a finales del siglo XX. Ensayos y propuestas, Alcalá, Universidad de Alcalá*, 2:779–787.

Corpas Pastor, Gloria. 2004. *Diez años de investigación en fraseología: análisis sintácticosemánticos, contrastivos y traductológicos*. Madrid: Vervuert.

Gurrutxaga, Antton. 2014. *Idiomatikotasunaren karakterizazio automatikoa: izena+aditza konbinazioak*. Ph.D. thesis, UPV-EHU.

Heylen, Dirk & Kerry Maxwell. 1994. Lexical functions and the translation of collocations. In *Proceedings of Euralex*.

Iñurrieta, Uxoa. 2013. Itzulpen automatikorako patroi-bilaketa. Master's thesis, UPV-EHU.

Lü, Yajuan & Ming Zhou. 2004. Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 167. Association for Computational Linguistics.

Mayor, Aingeru, Iñaki Alegria, Arantza Díaz De Ilarraza, Gorka Labaka, Mikel Lersundi, & Kepa Sarasola. 2011. Matxin, an opensource rule-based machine translation system for basque. *Machine translation*, 25(1):53–82.

Mel'ćuk, Igor A. 1998. Collocations and lexical functions. *Phraseology. Theory, Analysis, and Applications*, pages 23–53.

Rafel, Joan. 2004. Los predicados complejos en español. In *Las fronteras de la composición en lenguas románicas y en vasco*, pages 445–534. Servicio de Publicaciones.

Sanz Villar, Zuriñe. 2011. Alemanetik euskarara itzulitako unitate fraseologikoen azterketarako jarraibideak. *Senez: itzulpen aldizkaria*, 41:125–139.

Seretan, Violeta. 2013. On collocations and their interaction with parsing and translation. *Informatics*, 1(1):11–31.

Smadja, Frank, Kathleen McKeown, & Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38.

Urizar, Ruben. 2012. *Euskal lokuzioen tratamendu konputazionala*. Ph.D. thesis, UPV-EHU.

Wehrli, Eric, Violeta Seretan, Luka Nerima, & Lorenza Russo. 2009. Collocations in a rule-based mt system: A case study evaluation of their translation adequacy. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, pages 128–135.

Zabala, Igone. 2004. Los predicados complejos en vasco. In *Las fronteras de la composición en lenguas románicas y en vasco*, pages 445–534. Deustuko Unibertsitatea.

Extração de Relações utilizando Features Diferenciadas para Português*

Relation Extraction using Different Features in Portuguese

Erick Nilsen Pereira de Souza
Universidade Federal da Bahia
ericknilsen@gmail.com

Daniela Barreiro Claro
Universidade Federal da Bahia
dclaro@ufba.br

Resumo

A Extração de Relações (ER) é uma tarefa da Extração da Informação responsável pela descoberta de relacionamentos semânticos entre conceitos em textos não estruturados. Quando a extração não é limitada por um conjunto predefinido de relações, a ER é dita Aberta, cujo principal desafio consiste em reduzir a proporção de extrações inválidas no universo de relações identificadas. As soluções atuais, baseadas em aprendizado sobre um conjunto de features linguísticas específicas, embora consigam eliminar grande parte das extrações inválidas, possuem como desvantagem a alta dependência do idioma. Tal dependência decorre da dificuldade inerente à determinação do conjunto de features mais representativo para o problema, considerando as peculiaridades de cada língua. Neste sentido, o presente trabalho propõe avaliar as dificuldades da classificação baseada em features na extração de relações semânticas abertas em Português, com o objetivo de embasar novas soluções capazes de reduzir a dependência do idioma nesta tarefa. Os resultados obtidos indicam que nem todas as *features* representativas em Inglês podem ser mapeadas diretamente para a Língua Portuguesa com méritos de classificação satisfatórios. Dentre os algoritmos de classificação avaliados, o J48 apresentou os melhores resultados com uma medida-F de 84,1%, seguido pelo SVM (83,9%), Perceptron (82,0%) e Naive Bayes (79,9%).

Palavras chave

Extração de Relações Abertas, Seleção de Características

Abstract

Relation Extraction (RE) is a task of Information Extraction (IE) responsible for the discovery of semantic relationships between concepts in unstructured text. When the extraction is not limited to a predefined set of relations, the task is called Open Relation Extraction, whose main challenge is to reduce the proportion of invalid extractions in the universe of relationships identified. Current methods based on a set of specific machine learning features eliminate much of the invalid extractions. However, these solutions have the disadvantage of being highly language-dependent. This dependence arises from the difficulty in finding the most representative set of features to the Open RE problem, considering the peculiarities of each language. In this context, the present work proposes to assess the difficulties of classification based on features in open relation extraction in Portuguese, aiming to base new solutions that can reduce language dependence in this task. The results indicate that many representative features in English can not be mapped directly to the Portuguese language with satisfactory merits of classification. Among the classification algorithms evaluated, J48 showed the best results with a F-measure value of 84.1%, followed by SVM (83.9%), Perceptron (82.0%) and Naive Bayes (79,9%).

Keywords

Open Relation Extraction, Feature Selection

1 Introdução

Embora a quantidade de documentos não estruturados publicados na Web cresça a cada ano, a velocidade com que o ser humano consegue interpretar informações permanece constante. Por conta disso, técnicas de Extração da Informação (EI) vêm sendo desenvolvidas com o intuito de identificar conteúdo relevante em grandes quantidades de documentos (Brin, 1998; Feldman e Sanger, 2007; Lutz e Heuser, 2013). Métodos de reconhecimento de conceitos e seus relacionamentos são considerados cruciais em diversas aplicações de processamento linguístico, tais como na construção automática de ontologias e léxicos computacionais (Chaves, 2008),

*Agradecimentos à FAPESB pelo apoio parcial neste projeto.

em sistemas de respostas a perguntas (Hirschman e Gaizauskas, 2001) e na computação forense (Anyanwu, Maduko e Sheth, 2005). Porém, as principais soluções para extração de relações entre conceitos são limitadas por um conjunto predefinido de relações possíveis, o que reduz a aplicabilidade dos métodos a domínios e idiomas específicos.

Um exemplo de aplicação de EI onde a limitação de domínio e idioma constitui um fator proibitivo é no Reconhecimento de Entidades Mencionadas (REM) aplicado à computação forense. Autores em (Dalben e Claro, 2011) afirmam que a identificação de nomes de pessoas e organizações em mídias apreendidas pode reduzir em mais de 90% a quantidade de arquivos analisados manualmente por peritos criminais. Em aplicações deste tipo, é comum que a coleção de documentos contenha vocábulos de domínios e idiomas distintos, pois uma mesma investigação pode envolver organizações com atuações diferentes (como uma clínica médica e um órgão público) em mais de um país. Pelo mesmo motivo, o requisito de independência do domínio se mantém na tarefa de Extração de Relações (ER) entre as entidades identificadas nesses documentos.

Estudos recentes têm sido desenvolvidos com o intuito de contornar as limitações dos métodos tradicionais de ER (Souza e Claro, 2014). Nesse contexto, a Extração de Relações Abertas, derivada da *Open Information Extraction (Open IE)* (Banko e Etzioni, 2008), consiste na tarefa de extrair relações semânticas com vocabulário não-limitado a partir de *corpora* em larga escala. Entretanto, a ambiguidade inerente à linguagem natural tem ocasionado grande proporção de relações inválidas, exemplificadas nas sentenças da Tabela 1.

Uma relação é dita inválida quando é incoerente e/ou incompleta. Intuitivamente, uma extração incoerente ocorre quando a semântica do relacionamento entre as entidades, mesmo sendo completa, não condiz com a interpretação correta da sentença. A primeira linha da Tabela 1 mostra um exemplo de extração incoerente, já que a entidade *Defesa do Criciúma* rebate um objeto que está oculto na frase (a bola), e não a entidade *Maurinho*. Já na segunda linha, *vai emoldurar com* não denota uma relação com sentido completo entre as entidades *PT* e *Luiz Inácio Lula da Silva*.

Sentença	**Extração Inválida**
"Depois de a defesa do Criciúma rebater, Maurinho chutou e marcou."	*(Defesa do Criciúma, rebater, Maurinho)*
"A estrela símbolo do PT vai emoldurar com destaque o cenário dos programas do candidato Luiz Inácio Lula da Silva."	*(PT, vai emoldurar com, Luiz Inácio Lula da Silva)*

Tabela 1: Exemplos de extrações inválidas.

A distinção automática entre relações válidas e inválidas pode ser modelada como um problema de classificação. Trabalhos em (Banko e Etzioni, 2008) e (Fader, Soderland e Etzion, 2011) aplicam algoritmos de aprendizado de máquina sobre *features* extraídas das sentenças para elevar a precisão de classificação das relações. A principal desvantagem dessas abordagens é a dificuldade na seleção de *features* adequadas à tarefa. Além disso, o aprendizado baseado em *features* necessita de bases de treinamento relativamente grandes para gerar resultados satisfatórios. Recursos deste tipo são escassos ou inexistentes na maioria dos idiomas.

Neste trabalho é realizada uma análise do esforço necessário à identificação das *features* mais representativas para a classificação de relações semânticas abertas em textos redigidos em Português, já que a capacidade preditiva de um atributo pode sofrer grande variação em função da mudança de idioma, dificuldade que deve ser considerada em novas soluções capazes de reduzir tal dependência nesta tarefa. O presente artigo está organizado como segue. Na Seção 2 é descrita uma classificação dos métodos, além dos principais conceitos referentes à Extração de Relações (ER). A Seção 3 apresenta as características das abordagens mais recentes de ER Abertas. Na Seção 4 são descritos os experimentos realizados e analisados os resultados obtidos. A Seção 5 conclui este artigo e apresenta alguns trabalhos futuros.

2 Relações Semânticas

A Extração de Relações (ER) consiste na tarefa de descobrir relacionamentos semânticos entre conceitos em documentos não estruturados (Feldman e Sanger, 2007). Embora não exista uma categorização clara dos métodos de ER, é possível agrupá-los a partir dos principais trabalhos apresentados na literatura. Nesta seção são descritos dois tipos de classificação na tarefa de ER: i) Por técnica aplicada; ii) Por tipo de relação extraída.

2.1 Classificação por técnica aplicada

A classificação mais genérica dos métodos de ER distingue as abordagens baseadas em padrões textuais das que utilizam aprendizado de máquina (Taba e Caseli, 2012). A seguir é feita uma breve descrição das principais características de cada tipo de método.

Os métodos de padrões textuais extraem relações utilizando regras formadas por expressões regulares. Um exemplo deste tipo de regra, que pode ser encontrado em (Hearst, 1992), é dado por:

$$NP_1\{,\}\ especially\ \{NP_2, NP_3...\}\ \{or|and\}\ NP_n \quad (1)$$

Com este padrão é possível identificar relações de hiponímia do tipo *is-a* entre as frases nominais NP_i e NP_1, com $i \in \{2, 3, ..., n\}$. Tomando como exemplo a frase "*most countries, especially France, England and Spain*" ("a maioria dos países, especialmente Franca, Inglaterra e Espanha"), a aplicação da regra permite extrair as seguintes relações: *is-a(France, country)*, *is-a(England, country)* e *is-a(Spain, country)*.

É possível elencar uma série de deficiências e limitações nos métodos baseados em padrões textuais. Primeiro, a especificidade das regras resulta em alta precisão, mas baixa cobertura (Freitas e Quental, 2007; Snow, Jurafsky e Ng, 2005). Segundo, devido à grande diversidade das variações linguísticas, certos padrões podem ser associados a diversos tipos de relações, tornando inviável o mapeamento de todas as possibilidades (Girju et al., 2010). Por exemplo, o padrão "tais como" é comumente reduzido à palavra denotativa "como" em textos escritos em Português, que pode pertencer às seguintes classes morfológicas: conjunção, pronome relativo, substantivo, advérbio interrogativo, advérbio de modo, interjeição e preposição. Entretanto, o único sentido da palavra "como" que deve ser reconhecido pelo referido padrão é o equivalente a "por exemplo" (pronome relativo). Por conta disso, a criação de uma base de regras minimamente representativa para esse tipo de método consiste em uma tarefa altamente dispendiosa. Trabalhos recentes vêm apresentando resultados mais efetivos em termos de precisão e cobertura, através de técnicas de aprendizado de máquina.

As abordagens baseadas em aprendizado de máquina selecionam atributos (*features*[1]) a partir de um conjunto de treinamento, a fim de determinar se existe uma relação entre as entidades de uma nova instância (Kambhatla, 2004). Mais precisamente, dada uma sentença $S = w_1, w_2, ..., e_1, ..., w_j, ..., e_2, ...w_n$, onde e_1 e e_2 são entidades existentes entre as palavras $w_1, w_2, ..., w_n$, uma função de mapeamento f é definida por:

$$f_R(\Theta(S)) = \begin{cases} +1, & \text{se existe R entre } e_1 \text{ e } e_2, \\ -1, & \text{caso contrário} \end{cases} \quad (2)$$

Onde $\Theta(S)$ constitui o conjunto de *features* extraídas de S e R representa a relação semântica. Assim, a Equação 2 decide se existe uma relação semântica R entre as entidades e_1 e e_2.

Além das soluções baseadas em *features*, existem trabalhos que utilizam uma generalização da similaridade de subsequências de strings (*string-kernels* (Zelenko, Aone e Richardella, 2003)) para a realização de treinamentos. Considerando duas strings x e y, a similaridade $K(x, y)$ em *string-kernels* é calculada em função do número de subsequências que são comuns a ambas. Ou seja, quanto maior a quantidade de subsequências comuns entre x e y, maior a similaridade entre elas.

Partindo deste princípio, sendo A e B exemplos de sentenças com relação positiva e negativa entre duas entidades, respectivamente, no conjunto de treinamento, a função de similaridade que indica a classe de uma instância de teste T é calculada com base na seguinte equação:

$$f_R(K) = \begin{cases} +1, & \text{se } K({S_A}^+, S_T) > K({S_B}^-, S_T), \\ -1, & \text{caso contrário} \end{cases} \quad (3)$$

Onde ${S_A}^+$, ${S_B}^-$ e S_T representam os respectivos conjuntos constituídos pelos termos que cercam as entidades nas sentenças A, B e T. Como exemplo, considerando a sentença *"O campus da UFBA está situado em Ondina"*, as palavras *campus* e *situado* indicam uma relação do tipo *localidade* entre as entidades *UFBA* e *Ondina*, cujas similaridades com os termos que cercam entidades em outras sentenças podem ser utilizadas para extrair delas o mesmo tipo de relação.

A Figura 1(a) mostra a classificação dos métodos de ER considerando o tipo de método.

2.2 Classificação por tipo de relação extraída

A semântica das relações extraídas varia bastante nos trabalhos de ER. Entretanto, é possível identificar dois tipos de métodos: os que extraem

[1] As features representam propriedades léxicas, sintáticas ou semânticas dos termos de uma sentença.

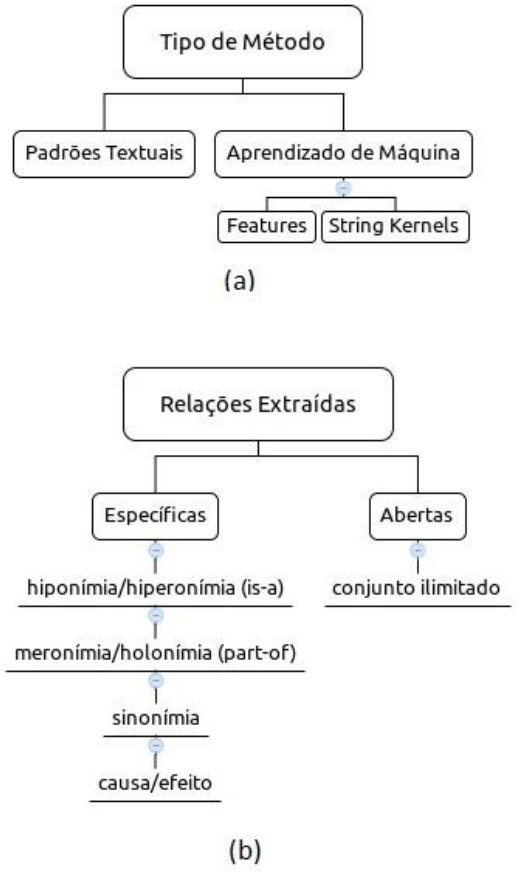


Figura 1: Classificação dos métodos de ER: (a) por tipo de método e (b) por tipo de relação.

relações específicas e os que extraem relações abertas. Um esquema que ilustra esta classificação é mostrado na Figura 1(b).

Na extração de relações específicas, um domínio finito de relações semânticas é definido para a tarefa de ER, conforme exemplos da Tabela 2.

A principal desvantagem dessa abordagem é a limitação da extração a um universo específico. Dessa forma, muitas relações semânticas importantes presentes no texto não são identificadas por não pertencerem ao domínio definido e nem ao conjunto predeterminado de relações.

Relação	location-of(algo/alguém, local)
Exemplo	Um aluno pode ser encontrado na escola
Extração	location-of(aluno, escola)
Relação	isa(subclasse, superclasse)
Exemplo	Salvador é uma cidade
Extração	is-a(Salvador, cidade)
Relação	part-of(todo,parte)
Exemplo	Roda é parte de um carro
Extração	part-of(roda, carro)

Tabela 2: Exemplos de relações específicas.

A descoberta de relações sem restrição de domínio representa um requisito essencial em diversas aplicações de EI. Por conta disso, estudos têm sido conduzidos no sentido de identificar relações de vocabulário não-limitado, caracterizando a Extração de Relações Abertas (do inglês, *Open Relation Extraction*) (Banko e Etzioni, 2008; Nakashole e Mitchell, 2014), tarefa abordada neste trabalho.

Como a categorização apresentada não é mutualmente exclusiva, os métodos de ER se enquadram em ambos os tipos de classificação, sendo possível identificar certas associações entre eles. Por exemplo, todas as abordagens de padrões textuais necessariamente extraem relações específicas (Hearst, 1992; Freitas e Quental, 2007; Girju et al., 2010). Por outro lado, existem abordagens de aprendizado de máquina utilizadas tanto na extração de relações específicas (Kambhatla, 2004; Zelenko, Aone e Richardella, 2003), quanto na extração de relações abertas (Banko e Etzioni, 2008; Fader, Soderland e Etzion, 2011). Nos métodos de extração de relações abertas investigados, as extrações são identificadas através de padrões morfológicos e classificadas utilizando aprendizado supervisionado.

Em relação à ER na Língua Portuguesa, percebe-se que a maioria dos trabalhos utiliza técnicas rudimentares baseadas em padrões textuais, sendo que as abordagem de aprendizado de máquina ainda são pouco exploradas. Isto se deve, possivelmente, à falta de recursos linguísticos em Português, dificultando a construção de bases de treinamento de forma automática ou semi-automática para a tarefa, que necessita de *features* representativas obtidas a partir de conhecimento especializado na língua. Dentre o universo de relações específicas extraídas em Português, as mais frequentes são as relações de hiponímia, meronímia e localidade (Oliveira, Santos e Gomes, 2010; Cardoso, 2008; Chaves, 2008; Bruckschen et al., 2008). Por outro lado, não foi identificada nenhuma pesquisa voltada para a Extração de Relações Abertas neste idioma.

Na próxima seção são descritas as principais características da ER Abertas.

3 Extração de Relações Abertas

Os métodos precursores de ER Abertas obtém extratos na forma *(e_1, frase relacional, e_2)* em três etapas (Fader, Soderland e Etzion, 2011):

1. **Etiquetação:** As sentenças são etiquetadas automaticamente através de heurísticas ou a partir de supervisão distante (treinamento semi-supervisionado);
2. **Aprendizado:** Um extrator de frases relacionais é treinado utilizando um modelo de etiquetação sequencial (e.g. CRF);
3. **Extração:** Um conjunto de argumentos (e_1, e_2) é identificado na sentença de teste. Em seguida, o extrator treinado na etapa 2 é utilizado para etiquetar as palavras contidas entre os argumentos e compor a frase relacional (caso ela exista), extraindo a relação no formato *(e_1, frase relacional, e_2)*.

Uma das desvantagens dessas abordagens reside no fato de que a etiquetação precisa ser realizada em uma quantidade muito grande de sentenças (na ordem de centenas de milhares) para que a etapa de aprendizado seja efetiva. Isto implica em alto custo de construção dos conjuntos de treinamento, além da demanda de recursos linguísticos sofisticados para viabilizar a etiquetação automática, dificilmente encontrados na maioria dos idiomas. Além disso, o método de extração por etiquetação sequencial é pouco eficaz em sentenças maiores, pois há um aumento da incerteza na associação de cada etiqueta a uma palavra à medida que a sequência cresce.

Abordagens mais recentes têm sido desenvolvidas para contornar algumas dessas limitações, por meio de modificações na metodologia e, consequentemente, nas estratégias adotadas nas etapas de extração (Fader, Soderland e Etzion, 2011; Banko et al., 2007; Banko e Etzioni, 2008). Assim, é realizada primeiramente a etapa de extração, seguida pelo aprendizado necessário à posterior classificação das relações, conforme descrito abaixo:

1. **Extração:** Inicialmente, um extrator baseado em padrões linguísticos (e.g. padrões verbais) seleciona uma sequência de palavras que representa a relação semântica entre e_1 e e_2, identificando frases relacionais que casam com esses padrões. Em seguida, se um conjunto de argumentos (e_1, e_2) for identificado na sentença de teste, então é gerada a relação na forma *(e_1, frase relacional, e_2)*;
2. **Aprendizado:** Um classificador de extrações é treinado por meio de um conjunto de *features* linguísticas;
3. **Classificação:** O classificador treinado na etapa 2 é utilizado para distinguir as relações válidas das inválidas geradas na etapa 1.

Essa nova abordagem substitui o aprendizado na etapa de extração pelo processamento de regras baseadas em padrões morfológicos. Em seguida, um classificador é utilizado na remoção das relações inválidas do conjunto que contém todas as relações extraídas. Esta metodologia permite uma redução significativa na cardinalidade do conjunto de treinamento, já que a complexidade do aprendizado para classificação das relações é inferior à do aprendizado para a identificação das relações. Por outro lado, a construção de conjuntos de treinamento a partir de *features* linguísticas eleva o custo de classificação, pois a identificação de *features* representativas requer uma análise mais aprofundada das características da língua no contexto do problema. Neste trabalho, é realizada uma análise do esforço necessário à identificação das *features* mais representativas para a classificação de relações semânticas abertas em textos redigidos em Português, a partir dos experimentos descritos na próxima seção.

4 Experimentos e Resultados

Os experimentos foram realizados utilizando o corpus CETENFolha[2] (Corpus de Extratos de Textos Eletrônicos NILC/Folha de S. Paulo), que contém cerca de 24 milhões de palavras em Português, extraídas de textos do jornal Folha de São Paulo. Foram selecionadas aleatoriamente 500 sentenças do corpus envolvendo diferentes temas, tais como política, economia, esportes e ciência. As classes morfológicas das palavras contidas nas sentenças selecionadas foram obtidas automaticamente pelo etiquetador morfossintático do Cogroo[3], um corretor gramatical acoplável a um editor de texto de código aberto.

Após a etiquetação morfológica, foram extraídas 582 relações do tipo (fn_1, rel, fn_2), onde fn_1 e fn_2 representam as frases nominais contendo entidades mencionadas, encontradas antes e depois da relação, e rel denota a frase relacional da extração. As frases relacionais foram obtidas a partir do procedimento para identificação de padrões morfológicos na etapa de extração descrita na Seção 3 e adaptado para a Língua Portuguesa, sendo as entidades inicialmente identificadas aquelas classificadas como nome próprio pelo Cogroo. Por fim, cada extração foi manualmente classificada como válida ou inválida para compor o conjunto de treinamento. Exemplos de relações válidas e inválidas são mostradas na Tabela 3.

[2] http://www.linguateca.pt/cetenfolha/
[3] http://cogroo.sourceforge.net/

Tipo	Exemplo
Válida	X *matou* Y.
Inválida	X o *matou* enquanto Y assistia.
Válida	X, após o trabalho, *veio buscar* Y.
Inválida	X *veio buscar* os documentos antes de Y retornar.
Válida	X correu, mas *negou ter roubado* Y.
Inválida	X *negou ter roubado*, mas Y confessou o crime.
Válida	No dia seguinte, X *negociou com* Y sobre a venda da empresa.
Inválida	X apresentou o seu sócio, que *negociou com* Y.
Válida	X ainda *deve contar com* Y.
Inválida	X *deve contar com* um novo jogador na partida contra Y.

Tabela 3: Exemplos de relações válidas e inválidas.

Para viabilizar a avaliação de classificação, foram selecionadas 12 *features* de treinamento, definidas originalmente em Inglês por (Fader, Soderland e Etzion, 2011) e adaptadas para a Língua Portuguesa neste trabalho (Tabela 4). Os valores das *features* foram extraídos automaticamente de todas as sentenças selecionadas do corpus e aplicados a quatro classificadores utilizando a ferramenta de mineração de dados WEKA[4].

A efetividade ou mérito das *features* é estimada pelo algoritmo *Correlation-based Feature Selection (CFS)* (Hall, 1999), que utiliza uma heurística baseada em correlação para avaliar a capacidade de cada atributo em predizer a classe de uma instância de teste, dado um conjunto de treinamento. A hipótese que embasa este algoritmo afirma que bons subconjuntos de atributos devem possuir alta correlação com a classe de predição e baixa correlação entre si, já que atributos que possuem alta correlação entre si são considerados redundantes e não contribuem para elevar a capacidade preditiva do subconjunto.

Formalmente, seja S um subconjunto contendo k atributos, o mérito de S é calculado pela Equação 4:

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (4)$$

Onde $\overline{r_{cf}}$ representa a correlação média entre cada atributo de S e o atributo de classe, e $\overline{r_{ff}}$ denota a correlação média entre todas as combinações de atributos em S. A correlação entre os atributos pode ser estimada por diversas heurísticas, como o coeficiente de incerteza simétrica (baseado nos conceitos de entropia e ganho de informação) (Kononenko e Bratko, 1991) e o algoritmo *Relief* (Kononenko, 1994) (que utiliza uma abordagem baseada em instâncias para associar pesos às iterações entre os atributos).

[4]http://www.cd.waikato.ac.nz/ml/weka

F_1	tamanho(sentença) - tamanho($fn_1 + rel + fn_2$) < 30 caracteres?
F_2	A última preposição em *rel* é "de"?
F_3	A última preposição em *rel* é "com"?
F_4	A última preposição em *rel* é "por"?
F_5	A última preposição em *rel* é "pela"?
F_6	A última preposição em *rel* é "pelo"?
F_7	A última preposição em *rel* é "para"?
F_8	A última preposição em *rel* é "em"?
F_9	A string $fn_1 + rel$ está contida na sentença?
F_{10}	A string $rel + fn_2$ está contida na sentença?
F_{11}	A string $fn_1 + rel + fn_2$ está contida na sentença?
F_{12}	Há menos de 30 palavras na sentença?

Tabela 4: Features utilizadas para a base de treinamento em Língua Portuguesa.

4.1 Resultados

A Figura 2 mostra o mérito das *features* descritas na Tabela 4, considerando todo o conjunto de dados (582 extrações obtidas de 500 sentenças). É possível notar que as *features* F_9, F_{10} e F_{11} são as que possuem as maiores capacidades de predição. Por outro lado, a *feature* F_1 pode ser eliminada do conjunto de atributos sem prejuízo à qualidade de classificação, já que possui mérito nulo.

Os resultados mostrados na Figura 2 foram obtidos a partir da execução do algoritmo CFS implementado no Weka, usando a estratégia de busca *BestFirst* com parâmetros $D = 1$ (*forward*

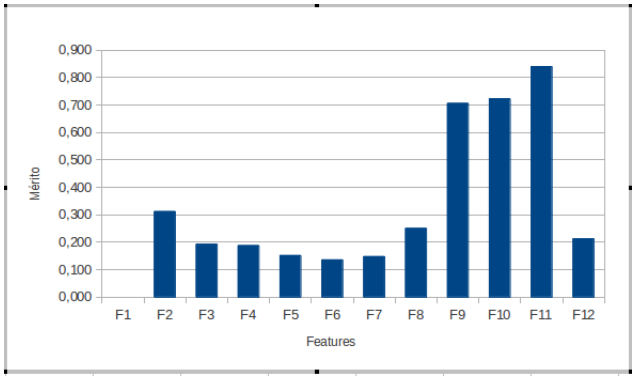


Figura 2: Representatividade das *features* no conjunto de dados.

Sub-conjunto	Feature avaliada	Elementos do melhor subconjunto
CF_1	-	$F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}$
CF_2	F_{11}	$F_9, F_{10}, F_{11}, F_{12}$
CF_3	F_{10}	$F_2, F_6, F_8, F_{10}, F_{11}$
CF_4	F_9	F_4, F_9, F_{11}

Tabela 5: Conjuntos de *features*.

search) e $N = 5$ (número de nós do critério de parada), com seleção de atributos usando todo o conjunto de treinamento. Diante desses resultados, foram selecionados quatro subconjuntos de *features* (Tabela 5) para avaliação. O grupo CF_1 é composto por todas as *features* que possuem mérito não nulo e os grupos CF_2, CF_3 e CF_4 correspondem aos subconjuntos obtidos a partir das melhores *features* avaliadas pelo algoritmo CFS.

É possível notar que nem sempre as *features* que possuem os maiores méritos formam o melhor subconjunto, já que pode haver alta correlação entre elas, redundância que não contribui para elevar a capacidade preditiva do subconjunto como um todo. Dessa maneira, as *features* F_2, F_9, F_{10} e F_{11} não formam um subconjunto com alta capacidade preditiva, devido à alta correlação entre F_2 e F_9.

Na Figura 3 são mostrados os valores médios da medida-F e da área sob a curva ROC (AUC) de quatro algoritmos de classificação avaliados (J48, SVM, Perceptron e Naive Bayes) em cada conjunto de *features*, utilizando o método de validação cruzada com 10 *folds*. Os resultados mostram valores aproximadamente iguais para os três grupos CF_1, CF_2 e CF_3, sendo a maior diferença equivalente a 0,7% para a medida-F e 1,4% para a AUC entre os grupos CF_1 e CF_3, indicando que a dimensionalidade dos atributos pode ser reduzida de 11 para 4 (CF_2) ou 5 (CF_3) *features*, com perdas mínimas na qualidade de classificação. Por outro lado, o grupo CF_4 apresentou

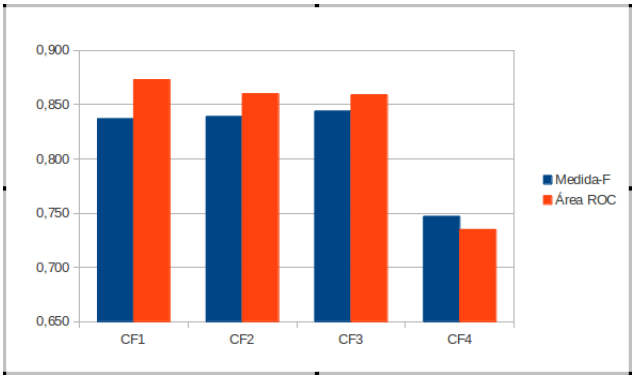


Figura 3: Avaliação dos conjuntos de *features*.

Método	Precisão	Cobertura	Medida-F
J48	0,848 ± 0,014	0,841 ± 0,018	0,841 ± 0,018
Lib SVM	0,848 ± 0,019	0,840 ± 0,018	0,839 ± 0,018
Perceptron	0,823 ± 0,038	0,820 ± 0,041	0,820 ± 0,040
Naive Bayes	0,800 ± 0,037	0,799 ± 0,039	0,799 ± 0,039

Tabela 6: Resultados médios obtidos por validação cruzada com 10 *folds*.

valores médios 9,5% inferiores para a medida-F e 13,8% para a AUC, sendo portanto o menos representativo dentre os conjuntos avaliados.

A Tabela 6 mostra os valores detalhados de precisão, cobertura e medida-F nos métodos de classificação testados em ordem decrescente de desempenho, a partir do conjunto de *features* CF_1. Os valores médios e desvios padrões correspondentes são obtidos pelo processamento de 10 conjuntos de sentenças com tamanhos distintos, que variam de 57 a 582 extrações.

Adicionalmente, as curvas no gráfico da Figura 4 ilustram as variações de precisão, cobertura e medida-F com o aumento do conjunto de treinamento e teste em cada algoritmo. É possível perceber que o algoritmo J48 obteve os melhores resultados na classificação de relações abertas em Português, tendo uma medida-F média 4,2% superior ao classificador bayesiano, que apresentou os piores resultados dentre os métodos testados. Além disso, nota-se um crescimento na medida-F dos algoritmos em função da cardinalidade do conjunto de treinamento.

Os resultados mostram as dificuldades encontradas na identificação de *features* linguísticas representativas para a tarefa de extração de relações abertas, já que o mérito de um atributo pode sofrer grande variação em função da mudança de idioma. Por exemplo, a *feature* F_1 apresentou mérito nulo para a Língua Portuguesa,

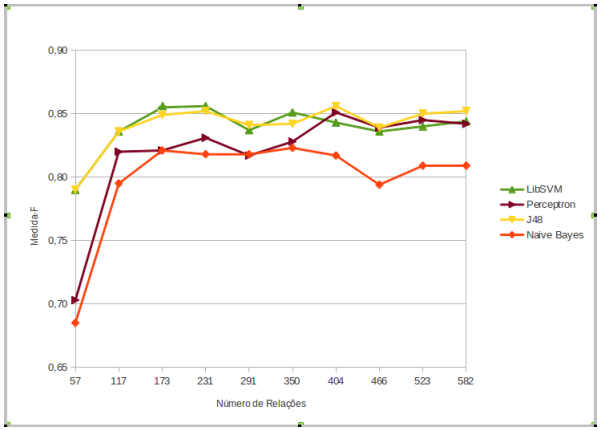


Figura 4: Avaliação da medida-F em função da quantidade de relações classificadas.

mas representa um dos atributos mais significativos para o mesmo problema na Língua Inglesa, como pode ser observado nos resultados obtidos em (Fader, Soderland e Etzion, 2011).

Como a análise do mérito das *features* não foi apresentada nos trabalhos voltados para a Língua Inglesa pesquisados, não foi possível realizar um estudo comparativo direto com os resultados obtidos no presente trabalho. Entretando, é possível comparar indiretamente o desempenho das *features* em Inglês e Português no problema de classificação de relações abertas tratado. Em (Fader, Soderland e Etzion, 2011), a utilização de um classificador de regressão logística treinado com um conjunto de *features* em Língua Inglesa apresentou uma medida-F cerca de 8% superior ao algoritmo de classificação com o melhor desempenho avaliado em Língua Portuguesa neste trabalho.

Essas observações permitem afirmar que o mapeamento direto de um conjunto de *features* de um idioma para outro não implica na seleção dos melhores atributos na classificação de relações abertas. Consequentemente, é necessária uma análise mais profunda das peculiaridades de cada idioma para a escolha de um conjunto representativo de *features*.

5 Conclusões e Trabalhos Futuros

A distinção automática entre relações válidas e inválidas representa um problema recorrente em sistemas de extração de relações em texto não estruturado. Quando as frases relacionais identificadas possuem vocabulário não limitado, a importância da tarefa de classificação na qualidade das extrações se torna mais evidente, já que a ambiguidade inerente à linguagem natural tem ocasionado grande proporção de relações inválidas nos métodos mais recentes que tratam desta tarefa.

Grande parte das soluções atuais extraem relações abertas exclusivamente a partir de textos redigidos em Inglês, idioma que possui os recursos linguísticos mais sofisticados, como etiquetadores morfossintáticos, extratores de entidades mencionadas, frases nominais e correferências, além de léxicos computacionais de alta granularidade e grandes bases de treinamento. Os principais trabalhos do estado da arte eliminam as relações inválidas por meio de classificadores treinados a partir de *features* linguísticas, altamente dependentes do idioma-alvo. Tal dependência decorre da dificuldade inerente à determinação do conjunto de *features* mais representativo para o problema, considerando as peculiaridades de cada língua. Em particular, o presente trabalho avalia esta dificuldade em textos redigidos em Português, que permite identificar duas limitações principais: 1) O mapeamento de um conjunto de *features* de um idioma para outro não implica na seleção dos melhores atributos de treinamento, dadas as especificidades de cada idioma, o que implica em novas análises para cada língua; 2) A qualidade de classificação das relações abertas baseada em *features* depende de conjuntos de treinamento extensos, que possuem alto custo de construção, conforme resultados descritos em trabalhos predecessores voltados para a Língua Inglesa (Wu e Weld, 2010; Banko e Etzioni, 2008). Neste trabalho, o crescimento da medida-F na classificação de relações abertas com o aumento da cardinalidade do conjunto de treinamento indica que esta característica também é válida para corpora redigidos em Língua Portuguesa.

Como trabalhos futuros, pretende-se investigar abordagens capazes de reduzir a dependência do idioma na tarefa de extração de relações abertas, por meio da eliminação da necessidade de construção de conjuntos extensos de treinamento baseados em *features* linguísticas específicas na etapa de classificação das relações.

Referências

Anyanwu, K., A. Maduko, e A. Sheth. 2005. Semrank: Ranking complex relationship search results on the semantic web. *Proc. of the 14th International World Wide Web Conference, ACM Press, 117-127.*

Banko, M. e O Etzioni. 2008. The tradeoffs between open and traditional relation extraction. *In Proceedings of ACL-08: HLT, pages 28-36, Columbus, Ohio, June. Association for Computational Linguistics.*

Banko, M., M. J. J. Cafarella, S. Soderland, M. Broadhead, e O. Etzioni. 2007. Open information extraction from the web. *In the Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 2670-2676, January.*

Brin, S. 1998. Extracting patterns and relations from the world wide web. *In International Workshop on The World Wide Web and Databases, p.172-183, March 27-28, 1998.*

Bruckschen, M., J. Souza, R. Vieira, e S. Rigo. 2008. Sistema serelep para o reconhecimento de relações entre entidades mencionadas. *In Cristina Mota; Diana Santos (ed.), Desafios na avaliação conjunta do reconhecimento de*

entidades mencionadas: O Segundo HAREM. Linguateca, cap. 14, p. 247-260.

Cardoso, N. 2008. Rembrandt - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto, 2008. *In:: Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. s.l.:Linguateca, pp. 195-211.*

Chaves, S. 2008. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem. *In Cristina Mota; Diana Santos (ed.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, cap. 13, p. 231-245.*

Dalben, O. J. e D. B. Claro. 2011. Uma análise do reconhecimento textual de nomes de pessoas e organizações na computação forense. *Proceeding of the Sixth International Conference on Forensic Computer Science - ICoFCS 2011, pp. 7-15.*

Fader, A., S. Soderland, e O. Etzion. 2011. Identifying relations for open information extraction. *In Proceedings of Conference on Empirical Methods in Natural Language Processing.*

Feldman, R. e J. Sanger. 2007. *The text mining handbook: advanced approaches analyzing advanced unstructured data.* New York: Cambridge University Press.

Freitas, C. e V. Quental. 2007. Subsídios para a elaboração automática de taxonomias. *Anais do XXVII Congresso da SBC. Rio de Janeiro, Rio de Janeiro: [s.n.], 2007. (V Workshop em Tecnologia da Informacao e da Linguagem Humana TIL), p. 1585-1594.*

Girju, R., B. Beamer, A. Rozovskaya, A. Fister, e S. Bhat. 2010. A knowledge-rich approach to identifying semantic relations between nominals. *Information Processing and Management,v. 46, n. 5, p. 589-610.*

Hall, M. 1999. *Correlation-based Feature Selection for Machine Learning.* Tese de doutoramento, University of Waikato, Hamilton, NewZealand.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th International Conference on Computational linguistics - Volume 2. Nantes, France, p. 539-545.*

Hirschman, L. e R. Gaizauskas. 2001. Natural language question answering: the view from here. *Natural Language Engineering 7 (4): 275-300.*

Kambhatla, N. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL.*

Kononenko, I. 1994. Estimating attributes: Analysis and extensions of relief. *In Proceedings of the European Conference on Machine Learning.*

Kononenko, I. e I. Bratko. 1991. Information-based evaluation criterion for classifiers performance. *Machine Learning, 6:67-80.*

Lutz, J. e C. Heuser. 2013. Descoberta de ruído em páginas da web oculta através de uma abordagem de aprendizagem supervisionada. *Simpósio Brasileiro de Banco de Dados (SBBD'13), Recife, PE, Brazil.*

Nakashole, N. e T. Mitchell. 2014. Language-aware truth assessment of fact candidates. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 2014.*

Oliveira, H., D. Santos, e P. Gomes. 2010. Extração de relações semânticas entre palavras a partir de um dicionário: o papel e sua avaliação. *Linguamática, v. 2, n. 1, p. 77-94.*

Snow, R., D. Jurafsky, e A. Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. *In Advances in Neural Information Processing Systems 17, pages 1297-1304. MIT Press.*

Souza, E. e D. Claro. 2014. Detecção multilíngue de serviços web duplicados baseada na similaridade textual. *Simpósio Brasileiro de Sistemas de Informação (SBSI'14), Maio 27-30, Londrina/PR, Brazil.*

Taba, L. S. e H. Caseli. 2012. Automatic hyponymy identification from brazilian portuguese texts. *In Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR).*

Wu, F. e D. S. Weld. 2010. Open information extraction using wikipedia. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pages 118-127, Morristown.*

Zelenko, D., C. Aone, e A. Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research 3 1083-1106.*

Projetos, Apresentam-Se!

O dicionario de sinónimos como recurso para a expansión de WordNet*

The dictionary of synonyms as a resource for expanding WordNet

Xavier Gómez Guinovart
Universidade de Vigo
xgg@uvigo.es

Miguel Anxo Solla Portela
Universidade de Vigo
miguelsolla@uvigo.es

Resumo

Neste artigo presentamos os alicerces dun experimento de extracción léxica deseñado no marco do proxecto de investigación SKATeR e orientado á ampliación do WordNet do galego mediante a explotación dos datos lexicográficos recollidos nun dicionario de sinónimos "tradicional" desta lingua.

Palabras chave

WordNet, adquisición de información léxica, dicionario de sinónimos, recursos lingüísticos

Abstract

In this paper, we present the foundations for a lexical acquisition experiment designed in the framework of the SKATeR research project and aimed to the expansion of the Galician WordNet using the lexicographical data collected in a "traditional" Galician dictionary of synonyms.

Keywords

WordNet, lexical acquisition, dictionary of synonyms, language resources

1 Introdución

Neste artigo[1] presentamos os alicerces dun experimento de extracción léxica deseñado no marco do proxecto de investigación SKATeR[2] e orientado á ampliación do WordNet do galego mediante a explotación dos datos lexicográficos recollidos nun dicionario de sinónimos "tradicional" desta lingua.

En principio, as características de WordNet (Miller et al., 1990) como rede léxico-semántica estruturada en nós conceptuais (*synsets*) compostos por *variantes* léxicas dun mesmo significado permitirían inferir sen dificultades a hipótese de que os dicionarios de sinónimos existentes dunha lingua constitúen fontes lexicográficas directas moi axeitadas para a expansión deste recurso.

Porén, os experimentos previos de extensión do WordNet do galego a partir dun dicionario de sinónimos realizados polos autores (Gómez Guinovart, 2014) (Solla Portela e Gómez Guinovart, 2014) demostran as dificultades de acadar mediante este recurso uns índices de precisión que permitan a validacion manual dos resultados da extracción nun tempo razoábel. As causas desta dificultade radican principalmente no distinto concepto de sinonimia utilizado por cada un destes dous recursos, moito máis estrito e delimitado pola glosa no WordNet (Gómez Clemente et al., 2013), moito máis laxo e abranguente doutras relacións semánticas no dicionario de sinónimos tradicional.

Deste xeito, mentres que o concepto de sinonimia manexado en WordNet fai referencia ao concepto máis específico de sinonimia contextual ou intercambiabilidade dos sinónimos limitada a un contexto (Miller, 1998), no dicionario de sinónimos tradicional as relacións semánticas entre o lema e os "sinónimos" que forman parte da entrada pode ser tanto de sinonimia, como de hiperonimia, hiponimia, holonimia, meronimia, troponimia ou implicación, entre outras.

Así, por exemplo, no dicionario de sinónimos usado neste experimento, a entrada para o lema *climaterio* ("período da vida do home e da muller en que se producen unha serie de cambios no organismo debidos á diminución da actividade das glándulas sexuais") está formada polos seus dous hipónimos *andropausa* (isto é, o climaterio mas-

*Esta investigación realízase no marco do proxecto *Adquisición de escenarios de conocimiento a través de la lectura de textos: Desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego (SKATeR-UVIGO)* financiado polo Ministerio de Economía y Competitividad, TIN2012-38584-C06-04.

[1]Queremos agradecer aquí sinceramente as valiosas contribucións para a mellora do artigo das tres persoas expertas –Hugo Gonçalo Oliveira, da Universidade de Coimbra; Álvaro Iriarte Sanromán, da Universidade do Minho; e Mercè Lorente Casafont, da Universitat Pompeu Fabra– que realizaron a revisión previa á súa aceptación por parte da revista.

[2]http://nlp.lsi.upc.edu/skater/

culino) e *menopausa* (ou climaterio feminino).

Por este motivo, estamos a traballar no desenvolvemento dunha metodoloxía de extracción que nos permita maximizar os esforzos dedicados á revisión humana dos resultados, aumentando a precisión dos resultados e controlando ao mesmo tempo a súa amplitude, e combinando aspectos tratados nos experimentos previos, como a categoría gramatical e a dispersión semántica das entradas lexicográficas, con outros aínda non tratados, como a frecuencia nas obras lexicográficas dos lemas procesados.

Un traballo semellante de enriquecemento dunha ontoloxía léxica semellante a WordNet a partir doutros recursos léxicos é o realizado para o portugués no proxecto Onto.PT[3] (Gonçalo Oliveira e Gomes, 2014), no que se obtiveron synsets a partir de dicionarios de sinónimos e outras fontes textuais (corpus, enciclopedias e dicionarios da lingua) usando como método de extracción diversos índices de semellanza (Gonçalo Oliveira, 2013).

Nos seguintes apartados, presentaremos os recursos procesados no experimento, a metodoloxía deseñada para a extracción e unha avaliación dos seus resultados.

2 Recursos

Galnet, a versión galega de WordNet, distribúese con licenza Creative Commons como parte do MCR[4] (González-Agirre e Rigau, 2013) (González-Agirre, Laparra e Rigau, 2012). A versión de Galnet desta distribución (de finais de 2012) alcanza a cobertura léxica que se amosa na Táboa 1 (onde a columna *Vars* indica o número total de *variantes* sinonímicas en cada categoría e a columna *Syns* o número total de *synsets* ou nós conceptuais recompilados) en comparación coa do WordNet 3.0 do inglés (na táboa, *EWN30*).

	EWN30		Galnet	
	Vars	Syns	Vars	Syns
N	146312	82115	18949	14285
V	25047	13767	1416	612
Adx	30002	18156	6773	4415
Adv	5580	3621	0	0
TOTAL	206941	117659	27138	19312

Táboa 1: Distribución actual de Galnet no MCR

A partir desta versión inicial de 2012, continuamos ampliando Galnet mediante técnicas de extracción léxica baseadas en recursos textuais monolingües e bilingües existentes (corpus paralelos e dicionarios) (Gómez Guinovart e Simões, 2013) (Gómez Guinovart, 2014) (Solla Portela e Gómez Guinovart, 2014) (Gómez Guinovart e Oliver, 2014). Os resultados das expansións en curso pódense observar na interface web de consulta de Galnet[5] realizando buscas sobre a versión de desenvolvemento do recurso. O experimento aquí presentado realizouse coa versión de desenvolvemento 3.0.4 de Galnet, cuxa cobertura se recolle na Táboa 2.

	WN30		Galnet	
	Vars	Syns	Vars	Syns
N	146312	82115	22186	16812
V	25047	13767	3996	1423
Adx	30002	18156	7884	4962
Adv	5580	3621	253	223
TOTAL	206941	117659	34319	23420

Táboa 2: Versión de desenvolvemento 3.0.4

O *Dicionario de sinónimos do galego* (Gómez Clemente, Gómez Guinovart e Simões, 2014) usado para a expansión de Galnet foi elaborado tamén no marco deste proxecto (Gómez Guinovart, 2014) (Gómez Guinovart e Simões, 2013). Este dicionario de sinónimos está dispoñíbel para a súa libre consulta na web[6], pódese descargar tamén como app para Android[7] e para iOS[8], e constitúe o único recurso electrónico existente para a lingua galega dentro desta categoría de dicionarios. No experimento presentado utilizouse a versión de desenvolvemento 0.7 deste recurso coa cobertura léxica que se indica na Táboa 3.

Entradas	27104
Acepcións	44849
Sinónimos	203251

Táboa 3: Extensión do dicionario

Nesta descrición dos contidos do dicionario de sinónimos, entendemos por *entradas* os artigos lexicográficos (na tradición escrita, ordenados alfabeticamente) nos que se divide a estrutura principal do dicionario; entendemos por *acepcións* os conxuntos dun ou máis sinónimos, agrupados por significado, nos que se dividen as entradas; e entendemos por *sinónimos* os elementos léxicos

[3] `http://ontopt.dei.uc.pt/`

[4] `http://adimen.si.ehu.es/web/MCR/`

[5] `http://sli.uvigo.es/galnet/`

[6] `http://sli.uvigo.es/sinonimos/`

[7] `https://play.google.com/store/apps/details?id=net.ayco.sinonimosgal`

[8] `https://itunes.apple.com/us/app/sinonimos-do-galego/id940045971?l=es&ls=1&mt=8`

(monoléxicos ou pluriléxicos) que compoñen as acepcións nas que se dividen as entradas.

3 Metodoloxía

3.1 Alicerces

Temos, por unha banda, un dicionario de sinónimos D formado polo conxunto de acepcións ($\{A_1, A_2, ... A_n\}$) que forman parte da microestrutura das entradas do dicionario

$$D = \{A_1, A_2, ..., A_n\}$$

e temos, por outra banda, un WordNet do galego G formado por un conxunto de synsets ($\{S_1, S_2, ... S_n\}$) que recollen os sinónimos (ou variantes sinónimicas) asignados a cada concepto desta rede léxico-semántica

$$G = \{S_1, S_2, ..., S_n\}$$

Cada acepción A_k de D está formada por un conxunto de formas léxicas sinonímicas ($\{s_1, s_2, ... s_n\}$) composto polo lema e os sinónimos dunha acepción

$$A_k = \{s_1, s_2, ..., s_n\}$$

e cada synset S_l de G está formado por un conxunto de variantes sinonímicas ($\{v_1, v_2, ... v_n\}$) para un concepto da rede léxico-semántica

$$S_l = \{v_1, v_2, ..., v_n\}$$

O método de extracción utilizado baséase na hipótese de que unha acepción lexicográfica A_k do dicionario representa probabelmente o mesmo valor semántico que un synset S_l de Galnet se A_k e S_l comparten cando menos un elemento idéntico da mesma categoría morfolóxica

$$\{s_x, v_y\} \in A_k \cap S_l \wedge s_x = v_y \wedge CAT(s_x) = CAT(v_y)$$

Se se cumpre esta condición, os sinónimos de A_k distintos de s_x (e ausentes de S_l) serán candidatos susceptíbeis de engadirse a S_l após un certo grao de revisión humana, incrementando deste modo o número de variantes do synset.

Considérese, por exemplo, a entrada para o adxectivo *aleuto* no dicionario, presentada no seu código fonte en XML na Listaxe 1, que está conformada por unha única acepción e catro sinónimos. Esta acepción estaría composta por cinco formas léxicas sinonímicas: {*aleuto, agudo, espelido, intelixente, listo*}.

Por outra banda, considérese o synset adxectivo identificado como glg-30-00061885-a formado no Galnet 3.0.4 polo conxunto de dúas variantes sinonímicas {*enxeñoso, aleuto*}.

De acordo coa metodoloxía utilizada, supoñemos que esta acepción do dicionario de sinónimos para a entrada *aleuto* e este synset glg-30-00061885-a do Galnet son susceptíbeis de representar o mesmo concepto semántico, xa que as dúas constelacións léxicas son de categoría morfolóxica adxectiva e as dúas comparten unha forma coincidente (*aleuto*). Deste xeito, os sinónimos da entrada *aleuto* do dicionario ausentes do synset glg-30-00061885-a do Galnet –isto é, os sinónimos {*agudo, espelido, intelixente, listo*}– serán variantes candidatas a integrar o synset glg-30-00061885-a do Galnet previa revisión humana.

Listaxe 1: Entrada de *aleuto*

```
<entry>
<form>
<orth>aleuto</orth>
</form>
<sense>
<gramGrp>adx</gramGrp>
<def n="1">
<syn><lemma>Agudo</lemma></syn>,
<syn><lemma>espelido</lemma></syn>,
<syn><lemma>intelixente</lemma></syn>,
<syn><lemma>listo</lemma></syn>.
</def>
</sense>
</entry>
```

3.2 Parámetros

O problema desta condición mínima é que produce moito ruído: ofrece moitos sinónimos candidatos para formar parte dun synset (alta cobertura) pero o seu índice de acertos non é moi elevado (baixa precisión), o que imposibilita a planificación dunha revisión humana dos resultados. Concretamente, cos recursos anteriormente descritos, o número de candidaturas elévase a 296.246 sinónimos.

Por esta razón, deseñamos un experimento que nos permitise maximizar os esforzos (sempre demasiado escasos) dedicados á revisión humana dos resultados, aumentando o máis posíbel a precisión dos resultados sen diminuír a súa cobertura a límites inútiles. Os parámetros que se tiveron en conta neste experimento foron os seis que se indican deseguido:

- P_1 Número de elementos idénticos e coa mesma categoría morfolóxica en $A_k \cap S_l$ (mínimo 1)
- P_2 Número de elementos en A_k
- P_3 Frecuencia absoluta do sinónimo candidato en D

P_4 Frecuencia absoluta do sinónimo candidato en G

P_5 Frecuencia absoluta en D das formas léxicas compartidas (idénticas e coa mesma categoría morfolóxica) en A_k e S_l

P_6 Frecuencia absoluta en G das formas léxicas compartidas (idénticas e coa mesma categoría morfolóxica) en A_k e S_l

P_1 determina a cantidade de formas compartidas entre o synset e a acepción. Se A_k e S_l comparten algún elemento coa mesma forma e coa mesma categoría gramatical existe a posibilidade de que compartan o significado. En principio, cantos máis elementos compartan A_k e S_l, maior será a seguridade de atinar na súa coincidencia semántica e, por tanto, a precisión dos resultados será maior. Por outra banda, a cobertura do experimento descende a medida que sobe P_1. Na Táboa 4 amósase o número de candidatos resultantes ao usarmos só esta condición.

P_1	candidatos
> 0	296.246
> 1	26.610
> 2	6.698
> 3	2.954
> 4	1.421
> 5	786
> 6	436
> 7	220
> 8	178
> 9	135

Táboa 4: Candidatos por P_1

Con P_2 tratamos de manexar o fenómeno da dispersión semántica propio do dicionario de sinónimos. O certo é que, durante a revisión das formas candidatas que se obtiveron nun experimento anterior (Solla Portela e Gómez Guinovart, 2014), detectouse que a precisión das candidaturas propostas para se incorporaren a Galnet diminuía a medida que se incrementaba o número de sinónimos na mesma acepción do dicionario, debido probabelmente á menor cohesión semántica entre as formas agrupadas na acepción. De xeito experimental, puidemos comprobar que os valores de P_2 entre 6 e 9 son os que ofrecen mellores resultados para incrementar a precisión sen que baixe demasiado a cobertura. Tomando como referencia un valor maior ca 1 para P_1, recollemos na Táboa 5 o número de candidatos que se obtiveron como resultado con diferentes valores de P_2.

P_{3-6} permítennos tratar no experimento o fenómeno da polisemia das formas léxicas que se manexaron. A idea é que cantas máis veces apareza unha mesma forma léxica no recurso léxico analizado (D ou G), maior será a posibilidade de que se trate dunha forma léxica polisémica e, polo tanto, maior será a posibilidade de incidir negativamente na selección das candidaturas. Por unha banda, P_{3-4} permiten controlar a frecuencia das formas candidatas en D e en G; e, por outra banda, P_{5-6} permiten controlar a frecuencia das formas compartidas entre D e G durante a selección das candidatas. En ambos os casos, o factor de control que ofrece maior rendemento é o que se aplicou sobre D (isto é, P_3 e P_5) e, asemade, a supervisión da frecuencia das formas candidatas (P_3 e P_4) móstrase máis eficaz que a das formas compartidas (P_5 e P_6). Por exemplo, tomando como punto de partida a táboa anterior, con $P_1 > 1$ e con $P_2 < 8$, a Táboa 6 recolle o número de candidaturas que se obtiveron cos diferentes valores usados no experimento para P_3.

P_2	candidatos
< 3	0
< 4	1.088
< 5	3.895
< 6	7.265
< 7	10.771
< 8	14.010
< 9	16.317
< 10	18.347
< 11	19.815
< 12	20.995
< 13	21.861

Táboa 5: Candidatos por P_2 para $P_1 > 1$

P_3	candidatos
< 12	7.262
< 11	6.716
< 10	6.141
< 9	5.588
< 8	4.944
< 7	4.292
< 6	3.585
< 5	2.870
< 4	2.070
< 3	1.292
< 2	474

Táboa 6: Candidatos por P_3 para $P_1 > 1$ e $P_2 < 8$

4 Avaliación

A avaliación da precisión dos resultados realizouse mediante a validación manual dos candidatos. Realizamos diversos experimentos de extrac-

experimento	P_1	P_2	P_3	P_4	P_5	P_6	candidatos	precisión
E_1	> 1	< 9	< 3	-	< 4	-	60	47 %
E_2	> 1	< 9	< 8	< 2	< 5	< 2	95	22 %
E_3	> 1	< 9	< 5	< 12	< 4	< 12	85	19 %
E_4	> 1	< 4	< 13	< 13	< 4	< 13	29	14 %
E_5	> 1	< 4	< 4	< 13	< 13	< 13	77	39 %
E_6	> 1	< 4	< 8	< 8	< 8	< 8	92	37 %
E_7	> 2	< 9	< 8	< 12	< 8	< 12	52	20 %
E_8	> 2	-	-	-	-	-	6.335	35 %
E_9	> 2	< 6	-	-	-	-	856	60 %

Táboa 7: Precisión

ción usando diferentes combinacións de parámetros, tratando sempre de acadar un compromiso entre cobertura e precisión que nos permita maximizar a rendibilidade das horas de dedicación humana á tarefa de revisión. A Táboa 7 presenta os datos de precisión en diversas combinacións de parámetros que se experimentaron cos seguintes criterios:

E_1 Dispersión moi baixa e frecuencia baixa no dicionario de sinónimos das formas compartidas, e sen restricións na frecuencia en Galnet

E_2 Frecuencia mínima en Galnet das formas candidatas e compartidas

E_3 Frecuencia baixa no dicionario das formas candidatas e compartidas

E_4 Dispersión baixa e frecuencia baixa no dicionario das formas compartidas

E_5 Dispersión baixa e frecuencia baixa no dicionario das formas candidatas

E_6 Baixa dispersión semántica

E_7 Coincidencia de 3 formas no dicionario e en Galnet

Para a súa comparación, inclúense os resultados previos de $[E_8]$ e $[E_9]$ que se obtiveron coa versión 3.0.2 de Galnet a partir da avaliación manual de 100 formas candidatas (Solla Portela e Gómez Guinovart, 2014).

Os datos que se analizan na Táboa 7 reflicten o impacto inicial da aplicación da metodoloxía e da comparación entre as parametrizacións expostas nun estadio concreto do desenvolvemento do proxecto; mais prevese unha mellora da precisión en futuras reutilizacións do procedemento baseándose en (a) a escolla de parámetros similares aos que ofreceron maior rendibilidade neste experimento, (b) a incorporación dun filtro de candidaturas non-desexadas a partir das formas desbotadas na revisión humana e (c) o cruzamento dos sinónimos cunha versión aumentada e revisada do Galnet respecto da que se utilizou nesta ocasión.

5 Conclusións

Os resultados deste experimento destacan a necesidade de lograr un compromiso entre cobertura e precisión que facilite a viabilidade da revisión humana.

A pesar de que a precisión que se obtivo foi máis ben baixa na maior parte dos experimentos con diversas combinacións de parámetros, prevese que a aplicación cíclica dos experimentos (sobre o Galnet mellorado cos candidatos validados e ampliado con variantes procedentes doutros experimentos) aumente a precisión dos seus resultados. Neste sentido, cómpre ter en conta tamén que o uso de filtros constituídos coas candidaturas que se rexeitan na revisión repercute nun melloramento inmediato da precisión da extracción nos ciclos posteriores de aplicación do experimento.

Así mesmo, queremos apuntar as posibilidades de aplicar a mesma metodoloxía para a expansión de wordnets doutras linguas que dispoñan dun recurso asimilábel a un dicionario de sinónimos.

Como liña futura de investigación, pretendemos experimentar coas posibilidades de ampliación do dicionario de sinónimos mediante técnicas de extracción léxica a partir de WordNet; é dicir, seguir a vía inversa do traballo que se presenta neste artigo.

Referencias

Gómez Clemente, Xosé María, Xavier Gómez Guinovart, Andrea González Pereira, e Verónica Taboada Lorenzo. 2013. Sinonimia e rexistros na construción do WordNet do galego. *Estudos de Lingüística Galega*, 5:27–42.

Gómez Clemente, Xosé María, Xavier Gómez Guinovart, e Alberto Simões. 2014. *Dicionario de sinónimos do galego*. Área de Normalización Lingüística, Universidade de Vigo, Vigo. URL: `http://sli.uvigo.es/sinonimos/`.

Gómez Guinovart, Xavier. 2014. Do dicionario de sinónimos á rede semántica: fontes lexicográficas na construción do WordNet do galego. En Ana Gabriela Macedo, Carlos Mendes de Sousa, e Vítor Moura, editores, *XV Colóquio de Outono - As humanidades e as ciências: disjunções e confluências*, Braga. CEHUM, Universidade do Minho.

Gómez Guinovart, Xavier e Antoni Oliver. 2014. Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit. *Procesamiento del Lenguaje Natural*, 53:43–50.

Gómez Guinovart, Xavier e Alberto Simões. 2013. Retreading dictionaries for the 21st century. En José Paulo Leal, Ricardo Rocha, e Alberto Simões, editores, *2nd Symposium on Languages, Applications and Technologies*, pp. 115–126, Saarbrücken. Dagstuhl Publishing.

Gonçalo Oliveira, Hugo. 2013. *Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*. Tese de doutoramento, Universidade de Coimbra. URL: `http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira_PhdThesis2012.pdf`.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2014. ECO and Onto.PT: a flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation*, 48(2):373–393.

González-Agirre, Aitor, Egoitz Laparra, e German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. En *Proceedings of the Sixth International Global WordNet Conference (GWC'12)*, Matsue, Japan.

González-Agirre, Aitor e German Rigau. 2013. Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual central repository. *Linguamática*, 5(1):13–28.

Miller, George A., 1998. *Nouns in WordNet*, pp. 23–46. The MIT Press, Cambridge, Massachusetts.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, e Katherine Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Solla Portela, Miguel Anxo e Xavier Gómez Guinovart. 2014. Ampliación de WordNet mediante extracción léxica a partir de un diccionario de sinónimos. En L. Alfonso Ureña López et al., editor, *Actas de las V Jornadas de la Red en Tratamiento de la Información Multilingüe y Multimodal*, volume 1199, pp. 29–32, Aachen. CEUR Workshop Proceedings (CEUR-WS.org).

Projetos sobre Tradução Automática do Português no Laboratório de Sistemas de Língua Falada do INESC-ID

Machine Translation Projects for Portuguese at INESC ID's Spoken Language Systems Laboratory

Anabela Barreiro
L2F - INESC-ID, Rua Alves Redol 9
1000-029, Lisboa
anabela.barreiro@inesc-id.pt

Wang Ling
L2F - INESC-ID Lisboa
Carnegie Mellon University
IST - Universidade de Lisboa

Luísa Coheur
L2F - INESC-ID
IST - Universidade de Lisboa

Fernando Batista
L2F - INESC-ID Lisboa
ISCTE-IUL

Isabel Trancoso
L2F - INESC-ID Lisboa
IST - Universidade de Lisboa

Resumo

As tecnologias da língua, de um modo especial as aplicações de tradução automática, têm o potencial de ajudar a quebrar barreiras linguísticas e culturais, apresentando um importante contributo para a globalização e internacionalização do português ao permitir que conteúdos linguísticos sejam partilhados 'a partir de' e 'para' esta língua. O presente artigo tem como objetivo apresentar o trabalho de investigação na área da tradução automática realizada pelo Laboratório de Sistemas de Língua Falada do INESC-ID, nomeadamente a tradução automática de fala, a tradução de microblogues e a criação de um sistema híbrido de tradução automática. Centraremos a nossa atenção na criação do sistema híbrido, que tem como objetivo a combinação de conhecimento linguístico, nomeadamente semântico-sintático, com conhecimento estatístico, de forma a aumentar o nível de qualidade da tradução.

Palavras Chave

Tradução Automática, Sistemas Híbridos, OpenLogos, Conhecimento Semântico-Sintático

Abstract

Language technologies, in particular machine translation applications, have the potential to help break down linguistic and cultural barriers, presenting an important contribution to the globalization and internationalization of the Portuguese language, by allowing content to be shared 'from' and 'to' this language. This article aims to present the research work developed at the Laboratory of Spoken Language Systems of INESC-ID in the field of machine translation, namely the automated speech translation, the translation of microblogs and the creation of a hybrid machine translation system. We will focus on the creation of the hybrid system, which aims at combining linguistic knowledge, in particular semantico-syntactic knowledge, with statistical knowledge, to increase the level of translation quality.

Keywords

Machine Translation, Hybrid Systems, OpenLogos, Semantico-Syntactic Knowledge

1 Introdução

A tradução automática, uma das mais complexas aplicações de língua natural, tem vindo a evoluir velozmente graças ao aumento quantitativo e qualitativo dos recursos linguísticos, nomeadamente dos dicionários electrónicos e corpos paralelos. Para além da evolução na construção de recursos, o desenvolvimento de algoritmos matemáticos que visam o mapeamento bilíngue e multilíngue de palavras, unidades lexicais multipalavra e expressões, e exploram a aprendizagem automática deste tipo de mapeamentos, vieram causar um impacto significativo nos sistemas estatísticos, especialmente em termos de velocidade e eficácia na aquisição de vocabulário. Com esta evolução de técnicas e recursos, a tradução automática tem vindo a afirmar-se no universo da tecnologia da língua, a conquistar a simpatia dos internautas e utilizadores das redes sociais e a estabelecer-se como ferramenta por excelência de globalização das línguas. No entanto, a tradução automática não é um problema resolvido e apesar da facilidade de acesso e utilidade das traduções produzidas por intermédio desta tecnologia, a sua qualidade é ainda limitada por falta de conhecimento linguístico inerente a um tradutor humano. Neste contexto, os utilizadores da tradução automática ainda a consideram "pouco fiável". A falta de precisão nas traduções obtidas através desta ferramenta tecnológica está na base do trabalho de investigação em tradução automática decorrente no INESC-ID, nomeadamente a tradução de fala (Secção 3.1), a tradução de microblogues (Secção 3.2) e a criação de um sistema híbrido de tradução automática, tema que será explorado com algum detalhe adicional (Secção 3.3). Este trabalho visa enriquecer os atuais sistemas de tradução automática de base estatística com conhecimento linguístico (nomeadamente morfológico e semântico-sintático), de modo a que o mapeamento de uma frase na língua-fonte numa frase na língua-alvo

reflita o modo como a linguagem é processada pelo cérebro humano.

2 Paradigmas da Tradução Automática

A tradução automática tem evoluído com base em dois paradigmas principais: um baseado em conhecimento linguístico e outro baseado em dados e métodos estatísticos de mapeamento desses dados.

Os sistemas construídos com base em conhecimento linguístico, conhecidos como sistemas por regras, era o paradigma vigente até ao final dos anos 90 (Nirenburg et al., 2003; Scott, 2003). Estes sistemas não precisam de corpos paralelos, produzem tradução de qualidade, e funcionam bem até com poucos dados e poucas regras em domínios especializados, desde que se baseiem em dicionários e terminologias de qualidade. Também funcionam bem em línguas com um sistema morfológico rico em que algumas regras de flexão são suficientes para traduzir um grande número de formas com o mesmo radical. No entanto, o desenvolvimento deste tipo de sistemas envolve um grande investimento de tempo e recursos humanos especializados, necessários ao desenvolvimento de recursos linguísticos avançados para cada par de línguas a ser traduzido.

Os custos e morosidade envolvidos na construção dos sistemas por regras conduziram ao aparecimento de um novo paradigma, nos finais dos anos 90, os modelos estatísticos de tradução automática. Inicialmente baseados no alinhamento de n-gramas, estes modelos foram evoluindo com o tempo e começando a estender-se a expressões, de natureza não linguística (Koehn, 2007)[*]. Os sistemas estatísticos de tradução automática criam-se com base em corpos paralelos aos quais são aplicados algoritmos e técnicas de alinhamento dos vários elementos ou expressões da frase. Os algoritmos permitem encontrar padrões e prever a probabilidade de uma palavra ser a tradução de outra baseado no número de ocorrências dessa palavra em contexto nas duas línguas. O tempo de desenvolvimento dos sistemas estatísticos é rápido, desde que existam corpos paralelos para os pares de línguas que se pretendam traduzir, e por este motivo, a sua construção é muito mais económica do que a dos sistemas por regras. Até recentemente, os sistemas estatísticos envolviam conhecimento linguístico muito limitado, resultando em erros crassos há muito resolvidos pelos sistemas por regras. Mesmo com uma quantidade avultada de dados, como a que é utilizada por sistemas como o Google Translate, é necessária a pós-edição de erros simples, como os de concordância entre substantivo e adjetivo qualificativo, entre sujeito e verbo, entre outros. Para além disso, esses sistemas estão totalmente dependentes da quantidade e qualidade dos corpos paralelos que utilizam para os seus alinhamentos. Há línguas para os quais os corpos paralelos abundam e são de qualidade aceitável, como é o caso do inglês-mandarim, mas para outras línguas, como o basco, os corpos paralelos são escassos ou de qualidade reduzida, o que torna difícil a extração de generalizações necessária à tradução (Labaka et al., 2007). Apesar de terem sido propostos modelos mais sofisticados para a tradução de línguas com sistemas morfológicos complexos (Chahuneau et al., 2013), os sistemas por regras apresentam-se como a solução mais viável para traduzir estas línguas por necessitarem de uma quantidade menor de dados para traduzir as diferentes formas flexionadas de uma palavra.

2.1 Os Sistemas OpenLogos e Google Translate

Um dos modelos de tradução mais antigos é o sistema Logos (Scott, 2003), um sistema comercial baseado em regras, desenvolvido entre 1970 e 2001, e agora explorado na sua versão em código aberto: OpenLogos (Barreiro et al., 2011), adaptado pelo DFKI e disponível no SourceForge. O OpenLogos é considerado um sistema de tradução automática de qualidade, que comporta oito pares de línguas, contemplando a tradução de inglês para alemão, francês, espanhol, italiano e português, e de alemão para inglês, francês e italiano. A qualidade da tradução do OpenLogos resulta da sua componente semântico-sintática e da análise da língua de forma a que esta seja "entendida" pelo sistema computacional, tal como será descrito na Secção 5. A aproximação Logos assemelha-se, em espírito, à aproximação estatística, na medida em que as regras de base gramatical (semântico-sintática) são aplicadas a padrões em contexto. O conhecimento linguístico envolvido no sistema permite colmatar dificuldades e fraquezas apresentadas pelos métodos estatísticos, colocando-o na posição de plataforma ideal para uma solução híbrida.

Um dos sistemas de tradução automática mais populares é o Google Translate, disponível gratuitamente através da internet. O Google Translate utiliza um método estatístico para traduzir, que tem como alicerce o sistema em código aberto Moses (Koehn et al., 2003), usado por uma larga comunidade de investigadores e por alguns sistemas comerciais, como o Asia Online, entre outros. O Go-

[*] A tradução automática estatística é um paradigma com diversas vertentes, que não exploraremos neste artigo. Aconselhamos a leitura do seguinte estudo: http://www.cs.jhu.edu/~alopez/papers/survey.pdf

ogle Translate tem a forte vantagem de aceder a uma quantidade muito grande de corpos paralelos recolhidos da web, o que lhe permite a tradução de um número elevado de pares de línguas (cerca de 80), que varia em qualidade dependendo de fatores como a proximidade entre a língua-fonte e a língua-alvo, ou a quantidade e qualidade dos corpos disponíveis para a tradução de cada par de línguas. O Google Translate é um sistema comercial, pelo que não se sabe que módulos integra e se algum desses módulos contém conhecimento semântico necessário à tradução de qualidade.

2.2 Modelos Híbridos

Como o objetivo último dos investigadores e desenvolvedores de sistemas de tradução automática é o de criar sistemas que produzam tradução de qualidade comparável à que é produzida por tradutores humanos, a necessidade de um paradigma mais robusto e linguisticamente mais avançado tem vindo a afirmar-se nos últimos anos. Deste modo, surgiram os sistemas híbridos, que têm a vantagem de poder usufruir do trabalho de investigação de várias décadas, donde resultou, por um lado, a invenção e aperfeiçoamento das técnicas estatísticas que aceleram o processo de aquisição lexical e de tradução, e por outro lado, o desenvolvimento de maior quantidade de recursos linguísticos de melhor qualidade e para mais línguas. Os progressos alcançados nos paradigmas de vertente linguística e matemática da tradução automática tornaram-se, assim, um campo fértil para o desenvolvimento da nova geração de sistemas de tradução que ambicionam uma tradução de qualidade mais próxima à do tradutor humano. Vários métodos têm sido propostos para combinar tradução automática baseada em regras com tradução automática estatística.

Alguns sistemas processam estatisticamente as traduções obtidas a partir de um sistema por regras, enquanto que outros utilizam regras de base gramatical para processar previamente os dados, regras essas que ajudam a guiar o sistema estatístico. Um modelo de hibridização simples é o que combina traduções do mesmo texto por dois sistemas conceptualmente distintos (Heafield e Lavie, 2011; Eisele et al., 2003). Para além da combinação de sistemas, os modelos híbridos podem assentar, por um lado, na aplicação de técnicas estatísticas de alinhamento de expressões ou exemplos para melhorar a cobertura num sistema de tradução por regras (Eisele et al., 2003; Sánchez-Martínez et al., 2009) ou no melhoramento da qualidade de tradução de um sistema por regras através da utilização de métodos estatísticos de pós-edição (Simard et al., 2007; Elming, 2006; Dugast et al., 2007; Terumasa, 2007). Por outro lado, também tem sido proposta a integração de conhecimento linguístico em sistemas de tradução automática estatística (Satoshi et al., 1997). O trabalho de Niessen e Ney (2004) usa modelos lexicais hierárquicos para induzir a forma base das palavras em alemão para a tradução de termos compostos. Por outro lado, Ueffing et al. (2003) usa conhecimento linguístico para obter a forma correta das palavras quando a tradução é feita para línguas morfologicamente ricas como é o caso do espanhol e do catalão. Finalmente, a informação linguística pode também ser usada para melhorar a qualidade dos alinhamentos através do uso de informação acerca da categoria gramatical, como foi feito por Koehn e Night (2001) para o par de línguas alemão-inglês.

Embora não exista ainda indicação de que abordagem híbrida seja mais eficaz e conduza a um aumento do nível da qualidade da tradução, uma junção dos pontos mais fortes de cada paradigma ajuda a melhorar a tradução alcançada pelos novos sistemas e, como tal, a hibridização de sistemas de tradução automática continua a representar uma linha de investigação promissora. É no âmbito dessa aposta que se enquadra o trabalho de investigação descrito na Secção 3.3, onde exploramos uma aproximação nova à tradução automática híbrida, partindo da integração do conhecimento semântico-sintático do sistema OpenLogos em modelos estatísticos. Mas, antes de chegarmos a esse trabalho, abordaremos outros desafios na área da tradução automática, que são o da tradução de fala para fala, apresentado na secção 3.1, e o da tradução da linguagem usada em microblogues, descrito na secção 3.2.

3 Tradução Automática no INESC-ID

3.1 Tradução Automática de Fala

A tarefa de tradução ganha uma utilidade adicional se tiver como objetivo traduzir fala para fala. No entanto, a tarefa em si torna-se igualmente muito mais complexa, pelo que a investigação na área da tradução de fala para fala enfrenta desafios adicionais para além dos usualmente associados às tarefas de tradução de texto para texto. Este processo pode ser visto como uma sequência de três etapas: reconhecimento automático de fala (passagem de fala para texto na língua-fonte), tradução (passagem do texto na língua-fonte para texto na língua-alvo) e síntese (passagem do texto na língua-alvo para fala). Assim, quaisquer erros que ocorram em cada um destes módulos dificultam seriamente as tarefas dos módulos subsequentes. Neste cenário, o primeiro grande desafio consiste na tradução de fala espontânea. Dado que os módulos de

reconhecimento e de tradução são normalmente treinados com base em textos escritos, torna-se extremamente difícil processar hesitações, repetições, pausas preenchidas e expressões não gramaticais, muito frequentes em fala espontânea.

O projeto PT-STAR (financiado pela FCT, no quadro do programa Carnegie Mellon – Portugal e recentemente terminado) teve como objetivo melhorar os sistemas de tradução de fala para fala 'de' e 'para' português, focando-se na interligação entre os três principais módulos destes sistemas. Fizeram parte deste consórcio o Laboratório de Sistemas de Língua Falada (L2F), do INESC-ID Lisboa, o Instituto de Tecnologias da Língua (LTI) da Universidade de Carnegie Mellon, o Centro de Linguística da Universidade de Lisboa e a Universidade da Beira Interior. O estado da arte atual em tradução de fala para fala mostra uma integração relativamente fraca entre os três módulos, não explorando as sinergias existentes entre o reconhecimento e a tradução, entre a tradução e a síntese e ainda entre o reconhecimento e a síntese. Por exemplo, o módulo de reconhecimento escolhe normalmente uma única hipótese de transcrição que será a entrada do módulo de tradução. Se, em alternativa a esta hipótese, for oferecida ao módulo de tradução uma lista de possíveis transcrições, este pode decidir qual a mais adequada aos modelos de tradução. Por outro lado, o módulo de síntese assume que receberá como entrada texto fluente, o que usualmente não acontece quando essa entrada resulta de um módulo de tradução automática. Assim, de maneira a que a interligação entre estes dois módulos seja mais robusta, a estratégia de síntese tem que ser modificada a fim de evitar a produção não compreensível de voz, por exemplo, eliminando palavras com baixa confiança. Ser capaz de transferir o foco principal (ou ênfase) de entrada da língua-fonte para a língua-alvo é outra tarefa extremamente desafiante e que obriga a que exista uma ligação entre os módulos de reconhecimento e de síntese.

3.1.1 Cenários de Aplicação

Dois grandes cenários foram palco do grosso dos avanços no projeto PT-STAR: a tradução de TED Talks e de notícias televisivas. Na tradução das TED Talks, a adaptação de domínio e de conversão de voz foram os principais desafios, tendo o sistema de lidar com aplausos ocasionais e risos da plateia. A tradução das notícias televisivas tornou-se também um cenário de grande interesse, pois permitiu, para além de testar técnicas de adaptação ao domínio, trabalhar sobre fala controlada (por exemplo, dos pivôs do telejornal), bem como fala espontânea. O leitor pode encontrar uma demonstração, feita em tempo real, do sistema de tradução de fala para fala desenvolvido no projeto PT-STAR no seguinte endereço: www.l2f.inesc-id.pt/wiki/index.php/Demos. Nesta demonstração o reconhecedor foi treinado com textos de jornais e o tradutor com os habituais dados do Europarl (Koehn, 2005). A qualidade da tradução do sistema PT-STAR deve-se à proximidade entre os dois domínios.

3.1.2 Desafios em Destaque

Dos vários trabalhos de investigação abordados no projeto PT-STAR, destacamos dois que ilustram a problemática em mãos: o primeiro na área do reconhecimento, o segundo na fronteira entre a tradução e a síntese.

Enriquecimento de transcrições

A saída de um reconhecedor consiste apenas numa sequência de palavras. A capacidade de enriquecer esta sequência com a pontuação apropriada afeta não só a qualidade da transcrição, mas também a possibilidade de melhorar o passo de tradução, pois uma segmentação adequada é fundamental para o sucesso da tarefa de tradução (por exemplo, conseguiu-se uma melhoria de 2 pontos BLEU numa experiência que consistiu em passar imediatamente ao tradutor todos os segmentos terminados com qualquer sinal de pontuação e não apenas os segmentos terminados com um ponto final (Grazina, 2010). Assim, uma das tarefas deste projeto consistiu no desenvolvimento de módulos capazes de inserir vários sinais de pontuações em transcrições automáticas. Duas estratégias diferentes foram exploradas no que diz respeito ao ponto final e à vírgula. O primeiro fez uso de fontes de informação que podem ser encontradas em textos; o segundo consiste na introdução de características prosódicas: além de pistas lexicais, foram utilizadas pistas baseadas no tempo e em características do falante. Ambas as estratégias melhoraram os resultados iniciais. Por exemplo, para a saída do reconhecedor para português, a pontuação foi melhorada em cerca de 5,6% (Batista et al., 2012).

Melhoria da síntese resultante de um texto traduzido automaticamente

A saída do módulo de tradução automática é muitas vezes inadequado para ser passada diretamente ao módulo de síntese, o que representa um problema na interface entre a tradução automática e a síntese. Como os modelos do sintetizador são geralmente treinados com texto fluente, este sintetizará a saída do tradutor assumindo a sua fluência,

o que tornará a fala resultante difícil de entender. Assim sendo, um dos desafios deste projeto prendeu-se com a tentativa de otimizar o sintetizador de modo a que a compreensão do texto fosse a melhor possível, apesar dos erros de tradução. Várias técnicas foram testadas (Parlikar et al., 2010), tais como a utilização de material de preenchimento de pausas que provaram ser de utilidade para melhorar a inteligibilidade da fala.

3.2 Tradução Automática de Microblogues

Outra problemática da tradução automática abordada em projetos desenvolvidos no INESC-ID é o da tradução de linguagem de microblogues. Na última década, os microblogues, como o Facebook, o Twitter, o Youtube ou o Sina Weibo (versão chinesa do Twitter), têm sido alvo de uma atenção especial pela comunidade científica por razões que se prendem com a quantidade de pessoas que as utilizam e com o volume de informação existente neste domínio. No entanto, os conteúdos textuais abrangidos pelos microblogues caracterizam-se por incluírem termos pouco formais e linguagem não padronizada. Alguns exemplos incluem a presença de abreviaturas como a da expressão em inglês *r u still following me or what?*. Estas expressões são geralmente problemáticas para os sistemas de tradução automática, por dois motivos principais, que passaremos a descrever.

No primeiro caso, os modelos de tradução não são treinados com dados deste domínio, simplesmente porque eles não existem. Em consequência, os sistemas de tradução treinados com dados fora do domínio dos microblogues não estão aptos para traduzir a linguagem nele usada. Este problema dá origem a erros de tradução, tais como os que são evidenciados na tradução para português do exemplo acima pelo Google Translate: *r u ainda me seguindo ou o quê?*, em que as abreviaturas *r* e *u* simplesmente não são traduzidas. Em resposta a este problema, foi construído um sistema de extração automática de traduções do Twitter e Sina Weibo. Este sistema é motivado pela observação que há alguns utilizadores que traduzem os seus tuítes e estes podem ser extraídos e usados para melhorar significativamente os sistemas de tradução no domínio. Por exemplo, o tuíte *Male Body Painting - Pintura Corporal Masculina (Essa Moda Pega?)* contém a tradução do nome composto em inglês *Male Body Painting* para o nome composto em português *Pintura Corporal Masculina*. Estas traduções no tuíte publicado podem ser encontradas através de um algoritmo de "emparelhamento baseado em conteúdo" (Resnik e Smith, 2003), que representa o estado da arte (Ling et al., 2013b). Este algoritmo explora técnicas para a extração de corpos paralelos da web, onde dois documentos são identificados como traduções um do outro se existir uma grande percentagem de palavras que são consideradas como boas traduções entre esses documentos. O contributo principal desse trabalho consiste na extensão do algoritmo aos casos em que a tradução se encontra no mesmo documento, como acontece no caso dos microblogues. Com este algoritmo é possível obter uma grande quantidade de traduções para várias línguas, incluindo o português. As experiências feitas com 3 milhões de frases paralelas para chinês-inglês mostram que a utilização desse corpo pode fazer aumentar consideravelmente a qualidade da tradução. A melhoria deve-se essencialmente ao facto de os sistemas existentes serem incapazes de traduzir termos frequentes como *u*, *thx* e *r*, responsáveis pela degradação da qualidade da tradução.

O segundo problema encontrado nos atuais sistemas de tradução automática estatística está relacionado com facto de estes sistemas modelarem a linguagem como sequências de palavras. Por exemplo, os modelos consideram as formas *hellllo* e *hello* como dois códigos (*tokens*) diferentes. Assim sendo, os modelos de tradução apenas conseguem traduzir a forma *hellllo* se esta constar no conjunto de dados de treino do sistema. Este problema não pode ser tratado através da extração de mais corpos por não ser possível obter traduções para todas as formas que aparecem no corpo (e.g. *helo*, *heeeeello*, *ello*, etc.). É necessário que o modelo aprenda a generalizar o processo de tradução de maneira a que reconheça todas estas formas como variantes da palavra *hello*. Para solucionar o problema, Ling et al. (2013a) propõem um sistema de normalização que aprende a converter as frases informais em paráfrases com a mesma informação, mas representada de forma padronizada. Este sistema de normalização converte, por exemplo, a frase *r u still following me or what??* em *Are you still with me or what?*. Este parafraseamento permite um processamento mais fácil destas mensagens, quer por sistemas de tradução, quer por outros sistemas de processamento de linguagem natural.

O sistema de tradução de microblogues assenta em tecnologias de parafraseamento construídas a partir da tradução (Callison-Burch, 2007). O sistema usa um corpo paralelo inglês-chinês, onde a porção chinesa é traduzida de forma automática para inglês, gerando um corpo de paráfrases que constitui uma versão alternativa do corpo original. É possível usar esse corpo recolhido automaticamente para obter um mapeamento das frases onde podem ocorrer variações estilísticas entre a frase original e a tradução. Uma vantagem que o método apresenta é tornar possível o uso de frases paralelas

	Fenómeno Linguístico	Original em Inglês	OpenLogos	Google Translate
(1)	Conc. SUJ-V PRON OI	Kennedy interviewed you.	Kennedy entrevistou-o.	*Kennedy entrevistei.
(2)	PE vs PB	Kennedy interviewed me.	Kennedy entrevistou-me. (PE)	Kennedy me entrevistou. (PB)
(3)	Conc. SUJ-V PRON OI	Kennedy interviewed us.	Kennedy entrevistou-nos.	*Kennedy nos entrevistaram.
(4)	Conc. SUJ-V	Me and her interviewed Kennedy.	Eu e ela entrevistámos Kennedy.	*Eu e ela entrevistou Kennedy.
(5)	Conc. N-ADJ	Kennedy has a bookcase that is heavy.	Kennedy tem uma estante que é pesada.	*Kennedy tem uma estante que é pesado.
(6)	V-Aux	Kennedy hired women who were competent.	Kennedy contratou mulheres que foram competentes.	*Kennedy contratou mulheres que estavam competente.
(7)	V-Sem	She manages whom?	Ela dirige quem?	*Ela consegue quem?
(8)	HOMO N-V HOMO V-N	Managers work.	Os gerentes trabalham.	*Gerentes de trabalho.
(9)	HOMO V-N	List women who have bookcases.	Enumere mulheres que têm estantes.	*Lista de mulheres que têm estantes.
(10)	V-PT vs V-PP PRON REFL	The women evaluated themselves.	As mulheres avaliaram-se.	*As mulheres avaliadas si.

Tabela 1 – Tradução de aspetos da gramática traduzidos pelos sistemas OpenLogos e Google Translate.

em línguas para as quais é viável extrair uma grande quantidade de dados, como é o caso do par inglês-chinês e criar sistemas de normalização para estas línguas, estendendo-a a pares de línguas para os quais existem poucos dados traduzidos, como é o caso do inglês-português ou do chinês-português. Outra vantagem consiste na possibilidade de utilizar as ferramentas de parafraseamento e normalização para outras tarefas de processamento de linguagem natural. Xu et al. (2014) apresentam a extração, por via de métodos estatísticos, de pares de tuítes semelhantes que constituem paráfrases. No entanto, para a finalidade de normalização não há garantia que os tuítes colecionados tenham o mesmo nível de língua não-padrão e as paráfrases podem não ser adequadas para a criação de um sistema de normalização.

3.3 Tradução Automática com Conhecimento Semântico-Sintático

O trabalho de investigação em tradução automática híbrida atualmente em vigor no INESC-ID consiste na criação de um novo modelo que combina conhecimento linguístico com tradução automática estatística. Para o efeito, partimos do sistema OpenLogos, um sistema com características peculiares que lhe permitem servir de plataforma para a criação desse novo modelo híbrido. Embora a hibridização exija um esforço significativo, os seus princípios basilares assentam na integração de conhecimento linguístico, que já deu provas de sucesso na tradução de muitos aspetos da gramática. A Tabela 1 apresenta a tradução de frases com diferentes níveis de gramaticalidade/naturalidade extraídas de um corpo construído para testar fenómenos linguísticos no sistema OpenLogos contrastando-a com a tradução obtida através do sistema Google Translate. Estas frases apresentam fenómenos variados como a concordância entre o sujeito e o verbo (exemplos (1), (3) e (4)) ou a concordância entre o nome e o adjetivo qualificativo numa construção relativa (exemplo (5)), os diferentes tipos de pronome (objeto indireto, reflexo, etc. (exemplos (1), (3) e (10)), a escolha do verbo auxiliar (*ser*, *estar* (exemplo (6)), a semântica do verbo (exemplo (7)), as palavras homógrafas, como nome-verbo (exemplos (8) e (9)), as formas terminadas em *-ed*, que podem ser formas do pretérito perfeito ou do particípio passado (exemplo (10)), entre outros. As variantes europeia (PE) e brasileira (PB) do português (exemplo (2)) também surgem em contraste nas traduções obtidas através dos dois sistemas. As frases traduzidas pelo sistema Google Translate refletem a dificuldade e imprevisibilidade dos sistemas estatísticos quando confrontados com alguns fenómenos gramaticais, principalmente em frases curtas e ambíguas para os quais não foram treinados. Em contrapartida, estes sistemas conseguem uma cobertura lexical mais abrangente, o que, por um lado representa uma grande vantagem, mas por outro, pode provocar nos utilizadores uma sensação de que os sistemas estatísticos traduzem melhor do que os sistemas por regras, mesmo que essa tradução se deva essencialmente ao acesso a grandes quantidades de dados que permitem treinar os sistemas estatísticos em domínios específicos de interesse e de utilidade para os seus utilizadores.

A avaliação desenvolvida em trabalho anterior (Barreiro et al., 2013) mostra que tanto os modelos linguísticos como os estatísticos apresentam uma baixa qualidade na tradução de unidades lexicais multipalavra e que este fenómeno linguístico continua a representar um desafio significativo para a tradução automática, independentemente do tipo de paradigma. Na mesma linha de pensamento, a avaliação quantitativa e qualitativa anteriormente realizada (Barreiro et al., 2014b) das traduções de construções com verbos-suporte pelos sistemas OpenLogos e Google Translate reforça a necessidade de criação de sistemas híbridos que tirem partido da robustez dos sistemas por regras para melhorar a tradução automática de unidades lexicais multipalavra. A proposta consiste na hibridização por via da integração de conhecimento semântico-sintático nos sistemas estatísticos.

O sistema OpenLogos foi o escolhido por oferecer duas importantes vantagens em relação a outros sistemas baseados em regras. Uma diz respeito às regras não serem dependentes de algoritmos (ou guiadas por meta-regras), libertando o sistema da saturação da lógica quando confrontado com um problema tão complexo como o da tradução. A outra, está relacionada com a representação simbólica da língua natural, que no sistema OpenLogos, é transposta para um nível mais abstrato do que as palavras. Normalmente, os sistemas algorítmicos tendem a estar limitados à capacidade do algoritmo, e qualquer sequência lógica deixa de funcionar a determinada altura ao processar a língua natural. Não tendo a limitação do algoritmo de supervisão, o OpenLogos tem capacidade ilimitada para melhorar a tradução e as melhorias são fácil e imediatamente implementáveis. Por outro lado, as regras do sistema OpenLogos são baseadas em padrões de língua natural organizadas numa taxonomia de segunda ordem, chamada linguagem de abstração semântico-sintática, de ora em diante, SAL[†], descrita em Barreiro et al. (2011) e no Tutorial SAL[‡]. Estas duas importantes características do sistema posicionam o OpenLogos num patamar semelhante, em espírito, ao da tradução automática estatística e, em relação a esta, lhe concedem a vantagem de não sofrer do problema típico de escassez dos dados de treino, possibilitando o processamento de texto e a sua tradução. O analisador gramatical do sistema permite gerar árvores em diferentes níveis de análise, como descrito em Barreiro et al. (2011). Em cada nível de análise, as sequências de língua natural (palavras ou expressões) são representados por unidades SAL, que podem substituir os elementos comuns de mapeamento dos sistemas estatísticos para a finalidade da tradução. Com esta representação do conhecimento linguístico, os n-gramas evoluem de palavras e expressões para sequências de elementos SAL (ou seja, sequências de palavras ou expressões com propriedades semântico-sintáticas), permitindo que este mapeamento ocorra a um nível mais abstrato. Esta técnica conduz a um aumento da capacidade de mapeamento de sinónimos e expressões semanticamente equivalentes, aumentando a capacidade de encontrar paráfrases e traduções mais adequadas. Em consonância com SAL, o sistema usa regras semântico-sintáticas (designadas por SEMTAB) para realizar transformações monolíngues e multilíngues que funcionam com elevada eficácia na tradução das áreas mais frágeis dos sistemas estatísticos, como a tradução de todo o tipo de homógrafos (Barreiro et al., 2005), de verbos com traduções diferentes dependendo dos seus argumentos (e.g. *to raise a child - criar/educar um(a) criança/filho*; *to raise awareness – consciencializar*; *to raise concerns - suscitar preocupação*; *to raise funds – angariar / arranjar / obter financiamentos/fundos*) e de unidades lexicais multipalavra (incluindo as unidades não adjacentes, ou descontínuas, tais como *was in no way related to - não estava de forma alguma relacionado com* ou *is falling far short of - está bem aquém de*), entre outros.

O modelo integra um módulo de parafraseamento que permite melhorar o mapeamento de termos e expressões semanticamente idênticos. Este modelo funciona independentemente da quantidade e qualidade de corpos disponível e assenta numa metodologia repetível, que pode ser usada em diferentes aplicações de processamento de linguagem natural e tarefas multilíngues. O módulo de parafraseamento do modelo híbrido ajuda a tradução automática, permitindo o mapeamento de unidades lexicais multipalavra com palavras ou expressões com o mesmo significado. Por exemplo, as construções com verbos-suporte podem ser transformadas em verbos (*fazer a apresentação de - apresentar* ou *dar um abraço a - abraçar*), porque os nomes predicativos *apresentação* e *abraço* estão semanticamente relacionados com os verbos *apresentar* e *abraçar* não apenas a um nível abstrato (SAL) independente da categoria gramatical, mas também ao nível morfossintático, através de regras de derivação.

Em suma, o modelo híbrido em desenvolvimento aplica conhecimento linguístico na análise e resolução de ambiguidades por meio de regras e técnicas estatísticas de mapeamento de unidades linguísticas em vez de n-gramas, ou seja, palavras ou sequências de palavras, onde a informação semânti-

[†] SAL é o acrónimo de Semantico-Syntactic Abstraction Language.

[‡] O Tutorial SAL está disponível no seguinte endereço da internet: http://www.l2f.inesc-id.pt/~abarreiro/openlogos-tutorial/new_A2menu.htm

```
<Entry source="pipe" target="tubo">
   <source head_word="1" homograph="no" word_type="01">
     <pos description="Noun" wclass="01"/>
     <morphology num_id="1" number="singular">
       <inflection description="like book, books" example="book" id="16"/>
     </morphology>
     <sal code="3,34,745" mnemonic="COcond" set="functional" subset="conduit"
superset="concrete"/>
   </source>
   <target head_word="1" word_type="01">
     <pos description="Noun" wclass="01"/>
     <morphology gen_id="1" gender="masculine" num_id="1" number="singular">
       <inflection description="plural adding -s" example="tinteiro" gender_code="1" id="99"/>
     </morphology>
   </target>
</Entry>
```

Figura 1 - Entrada lexical para o nome concreto, funcional, objeto condutor: EN *pipe* - PT *tubo*.

co-sintática está disponível em cada etapa da análise da frase. A técnica permite respeitar a representação científica da linguagem, como, por exemplo, a unicidade de unidades lexicais multipalavra, para as quais os atuais sistemas de alinhamento não apresentam uma solução científica e tecnicamente viável.

3.3.1 Dicionário Inglês-Português

Os dicionários bilíngues representam um recurso importante na tradução automática, e quanto mais conhecimento envolverem, melhor poderão contribuir para a qualidade da tradução. O dicionário de inglês-português, tal como os restantes dicionários da OpenLogos apresentados em Barreiro et al. (2014a), contém conhecimento semântico-sintático e conhecimento ontológico (SAL), desenvolvido ao longo de várias décadas pela equipa de linguistas da Logos. Estes dicionários integravam o antigo produto comercial de tradução automática LogosMT, agora disponível em código aberto no sistema OpenLogos[§]. Nos dicionários do OpenLogos existem mais de 1.000 categorias distribuídas por quatro níveis de abstração: o nível sintático (categoria gramatical) e três níveis de conceito abstrato SAL, designados de superconjunto (superset), conjunto (set) e subconjunto (subset). Essas categorias compreendem tanto classificações gramaticais, tais como verbo bitransitivo, nome próprio, etc.), como classificações conceptuais (método, instrumento, etc.), que estão sistematicamente relacionadas entre si.

Os verbos intransitivos, por exemplo, estão classificados semanticamente em três conjuntos: os *existenciais*, os *operacionais* e os *de movimento*, cada um destes, por sua vez, subdivididos em vários subconjuntos, de acordo com as suas propriedades sintáticas. Por outro lado, os substantivos *funcionais*, um conjunto do superconjunto *concreto*, subdivide-se em vários subconjuntos, tais como *ferramentas/dispositivos*, *objetos em tecido*, *objetos condutores*, entre outros.

Categoria	id	Frequência
Substantivo	1	28270
Verbo	2	34985
Advérbio (locativo)	3	458
Adjetivo	4	24503
Pronome	5	121
Advérbio (modo, grau, etc.)	6	2189
Preposição (não locativa)	11	140
Auxiliar/modal	12	34
Preposição (locativa)	13	148
Artigo definido	14	194
Artigo indefinido	15	66
Número em aposição	16	208
Negação	17	2
Pronome relativo e interrogativo	18	23
Conjunção	19	160
Pontuação	20	30
Total		91531

Tabela 2 - Total de entradas por categoria.

A Figura 1 apresenta a entrada lexical EN *pipe* – PT *tubo*, classificada como um nome concreto, funcional, objeto condutor. Para além da classificação SAL, a entrada lexical tem informação morfológica para a palavra em inglês e em português. *Pipe* segue o paradigma flexional #16 para substantivos que formam o plural em *-s*, como *book*. *Tubo* segue o paradigma flexional #99 para substantivos masculinos que formam o plural em *-s*, como *tinteiro*. Outro tipo de informação pode ser extraída dos dicionários, mas, por motivos de simplicidade, não será ilustrada neste artigo. A Tabela 2 apresenta o número de entradas por categoria gramatical. Estes recursos linguísticos estão disponíveis para integração em aplicações de processamento de linguagem natural, incluindo a tradução automática 'de' e 'para' português e serão usados no sistema híbrido em desenvolvimento.

[§] http://logos-os.dfki.de, openlogos-mt.sourceforge.net

4 Conclusão e Trabalho Futuro

A tradução automática veio para revolucionar a comunicação no mundo, permitindo o acesso à informação em várias línguas estar ao alcance de um simples clique. Resta ainda um trabalho significativo para melhorar a qualidade linguística dos atuais sistemas, para que esta tecnologia alcance a sua plenitude. Neste artigo, descrevemos o trabalho de investigação em tradução automática desenvolvido no INESC-ID. Este trabalho começa a revelar sólidos progressos em várias frentes, nomeadamente na tradução de fala para fala, na tradução de microblogues e na evolução da tecnologia de tradução automática híbrida com qualidade melhorada. Os esforços futuros continuarão a incidir sobre o desenvolvimento e aperfeiçoamento do modelo híbrido que visa ser linguisticamente mais avançado e tecnicamente mais rápido e fácil de explorar para outras aplicações. Integrará um módulo parafrástico resultante do projeto eSPERTo[**] (Sistema de Parafraseamento para Edição e Revisão de texto) que visa servir de ferramenta de pré-edição e de auxílio à tradução. Para além disso, o sistema híbrido ambiciona tirar proveito da computação em nuvem, de grandes volumes de dados e de técnicas de alinhamentos linguisticamente motivados, que contribuem, no seu conjunto, para um desenvolvimento mais fácil e rápido de novos pares de línguas. O modelo deverá também poder ser combinado com a tradução participativa ou colaborativa (*crowdsourcing*) que permite aumentar exponencialmente o volume dos corpos paralelos existentes para as várias línguas e posteriormente beneficiar do apoio de tradução coletiva especializada para aumentar a qualidade das traduções adquiridas colaborativamente. A língua portuguesa precisa de se manter viva e relevante para o mundo moderno, o que só acontecerá se se posicionar na linha da frente da tecnologia em tradução automática. Se os esforços forem canalizados nesse sentido e um maior investimento contemplar a melhoria da qualidade e quantidade dos recursos linguísticos orientados para a tradução, no futuro a tradução automática tornar-se-á a principal amiga da língua portuguesa.

Agradecimentos

Agradecemos aos organizadores do colóquio *Português, Lingua Global*, promovido pelo Centro de Estudos Lusíadas do Instituto de Letras e Ciências Humanas da Universidade do Minho, onde foi apresentada a comunicação *Contributos da Tecnologia da Língua na Globalização do Português*, e à sua respetiva audiência, a oportunidade de desenvolver o trabalho que esteve na base deste artigo.

O trabalho de Anabela Barreiro foi financiado pela FCT (bolsa de pós-doutoramento SFRH / BPD / 91446 / 2012). Este trabalho também contou com o apoio da FCT, através do projeto PEst-OE/EEI/LA0021/2013.

Referências

Barreiro, Anabela e Elisabete Ranchhod. 2005. Machine Translation Challenges for Portuguese, *Lingvisticæ Investigationes* 28:1, pp. 3-18. In Sylviane Cardey, Peter Greenfield, Séverine Vienney (eds), Machine Translation, Controlled Languages and Specialised Languages.

Barreiro, Anabela, Bernard Scott, Walter Kasper e Bernd Kiefer. 2011. OpenLogos Rule-Based Machine Translation: Philosophy, Model, Resources and Customization. *Machine Translation*, 25(2):107– 126.

Barreiro, Anabela, Johanna Monti, Brigitte Orliac e Fernando Batista. 2013. When Multiwords Go Bad in Machine Translation, in *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Machine Translation Summit XIV*.

Barreiro, Anabela, Fernando Batista, Ricardo Ribeiro, Helena Moniz e Isabel Trancoso. 2014a. OpenLogos Semantico-Syntactic Knowledge-Rich Bilingual Dictionaries, in *Proceedings of the 9th edition of the LREC conference*.

Barreiro, Anabela, Johanna Monti, Brigitte Orliac, Susanne Preuss, Kutz Arrieta, Wang Ling, Fernando Batista e Isabel Trancoso. 2014b. Linguistic Evaluation of Support Verb Construction Translations by OpenLogos and Google Translate, in *Proceedings of the 9th edition of the LREC conference*.

Batista, Fernando, Helena Moniz, Isabel Trancoso e Nuno J. Mamede. 2012. *Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts.* IEEE Transactions on Audio, Speech and Language Processing, Special Issue on New Frontiers in Rich Transcription, 20(2):474-485.

Brown, Ralf D. 1996. Example-Based Machine Translation in the Pangloss System, in *COLING* vol. 1, pp. 169-174.

Callison-Burch, Chris. 2007. Paraphrasing and Translation, PhD Thesis, University of Edinburgh.

** https://esperto.l2f.inesc-id.pt

Chahuneau, Victor, Smith, Noah A. e Dyer, Chris. 2013. "Knowledge-Rich Morphological Priors for Bayesian Language Models.", in *HLT-NAACL* (ACL), pp. 1206-1215.

Dugast, Löıc, Senellart, Jean e Koehn, Philipp. 2007. Statistical Post-editing on SYSTRAN's Rule-based Translation System, in *Proceedings of the Second Workshop on Statistical Machine Translation* (Stroudsburg, PA, USA: ACL), pp. 220-223.

Eisele, Andreas, Christian Federmann, Hans Uszkoreit, Hervé Saint-Amand, Martin Kay, Michael Jellinghaus, Sabine Hunsicker, Teresa Herrmann e Yu Chen. 2008. Hybrid Machine Translation Architectures within and beyond the EuroMatrix project, in J. Hutchins and W.V. Hahn, ed., *Hybrid MT Methods in Practice: Their Use in Multilingual Extraction, Cross-Language Information Retrieval, Multilingual Summarization, and Applications in Hand-Held Devices. Proceedings of the European Machine Translation Conference* (HITEC e. V 2008), pp. 27-34.

Elming, Jakob. 2006. Transformation-based correction of rule-based MT., in *Proceedings of EAMT 2006*, pp. 219--226.

Grazina, Nuno. 2010. Tradução Automática de Fala. Tese de Mestrado. IST.

Heafield, Kenneth e Lavie, Alon. 2011. CMU System Combination in WMT 2011, in *Proceedings of the Sixth Workshop on Statistical Machine Translation* (Stroudsburg, PA, USA: ACL), pp. 145-151.

Koehn, Philipp e Kevin Knight. 2002. ChunkMT: Statistical Machine Translation with Richer Linguistic Knowledge.

Koehn, Philipp, Franz Josef Och e Daniel Marcu. 2003. Statistical Phrase-Based Translation, in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL '03)* (Morristown, NJ, USA: ACL), pp. 48-54.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondřej, Constantin, Alexandra e Herbst, Evan. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (Stroudsburg, PA, USA: ACL), pp. 177-180.

Labaka, Gorka, Stroppa, Nicolas, Way, Andy e Sarasola, Kepa. 2007. "Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation", in *Machine Translation Summit XI* (Copenhagen, Denmark).

Ling, Wang, Dyer, Chris, Black, Alan W. e Trancoso, Isabel. 2013a. Paraphrasing 4 Microblog Normalization, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, Washington, USA: ACL), pp. 73-84.

Ling, Wang, Xiang, Guang, Dyer, Chris, Black, Alan e Trancoso, Isabel. 2013b. Microblogs as Parallel Corpora, in *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics* (ACL).

Niessen, Sonja e Hermann Ney. 2004. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information, *Computational Linguistics* 30, 2, pp. 181--204.

Nirenburg, Sergei, Harold Somers e Yorick Wilks. 2003. *Readings in Machine Translation* (Five Cambridge Center, Cambridge, MA: The MIT Press).

Parlikar, Alok, Alan W. Black e Stephan Vogel. 2010. Improving Speech Synthesis of Machine Translation Output, Interspeech (2010), Makuhari, Japan.

Resnik, Philip e Noah A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics* 29, 3 (September 2003), 349-380.

Sánchez-Martínez, Felipe, Forcada, Mikel L. e Way, Andy. 2009. Hybrid rule-based - example-based MT: feeding Apertium with sub-sentential translation units", in *EBMT 2009 - 3rd Workshop on Example-Based Machine Translation* (Dublin, Ireland: DORAS).

Scott, Bernard. 2003. The Logos Model: An Historical Perspective, *Machine Translation* 18, 1, pp. 1-72.

Shirai, Satoshi, Francis Bond e Yamato Takahashi. 1997. A Hybrid Rule and Example-based Method for Machine Translation, in *Recent Advances in Example-Based Machine Translation* (Kluwer Academic Publishers), pp. 211-224.

Simard, Michel, Ueffing, Nicola, Isabelle, Pierre e Kuhn, Roland. 2007. Rule-Based Translation with Statistical Phrase-Based Post-Editing", in *Proceedings of the Second Workshop on Statisti-*

cal Machine Translation (Prague, Czech Republic: ACL, pp. 203-206.

Terumasa, Ehara. 2007. Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation, in *Proceedings of the MT Summit XI Workshop on Patent Translation* vol. 11, pp. 13--18.

Ueffing, Nicola e Ney, Hermann. 2003. Using POS Information for Statistical Machine Translation into Morphologically Rich Languages, in *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1* (Stroudsburg, PA, USA: ACL), pp. 347--354.

Xu, Wei, Ritter, Alan, Callison-Burch, Chris, Dolan, William, e Ji, Yangfeng. 2014. Extracting Lexically Divergent Paraphrases from Twitter. *Transactions Of The Association For Computational Linguistics, 2*, 435-448.

Volume 6, Número 2 - Dezembro 2014

linguaMATICA

ISSN: 1647-0818

Artigos de Investigação

Projetos, Apresentam-se

http://linguamatica.com