Volume 3, Número 1- Junho 2011

linguaMATICA

ISSN: 1647-0818

Universidade de Vigo

Universidade do Minho

APPIA

Volume 3, Número 1 – Junho 2011

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Editorial

Esta é a sexta edição da revista Linguamática, e a primeira do seu terceiro ano de vida. Para comemorarmos este aniversário oficializamos a abertura da revista a artigos escritos em basco (euskera), publicando como artigo convidado uma revisão sobre o trabalho que tem sido feito para o processamento desta língua por parte do Grupo Ixa da Univesidade do País Basco.

Este sexto número da Linguamática inclui artigos en português, espanhol, galego e basco. Convidamos os investigadores em PLN das línguas ibéricas a enviar propostas à revista em qualquer uma das línguas oficiais e originárias da península, seja português, galego, catalão, castelhano ou basco (e no futuro, quem sabe não possamos também aceitar artigos em mirandês ou aranês).

Para além de incorporar o basco às línguas de publicação da revista, a partir deste número acolhemos também quatro novos membros na Comissão Científica, a quem muito agradecemos pelo interesse e disponibilidade para ajudar nesta tarefa.

Como habitual, agradecemos a todos os revisores envolvidos nesta publicação, bem como aos autores do artigo convidado, e todos os restantes autores, de artigos publicados ou não, pelo seu interesse na revista Linguamática.

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, ISLA e Universidade Nova de Lisboa
Anselmo Peñas, Universidad Nacional de Educación a Distancia
Antón Santamarina, Universidade de Santiago de Compostela
Antonio Moreno Sandoval, Universidad Autónoma de Madrid
António Teixeira, Universidade de Aveiro
Arantza Díaz de Ilarraza, Euskal Herriko Unibertsitatea
Belinda Maia, Universidade do Porto
Carmen García Mateo, Universidade de Vigo
Diana Santos, Linguateca/FCCN
Ferran Pla, Universitat Politècnica de València
Gael Harry Dias, Universidade Beira Interior
Gerardo Sierra, Universidad Nacional Autónoma de México
German Rigau, Euskal Herriko Unibertsitatea
Helena de Medeiros Caseli, Universidade Federal de São Carlos
Horacio Saggion, University of Sheffield
Iñaki Alegria, Euskal Herriko Unibertsitatea
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, Laboratoire Informatique d'Avignon - UAPV
Kepa Sarasola, Euskal Herriko Unibertsitatea
Lluís Padró, Universitat Politècnica de Catalunya
Maria das Graças Volpe Nunes, Universidade de São Paulo
Mercè Lorente Casafont, Universitat Pompeu Fabra
Mikel Forcada, Universitat d'Alacant
Patrícia Cunha França, Universidade do Minho
Pablo Gamallo Otero, Universidade de Santiago de Compostela
Salvador Climent Roca, Universitat Oberta de Catalunya
Susana Afonso Cavadas, University of Sheffield
Tony Berber Sardinha, Pontifícia Universidade Católica de São Paulo
Xavier Gómez Guinovart, Universidade de Vigo

Dossier

Teknologia garatzeko estrategiak baliabide urriko hizkuntzetarako: euskararen eta Ixa taldearen adibidea

Iñaki Alegria, Xabier Artola,
Arantza Diaz de Ilarraza, Kepa Sarasola
Universidad del Pais Vasco -
Euskal Herriko Unibertsitatea
i.alegria@ehu.es

Itziar Aduriz
Universitat de Barcelona
itziar.aduriz@ub.edu

Resumen

El artículo comienza presentando varios datos que muestran la situación de la lengua vasca, y a continuación proponiendo una clasificación para las lenguas del mundo según sea su presencia en Internet y en la tecnología de la lengua. El cuerpo del artículo presenta el trabajo hecho por el grupo Ixa en el campo del procesamiento automático del euskara, identificando sus siete hitos principales y describiendo la estrategia que ha guiado este desarrollo. Se plantea que esta estrategia puede servir como referencia para 190 lenguas que según la clasificación propuesta no poseen recursos de tecnología de la lengua pero si poseen una mínima presencia significativa en Internet.

Laburpena

Euskararen egoeraren inguruan hainbat datu ematen dira labur-labur, eta horrekin batera munduko hizkuntzak sailkatzeko proposamen bat aurkezten da Interneten eta hizkuntz teknologian duten egoeren araberakoa. Euskararen prozesaketa automatikoan Ixa taldeak izan duen bilakaeraren nondik norakoak zehazten dira gero, hainbat mugarri azpimarratuz eta ibilbide hori jarraitzeko erabili den estrategia deskribatuz. Munduko 190 hizkuntzentzat erreferentzia izan daiteke estrategia hori, hain zuen, Interneten presentzia minimo eduki bai baina oraindik hizkuntza-teknologia mota hau landu ez duten hizkuntzentzat

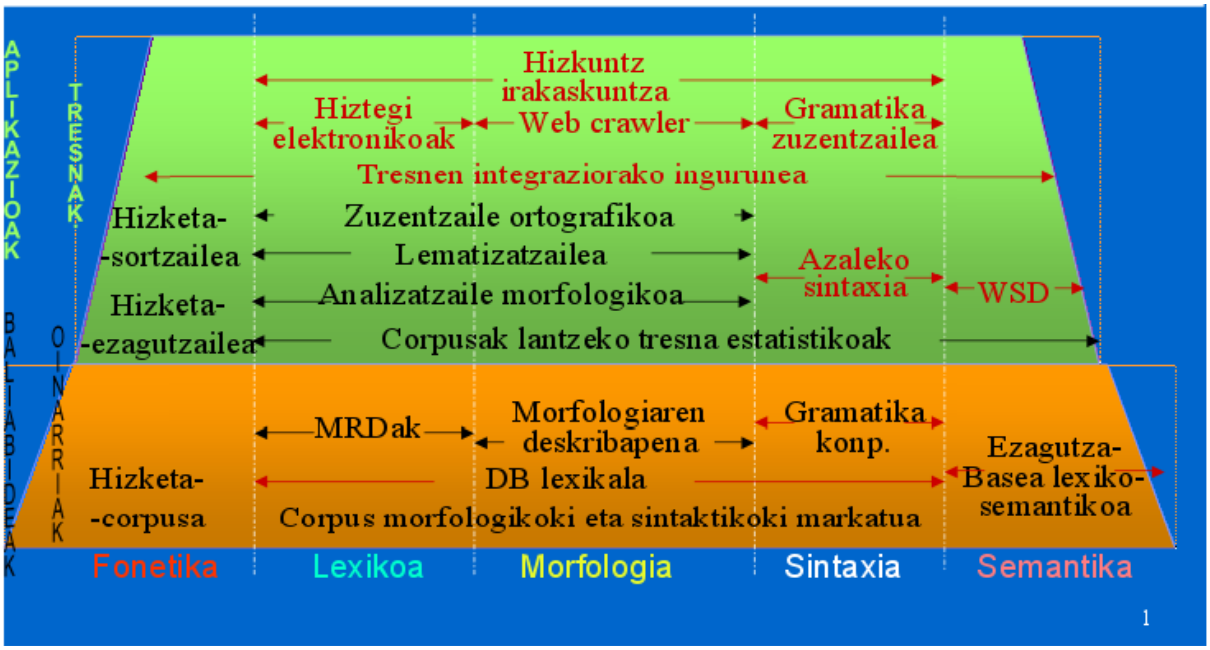
1 Sarrera

1988an Ixa taldea[1] sortu zenean Hizkuntzaren Prozesaketa eta Hizkuntz Ingeniaritza mundu akademikotik kanpo kontzeptu erabat ezezagunak ziren. Hala ere jakintza-arlo horretan oinarritutako hainbat produktu baziren merkatuan, hizkuntza gutxi batzuetarako. Adibidez, ortografia-zuzentzaileak ibiltzen ziren garaiko *MS-Word* eta *WordPerfect* testu-prozesadore ezagunetan, eta lehenengo itzultzaile automatikoak martxan zeuden erakunde handi batzuetan. Informatikaren sorreraren garaitik ametsa izan zena errealitate bihurtzen hasia zen. Edozein kasutan ere, argi zegoen bide luzea geratzen zela egiteko, are luzeago hedadura urriko hizkuntzetarako. Geroago etorriko zen Interneteko zabalkundeak areagotu egin zuen hizkuntza-teknologiaren beharra.

UPV/EHUko Informatika Fakultateko irakasle batzuek egoera horretan aukera ezin hobea ikusi genuen arnas luzeko ikerketa-lerro bat zabaltzeko. Idatzi gabe bazeuden ere, buruan argi izan genituen hainbat funts metodologiko hasiera-hasieratik:

- Euskara izango zen gure ikerkuntzaren zutabeetako bat. Guk heltzen ez bagenion, seguruen urte luzeetan beste inork ez zion helduko. Gainera euskararen ezaugarri linguistiko eta soziolinguistikoek aztergai berezia eta interesgarria eskaintzen zuten zientziaren ikuspuntutik.
- Anbizioa eta nazioarteko erreferentzia. Euskara erreferentzia izateak ez zuen ekarri beharko isolamendurik edo txokokeriarik. Nazioarteko aldizkari eta kongresuak izan behar ziren gure lanerako inspirazioa eta bertan argitaratu nahi genituen gure emaitzak.
- Berrerabilpena. Ikerketa eta aplikazioa uztartu nahi genuen, eta uztarketa horretan arrakasta izateko berrerabilpena funtsezkoa izango zen eman beharreko urrats bakoitzean.

1 http://ixa.si.ehu.es



1. irudia. Euskararako dauden zenbait baliabide linguistiko, tresna eta aplikazioak.

- Lankidetza-giroa eta diziplinarteko ekinbidea. Aurrekoarekin lotuz, ezin genuen aparte utzi Euskal Herrian gure lankide izan zitezkeen ikerlariak eta erakundeak: UZEI, Elhuyar, Euskaltzaindia, EHUko *Euskal Filologia* saila, eta abar luze batekin saiatu behar ginen koordinatzen. Horrez gain informazio eta Komunikazioetako Teknologien (IKT) arloko hainbat enpresa ere bidelagun gisa hartu nahi genituen.

Hau guztia azaleratu zen gure lehen proiektuan. Itzulpen automatikoaz egindako azterketa bat izan zen lan hura, eta bere ondorioa hau izan zen: *egunen batean heldu beharreko eginkizuna izango zen itzulpen automatikoa, baina hori baino lehen oinarri sendoak behar ziren, artean euskararako oinarrizko baliabiderik gabe itzulpen automatikoaren proiektua ariketa akademiko hutsean geratuko baitzen asmoa*. Eta hortik irten zen gure estrategia, 1. irudian azaltzen den piramideko produktuak urratsez-urrats eraikitzeko bidea markatu diguna (Aduriz et al., 1998). Produktu bakoitza, produktu berrien garapenean ahalik eta modu zabalenean berrerabiltzea izan da helburua.Artikulu honetan 23 urtetan zehar Ixa taldean egin den lanaren laburpena egiten da, antzeko egoera soziolinguistikoan dauden hainbat hizkuntzatarako baliagarria izango delakoan.

2. atalean euskararen egoeraren inguruan hainbat datu emango dira labur-labur, eta horrekin batera Internet eta hizkuntz teknologien arloan hizkuntzek duten egoera desberdinak tipifikatzen saio bat egingo dugu.

3. atalean sarrera honetan perfilatutako estrategiaz eztabaidatuko dugu, gure ideiez eta bideaz gain azken urteetan arlo honetan nazioartekoan egin diren ekarpenak hona ekarriz.

4. kapituluan taldearen bilakaeraren nondik norakoak zehazten dira, hainbat mugarri azpimarratuz. Bukatzeko, baliabide urriko hizkuntza baten ikuspuntutik teknologia garatzeko estrategiari buruz hainbat ondorio azpimarratu nahi izan ditugu.

2 Euskararen egoera soziolinguistikoa eta bere posizioa teknologia linguistikoan

2.1 Euskararen egoera

Mendeetako erregresio-prozesu batean sartuta ibili da euskara. Amorrorturen (2002) arabera horren arrazoi nagusiak hauek izan dira: batetik hizkuntza ofiziala ez izatea, eta bestetik hezkuntza sistematik kanpo, komunikabideetatik kanpo, eta industri guneetatik kanpo egotea. Gainera hainbat euskalki diferente izateak ez zuen laguntzen euskara idatziaren zabalkuntzan.

Azken hamarkadetan, baina, urrats kualitatibo oso esanguratsuak egin ditugu egoera horri buelta emateko. Berpizkunde moduko hori honako urratsetan nabaritzen da:

- Euskara hizkuntza koofiziala da Hegoaldean (Nafarroa osoan ez baina).
- Hizkuntza-sisteman txertatua izan da Hegoaldean eta Nafarroako lurralde mistoan.
- Euskarazko komunikabideak daude (EITB telebista, Berria egunkaria...)
- Euskara estandarraren oinarria definitu zuen Euskaltzaindiak 1966an. Morfologia guztiz definituta dago orain, baina lexikoa oraindik ez. Euskara batua da gaur egun irakaskuntzan eta komunikabideetan erabiltzen dena.
- Egun 700.000 hiztun ditu euskarak, biztanleagoaren %25 gutxi gora-behera.

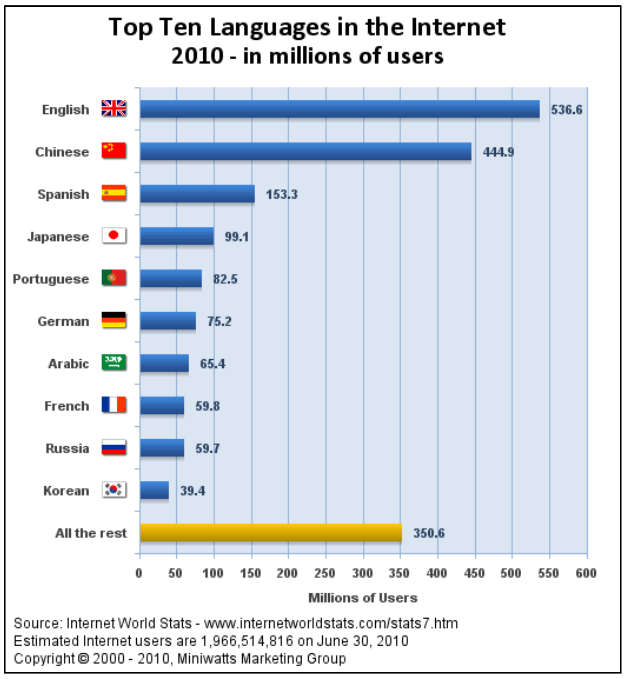
Baina ahalegin guzti horiek eginda ere euskararen etorkizuna oraindik ez dago ziurtatuta. Aipatu urrats horiek guztiz orokorrak ez izateaz gain, euskarak industri guneetatik kanpo jarraitzen du hein handi batean, baita Informazioa eta Komunikazioaren Teknologia (IKT) berriekin lotuta dauden industri guneetatik ere.

Ixa taldea hasieratik saiatu da Hizkuntzaren Teknologiaren arloa ikertzen eta produktuak gizarteratzen, beti ere IKT arloko euskararen erabileran normalizazioa sustatzeko. Adibidez, testuak errazago eta txukunago sortu ahal izateko, edo sarean edukiak zabaldu edo bilatu nahi dituenak tresna egokiak izan ditzan. Helburu horretan ere euskaldunak bere hizkuntzarekin errazago lan egin dezan eta norberaren hizkuntzarekin ere gozatu ahal izateko. Bide horretan gogoeta ugari egin ditugu taldean, gure indar mugatuei ahalik eta probetxu handiena atera ahal izateko. Gogoeta horrek ekarri zuen estrategia bat definitzea eta gero hainbat urtetan horri jarraitzea.

2.2 Hizkuntzen sailkapena tresnak eta baliabideen arabera

Hizkuntza automatikoki lantzeko tresnak errealitatea dira gaur egun, hizkuntz teknologia edo HLT (Human Language Technology) izenarekin ezagutzen den arloaren barruak sortu dira. Gaur egun badira testua edo hizketa lantzeko zenbait aplikazio eskuragarri arlo honetan, hala nola, ortografia-zuzentzaileak, estilo-zuzentzaileak, hiztegi-kontsultak on-line, itzulpen automatikoa eta itzulpen-laguntzak, hizketa testua bihurtzen duten sistemak, testuak irakurtzen dutenak, bigarren hizkuntza ikasteko sistemak, aplikazio informatikoak, gure hizkuntzan erabiltzeko interfazeak, galderetarako erantzunak bilatzeko sistemak (*Question Answering*), dokumentu-bilatzaileak (IR, *Information Retrieval*), informazio-erauzketa dokumentuetatik (IE, *Information Extraction*).

Baliabide urriko hizkuntzen artean ikusten dugu guk euskara (*less-resourced language* termino egokiena iruditzen zaigu[2]). Baina hori erlatiboa da, askoz baliabide urriagoak dituzten beste hizkuntza batzuekin konparatuta baten batek zalantzan jar dezake hori.



2. irudia. Interneteko erabileran hamar hizkuntza nagusiak (*Internet World Stats, 2010)*

Datu estatistikoak eskuratu nahi izan ditugu hizkuntzen arteko sailkapen bat zirriborratze aldera. Honakoak aurkitu ditugu IKT baliabideei buruz:

- *Internet World Stats*[3] webgunean Interneteko erabiltzaileen datuak jasotzen dira. 2010an bertan azaltzen ziren lehen 10 hizkuntzak hauek dira: ingelesa, txinera, espainiera, japoniera, portugesa, alemana, arabiera, frantsesa, errusiera eta koreera. Zoritxarrez ezin da datu zehatz gehiago jaso beste hizkuntzei buruz, baina webgune

2 Honetaz eztabaidatzen da artikulu honetan: Forcada, 2006

3 http://www.internetworldstats.com/stats7.htm

horrek ziurtatzen du gainontzeko hizkuntza guztien artean Interneteko %17,8a baino ez dutela osatzen.

- Dokumentu kopuruari dagokionez datu fidagarriak ez dira lortzen errazak. Hizkuntza erromantzeek Interneten duten hedadurari buruzko 2007ko azterketa batean[4] hauek dira lortutako datuak: dokumentuen %45 ingelesez dago, %5,9 alemanez, %3,80 espainieraz, %4,41 frantsesez, %2,66 italieraz, %1,39 portugesez, %0,28 errumanieraz eta %0,14 katalanez.
- Wikipediaren datuak[5] zehatzagoak dira. 2011ko ekaineko datuak hartuta, 281 hizkuntzetan daude artikuluak. Artikulu kopuruaren arabera lehen hamar hizkuntzak hauek dira: ingelesa, alemana, frantsesa, poloniera, japoniera, italiera, holandesa, espainiera, portugesa eta errusiera. Aurreko zerrendarekin konparatuta txinera, arabiera eta koreera desagertu dira. Beste hizkuntza iberikoei dagokienez, katalana 13. postuan agertzen da, euskara 37.ean eta galegoa 41.ean.

Hiru datu-iturri horien artean azkena da esangurasuena baina tamalez Interneten jokaera aktiboa duten hiztunen emaitza baino ez du eskaintzen.

Bestalde, IKT orokorreko datuetatik hizkuntz teknologia arloko datuetara salto eginez, hainbat webgune interesgarri kontsulta daitezke hizkuntzen egoera aztertzeko:

- **ELRA**:*European Language Resources Association*[6]. Batez ere Europakoak diren hizkuntza-baliabideak biltzen ditu (corpus eta lexikoiak). 60 hizkuntza baino gehiagoko baliabideak biltzen ditu, horien artean sei dira euskararako.
- **LDC**: *Linguistic Data Consortium*[7]. Aurrekoaren parekoa da, baina Amerikako Estatu Batuetako produktuetan espezializatua. 68 hizkuntzatako 450 bat baliabide katalogatu dira bertan, baina euskararakorik ez da agertzen.
- **ACLWiki**[8]: Hizkuntzalaritza konputazionalerako elkarteko wikia da (ACL, *Association for Computational Linguistics*). 58 hizkuntzatan dauden baliabideen berri jasotzeko gunea da. Euskararako 15 produkturen berri jaso du.
- **NLSR**: *Natural Language Soft Registry*[9]. DFKIk kudeatzen duen datu-base honetan 30 hizkuntza agertzen dira, euskararako hiru baliabide daude zerrendan, eta edozein hizkuntzatarako 59.
- ***yourdictionary.com:***[10]Hiztegi-kontsultak on-line eta itzulpen automatikoko doako zerbitzuak eskaintzen dira bertan. 307 hizkuntzarako zerbitzuak daude hor. Argi dago, baina, munduko hiztegi-zerbitzu guztiak ez daudela bertan, euskararako dagoena aztertzea aski da hori egiaztatzeko: daudenak ez dira hamarrera ailegatzen eta www.hiztegia.net gunean 50 baino gehiago bildu baitituzte. Dena dela webgune hori erreferentzia egokia izan daiteke hizkuntzen artean baliabide lexikalen azterketa konparatiboak egiteko.
- Itzulpen automatikoko sistemak eta sareko hainbat zerbitzuren berri biltzen dira gune hauetan: Translation Directory[11] eta Traduzione e computer[12]

Corpus linguistics around the world (Wilson et al., 2006) liburua ere erreferentzia interesgarria da teknologian hizkuntzek duten presentzia erlatiboa neurtzeko.

Beste adierazle interesgarri bat programen lokalizazioa da eta are gehiago oinarrizko *plug-in* linguistikoena.

- Testu-prozesadorerik hedatuena 85 hizkuntza-dialektotan dago lokalizatuta[13]. *OpenOffice*, berriz, 159tan[14] gutxienez. Euskara bietan dago.
- Bilatzaile ezagunenaren[15] interfazea 145 hizkuntzatan dago baina bilaketa aurreratuan 46 hizkuntza baino ez du bereizten du. Euskara ez da agertzen azken aukerarako.
- Itzulpen automatikoko tresna erabilienetan, BabelFish[16] eta Google[17], mugak dira hizkuntzen aldetik. Google-k ia 60 hizkuntza eskaintzen du, eta euskara alfa moduan agertzen da.

4 http://dtil.unilat.org/LI/2007/ro/resultados_ro.htm

5 http://meta.wikimedia.org/wiki/List_of_Wikipedias

6 http://www.elra.info/

7 http://www.ldc.upenn.edu/Catalog/catalogSearch.jsp

8 http://aclweb.org/aclwiki/index.php?title=List_of_resources_by_language

9 http://registry.dfki.de/

10 http://www.yourdictionary.com/languages.html

11 http://www.translation-directory.com/machine.html

12 http://www.federicozanettin.net/sslmit/cattools.htm#publications

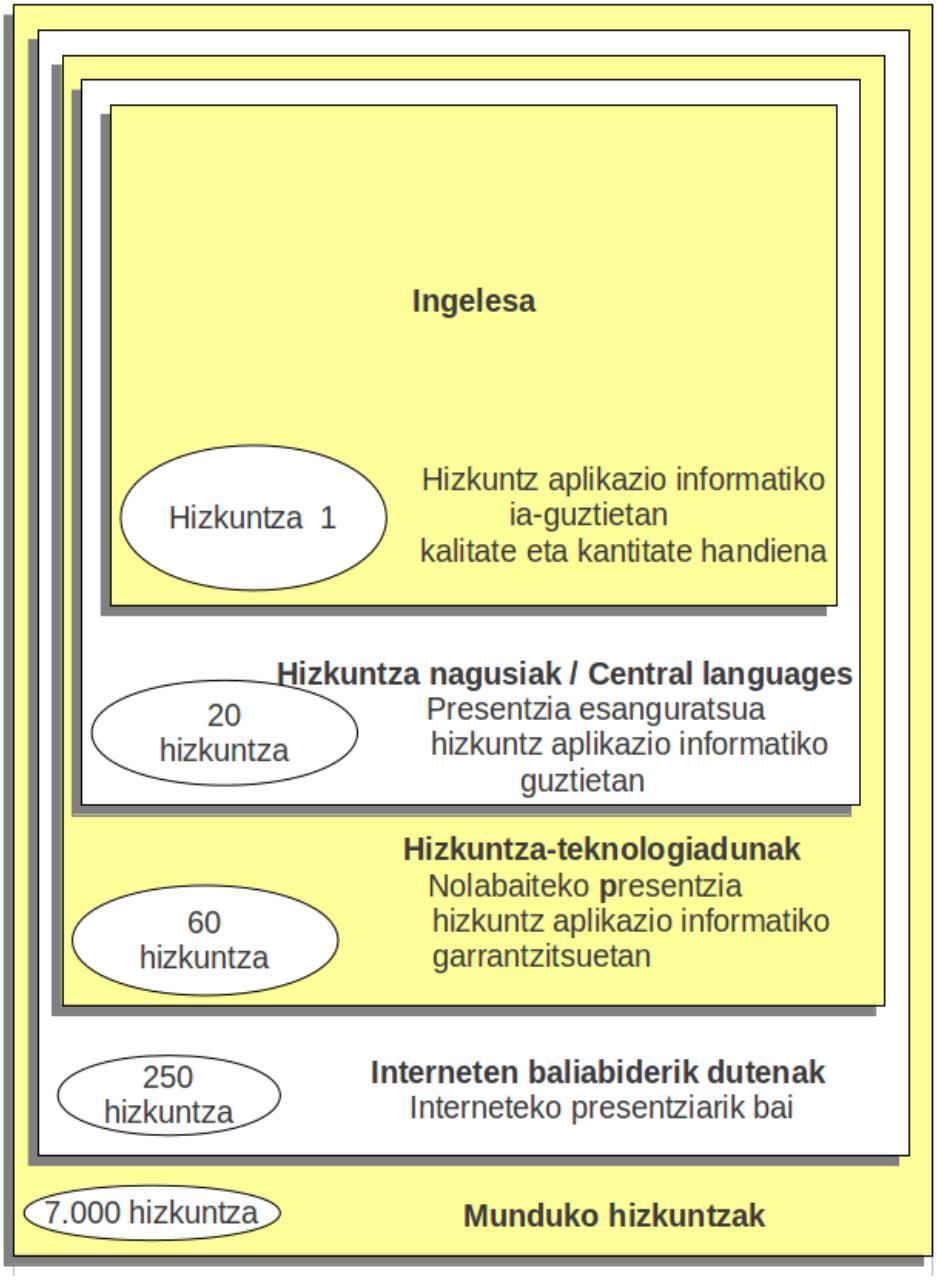
13 http://www.microsoft.com/unlimitedpotential/programs/llp.mspx

14 http://blogs.sun.com/GullFOSS/entry/locale_data_for_159_locales

15 http://www.google.com

16 http://babelfish.yahoo.com

17 http://translate.google.com



3. irudia. Hizkuntzen sailkapena hizkuntza-teknologian eta Interneten duten presentziaren arabera.

Aurreko datuak eta webguneak aztertuta ondoko sailkapena egitera ausartzen gara, beti ere hizkuntz teknologiari begira:

- Lehen maila: **Ingelesa**. Aipatutako zerrenda gehienetan %50a edo gehiago hartzen du. Berarekin alderatuta gainontzeko hizkuntza guztiak baliabide urrikoak dira.
- Bigarren maila: **10-20 bat hizkuntza**, aipatutako zerrendetan goialdean daudenak. Streiter et al. (2006) "central languages" deritze.
- Hirugarren maila: **60 bat hizkuntza gehiago**, HLT baliabideak dituztenak. Kopuru horien bueltan dabiltza ELRA, LDC eta ACLWiki.
- Laugarren maila: **250 hizkuntza inguru**, on-line baliabideren bat dutenak. Wikipedia ingurua horretan dabil.

Sailkapen hau ikusita, atera dezakegun ondorio begi-bistakoena hau da: hizkuntza asko geratzen direla sailkapen honetatik kanpo Interneten presentziarik ez dutelako. Aipatu izaten da

munduan 7.000 inguru hizkuntza daudela, eta horietatik 250 bat baino ez direla "elektronikoki alfabetatuta" daudenak. Ikus 3. irudia.

Kontuan hartu behar da, bestetik, sailkapena ez dela guztiz zehatza. Adibidez, katalana hirugarren mailan koka genezake baina aipatu adierazleren baten arabera (Wikipediako artikulu kopuruaren arabera) bigarren mailako *hizkuntza nagusi* moduan ere ikus daiteke. Euskara berriz, hirugarren mailan dago sailkapen honen arabera, hiztun kopuruaren arabera dagokiona baino gorago; hori horrela da, Ixa taldeak eta beste eragile batzuek alor honetan egindako lanari esker. Dena den bi hizkuntza horietan egindako lanak aurkezten dira SALTMIL workshop-era[18] (*HLT for minority language*s).

Oso baliagarria litzateke honen inguruan behatoki bat egitea.

3 Hizkuntza-Teknologiak lantzeko estrategiak

3.1 Estrategia taldearen esperientzian oinarrituta

Aurreko atalean ikusi dugunez argi dago ingelesa dela nagusia teknologia berri hauetan. Ingelesa batez ere, baina beste hizkuntza nagusiek ere, bigarren maila batean, hainbat produktu eta baliabide garatu dituzte. Argi dago beste hizkuntzek ahalegin handia egin behar dutela atzean ez gelditzeko, are gehiago euskara bezalako hizkuntza txikiek (Petek, 2000; Williams et al., 2001). Zer egin daiteke atzean ez geratzeko? Nola ekin erronka honi? Ixa taldean urteetan jarraitu izan dugu estrategia bat, urrats-kate bat hizkuntzaren teknologiari metodologia batekin ekiteko.

Sarreran esan dugun bezala orain dela 23 urte euskara lantzeko gure lehenengo proiektua itzulpen-sistema bat sortzearen ingurukoa izan zen, bere bideragarritasuna aztertzekoa. Orduan lau irakasle baino ez ginen eta konturatu ginen gure orduko indarrekin askoz zentzuzkoagoa zela eta ez itzulpen-sistema oso mugatu bat egitea, jostailuzko lexikoia eta gramatikak izango zituena. Euskararen morfologia lantzen hasi ginen konturatu ginen zein diferentea zen euskara inguru erdaretatik, baita konturatu ere, beste hizkuntzetarako produktuak gurera egokitzerakoan arazo larriak aurkituko genituela beti. Gauzak horrela, ondorioztatu genuen hoberena zela lehenbailehen sakonki ekitea morfologiaren azterketari. Beraz, itzulpen automatikoaren inguruko kontuak gerorako utzi eta lexikoa eta morfologiari ekin genien modu sakonean, eta tresna horiek geroago itzulpen automatikorako erabili ahal izango ziren, baita beste hainbat aplikaziotarako ere. Geroago etorri ziren morfologiaren gaineko aplikazio informatikoak, gorago aipatu ditugunak, geroago etorri ziren ere beste tresna eta aplikazio konplexuagoak.

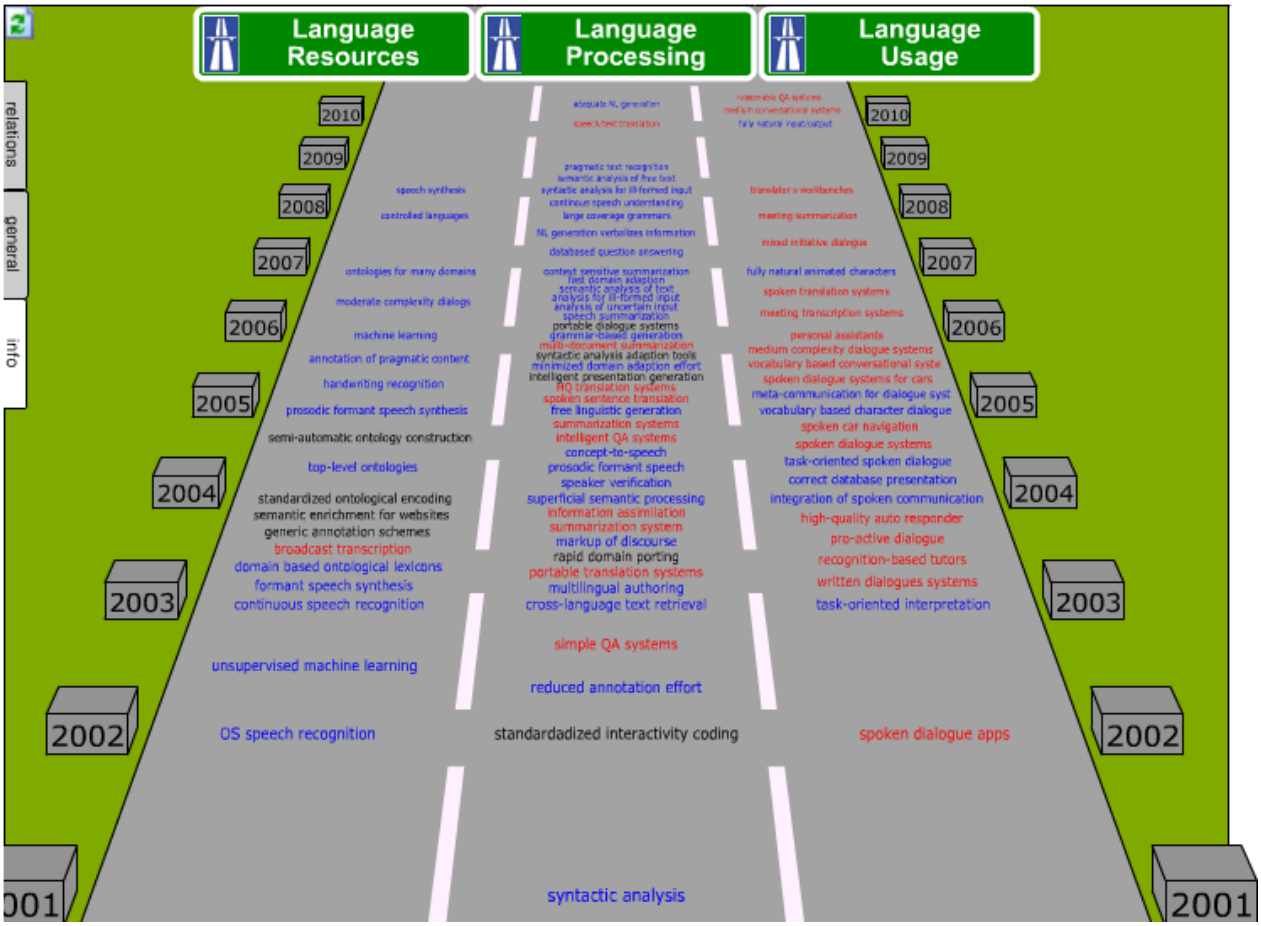
Taldearen 23 urteko ibilbidea estrategia horren arabera egin dugu. Nazioarteko forotan ere aurkeztu eta kontrastatu dugu beste ikerlari batzuekin (Sarasola, 2007). Ideia nagusiak ondokoak dira:

- **Hasieran oinarrizko baliabide eta tresna sendoak sortu behar dira**, eta geroago sortu merkatu-aplikazioak. Alderantziz ez dela egin behar. Produktuen artean bereizi izan dugu zein diren hizkuntza-baliabideak, zein tresna, eta zein aplikazioa. Tresna eta aplikazioak bereizten ditugu, biak produktu informatikoak izan arren, tresnak ez baitira erabiltzaile arruntarentzat eta aplikazioak bai; tresnak hizkuntza-teknologiako teknikariek erabil ditzaten definitu dira. Baliabide, tresna edo aplikazio bakoitza noiz egin behar den aurreikusi izan dugu, ekoizpen prozesu hori optimizatu nahian.
- Horrez gain 1. irudian islatzen den **lexiko-morfologia-sintaxia-semantika progresioa aplikatu behar da,** hein handi batean behintzat.
- **Formatu estandarrak erabili behar dira**. Produktu bakoitza geroagoko produktu berrien garapenean ahalik eta modu zabalenean berrerabilia izatea da gure helburua. Sortzen diren produktuak formatu estandarren arabera[19] definitu behar ditugu, bai hartuko dituzten datuetan, bai itzuliko dituzten emaitzetan. Horrela berrerabilgarriak izango dira beste hainbat produktutan eta haien garapena modu inkrementalean egin ahal izango da.
- **Ahal den guztietan saiatu behar da software librea erabiltzen eta sortzen.** Berrerabili ahal izateko, noski, oso bide interesgarria da produktuak software libre moduan plazaratzea.

Badakigu puntu horiek "oso sinpleak" diruditela, informatikako edozein aplikazio garatzeko erabili behar direnak direla, baina gure eskarmentuak dio hainbat hizkuntzatarako proiektutan ez dela horrela jokatu.

18 http://ixa2.si.ehu.es/saltmil/

19 XML and TEI dira estandar egokienak.



4. irudia. Baliabideen garapenari eta ebaluazioari begirako errepide-mapa (Busemann & Uszkoreit, 2004).

Egun euskarak hizkuntza-teknologiako lehenengo 60-80 hizkuntzen artean baldin badago, neurri handi batean estrategia horri jarraitu izan zaiolako dela uste dugu. Estrategiaren erabileraren adibide gisa esan dezakegu itzulpen automatikoan denbora laburrean garatu ahal izan bada lehenengo prototipoa, hainbat tresna berrerabili direlako izan dela, bai taldean aurretik sortutakoak (baliabide lexikalak eta morfologia), baita software libreko beste batzuk ere (erdaretarako analizatzaileak), modulu berri bakar batzuk soilik sortu behar izan ditugu eta horiek beste talde batzuekin garatu ditugu elkarlanean.

3.2 Nazioarteko erreferentziak

Atal honetan hizkuntz teknologien garapenaren inguruan argitaratu diren hainbat hausnarketaren berri eman nahi dugu. Kasu batzuetan baliabide urriko hizkuntzekin ere lotuta daude.

Hizkuntza bat gaurko behar teknologikoetan egokikuta egoteko estrategia edo jarraibide bat baino gehiago proposatu izan da nazioartean.

Gure taldean hasierako proposamen bat egin zen hizkuntza teknologiarako oinarrizko tresneria defintzeko (Agirre et al., 2002), *Basic toolkit for HLT* deitu genuena.

Krauwer-ek (2003) BLARK kontzeptua (*Basic LAnguage Resource Kit*) proposatu zuen, arrakasta lortuz, aurretik ELSNET (European Network of Excellence in Language and Speech) eta ELRA (European Language Resources Association) elkarteek proposatutakoari helduz. Ildo honetatik Maegaard-en taldeak (2004) BLARK bat proposatzen dute arabierarako eta Simov-ek eta lankideek (2004) beste bat bulgarierarako. Geroztik BLARK terminoa maiz erabili izan da literatura zientifikoan.

Streiter-ek beste batzuen lankidetzarekin 2006an argitaratutako lan batean (Streiter et al., 2006) hizkuntza ez-zentraletarako hainbat HLT

proiekturen berri ematen da. Testuinguru horretan, hainbat gomendio ematen dira erakundeek hizkuntza teknologia garatzeko banatzen dituzten laguntzak bideratzeko. *Non-central* terminoaren erabilerarekin batera azpimarratzekoa da software librearen alde egiten duen hautua. Forcada-k (2006) ere azpimarratzen du kode irekia erabiltzearen egokitasuna hizkuntza hauetarako itzulpen-sistemak garatzean.

ELSNET sareak, 2004an, baliabideen garapenari eta ebaluazioari begirako errepide-mapa[20] bat ere eskaini zuen (Busemann & Uszkoreit, 2004). Ikus

Testuinguru honetan mapa anitz argitaratu dira azken urteotan[21]. 2002ko gure proposamenean bezala diagramako elementuak hiru multzotan banatzen dira: *Language Resources / Language Processing / Language Usage* aipatutako mapetan eta *Language resources / Language Tools / Language Applications*. Gure estrategia hizkuntza baten garapen teknologikorako lehen urratsak modu sendoan emateko bideratuta dagoen bitartean ELSNETeko ekarpenean zerrenda askoz zehatzagoa da eta hizkuntza zentraletan jartzen da fokua (Europako hizkuntza ofizialak).

Borin-ek (2006 eta 2009) HLTek eremu urriko hizkuntzei irekitzen dizkieten aukerak aztertzen ditu, informazioaren gizartean hizkuntza aniztasunak duen garrantzia azpimarratuz. Ostler-en aipamen hau ere "a *language will not get by in the world of today unless it is equipped with a parser and a multi-million-word corpus of text*" ekartzen du.

Azken urteetako Europako proiektu berri batzuek, Clarin[22] eta Flarenet[23] esaterako, Europako hizkuntzetarako baliabideen eta tresnen garapena eta ustiapena lankidetzan eta koordinazio handiagoz egitea bultzatu nahi dute.

Eta azkenik aipatzeko, SALTMIL (*Speech And Language Technology for Minority Languages*) elkarteak hainbat batzar[24] antolatzen ditu HLT eta baliabide urriko hizkuntzak uztartuz.

4 Mugarriak eta etapak Ixa taldearen jardunbidean

Ixa taldearen 23 urteko ibilibidea errazago aurkeztearren etapaka banatu dugu denbora tarte hori. Ibilbide horietako mugarri nabarmenenak aukeratu ditugu etapak bereizteko, eta etapa bakoitzaren deskribapenean garatu diren produktu eta landu diren ikerketa-proiektu garrantzitsuenak aipatuko ditugu. Hauek izan dira aukeratu ditugun mugarriak:

- 1993: Morfologia eta Xuxen zuzentzaile ortografikoa
- 1996: Lexikoa eta EDBL datu-base lexikala
- 1999: Lematizatzailea
- 2002: Sintaxia
- 2005: Lexiko-semantika eta EuskalWornet
- 2009: Eleaniztasuna eta Matxin itzultzaile automatikoa
- 2010: Aplikazio aurretatuak: *Ihardetsi* galderak erantzuteko sistema

4.1 1988-1993: Morfologia eta Xuxen zuzentzaile ortografikoa

Esan bezala, hasieran gure lehenengo proiektuaren helburua euskararako itzulpen-sistema bat sortzea izan zen. Baina gure ahaleginei probetxu handiagoa ateratzearren laster itzulpen-kontuak geroago egiteko utzi eta momentuan morfologiari ekin genion modu sakonean. Ahalegin horretatik sortu zen urte gutxiren buruan zuzentzaile ortografikoa.

	1988-1993
Proiektuak Gipuzkoan (GFA)	Itzulpena Xuxen
Produktuak Morfologia	**Xuxen**

1. taula: Proiektu eta produktuak, 1988-1993.

Euskaldunak bere hizkuntzaz idatzi gura duenean zalantza ugari aurkitzen ditu. Batetik, eskolak toki guztietan oraindik idazteko gaitasuna bermatzen ez duelako, edo bestetik, belaunaldi zaharragoek euskaraz ikasteko aukerarik izan ez dutelako, sarritan euskaldunak badaki esaten hitz bat baina ez daki nola idatzi behar den batuaz. Esate baterako, hau izan daiteke duda bat idazterakoan: “Ondoko hitzen artean zein da batuaz erabili behar dudana *arbola* adierazteko*? Zuhaitz? Zuhatz? Zugaitz? Zugatz? Zuhaits? Sugatx?”*. Bestetik, euskara estandarraren definizioa berri samarra denez (beste inguruko erdaren estandarrekin konparatuta),

20 http://elsnet.dfki.de/roadmap.php

21 http://elsnet.dfki.de/roadmap.phpversion=LREC_2004

22 http://www.clarin.eu/

23 http://www.flarenet.eu

24 http://ixa2.si.ehu.es/saltmil/eu/activities/workshops/workshops.html

lexikoaren estandarizazioa oraindik bukatzear dagoenez, eta batzuetan estandarizazioan aldaketak gertatzen direnez (esate baterako, hasieran *eritzi*, eta *iharduera* hitzak erabili behar zirenak gaur egun *iritzi* eta *jarduera* idatzi behar dira) beste hainbat duda sortzen dira.

Horrelakoetan XUXEN zuzentzaile ortografikoak (Aduriz et al., 1997) laguntza paregabea eskaintzen dio erabiltzaileari testuaren kalitatea hobetzeko eta forma estandarrekin ohitzen joateko apurka-apurka. Horrela, esan dezakegu euskararen estandarizazio-prozesuaren aliatu indartsua dela XUXEN programa. Eta gainera dohainik jaitsi daiteke www.euskara.euskadi.net webgunetik. Bere erabilera orokortuz doala erakusteko esan daiteke gune horretatik 20.000 erabiltzailek jaso duela honezkero. Gainera azken urteetan atera diren egokitzapen berriei esker XUXEN eskuragarriago dago. Lehen Word editorearekin bakarrik erabil zitekeen, orain erraz jar dezakegu martxan Mozilla Thunderbird-ekin, nabigatzailearekin Interneten bidez edozein mezu edo inprimaki betetzen ari garenean, edo Openoffice-ekin. Posible da beste edozein aplikaziorekin ere testua zuzentzeko www.xuxen.com zerbitzarira jotzen badugu.

Espainiera, frantsesa edo ingeleserako zuzentzaileak baino dezente konplexuagoa da XUXEN, hitz posibleak askoz gehiago direlako, eta ondorioz, hitzen analisi morfologikoa egin behar delako.

Xuxen-en inplementazioa hasieran programa propio bat izan zen. Geroago exekuzio azkarragoa lortzearren Xeroxeko tresnetara egokitu zen[25], eta zken urteetan software librera jauzi ahal izateko *hunspell* eta *foma* tresnak erabili izan dira (Alegria et al., 2009).

Oraindik lexiko eta morfologiako erroreak baino ez ditu harrapatzen, baina hitz maila horretan oso praktikoa da. Sintaxiko edo estiloko zenbait errore harrapatzeko ikerketak egiten ari gara orain, eta lehen bertsio bat integratu da XuxenIV bertsioan.

Xuxen programaren erabilera guztiz hedatuta dago gaur egun, datu hauetan ikus daitekeenez:

- 1998z geroztik Microsoft Officeko banaketa ofizial guztiek barruan daukate.
- www.euskara.euskadi.net webgunetik egin diren deskargak 20.000 baino gehiago izan dira.
- Firefoxerako deskargak 120.000 baino gehiago 2007-2011 tartean.
- OpenOffice-rako deskargak 7.000 baino gehiago izan ziren 2010. urtean.

25 www.stanford.edu/~laurik/fsmbook

4.2 1993-1996: Lexikoa eta EDBL datu-base lexikala

Xuxen zuzentzaile ortografikoaren garapenean erabili ziren lexikoa eta analizatzaile morfologikoa ondo antolatu ziren geroago bere mantentze-lanak errazteko eta beste aplikazioetan erabili ahal izateko.

	1993-1996
Proiektuak Eusko Jaurlaritzan	Xuxen
Proiektuak Gipuzkoan (GFA)	HAIN
Produktuak Lexikoa	**EDBL**
Produktuak Morfologia	Xuxen... Morfeus

2. taula. Proiektu eta produktuak, 1993-1996.

Hizkuntzaren lexikoaren biltegi orokorra da datu-base lexikala. Hiztegi elektroniko moduko bat da, hizkuntzaren tratamendu automatikoari begira eraikia, eta, beraz, hizkuntzaren tratamendua automatizatu nahiak dituen eskakizunak kontuan harturik antolatua. Horrek lexiko-deskribapenaren sistematizazio bat eskatzen du: sarreren kategoria-sistema bateratu eta homogeneoa, kategoria bakoitzeko elementuak behar den bezala deskribatzeko beharrezko diren ezaugarriak zehaztea, etab. *EDBLk* (euskararen datu-base lexikala) lehen bertsioan 60.000 sarrera zituen eta 147.700 inguru biltzen ditu egun —120.000 hiztegi-sarrera, 20.000 adizki eta 700 morfema ez-independente—, eta Ixa taldea arduratzen da egunean mantentzeaz (Aldezabal et al., 2001). Internet bidez kontsulta daiteke[26]. Hasieran zuzentzailearen oinarri lexikal gisa pentsatu bazen ere, gaur egun oinarri lexiko orokorra da eta hainbat tresna elikatzen ditu: analizatzaile morfologikoa, lematizatzailea, hitz anitzeko espresioen errekonizatzailea, entitate-espresioen errekonizatzailea, etab.

Informazio ez-estandarra ere gehitu zaio morfemetan eta hitzetan, hala nola forma dialektalak, errore tipikoak etab., beti ere dagozkien erabilpen estandar eta zuzenarekin lotuta. Gainera,

26 http://ixa2.si.ehu.es:7777/forms/frmservlet?config=lbdbl
Kontsulta sinplifikatua: http://ixa2.si.ehu.es/edbl/

hedapen dialektala eraman da aurrera, aldaerak integratuz, eta horixe izan da abiapuntua beste produktu berri batzuk sortzeko, esate baterako, Xuxen-B[27] bizkaieraren zuzentzaile ortografikoan, eta batua-bizkaiera bihurtzaile automatikoan[28].

Informazio sintaktikoa eta semantikoa (azpikategorizazioa, atributu semantikoak, etab.) barneratzeko ere diseinatuta dago, eta kasu askotan informazio hori osatuta dago.

Azken bi urteetan Euskaltzaindiak lideratutako proiektu baten barruan dago taldea (Lexikoaren Behatokia1) eta UZEI eta Elhuyar erakundeekin elkarlanean datu-basea aberastu egin da.

Aurreko baliabideetan oinarrituta analizatzaile morfologikoa eraiki genuen. Analizatzaile morfologikoa hizkuntza guztietan beharrezkoa izanda euskara bezalako hizkuntza eranskarien kasuan ezinbestekoa gertatzen da. Analizatzaile (eta sintetizatzaile) morfologikoaren zeregina hitz-forma osatzen duten morfemak ezagutzea (eta konposatzea) da, eta morfema bakoitzari dagokion informazio morfologiko-lexikala ematea. Erreminta hau oinarri da hainbat aplikaziotan, hala nola, zuzentzaile ortografikoan, karaktere-ezagutze optikoan (OCR), eta aplikazio sofistikatuago guztietan —itzulpen automatikoa, adib.—. Interneten erabil daiteke demo[29] bat.

4.3 1996-1999: Lematizatzailea

Lematizatzaile/etiketatzailea analizatzaile morfologikotik eratortzen da, eta hitz-forma baten lema eta kategoria ematen ditu, anbiguotasuna saihestu edo gutxitzearren testuingurua aintzat hartuz (Ezeiza et al., 1998). Garaian berritasun handia izan zen desabiguaziorako teknika estatistikoekin batera murriztapen-gramatikaren formalismoa erabiltzea (Constraint Grammar ingelesez), sistema konbinatua garatuz. Zeregin nagusia desanbiguazioa bada ere, beste egitekorik ere badu halako tresna batek, esate baterako, hitz anitzeko unitate lexikalen identifikazioa (lokuzioak, hitz-elkarketak, pertsonaizenak, etab.). Oso aplikazio interesgarriak dituzte lematizatzaileek, esate baterako, dokumentubilatzaileak, informazio-eskurapena, terminologia, lexikografia, etab.

	1996-1999
Proiektuak Europan	
Proiektuak Madrilen (MEC)	Item
Proiektuak Jaurlaritzan	Xuxen, EDBL, Item, Lematizatzailea
Proiektuak Gipuzkoan	Xuxen Idazkide
Produktuak.Apl. orokorra	Multimeteo
Produktuak Semantika	
Produktuak Sintaxia	
Produktuak Lexikoa	EDBL
Produktuak Morfologia	Xuxen Eustagger

3. taula: Proiektu eta produktuak, 1996-1999.

Geroko hainbat proiekturen atea zabaldu zuen programa honek, adibidez Espainiako zenbait talderekin lankidetzan burutu ziren Item eta Hermes proiektuak. Demo[30] bat erabil daiteke.

4.4 1999-2002: Sintaxia

Morfologia eta lematizazioa bideratuta hurrengo urratsa perpaus sinpleak sintaktikoki analizatzeko tresna izan zen. Baterakuntza gramatikako erregelatan oinarrituta zegoen Patr-Ixa (Aldezabal et al., 2003). Hitzen barruko analisi morfosintaktikoa ere egiten zuen.

Geroago sintaxirako beste tresna landuago batzuk garatu dira perpaus konplexuetatik ere informazio sintaktikoa hobeto atera ahal izateko:

- *Zatiak* (edo *Ixati* ere deitua) azaleko analizatzaile sintaktikoa. Esaldiko sintagmak edo chunkak bereizten dituena.
- *EDGK:* Dependentzia Gramatika. Esaldiko buruak (aditzak izaten dira normalean) mendeko elementuekin dituzten erlazioak markatzen dira dependentziazko gramatiketan.
- *Maltixa*[31]*:* Analizatzaile sintaktiko estatistikoa

27 http://www.azkuefundazioa.org/lan-tresnak/xuxen-bizkaieraz

28 http://www.eleka.net/berriak/berria.php?id=eu&a=1&b=1303197461

29 http://ixa2.si.ehu.es/demo/analisianali.jsp

30 http://ixa2.si.ehu.es/demo/analisimorf.jsp

31 http://sisx04.si.ehu.es:8080/maltixa/index.jsp

- Eihera[32]: Testuetan entitateak ezagutzeko tresna (pertsonak, tokiak, erakundeak).
- EPEC[33] eta *Ancora*[34] corpusak: EPEC zuhaitz sintaktikoen bankua da. Prozedura erdiautomatiko bat erabili zen etiketatzeko. Guztira 50.000 hitz dauzka. Ancora ingurunean espainiera, katalanera eta euskarazko treebank-ak biltzen dira.
- Erreus corpusa[35]: Ikasleen idazketa-erroreen korpusa.

	1999-2002
Proiektuak Europan	
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)	Hermes
Proiektuak Jaurlaritzan	Xuxen, Sintaxi lexikoa, Ixa taldea UZEI sinon-hizt.
Proiektuak Gipuzkoan	Berbasare, Gainternet
Produktuak. Apl. orokorra	
Produktuak Semantika	
Produktuak Sintaxia	Zatiak-Ixati
Produktuak Lexikoa	Elhuyar-Word
Produktuak Morf.	Xuxen Elhuyar-Word

4. taula: Proiektu eta produktuak, 1999-2002.

Urte horietan Elhuyar-Word hiztegi-kontsultarako programa sortu zen. Word testu-prozesadorean plugin gisa integratu zena. Tresna honek euskarazko edo gaztelaniazko edozein hitz hartuta, bere lemari dagozkion itzulpenak eskaintzen dizkio erabiltzaileari; Elhuyar Hiztegi Txikia (Euskara-Gaztelania/Castellano-Vasco) elebidunean agertzen diren itzulpenak hain zuzen. Kontsultatzen den hitzaren lema eta kategoria konbinazio posible guztiak erakutsiko ditu, eta bikote bakoitzari beste hizkuntzan dagozkion ordainak ere. Ildo horretatik, geroago euskara-frantserako bertsioa ere sortu du Elhuyarrek, eta UZEIk Word-erako plugin bat garatu zuen sinonimo-hiztegia erabiltzeko.

Garai hartako Hermes eta Gainternet proiektuetan lematizazioaren erabileraren aukerak aztertzen hasi ziren (Hermes, hemeroteka elektronikoak: bilaketa eleanitza eta erauzketa semantikoa).

4.5 2002-2005: Lexiko-semantika eta EuskalWornet

Hizkuntza ulertzea xede denean, eta morfologia eta sintaxiari buruzko informazioarekin aski ez denean, semantikari buruzkoarekin aberastu behar da programa. Anbiguotasun linguistikoa ebatzi ezina da, askotan, semantikaz baliatu ezean. Hizkuntza baten tratamendurako azpiegituran, osagai semantikoak ere behar du bere lekua, beraz. Eta semantika lexikala da, beharbada, osagai horren prestakuntzan landu beharreko lehenengo alderdia. Semantika lexikalak lexikoko elementuen artean dauden erlazio lexiko-semantikoak biltzen ditu: sinonimia, antonimia, hiperonimia/hiponimia (klase/azpiklase erlazioak), eta beste.

	2002-2005
Proiektuak Europan	Meaning
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)	EuropenTrad, Hiztegia2002, Hizking21, Bilatzailea, RICOTERM2
Proiektuak Jaurlaritzan	Ixa taldea Hizking21-ETORTEK
Proiektuak Gipuzkoan	Hermes
Produktuak Semantika	**EuskalWordnet**
Produktuak Sintaxia	Erreus corpusa
Produktuak Lexikoa	UZEI sinon-hizt
Produktuak Morf.	Xuxen Eihera

5. taula: Proiektu eta produktuak, 2002-2005.

Erlazio lexiko-semantiko horiek sare semantiko moduko batean adierazten dira esplizituki. Ingelesezko sare semantikoen artean ezagunena-edo

32 http://ixa2.si.ehu.es/demo/entitateak.jsp

33 http://ixa.si.ehu.es/Ixa/resources/Treebank

34 http://clic.ub.edu/ancora/

35 http://ixa.si.ehu.es/Erreus

WordNet izenekoa dugu, eta haren euskararako egokitzapenari *Euskal WordNet* edo *EusWN*[36] deitzen diogu (Agirre et al., 2006).

EusWN hori *EuroWordNet*-en markoan garatu zen, euskal hitzak ingelesezko *WordNet*-era metodo erdiautomatikoz lotuz (ezagutzaren eskurapen automatikoa). Ingelesaren gehiegizko eragina saihesteko eta kalitate linguistikoa babesteko euskal *synset*-ak eskuz orraztu ziren geroago.

EusWN garatzeko taldeak egin zituen ahaleginak indartu egin ziren *MEANING* europar proiektuan parte hartzearekin (*MEANING*: Amaraun mailako hizkuntza-teknologia eleanitzen garapena). Lau hizkuntzatarako wordnet diferenteak lotu egin ziren *MCR* egitura lexiko-semantiko eleanitza[37] sortzeko.

Behin hitzen adierak zein diren definituta zegoela, esaldien ulerkuntzari ekin ahal izateko garrantzitsua zen jakitea bereizten esaldi konkretu batean zein den hitz bakoitzerako erabili den adiera. Horregatik Hitz-Adieren Desanbiguatze (HAD) sistema bat[38] garatu zen geroago, *Support Vector Machine* metodo ezagunean oinarrituta. Hasieran ingeleseko corpusen gainean bakarrik aplikatu zen, baina gaur egunean, EuSemCor semantikoki etikatuta dagoen corpusa[39] sortu denez gero, posible izan da euskararako ere entrenatzea.

2002an ELEKA enpresa sortu zen, spin-off moduan, Ixa eta Elhuyar fundazioaren lankidetzaren fruitua da. Aurretik sortutako produktuen kudeaketa informatikoa eta merkatu-garapenaz arduratuko zena, ixakideen jarduera ikerketan hobeto zentratzearren. Aldi berean Elhuyar fundazioak bere I+G atala antolatu zuen.

2002.ean ETORTEK deialdiko *Hizking21* proiektua abiatu zen. Eusko Jaurlaritzako ikerketa estrategikorako proiektu horretan lankidetzan hasi ginen arloko beste ikerketa zentro batzuekin: Elhuyar Fundazioa, EHUko Aholab ikerketa-taldeak eta Vicomtech eta Robotiker teknologia-zentroak. Hizkuntza-, hizketa- eta multimedia-teknologien alorrean ezagutza eta eskarmenturik handiena duen euskal taldea osatzen dugu elkarrekin. Hizking21 proiektuaren jarraipena izan dira gero *Anhitz* (2006-2008) eta *Berbatek* (2010-2012) proiektuak.

Hizking21 proiektuko emaitzen artean *ZT corpusa*[40] azpimarratu behar da. Zientzia eta teknologiaren alorreko euskarazko testu-bilduma egituratu eta etiketatua da, eta alor horietako euskararen erabilera ikertzeko baliabidea izatea du helburu nagusia. Corpus berezi edo espezializatua da, eta Ixa taldeak eta Elhuyar Fundazioak elkarlanean eratu dute. Corpus etiketatua da, bai testuaren egiturari eta formatuari dagokionez, baita linguistikoki ere. Testuko hitz bakoitzaren lema eta kategoria/azpikategoria etiketatu dira. Corpusaren lehen bertsio honetan, 8,5 milioi hitz daude, eta horietatik 1,9 milioi hitz eskuz berrikusi, desanbiguatu eta zuzendu dira.

4.6 2006-2009: Eleaniztasuna eta Matxin itzultzaile automatikoa

2006. urtean Matxin sortu zen, euskararako lehen itzultzaile automatikoa (Alegria et al., 2007). Aurreko etapako *Opentrad* eta *Europentrad* proiektuen barruan garatu zen, estatu espainiarreko lau hizkuntza ofizialen arteko itzulpena landu baitzen proiektu horietan. Lau unibertsitateren eta hainbat enpresaren arteko elkarlanaren emaitza izan ziren proiektu horiek. Parte hartu zuten unibertsitateak ondoko hauek izan ziren: Euskal Herriko Unibertsitateko Ixa taldea, Alacanteko Unibertsitateko Transducens taldea, Vigoko Unibertsitateko Linguistika Informatikoko Mintegia eta Kataluniako Unibertsitate Politeknikoko TALP taldea. Enpresa arduraduna Eleka Ingeniaritza Linguistikoa izan zen, Elhuyar Fundazioaren zein Galiziako Imaxin Software enpresaren laguntzarekin. Alacanten *Prompsit* izeneko enpresa bat sortu zen, proiektuaren emaitzak Herrialde Katalanetan zabaltzeko asmoz. Espainiako Industria, Turismo eta Merkataritza Ministerioaren laguntzaz garatu zen proiektua.

	2006-2009
Proiektuak Europan	Kyoto
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)	Know, OpenMT, IMLT, Praxem, Avivavoz
Proiektuak Jaurlaritzan	Ixa taldea Anhitz-ETORTEK

36 http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl

37 MCR, kontsulta on-line: http://garraf.epsevg.upc.es/cgi-bin/wei4/public/wei.consult.perl

38 HAD/WSD demoa: http://ixa3.si.ehu.es/wsd-demo

39 Eusemcor corpusaren demoa: http://sisx04.si.ehu.es:8080/eusemcor/

40 http://www.ztcorpusa.net/aurkezpena.htm

Proiektuak Gipuzkoan	Remixee, Prest
Produktuak. Aplikazio orokorra	Anhitz **Matxin**
Produktuak Semantika	MCR, WSD-Ixa
Produktuak Sintaxia	Ancora corp.
Produktuak Lexikoa	EDBL
Produktuak Morfologia	ZT corpusa Eulia

6. taula: Proiektu eta produktuak, 2006-2009.

Proiektu horien barruan bi teknologia sortu ziren, bat A*pertium* izenekoa, antz handia duten hizkuntzen artean itzultzeko; eta bestea M*atxin* izenekoa, egitura desberdineko hizkuntzen artean itzultzeko.

Matxin erabilera publikoko programa bihurtzen zen bitartean Itzulpengintza automatikoko ikerketak hurbilketa estatistikoan kokatzen ziren. Horrela OpenMT (2006-2009) proiektuaren barruan EUSMT itzultzaile estatistikoa sortu zen eta lehenengo sistema hibridoak (erregelak eta estatistika batuz) proposatu ziren. Proiektu horren segida den OpenMT2 (2009-2012) proiektu berrian sistema hibridoetan sakondu nahi da, ebaluazio-metodoetan, aurre-edizioan eta postedizio automatikoan.

Etapa honetan, 2006. urtean, europar mailan *Kyoto* proiektua abiatu zen. Aurreko etapako *Meaning* proiektuan sortutako hizkuntza prozesatzaileak hobetu ziren proiektu berri honetan eta domeinu espezifikoetako dokumentuetan kontzeptuak eta gertakizunak automatikoki erauzteko erabili ziren gero. Proiektuaren bukaeran informazioa eta dokumentuak bilatzeko teknikak garatu ziren horrela lortutako ezagutza-baseen gainean, eta, helburu horrekin erabilita, emaitza onak eman ditu hitzen adieren artean desanbiguatzeko balio duen UKB algoritmoak[41]. Espainiako MICINN ministerioak finantzaturiko *KNOW* (2006-2009) proiektu koordinatuan ere antzeko helburuak landu ziren, baita honen jarraipena den *KNOW2* (2009-2012) proiektuan, non eleaniztasunaren ikuspuntua ere lantzen den. Bestalde, MICINN ministerioko *IMLT* proiektuaren barruan sortutako tresna eta baliabide linguistiko horiek guztiak batera integratzeko eredu orokor bat eskaintzeko izan da. XML estandarrean oinarritutako proposamen sendo bat lortu da eta berau inplementatzeko balio duen *LibXml* programa-liburutegia sortu da. Proiektu hauen arteko elkarlana ikus daiteke puntu honetan: *IMLT* proiektuko XML eredu horren sinplifikazio bat onartua izan da Kyoto proiektuan ezagutza adierazpide gisa, *KAF eredu* deritzoguna.

Proiektu horietan alde batetik itzulpen automatikoan eta bestetik informazio-bilaketan lortutako sistemak prototipo batean erabili ahal izan ziren etapa honen bukaeran. ETORTEK deialdiko *AnHitz* proiektuan euskaraz hitz egiten duen 3D *avatar* bat sortu zen prototipo mailan. Zientzia eta teknologian aditua denez gai horien inguruko galderak erantzun ditzake, edo gai horietako termino bilaketa eleanitza egin eta emaitzak automatikoki euskarara itzuli. Ixa taldearen ekarpena batez ere galderak erantzuteko sisteman eta itzulpen automatikoan egon da. Baina prototipoan beste modulu batzuk integratu dira: 3D avatarra (VICOMTech), testu-ahots bihurgailu eleanitza (Aholab), euskarazko ahots-ezagutza (Robotiker, Aholab) termino-bilaketa eleanitza (Elhuyar), zientzia eta teknologiazko corpus eleanitzak (Elhuyar), eta azkenik modulu guztiak integratzeko sistema (Elhuyar).

4.7 2009tik gaurdaino: Aplikazio aurretatuak: *Ihardetsi*, galderak erantzuteko sistema

Anhitz proiektuan euskarazko galderak erantzuteko erabili zen modulua *Ihardetsi* sistema bihurtu da azken garai honetan, alegia, euskaraz egindako galderen erantzun zehatzak testu-bildumatan aurkitzen dituena. Aplikazio konplexu honek ohiko "Question Answering" (QA) sistemen ezaugarriak ditu (Ansa, 2006).

	2009...
Proiektuak Europan	Paths
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)	Know2, OpenMT2, Hybridosint, RTTH, TIMM, Ancora-corpus
Proiektuak Eusko Jaurlaritzan	Ixa taldea Berbatek-ETORTEK
Proiektuak Gipuzkoan	Langune
Produktuak	**Ihardetsi**

41 UKB algoritmoa deskargatu: http://ixa2.si.ehu.es/ukb/

Aplikazio orokorra	BASYQUE EUSMT
Produktuak Semantika	Eusemcor UKB
Produktuak Sintaxia	Maltixa EDGK
Produktuak Lexikoa	Lexkit Dicc. Escolar
Produktuak Morfologia	BertsolariXa LibiXaml

9. taula: Proiektu eta produktuak, 2009. urteaz geroztik.

Paraleloan beste aplikazio aurreratu baten eraikuntzan parte hartuko du taldeak Europa mailako PATHS proiektu berrian, Meaning eta Kyoto proiektuen ondorioa dena. Aurreko proiektuetan sortutako hizkuntza prozesatzaileak erabiliko dira gero domeinu espezifikoetako dokumentuetan kontzeptuak eta gertakizunak automatikoki erauzteko. Eta horrela lortutako ezagutza-baseetan informazioa eta dokumentuak bilatzeko teknikak garatuko dira, Europeana liburutegiaren esparruan.

Europeana Europako eduki digitalen liburutegi erraldoia da. Hainbat museo, liburutegi, agiri eta ikus-entzunezko bildumatarako sarbide irekia da. Bere helburu nagusia Europaren aniztasun kultural eta zientifikoaren zabalkundea erraztea da. Liburutegiak 15 milioi ale biltzen ditu hainbat formatutan (irudia, testua, audioa eta bideoa).

Interneteko liburutegi digitalei esker kultur-ondare diren material ugari daude eskuragarri gaur egun. Hala ere, kopuru erraldoi horiek nahasgarriak ere izan daitezke erabiltzaile arruntarentzat, zailtasunak izan baititzake aurkitutako informazio guztia interpretatzen. PATHS proiektuak pertsonalizatutako bisita gidatu interaktiboak eskaini nahi ditu, eta horri esker erabiltzaile arruntak ere eroso mugituko dira liburutegi digital horien barruan. Erabiltzaileari maiz proposatuko zaizkio berarentzat interesgarri izan daitezkeen antzeko edukiak eta, gainera, aurkitutako informazioa erraz interpretatzeko laguntza emango zaio.

PATHSek proposatzen duen nabigazio gidatu berri honek kontuan hartuko ditu bilduma digitalean zehar egin daitezkeen hainbat *ibilbide* ("path", ingelesez). Edozein gairi buruzkoa izan daiteke ibilbidea, adibidez, artista eta bere medioei buruz ("Picassoren margolanak"), garai historikoei buruz ("Gerra hotza"), lekuei buruz ("Venezia"), edota pertsonaia ezagunei buruzkoa ("Muhammad Ali"). Ibilbideak sortzeko eta jarraitzeko moduak hainbat izango dira, hala nola, aldez aurretik adituek definitutakoak, PATHS sistemak berak automatikoki proposatuak, edo, nahi izanez gero, erabiltzaileak sortutakoak ere. Beraz, eduki digitaletara iristeko era berritzailea eskainiko dio PATHSek erabiltzaileari, eta gainera erabiltzaileekin izandako esperientzia baliagarri izango zaio sistemari liburutegi digitala bera ere aberasteko.

Beste aplikazio bat hiztegiak editatzeko leXkit tresna[42] da (Alegria et al., 2001), bezero-zerbitzari arkitektura darabilena. Ezagutza teknikoren beharrik gabe, edizio-lana errazten dio lexikografoari. Sarrerei buruzko meta-informazioa baliatzen du funtzionalitate aurreratua eskaintzeko, esate baterako, testuinguruaren araberako atazak. Kubako *Diccionario Escolar*[43] hiztegi-aplikazioa (Miyares et al., 2010) tresna honekin sortu izan da.

BertsolariXa[44] aplikazioaren helburua (Arrieta et al., 2001) xumeagoa da, baina oso praktikoa: bukaera bat emanda, hitz errimatuak aurkitzen ditu. Lemak ez ezik, BertsolariXa gai da hitz deklinatuak eta aditz-formak ere eskaintzeko. Arloka iragaz daitezke emaitzak. Arau fonetikoak aplikatzeko aukera ere ematen du taldeko webgunean erabil daitekeen aplikazio honek.

Azken garai honetan nabarmena da taldea egiten ari den ahalegina nazioarteko sareetan parte hartzeko. Horrela sare hauetan integratu da:

- Clarin[45], Bere helburua hizkuntza-baliabide konputazionalak zabaltzea da giza zientzietako ikerketetan erabiliak izan daitezen.
- Flarenet[46] Fostering Language Resources Network.
- RTTH[47],Red Temática en Tecnologías del Habla.
- TIMM[48], Red Temática en Tratamiento de la Información Multilingüe y Multimodal.

42 http://sourceforge.net/projects/lexkit/

43 http://www.unibertsitatea.net/blogak/ixa/aaa

44 http://ixa3.si.ehu.es/tresnak/bertso/nagusia.html

45 http://www.clarin.eu/external/

46 http://www.flarenet.eu/

47 http://lorien.die.upm.es/~lapiz/rtth/

48 http://ararat.ujaen.es/timm

Bukatzeko, azpimarragarria da *Langune*[49], Hizkuntzen Industriaren alorreko Euskal Herriko enpresen elkartea sortu egin dela eta gure taldea tartean egon dela. Elkarte hau 2010an sortu da eta itzulpengintza, edukiak, irakaskuntza eta hizkuntzen teknologiaren alorreko 30 enpresatik gora elkartzen ditu.

5 Ondorioak

5.1 Teknologia garatzeko estrategia bat baliabide urriko hizkuntzetarako

Ixa taldearen jardunbidea oinarri hartuta baliabide urriko hizkuntzetarako baliagarri izan daitekeen estrategia azaldu dugu teknologia garatzeko. Orain dela 23 urte hasi ginen definitzen estrategia hori, eta beti izan da gure iparra jardunbidea planifikatzeko. Hori hobeto azaltzearren taldearen ibilbidearen traza erakutsi nahi izan dugu bi tauletan (ikus 10. eta 11. taulak). Taula horiek taldeak sortu dituen produktu eta proiektu nagusiak biltzen dituzte urteen eta ezagutza linguistikoen arabera ordenatuta. Ikus daitekeenez, etapa horien edukia eta ordena guztiz bat dator orain dela aspaldi definitu genuen estrategiarekin:

- **Hasieran oinarrizko baliabide eta tresna sendoak sortu ditugu**, eta geroago merkatu-aplikazioak. 10. taulan argi ikus daiteke hori.
- **Produktuen garapenean lexiko-morfologia-sintaxia-semantika progresioa erabili dugu**: hasierako produktuak oinarri-oinarri den morfologiaren gainean sortu ziren, geroago morfologiako produktuak hobetzen joan ziren bitartean lexikoan oinarritutakoak sortu ziren, geroago sintaxikoak, semantikakoak eta azkenik aplikazio aurreratuekin lotuta daudenak (itzulpen automatikoa, eta galderak erantzuteko sistemak batez ere).
- **Formatu estandarrak erabili ditugu** produktuen berrerabilpena erraztearren, XML formatuen erabilera zabalak eta LibiXa liburutegiaren sorkuntzak frogatzen dute hori.
- **Ahal izan den guztietan software librea erabili eta sortu dugu**. Hiru produktu ditugu eskuragarri Sourceforge biltegian, eta taldeak sortu dituen hainbat aplikazio publikoki erabil daitezke.

Estrategia horri jarraitu izanaren ondorio nabarmenak dira euskararen prozesamenduan egin diren aurrerapauso esanguratsuak.

Hizkuntzak sailkatzeko irizpide batzuk definitu ditugu hizkuntza-teknologian eta Interneten duten presentziaren arabera. Sailkapen horren arabera, duen hiztun kopuruagatik eta bere egoera soziolinguistikoarengatik, logikoena euskara laugarren mailan egotea litzateke, azkenaurreko mailan alegia.. Baina Ixa taldeak (beste eragile batzuen laguntzarekin) alor honetan egindako lanari esker euskara ez dago laugarren mailan, hirugarren mailan baizik. Estrategia horri jarraitu izanak aukera eman du horretarako. Gaur egun euskarak nolabaiteko presentzia du hizkuntz aplikazio informatiko gehienetan. Munduan dauden 7000 hizkuntzetatik 60 bat bakarrik dira maila horretara iritsi direnak, eta euskara horietako bat da.

Gure ustez gure estrategia eta ibilbidearen erreferentzia lagungarria izan daiteke oraindik hizkuntz teknologian sartuta ez dauden hizkuntzentzat, eta bereziki hizkuntz teknologian sartu ez baina IKTetan hasierako urratsak egin duten 190 hizkuntzentzat, alegia, sailkapeneko laugarren mailan sartzen diren hizkuntzentzat.

5.2 Aurrerapausoak alorrez-alor

Ikus ditzagun orain zein izan diren euskararen prozesamenduan egin diren aurrerapausoak alorrez alor (morfologia, lexiko, sintaxia, semantika eta pragmatika). Hizkuntzaren prozesaketaren bidean geratzen diren hutsuneak eta eginkizunak ere markatuko ditugu ildo horretan.

Euskararen **morfologia**ren azterketa ia osorik dago eginda eta inplementatuta. Morfologiaren alorrean gaur egun gure erronka inplementazio horien bertsio eraginkorragoak eta libreak eraikitzea da.

Sintaxi konputazionalaren tratamendua oraindik ikergai irekia da. Egun ezinezkoa da euskarazko edozein esaldi luze sintaktikoki ondo analizatzea. Ingelesarentzat gehiago landu da eta analizatzaile sintaktiko aurreratuagoak dituzte, baina beste hizkuntz nagusietan ere oraindik arazoak agertzen dira “esaldi errealak” osorik analizatu nahi izaterakoan. Egoera horretan, sintaxiari buruzko alorrean, hiru ikerlerro jorratzen ditugu: azaleko sintaxia (*Ixati* tresnak eramaten du aurrera eta testua zati sintaktikoetan banatzea du helburu), hitzen arteko dependentzia-erlazioak atzematea (EDGK), eta parser estatistikoa (Maltixa). Bitartean sintaktikoki etiketatuta dagoen EPEC corpusaren tamaina 10 edo 100 aldiz handiagoa egitea da helburua, prozedura semiautomatikoak erabiliz.

49 http://www.langune.com/

PRODUKTUAK	1988-1993	1993-1996	1996-1999	1999-2002	2002-2005	2006-2009	2009...
Produktuak. Aplikazio orokorra			Multimeteo			Anhitz **Matxin**	**Ihardetsi** BASYQUE EUSMT
Produktuak Semantika					**EuskalWN**	MCR WSD-Ixa	Eusemcor UKB
Produktuak Sintaxia				**Zatiak-Ixati**	Erreus corp.	Ancora corp.	Maltixa EDGK
Produktuak Lexikoa		**EDBL**	EDBL	Elhuyar-Word	UZEI sinon-hizt	EDBL	Lexkit Dicc. Escolar
Produktuak Morfologia	**Xuxen**	Xuxen... Morfeus	Xuxen **Eustagger**	Xuxen Elhuyar-Word	Xuxen Eihera	ZT corp. Eulia	BertsolariXa LibiXaml

10. taula: Produktu garrantzitsuenak urtea eta finantzazioaren arabera.

PROIEKTUAK	1988-1993	1993-1996	1996-1999	1999-2002	2002-2005	2006-2009	2009...
Proiektuak Europan					Meaning	Kyoto	Paths
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)			Item	Hermes	Hizking21 EuropenTrad Bilatzailea RICOTERM2 Hiztegia2002	Know OpenMT IMLT Praxem Avivavoz	Know2 OpenMT2 Hybridosint RTTH, TIMM Ancora
Proiektuak Eusko Jaurlaritzan		Xuxen	Xuxen EDBL Lematiz. Item	Xuxen Ixa taldea Sintaxi lexikoa UZEI sinon-hizt	Ixa taldea Hizking21 ETORTEK	Ixa taldea Anhitz ETORTEK	Ixa taldea Berbatek ETORTEK
Proiektuak Gipuzkoan (GFA)	Itzulpena Xuxen	HAIN	Xuxen Idazkide	Berbasare Gainternet	Hermes	Remixee Prest	Langune

11. taula: Proiektu garrantzitsuenak urtea eta finantzazioaren arabera.

Semantikaren bidean askoz ere lan gehiago gelditzen da egiteko. EuskalWornet taxonomia oinarri ikaragarria da, baina segitu behar da osatzen eta hobetzen. Halaber, Wikipedia iturburu oparoa da ezagutza semantikoa aberasteko. Ildo horretan, metodo automatikoak definitu behar ditugu hortik informazioa eraузteko. Hala ere, hitzen adieretatik harantzago joanda, perpausen interpretazio semantikoa lortu ahal izateko oraindik urrats asko eman behar dira, hasieraren hasieran baikaude oraindik. Jarraitu behar da aditzen azpikategorizazioa eta rol tematikoak aztertzen, batez ere informazio giltzarria eskaintzen dutelako gainontzeko arloen erresoluzio egokian, hala nola, desanbiguazio sintaktikoan, erreferentzia-kidetasunaren azterketan, etab. Corpus semantikoki etiketatuak behar dira gero ikasketa automatikoa erabili ahal izateko. Horrekin batera, hitz mailan adierak etiketatu behar dira, eta esaldi mailan aditzen azpikategoriak eta rol tematikoak.

Semantika konputazionalaren egoera oso hasierakoa bada ere, **pragmatika**rena askoz gordinagoa da, ohian basati landugabea dela esan daiteke. Diskurtsoaren egituraren azterketarekin hasi gara lanean berriki eta horretan bi bide nagusi definitu ditugu. Alde batetik, hizkuntzalaritzaren ikuspegitik aztertzen ari gara. Eta bestetik, saioak egiten ari gara hori bera lantzen ikasketa automatikoko sistemak erabiliz. Azkenik, bidean dagoen beste ikerlerro batek diskurtsoaren erlaziozko egitura zehazten dihardu. Elipsia, erreferentzia, hizketa-ekintzak (*speech acts*), hizketaren planifikazioa, hizlari-ereduak erabiltzea... Gai horiek guztiak zain daude.

Hizkuntzaren erabateko ulerkuntza automatikoa oraindik urruti dago. Oraingo ezagutza mugatua da, baina azken urteetan argi frogatu da teknologia ez-oso (ala partzial) hori gauza dela aplikazio praktikoak sortzen. Eta helburu horrekin jarduten dugu Ixa taldean. Hasieran lau partaide ginenak, orain 33 informatikari, 10 hizkuntzalari eta 3 teknikari gara. Euskal Herriko zazpi enpresekin lankidetzan gabiltza eta atzerriko beste bostekin. Spin-off erako bi enpresen sorkuntzan parte hartu dugu. 2002. urtetik Eusko Jaurlaritzak definitu zuen *Ingeniaritza linguistikoa* ikerlerro estrategikoa parte hartu dugu (Hizking21 eta Anhitz proiektuak) beste ikerketa-zentrorekin batera (Aholab, Elhuyar, Vicomtech eta Robotiker).

5.3 Oraingo ikerlerroak

Oinarrizko baliabide orokorrak sortzeko lerroaz gain, hauek dira gure azken proiektu estrategikoan (BerbaTek) ezarri ditugun ikerketa-lerro garrantzitsuenak:

- **Oinarrizko baliabideak**: hizkuntzen industriaren zenbait esparrutan erabil daitezkeen baliabideak eta teknologiak dira, edo gainerako alorren batean baliagarriak izan daitezkeen tresnetarako lehengai izan daitezkeenak. Esate baterako, testu-edo ahots-corpusa, lexikoiak, hiztegiak, ontologiak, gramatika konputazionalak, analizatzaile morfosintaktikoa, ahots-ezagutzea, ahots-sintesia, elkarrizketa-sistemak...
- **Itzulpengintza**: itzulpengintzaren sektorean erabil daitezkeen sistemak, itzulpen automatikoa, itzulpen-memoriak, ahots-ahots itzulpen-sistemak eta bikoizketa automatikoa kasu.
- **Edukiak**: edukien sektorea hobetzen lagundu dezaketen sistemak, hala nola, informazio-bilaketa (elebakarra, eleaniztuna, semantikoa, multimedia...), informazio-erauzketa, idazketan laguntzeko sistemak (zuzentzaileak, adibidez), ezagutzaren kudeaketa, galderei erantzuteko sistemak...
- **Irakaskuntza**: irakaskuntzaren eremuan erabiltzeko sistemak dira; adibidez, tutore pertsonalak, e-learning sistemak, ahoskera zuzentzeko sistemak, ariketen eta adibideen eraikitze automatikoa...

5.4 Giza talde baten ilusioa

Euskararen erronka honi aurre egiteko pertsona trebatuak behar zirela jakinda, hasieratik ere saiatu gara heziketa egokia zabaltzen eta teknologia honen protagonistak izango diren teknikari eta ikerlariak trebatzen, beti ere alde informatikaria eta alde linguistikoa uztartuz. 1989an doktorego-ikastaroak ematen hasi ginen, 2002an Hiztek titulu propioa sortu zen UEUren lankidetzarekin, 2005ean doktorego-programa bat (Hizkuntzaren azterketa eta prozesamendua) eta 2008tik abiatu zen izen bereko Europako master ofiziala. Unibertsitate mailako euskarazko master ofizial bakanetakoa da.

Hauek guztiak hamaika ahalegin eta ilusioren fruituak dira. Zenbat irakasle eta ikasle ibili garen, hor, lanean elkarrekin, heziketa-aukera hau euskaraz egin ahal izateko! 75 baino gehiago dira bide horietatik titulua eta heziketa berezitua jaso

duten teknikari/ikerlari berriak. 24 doktorego-tesi[50] sortu dira iturri horretatik. Ingeniaritza linguistikoan I+G horretan (Ikerketan eta Garapenean) arituko den komunitate zabal bat sortu dugu, baina aurrera egingo badugu, zabaldu egin behar dugu komunitate zientifiko hau. Hala biz!

6 Erreferentziak

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M. 1997. A spelling corrector for Basque based on morphology. Literary & Linguistic Computing, Vol. 12, No. 1. 31-38. Oxford University Press. Oxford.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., Urkia M. 1998. A framework for the automatic processing of Basque. Proceedings of the Workshop on Lexical Resources for Minority Languages. First LREC Conference. Granada.

Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Sarasola K., Soroa A. 2002. Towards the definition of a basic toolkit for HLT. LREC 2002.Workshop on Portability Issues in HLT. Las Palmas, Canary Islands.

Agirre E., Aldezabal I., Pociello E. 2006. Euskararako ezagutza-base lexiko-semantikoaren eredu-hautaketa eta garapena: EuskalWordNet. GOGOA aldizkaria ISNN 1577-9424 (pp. 237-266).

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Sarasola K. 2003. Patrixa: A unification-based parser for Basque and its application to the automatic analysis of verbs. In Bernard Oyharcabal (ed.), Inquiries into the lexicon-syntax relations in Basque, Anuario de Filología Vasca "Julio de Urquijo" nº XLVI, pp 47-73. University of the Basque Country.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., Lersundi M. 2001. EDBL: a General Lexical Basis for the Automatic Processing of Basque. IRCS Workshop on linguistic databases. Philadelphia (USA).

Alegria I., Arregi X., Artola X., Astiz M. , L. Ruiz Miyares. 2001. A Dictionary Content Management System. Proceedings EURALEX 2006 I, 105-109 (Turin, Italy). (ISBN 88-7694-918-6).

50 http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak

Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. 2007. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. LNCS 4394. pp. 374-384. Cicling 2007.

I. Alegria, I. Etxeberria, M. Hulden, M. Maritxalar 2009. Porting Basque Morphological Grammars to foma, an Open-Source Tool. FSMNLP2009. Pretoria. South Africa.

Ansa O., Arregi X., Otegi A., Valverde A. 2006. An XML Framework for a Basque Question Answering System. 7th International Conference on Flexible Query Answering Systems. Milano, Italia.

Arrieta B., Alegria I., Arregi X. *An assistant tool for Verse-Making in Basque based on Two-Level Morphology. Literary and Linguistic Computing. Online ISSN 1477-4615 - Print ISSN 0268-1145 . Vol. 16, No. 1; pag 29-43; 2001 (Oxford University press).*

L Borin. Linguistic diversity in the information society. 2009 SALTMIL2009 Workshop: IR-IE-LRL Information Retrieval and Information Extraction for Less Resourced Languages. University of the Basque Country. ISBN 978-84-692-4940-6.

S. Busemann, and H. Uszkoreit (2004) Predicting the Future: Technology Roadmapping. In: ELSNews, (3) 2004.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. COLING-ACL'98. Pgs. 380 - 384. Vol 1. Montreal (Canada).

Forcada M. Open source machine translation: an opportunity for minor languages. 5th SALTMIL Workshop on Minority Languages: Strategies for developing machine translation for minority languages. Genoa. 2006.

S. Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. International Workshop Speech and Computer. 27-29 October 2003, Moscow.

Maegaard B., Krauwer S., Choukri, K. and Jorgensen, L.D. The BLARK concept and BLARK for Arabic. Fifth International Conference on Language Resources and Evaluation, LREC. 2006.

Miyares Bermúdez, Eloína, Leonel Ruiz Miyares, Cristina Álamo Suárez, Celia Pérez Marqués, Xabier Artola Zubillaga, Iñaki Alegria Loinaz,

Xabier Arregi Iparragirre. 2010. *La segunda y tercera ediciones del Diccionario Básico Escolar*. Euralex2010. Leeuwarden (Herbehereak)

B. Petek. 2000. Funding for research into human language technologies for less prevalent languages, Second International Conference on Language Resources and Evaluation (LREC 2000). Athens, Greece.

Sarasola K. 2007. Technology is an effective tool to promote use of Basque. ICML Colloquium on Language Revitalisation through Multimedia Technology, Pecs, Hungary.

Simov K., Osenova P., Kolkovska S., Balabanova E. and Doikoff D. A language resources infrastructure for Bulgarian. Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC. 1685--1688. 2004.

O. Streiter, K.P. Scannell, M. Stuflesser (2006) Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for developers. Machine Translation Journal. Volume 20, Number 4, pp. 267-289.

B. Williams, K. Sarasola, D. Ó´Cróinin, B. Petek. 2001. Speech and Language Technology for Minority Languages. Proceedings of Eurospeech 2001.

Wilson A., Archer D., and Rayson P. Corpus linguistics around the world. Rodopi. 2006.

Artigos de Investigação

BASYQUE: Aplicación para el estudio de la variación sintáctica

Larraitz Uria
Centro IKER (UMR5478) – IKERBASQUE
Baiona-Bayona (Francia)
larraitz.uria@ehu.es

Ricardo Etxepare
Centro IKER (UMR5478)
Baiona-Bayona (Francia)

Resumen

En este artículo presentamos BASYQUE, la aplicación que hemos desarrollado para el estudio de la variación sintáctica. Aunque este proyecto se centra fundamentalmente en los dialectos del País Vasco francés (*Iparralde*), BASYQUE es una aplicación multilingüe útil también para el análisis de otras lenguas y/o dialectos. Además de presentar las opciones que ofrece BASYQUE, abordaremos los aspectos metodológicos y los criterios que hemos seguido para realizar esta aplicación. El desarrollo de este proyecto cuenta con la colaboración del Centro IKER con sede en Bayona (Francia) y el Grupo IXA de la Universidad del País Vasco (UPV).

1. Introducción

Durante los últimos años ha ido cambiado considerablemente el punto de vista analítico sobre el uso del código estándar de una lengua y sus dialectos: en contra de la tesis que defendía un único modelo lingüístico, basado en el uso exclusivo y generalizado del código estándar, en la actualidad cobra fuerza la idea de que el código estándar no garantiza una comunicación eficaz en todos los contextos o situaciones comunicativas y que, por lo tanto, conviene abrir un espacio de uso también para los códigos comunicativos propios de cada lugar, es decir, para las variantes dialectales territoriales.

Desde el reciente interés suscitado por los registros dialectales en la constitución de corpus lingüísticos y con la estimable ayuda de los recursos que nos ofrecen las nuevas tecnologías en el campo del Procesamiento del Lenguaje Natural (PLN), hemos desarrollado la aplicación denominada BASYQUE[1]. Esta aplicación nos brinda la posibilidad de realizar diferentes tipos de consulta y obtener así información precisa sobre estructuras sintácticas concretas que presentan alternancia interdialectal. Se trata, por tanto, de un soporte que podemos definir como un atlas sintáctico básico.

La lengua vasca (*euskara*) constituye en la actualidad un campo de investigación privilegiado para el estudio de la variación sintáctica, ya que presenta un entorno lingüístico muy rico en la que aún concurren y conviven diversas variantes dialectales.

En este proyecto nos centramos en el análisis de la variación sintáctica que existe entre los dialectos del País Vasco francés (*Iparralde*). Precisamente en *Iparralde*, el carácter dialectal de los textos escritos es, con diferencia, más acusado que en cualquier otra zona del País Vasco. Ello se debe, entre otros factores, a que a pesar de que el código estándar (*euskara batua*) viene incrementando su presencia progresivamente, éste se introdujo más tardíamente en un contexto que además no conoce, de momento, una escolarización masiva en la lengua vasca. En este contexto, las variantes dialectales siguen teniendo un vigor considerable, tanto en el ámbito de la literatura escrita como en los distintos medios de comunicación que operan en *Iparralde*.

Por otra parte, la variación sintáctica presente en los dialectos de *Iparralde* no ha sido estudiada detenidamente desde el punto de vista gramatical, por lo que no disponemos ni de textos lingüísticamente etiquetados ni bases de datos, infraestructuras o entornos informáticos que proporcionen información detallada sobre las características de tales dialectos.

Consideramos, por tanto, que el desarrollo de herramientas que faciliten el análisis de estas variantes dialectales constituye una aportación interesante desde el punto de vista de los corpus disponibles en soporte informático y contribuye a una representación más adecuada de la riqueza lingüística existente en el ámbito del euskara. Con

[1] http://ixa2.si.ehu.es/atlas2/index.php?lang=es

ese objetivo hemos desarrollado BASYQUE. Y aunque este estudio se centra en los dialectos del País Vasco francés, BASYQUE es una aplicación multilingüe útil también para el tratamiento y análisis de otras lenguas y/o dialectos.

Este proyecto viene desarrollándose con la colaboración de dos grupos de investigación de amplia experiencia en este campo: el centro IKER (UMR5478) de Bayona (Francia), que trabaja en el ámbito de la lingüística y la dialectología de la lengua vasca, y el grupo IXA de la Universidad del País Vasco (UPV), que lleva años trabajando en la creación de recursos informáticos básicos para el PLN en general y para el análisis del *euskara*, en particular[2].

Tras esta breve introducción, nos centraremos primeramente en la metodología que hemos aplicado y en los criterios que hemos definido para la recogida de datos (apartado 2). En el apartado 3, presentaremos BASYQUE y daremos a conocer el tipo de información que puede consultarse en esta aplicación. Finalmente, en el apartado 4, trataremos de exponer unas primeras conclusiones y abordaremos las líneas futuras de trabajo.

2. Metodología y criterios para la recogida de datos

Para poder efectuar consultas en BASYQUE y obtener así información en torno a la variación sintáctica, es imprescindible alimentar primero la base de datos. Para ello, hay determinados aspectos metodológicos que hay que tener en cuenta en este tipo de atlas sintácticos: la delimitación del campo de trabajo, la selección de informantes, la elección de las variables lingüísticas que serán objeto de análisis, la elaboración de cuestionarios y, por último, la recogida de datos.

En lo que respecta a la delimitación del campo de trabajo, nuestro proyecto abarca los dialectos del País Vasco francés. *Iparralde* cuenta con una comunidad lingüística vasca relativamente modesta de unos 55.000 hablantes (para un total de 150.000 habitantes, aproximadamente) y donde el proceso de estandarización de la lengua vasca avanza a un ritmo más lento que en el resto del País Vasco. Así, el carácter dialectal de los textos escritos en dicho territorio es más marcado que en cualquier otra zona del país, por lo que nos encontramos en un entorno que podemos calificar de excepcional para el estudio de la variación sintáctica.

Para llevar a cabo este proyecto hemos seleccionado 20 localidades, tomando en cuenta la subdivisión dialectal del País Vasco francés.

En cada una de las localidades, seleccionamos cuatro informantes que deben de cumplir los siguientes requisitos:

- Tener como lengua materna el dialecto de su localidad.
- Haber nacido, tanto el informante como sus padres, en esa misma localidad.
- No haber residido durante un período superior a siete años fuera de la localidad natal.
- Tratarse de informantes hablantes activos de su dialecto.

De estos cuatro informantes, hemos de contar con la colaboración de una persona mayor de entre 50 y 70 años, otra de entre 36-50 años y dos jóvenes comprendidos en la franja de 18-35 años, de los cuales uno realiza o ha realizado sus estudios en *euskara* y el segundo no. Este último perfil busca la posibilidad de analizar i) hasta qué punto las formas dialectales locales siguen vigentes entre los jóvenes, ii) la incidencia potencial de los procesos de escolarización en el uso/abandono de las variantes dialectales, y iii) la influencia que puede ejercer en la juventud el contacto con una lengua dominante como es el francés.

Somos conscientes de que este criterio de selección tan riguroso puede dificultar la búsqueda de informantes apropiados, pero resulta un factor imprescindible para que los datos recopilados ofrezcan una visión lo más uniforme posible y para que las conclusiones sean debidamente contrastadas.

La elección de los fenómenos lingüísticos a analizar depende fundamentalmente de las características de

[2] El origen de este trabajo se encuentra en el proyecto TSABL (*Towards a Syntactic Atlas of the Basque Language*: http://www.iker.cnrs.fr/-tsabl-towards-a-syntactic-atlas-of-.html?lang=fr) y esta ligado al proyecto BasDiSyn (*Basque Dialect Syntax*: http://basdisyn.net/en/). BasDiSyn es un grupo compuesto por investigadores que trabajan en el estudio de la micro-variación sintáctica del *euskara*. Este grupo, a su vez, se encuentra en la red europea de investigación Edysin (http://www.dialectsyntax.org/index.php/home-mainmenu-1), en colaboración con otros investigadores que también se dedican a investigar en el campo de la dialectología (Carrilho, 2010; Barbiers and Bennis, 2007; entre otros).

cada lengua y/o dialecto. Las alternancias sintácticas en las que nos centramos en este proyecto están ligadas a la concordancia, la determinación, las estructuras posposicionales, la nominalización, las expresiones de modalidad y la evidencialidad. Todos ellos son aspectos centrales de la gramática del *euskara* en los que tiende a producirse alternancia o variación sintáctica interdialectal. A modo de ejemplo, presentamos algunas estructuras donde ocurre variación sintáctica:

- Concordancia entre el argumento dativo y el verbo:

Una regla gramatical de aplicación general en la lengua vasca exige que el argumento dativo concuerde en número y persona con el auxiliar. Esta regla general, que no conoce excepciones en el habla del País Vasco peninsular, tiene un estatus opcional en el *euskara* de *Iparralde*. Por tanto, es posible encontrarse con alternancias del tipo representado en (1a, b):

(1) *a. Ez* ***dio*** ***neori*** *deus igorri*
Ez-NEG dio-AUX neori-DAT deus-ABS igorri-V
abs/erg/dat
No ha enviado nada a nadie

b. Ez ***du*** ***neori*** *deus bota*
Ez-NEG du-AUX neori-DAT deus-ABS bota-V
abs/erg
No ha enviado nada a nadie

Esta opcionalidad esta sujeta a varios condicionantes, que tienen que ver con la naturaleza referencial del argumento dativo y las propiedades léxicas del verbo.

- Posición de la partícula evidencial *omen* ("según parece", "por lo visto"):

En algunas localidades la partícula *omen* aparece entre el verbo y el auxiliar (2a), mientras que en otras zonas el auxiliar se antepone al verbo principal y *omen* se coloca entre el auxiliar y el verbo (2b):

(2) *a. Langonen* ***bizi*** ***omen*** ***zen***
Langonen-INE bizi-V omen-PRT zen-AUX
Según parece residía en Langon

(2) *b. Langonen* ***zen*** ***omen*** ***bizi***
Langonen-INE zen-AUX omen-PRT bizi-V
Según parece residía en Langon

Estas diferencias parecen estar relacionadas con la categoría léxica de la partícula evidencial: en el primer caso, ésta se comporta como una cabeza funcional; en el segundo caso, como un adverbio.

- Alternancia en el uso del determinante:

En algunas zonas de *Iparralde* se omite el determinante del sustantivo en posición de objeto y el sustantivo tiene una lectura plural (3a), mientras que en otras localidades tal omisión es imposible (3b):

(3) *a. Hor bada* ***sagar***
Hor-ADV bada-V sagar-N
Ahí hay manzanas

(3) *b. Hor badira* ***sagarrak***
Hor-ADV badira-V sagar-N ak-DET
Ahí hay manzanas

Ejemplos de alternancia sintáctica como los arriba indicados son precisamente los que pretendemos recopilar y clasificar en BASYQUE. Así, los cuestionarios empleados indagan sobre la distribución geográfica de estas variantes.

Para la recopilación de los datos que nos interesan, disponemos de tres fuentes de información. La fuente principal de información la constituyen los cuestionarios. Cada cuestionario aborda un fenómeno lingüístico concreto y son cuestionarios específicos diseñados para elicitar ejemplos concretos de variación sintáctica interdialectal. Preparamos diferentes tipos de tests, que consisten principalmente en frases para traducir y tareas de selección de variantes basadas en una escala de aceptabilidad relativa. En algunos casos, pedimos al encuestado que complete una forma lingüística.

Pasamos los cuestionarios a los informantes previamente seleccionados, grabamos las respuestas y guardamos los resultados obtenidos (tanto las respuestas como la información correspondiente a cada una de ellas) en una base de datos a la que podemos acceder mediante la aplicación BASYQUE.

Aparte de las encuestas diseñadas específicamente para BASYQUE, existe en el País Vasco una amplia tradición investigadora en el campo de la dialectología, en la que abundan ejemplos característicos y significativos de las variantes dialectales del *euskara.* Además, en los textos literarios también están presentes las múltiples

alternancias sintácticas. Por tanto, y aunque los cuestionarios constituyan la principal fuente de información para la materialización de este estudio, hemos considerado interesante reunir en una misma aplicación ejemplos de origen diverso. Ello nos proporciona, por una parte, mayor abundancia de ejemplos y, por otra, la posibilidad de establecer comparaciones entre los datos obtenidos mediante cuestionarios específicos con los ejemplos obtenidos por medio de otras fuentes de información. De hecho, recopilar ejemplos de origen diverso es importante debido a que ninguna de las fuentes mencionadas tiene un carácter óptimo para el estudio de la variación sintáctica, es decir, todas presentan algunos pros y contras (véase, por ejemplo, Cornips, L. & C. Poletto, 2005). Obviamente, la base de datos consigna el origen distinto de cada una de las fuentes.

El proceso que conduce desde la recopilación de datos hasta su consulta, puede resumirse de la manera siguiente:

- Preparación previa de cuestionarios específicos que nos permitan obtener ejemplos de fenómenos lingüísticos concretos y búsqueda de ejemplos de dichas variantes dialectales tanto en grabaciones realizadas en otros proyectos como en textos literarios recientes.
- Introducción de los ejemplos obtenidos en una base de datos, junto con la correspondiente información adicional[3].
- Consulta sistemática de todos los datos almacenados con la ayuda de la aplicación BASYQUE.
- Análisis de datos y conclusiones.

Gracias a la información recopilada en BASYQUE, contamos con una base muy sólida (y en este caso única) que nos permite continuar progresando en el estudio de la variación sintáctica dialectal de *Iparralde*, tanto desde el punto de vista lingüístico como en el campo del PLN.

3. La aplicación BASYQUE

BASYQUE es la aplicación que hemos desarrollado para el estudio de la variación sintáctica que se produce entre dialectos.

En la imagen 1 (última página) puede verse la interfaz de la aplicación, donde se encuentran las principales opciones disponibles. En el apartado *Información* se recogen las aportaciones relacionadas con este trabajo o proyectos similares. La sección *Administración* está destinada a aquellos usuarios que se registran por primera vez (el registro no es condición necesaria para poder realizar consultas). En *Contacto* se especifican nuestros datos para que los usuarios puedan contactarnos para aclarar dudas, resolver problemas o realizar nuevas propuestas. En el apartado *Ayuda* el usuario dispone de un manual en el que se describen detalladamente las formas de uso y las posibilidades que ofrece la aplicación. Junto al apartado de *Ayuda*, se puede seleccionar el *idioma* en el que se quiere realizar las consultas: euskara, español, inglés o francés. En la parte superior de la interfaz se localiza el *Campo de Registro* desde el cual los anotadores y el administrador acceden a la sección privada de la aplicación para así introducir, modificar o eliminar datos. En la parte derecha se encuentra el *Menú de búsqueda* donde están disponibles diversas opciones de consulta. Las respuestas obtenidas en cada búsqueda se visualizan en el *Mapa* que se encuentra localizado en el centro mismo de la interfaz.

En lo que se refiere a la búsqueda de datos, BASYQUE ofrece la posibilidad de realizar diferentes tipos de consulta y de buscar ejemplos que provienen de tres fuentes de información diferentes: cuestionarios, conversaciones libres grabadas en otros proyectos y corpus literarios. Cada fuente de información cuenta con una serie de opciones y atributos de búsqueda que detallamos a continuación.

3.1. Consulta de ejemplos obtenidos mediante cuestionarios

Los ejemplos obtenidos mediante cuestionarios específicos constituyen la fuente de información principal de este proyecto. Las respuestas obtenidas pueden consultarse a distintos niveles: por localidades, por meta-categorías lingüísticas o por clases de verbos, por cuestionarios, por propiedades lingüísticas y también por edades de los informantes.

[3] Como detallaremos más adelante, junto a cada ejemplo, especificamos la escritura formal de la forma dialectal transcrita, el texto normalizado, la información gramatical de cada elemento de la frase (glosas), las propiedades lingüísticas que le corresponden, el cuestionario al que pertenece, localidad en la que se ha recogido el ejemplo, etc.

Al seleccionar una de las opciones mencionadas, en el menú de búsqueda se muestran los atributos disponibles que le corresponden a cada una de ellas. Por ejemplo, si seleccionamos la opción *Localidad*, se visualizan las localidades en las que se han realizado los cuestionarios o si optamos por las *Propiedades lingüísticas*, se muestran las propiedades o características lingüísticas que se han asignado a las respuestas de los informantes[4].

Es posible combinar varias opciones y atributos a la hora de buscar datos. Así, se pueden consultar:

- Ejemplos recogidos en una sola localidad.
- Ejemplos recogidos en una o más localidades que contengan una o más propiedades lingüísticas.
- Ejemplos proporcionados por informantes de una(s) edad(es) concreta(s) que pertenecen a una propiedad lingüística específica y que han sido recogidos en una o más localidades.
- Etc.

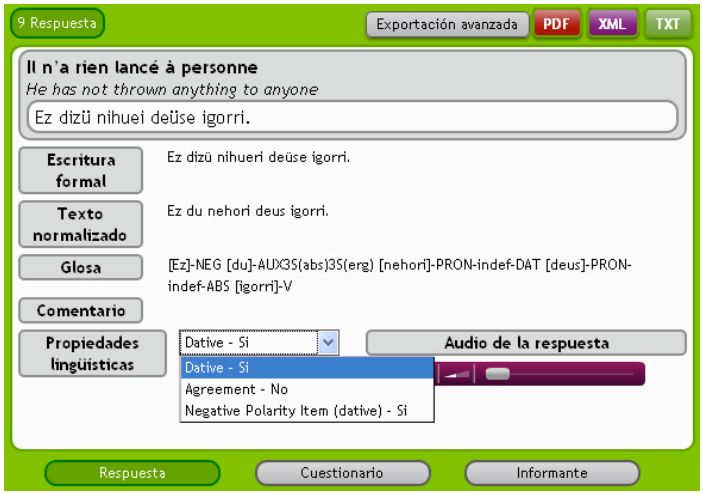
El abanico de posibilidades de consulta que ofrece BASYQUE es, por tanto, muy amplio.

Los resultados de la búsqueda se visualizan en el mapa central (imagen 1). Para ello, se hace uso de las tecnologías que ofrece Google Maps. De este modo, podemos observar cuales son las localidades en las que se detectan las alternancias sintácticas cuyo análisis consideramos de interés. El mapa es una característica propia de los atlas sintácticos y cada usuario puede optar por el tipo de mapa que más le guste: mapa, foto de satélite o la combinación de ambas opciones. Además, las localidades se pueden visualizar dentro de las provincias o las zonas dialectales especificadas por Zuazo (2008).

En la parte inferior del mapa, se indica el total de los resultados obtenidos en cada búsqueda y se detalla, en tres ventanas, la información correspondiente a cada una de las respuestas.

En la primera ventana se muestran:

- La pregunta que se ha formulado al informante.
- La traducción de la pregunta al inglés.
- La transcripción de la respuesta que ha proporcionado el informante, tal y como lo ha pronunciado literalmente.
- La escritura formal que corresponde a la respuesta.
- El texto normalizado que corresponde a la respuesta.
- Las glosas (información lingüística detallada) de cada elemento de la frase[5].
- Las propiedades lingüísticas asignadas a la respuesta.
- Posibles comentarios tanto del informante como del responsable de la encuesta.
- El reproductor de audio para poder escuchar la respuesta proporcionada por el informante.



[4] A todas las respuestas que recolectamos asignamos ciertas propiedades lingüísticas. Dichas propiedades sirven para indicar el tipo de variación sintáctica que conlleva cada respuesta. Por ejemplo, para el análisis de la variación que existe en lo que se refiere a la concordancia entre el verbo y el dativo, indicamos en cada ejemplo si el argumento dativo esta presente o no, el tipo de argumento dativo (nombre propio, nombre común, dativo reflexivo o recíproco…) y si existe o no concordancia con el verbo. O en el caso de la alternancia que ocurre con el evidencial *omen*, especificamos su posición: V + PRT-*omen* + AUX, AUX + PRT-*omen* + V, … Las propiedades lingüísticas actúan como atributos de búsqueda.

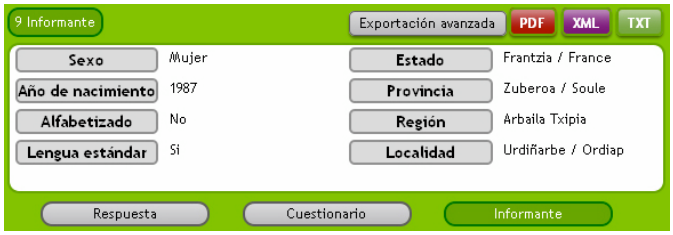
[5] Las glosas se asignan automáticamente, gracias al analizador morfosintáctico y un programa informático desarrollados en el grupo IXA de la UPV. El texto normalizado es precisamente necesario para poder glosar las frases automáticamente, ya que de momento no disponemos de analizadores sintácticos que analizan los textos escritos en los dialectos de *Iparralde*. Esta es, sin duda, una aportación muy novedosa a este tipo de Atlas sintácticos, no solo porque facilita el trabajo manual sino también porque toda esta información será una base muy importante para poder crear, en un futuro próximo, herramientas que posibiliten el análisis y procesamiento de las variantes dialectales.

En la segunda ventana se detallan:

- El cuestionario al que pertenece la respuesta obtenida.
- La sección (categoría y subcategoría) dentro del cuestionario al que pertenece la respuesta.
- El tipo de pregunta (traducción, ejercicio de complementación, pregunta de opción múltiple…).
- El responsable de la encuesta.
- Año en el que se ha obtenido la respuesta.
- La meta-categoría y/o la palabra clave (*keyword*) relacionada con esa categoría o subcategoría del cuestionario.

Por último, en la tercera ventana se especifican:

- Los datos del informante (año de nacimiento, sexo y la competencia lingüística en el estándar).
- La localidad a la que pertenece el informante.



Toda la información obtenida por cada respuesta (tanto los ejemplos como los datos correspondientes a los mismos) se puede exportar a documentos de formato pdf, xml o txt. Incluso existe la posibilidad de hacer una exportación avanzada para obtener solamente aquellos datos que más nos interesan (por ejemplo, solamente las respuestas proporcionadas por los informantes eliminando el resto de los datos, o las respuestas con las propiedades lingüísticas asignadas, o únicamente las respuestas con las glosas, o los cuestionarios conducidos, etc.). Esta posibilidad nos permite poder conservar esos datos en nuestro ordenador y poder utilizarlos para otros estudios.

3.2. Consulta de ejemplos recopilados en otros proyectos

BASYQUE ofrece también la posibilidad de poder consultar ejemplos obtenidos de las conversaciones grabadas en otros proyectos. En realidad, existen varios proyectos en el País Vasco que tienen como objetivo recoger el habla tradicional[6]. En tales proyectos se graban y se recogen conversaciones y/o testimonios de personas que se expresan en su propio dialecto. Normalmente son conversaciones que no están sujetas a un cuestionario específico; por el contrario, en cada grabación los hablantes versan libremente sobre un tema concreto (vacaciones, familia, tradiciones locales, etc.).

En dichas conversaciones se producen ejemplos de variación sintáctica, es decir, los ejemplos afloran sin tener que recurrir a la elicitación de los mismos. Esta nos parece una circunstancia interesante que podemos aprovechar para nuestros estudios. De este modo, además de obtener una mayor cantidad de ejemplos representativos, vamos creando redes de colaboración que nos permitirán ampliar nuestra base de informantes y profundizar en el estudio de las variantes dialectales de la lengua vasca.

En este apartado, las consultas pueden realizarse por medio de los siguientes parámetros diferenciadores: localidades, proyectos, fenómenos lingüísticos[7], propiedades lingüísticas y por edades de los informantes. Es decir, podemos consultar:

- Ejemplos recogidos en uno o más proyectos.
- Ejemplos que corresponden a una o más localidades.
- Ejemplos que tienen asignada(s) una(s) propiedad(es) lingüística(s) concreta(s).
- Ejemplos proporcionados por hablantes de una o más generaciones.

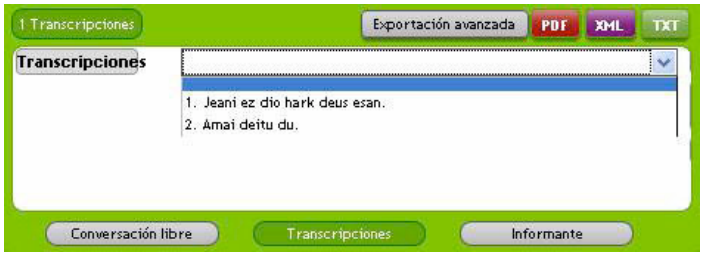
[6] *Ahotsak.com* o *EKE.org* son, por ejemplo, dos proyectos importantes que tienen como objetivo recopilar y difundir testimonios orales de vasco-parlantes.

[7] En el apartado de las conversaciones libres y de los corpus literarios, puesto que los ejemplos recopilados no están ligados a un cuestionario específico, indicamos el fenómeno lingüístico (concordancia, uso del determinante, evidencialidad, nominalización…) que corresponde a cada uno de los ejemplos.

- Etc.

Al igual que ocurre con las respuestas obtenidas mediante los cuestionarios, estos resultados también se visualizan en el mapa, ya que cada conversación esta grabada en una localidad concreta. Y la información correspondiente a cada ejemplo también se ofrece detallada en tres ventanas. En la primera ventana se indica el tema sobre la que versa la conversación, el responsable o el proyecto al que pertenece, el año de la grabación y el vídeo para poder escuchar el testimonio completo.

En esos testimonios grabados podemos escuchar frases características y/o ligadas a la variación sintáctica. Esas frases que nos interesan las transcribimos. En la segunda ventana encontramos una lista de los ejemplos que han sido extraídos y transcritos de la conversación global.



Seleccionando cada ejemplo, uno por uno, podemos ver toda la información que le corresponde: la forma dialectal del ejemplo, su transcripción o escritura formal, el texto normalizado, las glosas, el fenómeno lingüístico que le corresponde, las propiedades lingüísticas asignadas y el audio de la frase transcrita, al igual que en las respuestas obtenidas mediante los cuestionarios.

Finalmente, en la tercera ventana, se detallan los datos que corresponden al informante y a su localidad.

Podemos exportar, además, toda esa información a documentos de formato pdf, xml o txt, mediante los botones que se encuentran en la parte superior de cada ventana.

3.3. Consulta de ejemplos localizados en corpus literarios

Por último, en lo que se refiere a la búsqueda de los ejemplos de variación sintáctica encontrados en corpus o textos literarios, las opciones de consulta disponible son: por autores, por editoriales[8], por años de publicación, por fenómenos lingüísticos o por propiedades lingüísticas. Así, podemos obtener:

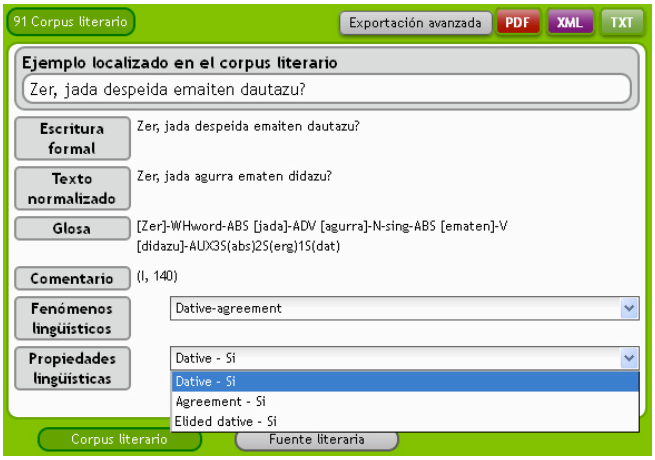
- Solamente los ejemplos localizados en los trabajos de un autor concreto.
- Ejemplos que corresponden a uno o más autores.
- Ejemplos relacionados a fenómenos lingüísticos concretos.

[8] Existen editoriales (*Maiatz*, entre otras) que tienen como objetivo exclusivo la promoción de la literatura escrita en *Iparralde* sin ninguna exigencia de estandarización.

- Ejemplos que corresponden a uno o más autores y a un año de publicación concreto, y que tienen asignada(s) una(s) propiedad(es) lingüística(s) concreta(s).
- Etc.

La diferencia, en este caso, con las dos anteriores fuentes de información, consiste en que estos ejemplos ni se visualizan en el mapa ni se pueden escuchar puesto que no se trata de ejemplos grabados en una localidad concreta sino que están ligadas a textos escritos. Los resultados y la información correspondiente a cada uno de los ejemplos se detallan en dos ventanas.

En la primera ventana podemos visualizar el ejemplo localizado y su correspondiente información (la forma dialectal del ejemplo, su escritura formal, el texto normalizado, las glosas asignadas automáticamente, posibles comentarios del anotador, el fenómeno lingüístico que le corresponde y las propiedades lingüísticas asignadas).



Y en la segunda ventana se especifican los datos que corresponden al texto literario en el que se ha localizado el ejemplo (el autor, el título del libro, la editorial, el año de producción, el año de publicación del texto y el área al que pertenece, en caso de haberlo especificado).

Los datos obtenidos también se pueden exportar a los formatos anteriormente mencionados (pdf, xml y txt).

Por tanto, si bien la fuente principal de información lo constituyen los cuestionarios, BASYQUE también ofrece la posibilidad de importar y consultar ejemplos de alternancia sintáctica que hayan sido localizados en conversaciones grabadas en otros proyectos que trabajan en el campo de la dialectología o en diversos corpus literarios.

De este modo, BASYQUE se convierte en una aplicación muy significativa y valiosa como instrumento de trabajo orientado al análisis de la variación sintáctica que se produce entre los dialectos de una misma lengua.

4. Conclusiones y trabajo futuro

En este artículo hemos presentado BASYQUE, la aplicación desarrollada para el análisis de la variación sintáctica intra-dialectal. Aunque desde nuestra propia perspectiva nos hemos centrado en los dialectos del País Vasco francés, BASYQUE es una aplicación multilingüe que también puede ser útil para el análisis de otras lenguas y dialectos.

Tal como lo venimos señalando a lo largo del artículo, la aplicación nos ofrece la posibilidad de consultar ejemplos que pertenecen a tres fuentes de información: cuestionarios específicos, grabaciones realizadas en otros proyectos y corpus literarios. De este modo, pretendemos i) ofrecer la oportunidad de consultar ejemplos de origen y carácter diferente, ii) aprovechar para nuestros estudios los datos que se han recogido en otros proyectos y textos literarios y iii) crear redes de trabajo en común con los grupos de investigación y fundaciones de tipo cultural que trabajan en el campo de la dialectología.

BASYQUE ofrece, por tanto, diferentes modos de consulta para obtener un tipo de información u otro. El hecho de poseer toda esa información recopilada, organizada y etiquetada en una base de datos, facilita en gran medida los análisis relacionados con la variación sintáctica, ya que de este modo está a disposición de los usuarios una gran cantidad de ejemplos de estructuras sintácticas concretas localizadas geográficamente, clasificadas en relación a determinados parámetros sociolingüísticos y acompañadas de un análisis gramatical básico.

Del conjunto de los datos almacenados, no solo obtenemos ejemplos concretos de alternancias sintácticas sino que, además, tenemos la posibilidad de realizar estudios comparativos, de poder analizar la evolución o la trayectoria por las que han atravesado las diversas variantes dialectales.

Esta aplicación puede resultar, por tanto, un primer paso adelante para aquellos investigadores interesados en el estudio de las variaciones sintácticas en lenguas o dialectos que carecen de bases de datos, plataformas o entornos informáticos que ofrezcan ejemplos e información sobre esas estructuras que sean objeto de estudio. Y en ese contexto encaja perfectamente la situación de los dialectos de *Iparralde*, ya que no existen textos lingüísticamente etiquetados ni bases de datos que faciliten información detallada sobre la variación sintáctica intra-dialectal de esa parte del País Vasco.

Además, toda la información que BASYQUE nos permite almacenar puede servir en un futuro próximo para canalizar el tratamiento automático de las estructuras sintácticas analizadas. Partiendo de la información almacenada en esta aplicación, nuestro primer objetivo es el de abrir nuevas líneas de investigación y dar pasos que permitan progresar en el campo del análisis y procesamiento de las variantes dialectales de *Iparralde*. Para ello, será necesario también estudiar las características léxicas y morfológicas de las variantes dialectales, asociándolos a los ejemplos de las estructuras sintácticas que recogemos en BASYQUE.

Una vez recopilada toda la información arriba referida y tomando como base los sistemas informáticos que el grupo IXA ha desarrollado para el análisis y procesamiento del *euskara*, estaremos en condiciones de crear aplicaciones, tales como marcadores de variantes dialectales, analizadores de textos escritos en diversos dialectos, aplicaciones que posibiliten consultar la forma estándar de una variante dialectal o las posibles variantes dialectales de una forma estándar. Ese tipo de aplicaciones pueden resultar realmente eficaces para seguir investigando no solo en el ámbito de la dialectología sino también en el campo de la lingüística teórica o la enseñanza de idiomas, por mencionar algunas de sus potencialidades.

Desde nuestro punto de vista, BASYQUE es una herramienta práctica que tiene un interés indudable para la comunidad lingüística vasca. Ese interés comprende tres áreas distintas pero complementarias: por una parte, la constitución de corpus específicos con un soporte informático que den cuenta de la variación lingüística existente en el dominio vasco; por otra parte, la creación de infraestructuras informáticas específicas para el procesamiento de la variación lingüística dialectal; y por último, el desarrollo de bases de datos directamente utilizables para el análisis gramatical de las formas dialectales.

References

Barbiers and Bennis. 2007. The Syntactic Atlas of the Dutch Dialects. A discussion of choices in the SAND-project. Nordlyd 34, 53-72. Internet: http://www.ub.uit.no/baser/nordlyd/.

Carrilho, Ernestina. 2010. Tools for dialect syntax: the case of CORDIAL-SIN (an annoated corpus of portugues dialects). In Tools for Linguistic Variation, Gotzon Aurrekoetxea & Jose Luis Ormaetxea (eds.), Anejos de ASJU, LII. Bilbo, UPV/EHU. ISBN: 978-84-9860-429-0.

Cornips, L. & C. Poletto. 2005. On standardising syntactic elicitation techniques. Part 1. Lingua 115 (7), 939-957.

Etxepare, Ricardo. 2009. On (some of) the Uses of a Syntactic Atlas. Tinta. E-journal of Hispanic and Lusophone Studies. Department of Hispanic and Portuguese Studies. Santa Barbara, University of California Santa Barbara 8(2), 1-24.

Etxepare, Ricardo. (aparecerá). The emergence of a dative alternation in north eastern varieties of Basque. En Carme Picallo y Jose María Brucart (eds). *Linguistic Variation in Minimalism*. Oxford: Oxford University Press.

Etxepare, Ricardo. 2010. *Omen* bariazioan [La partícula evidencial *omen* en sus variantes]. En Beatriz Fernandez, Pablo Albizu y Ricardo Etxepare (eds). Euskara eta euskarak. Suplementos de ASJU. Donostia: Universidad del País Vasco-Diputación Foral de Gipuzkoa.

Fernandez, Beatriz y Ortiz de Urbina, Jon. 2010. Datiboa hiztegian [El dativo en el léxico]. Euskal Herriko Unibertsitatea.

Haddican, William. 2005. Aspects of Language Variation and Change in Contemporary Basque. Doctoral Dissertation. New York University.

Holmer, Nils M. 1964. El idioma vasco hablado. Un estudio de dialectología euskérica. Donostia. ASJU.

Ortiz de Urbina, Jon. 1994. Datibo komunztaduraren inguruan [On dative agreement]. In Ricardo Gomez and Joseba Lakarra (eds). Euskal Dialektologiako Kongresua [Congress of Basque Dialectology]. Donostia: Supplements of the International Journal of Basque Language and Linguistics. 579-588.

Zuazo, K. 2008. Euskalkiak, euskararen dialektoak [The Dialects of Basque]. Donostia. Elkar.

Imagen 1: Interfaz de la aplicación BASYQUE.

O passar do TEMPO no HAREM

Cristina Mota
Linguateca (FCCN)
cmota@ist.utl.pt

Paula Carvalho
XLDB (FCUL)
pcc@di.fc.ul.pt

Resumo

O presente artigo apresenta um estudo contrastivo entre as propostas de análise do tempo na primeira e segunda edições do HAREM. Discutimos, entre outros aspectos, as principais vantagens e inconvenientes de uma e outra, tendo em linha de conta os princípios teórico-metodológicos subjacentes ao modelo geral do HAREM. A discussão é feita com base nas expressões temporais compreendidas nas colecções douradas do Primeiro e do Segundo HAREM, que reanalisámos de acordo com, respectivamente, as directivas do TEMPO adoptadas no Segundo e no Primeiro HAREM. Em forma de um balanço final, apresentamos algumas sugestões de melhoria no tratamento do tempo num próximo HAREM.

1 Introdução

A extracção automática de expressões temporais consiste na identificação e classificação de expressões que ajudam a localizar temporalmente os eventos descritos num texto. A primeira avaliação conjunta que se dedicou a avaliar sistemas que reconhecem este tipo de expressões foi a MUC (Message Understanding Conference) para textos em inglês. Inicialmente, o reconhecimento de expressões temporais estava integrado no reconhecimento de eventos (uma das informações que os sistemas precisavam de fornecer sobre o evento era o momento da ocorrência), e mais tarde, a partir da MUC-6 (Grishman e Sundheim, 1996), a sexta edição dessa avaliação, passou a ser uma sub-tarefa da nova tarefa de reconhecimento de entidades mencionadas. Essa opção pode ser discutível (veja-se, por exemplo, Hagège, Baptista e Mamede (2010), que argumentam a favor da separação das duas tarefas), mas dado que (i) várias dessas expressões são constituídas por maiúsculas e (ii) a intersecção do conjunto destas expressões e das entidades mencionadas não é vazio, ter as tarefas em conjunto não é completamente indefensável, uma vez que para poder reconhecer umas é preciso saber excluir as outras.

Assim, para a língua portuguesa, a Linguateca, tendo como uma das linhas directoras a promoção de avaliações conjuntas em diversas áreas do processamento computacional do português (Santos e Rocha, 2003), organizou o HAREM.

O HAREM é, então, uma avaliação conjunta de sistemas de reconhecimento de entidades mencionadas em português. A primeira iniciativa desta avalição, o Primeiro HAREM, consistiu em dois eventos: um que decorreu entre 2004 e 2005, e outro, designado Mini-HAREM, que decorreu no ano de 2006 e que teve por objectivo medir o progresso dos sistemas participantes no evento anterior[1]. A segunda edição desta avaliação, o Segundo HAREM, teve início em Setembro de 2007, culminando com o Encontro do Segundo HAREM, um ano depois.

O Segundo HAREM, no entanto, foi uma avaliação mais abrangente do que a anterior, que não só corrigiu e aperfeiçoou algumas arestas em relação ao Primeiro, como incluiu duas novas pistas, concretamente o reconhecimento e normalização de expressões temporais (Hagège, Baptista e Mamede, 2008b; Hagège, Baptista e Mamede, 2008a) e a detecção de relações semânticas entre entidades mencionadas (EM), o ReRelEM (Freitas et al., 2008b; Freitas et al., 2008a). A avaliação do reconhecimento de entidades mencionadas excluindo as da categoria TEMPO, que têm no Segundo HAREM uma pista exclusivamente a elas dedicada, foi então designada HAREM clássico[2].

[1] Apesar de as directivas terem sofrido de um evento para o outro pequenas alterações em algumas categorias (consulte-se Santos e Cardoso (2007) para mais informação sobre os dois eventos), iremos tratar indistintamente os dois eventos do Primeiro HAREM, a não ser nos casos específicos em que importa distingui-los, uma vez que a categoria TEMPO, sobre a qual este artigo se debruça, não foi afectada por alterações entre os dois eventos do Primeiro HAREM.

[2] Na verdade, a avaliação feita aos sistemas no HAREM clássico incluiu também as entidades com a categoria TEMPO sem ter em conta os atributos de normalização, mas neste artigo quando nos referimos ao HAREM clássico é para designar apenas a proposta de avaliação das restan-

Se em relação ao HAREM clássico e ao ReRelEM foi mantida a filosofia subjacente ao Primeiro HAREM, nomeadamente o modelo semântico (Santos, 2007) e modelo geral de avaliação (Santos, Cardoso e Seco, 2006), o mesmo não aconteceu com a pista do TEMPO, cuja proposta assentou numa abordagem diferente da da organização do HAREM ao problema do reconhecimento de entidades temporais. Saliente-se, a este respeito, que (i) embora o HAREM se tenha inspirado inicialmente na MUC, estas duas avaliações seguiram abordagens diferentes, como é discutido em detalhe em Seco (2007), e (ii) a proposta de avaliação do TEMPO no Segundo HAREM foi inspirada na TimeML (consulte-se, por exemplo, Pustejovsky et al. (2003)), uma avaliação dedicada exclusivamente ao reconhecimento de entidades temporais, eventos e relações entre ambos.

Neste artigo, discutimos então a proposta de reconhecimento e normalização das expressões temporais no âmbito do Segundo HAREM, procurando colocá-la em confronto com a proposta de classificação destas expressões, implementada na primeira edição desta avaliação conjunta. Começaremos por dar uma panorâmica da avaliação do TEMPO nas duas edições do HAREM; depois, faremos uma análise crítica contrastiva entre as duas propostas, destacando primeiro os aspectos que consideramos positivos na nova proposta e, em seguida, quais as características importantes do Primeiro HAREM que se perderam. Ilustraremos igualmente o impacto da aplicação das novas directivas no reconhecimento das expressões temporais anotadas no Primeiro HAREM, bem como a situação inversa de aplicação das directivas do Primeiro HAREM à colecção dourada do Segundo HAREM. Em forma de um balanço final, sumarizamos, na secção 6, algumas sugestões de melhorias no tratamento do TEMPO num próximo HAREM.

2 *Panorâmica sobre o tempo no HAREM*

No Primeiro HAREM, a proposta de análise de entidades temporais encontrava-se integrada com a das restantes entidades e totalmente a cargo da Linguateca. Porém, no Segundo HAREM, a proposta de reconhecimento de entidades temporais, juntamente com a nova tarefa de normalização dessas expressões, constituiu uma pista independente, proposta por um dos grupos participantes (Hagège, Baptista e Mamede, 2008b). Os proponentes da nova proposta pretendiam completar, enriquecer e alargar a definição desta categoria tal como proposta no Primeiro HAREM (cf. (Cardoso e Santos, 2007)), a uma noção mais geral de expressão temporal (Hagège, Baptista e Mamede, 2008b). Convém, no entanto, salientar que a implementação (criação da colecção dourada e desenvolvimento dos programas de avaliação) foi levada a cabo pela Linguateca. Tal como discutido em Mota et al. (2008a), não ter sido um mesmo grupo responsável pela proposta e implementação da avaliação do TEMPO levantou algumas dificuldades, que não serão aqui discutidas.

Uma das principais diferenças entre a prosposta de tratamento do TEMPO no Primeiro e no Segundo HAREM diz respeito aos critérios de identificação utilizados no reconhecimento de EM, o que leva a que o próprio conceito de entidade possa diferir (e difere, efectivamente na maioria dos casos) nessas propostas.

Assim, de modo a não deixar de fora expressões como, por exemplo, *ontem*, *depois de amanhã*, *há muitos anos atrás*, no Segundo HAREM as entidades temporais não ficaram limitadas a ter de conter maiúsculas[3] ou números. Como consequência natural, o total de entidades anotadas como TEMPO no Segundo HAREM cresceu substancialmente. A Tabela 1[4] mostra que o número de entidades com essa categoria aumentou cerca de 50% (806 entidades temporais no Primeiro HAREM e 1200 no Segundo HAREM), embora o total de entidades no Segundo HAREM até tenha decrescido.

Além disso, a classificação também sofreu alterações. Como se pode observar na Tabela 2, as categorias DATA e HORA foram consideradas tipos da nova categoria TEMPO_CALEND, deixou de se considerar as categorias PERIODO e CICLICO, e passaram a existir as categorias DURACAO, FREQUENCIA e GENERICO (na secção seguinte iremos detalhar melhor as noções que estas classificações tentam cobrir).

Ilustramos, com os exemplos (1) e (2), o formato das anotações para a categoria TEMPO no Primeiro e Segundo HAREM, respectivamente. Estes exemplos servem igualmente para ilustrar que no Segundo HAREM, as expressões temporais são reconhecidas e classificadas essencialmente com base num conjunto de critérios linguísticos frequentemente utilizados para iden-

tes categorias.

[3]Notamos, no entanto, que, por exemplo, a anotação dos meses do ano em português do Brasil, que são escritos em minúsculas, estava prevista nas directivas do Primeiro HAREM

[4]Estes valores foram calculados a partir das colecções douradas do Primeiro HAREM com as entidades anotadas no formato do Segundo HAREM, as quais foram disponibilizadas como material de treino no Segundo HAREM.

	CDPH	(CDPH-pe; CDPH-mh)	CDSH
Total de entidades	8734	(5065; 3669)	7845
Total de entidades tempo	806	(441; 365)	1200
Percentagem de entidades tempo	9,23%	(8,71%; 9,95%)	15,30%

Tabela 1: Total de entidades classificadas como TEMPO na colecção dourada do primeiro evento do Primeiro HAREM (CDPH-pe), na CD do Mini-HAREM do Primeiro HAREM (CDPH-mh) e na CD do Segundo HAREM (CDSH); CDPH reune as duas colecções do Primeiro HAREM.

	TIPO	SUBTIPO	Exemplo
PH	DATA	-	*Nasci em Angola, a **2 de Junho de 1968**.*
	HORA	-	*Chegamos a Presidente Figueiredo às **19h30***
	PERIODO	-	*actividades ligadas à campanha eleitoral de **Setembro***
	CICLICO	-	*quando vinha o padre benzer na **Páscoa***
SH	TEMPO_CALEND	DATA	*rocha ornamental utilizada **desde antes de 7000 a.C.***
		HORA	*o impacto dos destroços teria ocorrido **às 18h06***
	DURACAO	-	*tocando um vinil **por mais de um minuto***
	FREQUENCIA	-	*vinha **três vezes por semana**, dar-me lição de alemão*
	GENERICO	-	*percorreu **todo o século vinte***

Tabela 2: Classificação do TEMPO no Primeiro HAREM (PH) e no Segundo HAREM (SH).

tificação de constituintes sintácticos, em particular, grupos nominais com valor temporal (eventualmente antecedidos de preposição), e, portanto, os limites das entidades temporais foram alargados para incluir a preposição e outros modificadores que com elas formam o complemento adverbial de tempo. Veja-se que, por exemplo, a anotação do Primeiro HAREM delimita **1880** enquanto a do Segundo HAREM delimita **A partir de 1880**.

(1) A partir de <TEMPO TIPO="DATA≫ **1880** </TEMPO> ensinou psicologia e filosofia em Harvard, universidade que abandonou em <TEMPO TIPO="DATA≫**1907**</TEMPO>, proferindo conferências nas universidades de Columbia e Oxford. Morreu em Chocorua, New Hampshire, a <TEMPO TIPO="DATA≫**26 de Agosto de 1910**</TEMPO>.

(2) <EM ID="12" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+1880----T----E-- LMP≫**A partir de 1880**</EM> ensinou psicologia e filosofia em Harvard, universidade que abandonou <EM ID="14" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+1907----T----E-- LM-≫**em 1907**</EM>, proferindo conferências nas universidades de Columbia e Oxford. Morreu em Chocorua, New Hampshire, <EM ID="19" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+19100826T----E-- LM-≫**a 26 de Agosto de 1910**</EM>.

3 O tempo no bom caminho

No Primeiro HAREM, um dos critérios de base para a identificação de uma dada palavra ou expressão como potencial entidade mencionada prendia-se com o uso de maiúsculas ou algarismos. Neste contexto, apesar de as expressões abaixo assinaladas serem semanticamente equivalentes, apenas a expressão ilustrada em (3) seria reconhecida, dado que a ilustrada em (4) não obedece a nenhum dos requisitos formais explicitados naquelas directivas.

(3) Essa lei entrou em vigor a **1 de Janeiro de 2008**

(4) Essa lei entrou em vigor no **primeiro dia do ano de dois mil e oito**

Este critério formal exclui, pois, à partida, um número significativo de expressões temporais que seria igualmente importante reconhecer no âmbito da extracção de informação temporal de um dado texto. Neste sentido, parece inquestionável que a adopção das novas directivas do TEMPO, que não impõem tal requisito, permite um reconhecimento mais alargado (e, na nossa opinião, completamente justificado) das expressões temporais. De acordo com as directivas do Segundo HAREM, as duas expressões exemplificadas em (5) e (6) seriam, portanto, adequadamente reconhecidas e classificadas de forma idêntica:

(5) a 1 de Janeiro de 2008 [TEMPO TEMPO_CALEND DATA]

(6) no primeiro dia do ano de dois mil e oito [TEMPO TEMPO_CALEND DATA]

Além da melhoria em termos de cobertura, as novas directivas possibilitaram também, em alguns casos, uma melhor definição das próprias expressões temporais. Referimo-nos, em particular, ao alargamento da extensão da noção de entidade aos seus eventuais quantificadores e modificadores. Por exemplo, de acordo com o Primeiro HAREM, apenas a data (*2009*) deveria ser reconhecida como EM na construção ilustrada em (7).

(7) O Tratado foi ratificado antes de **2009**.

Se procurássemos estabelecer/compreender a relação que existe entre esssa data e a entidade a que a mesma se refere[5], estaríamos a representar/inferir uma relação errada, uma vez que o tratado em questão foi ratificado não em 2009, mas num ano anterior àquele (uma informação veiculada por todo o complemento adverbial, isto é, pela locução prepositiva *antes de* e a data, propriamente dita).

No Segundo HAREM, a EM temporal (data e respectivo modificador/localizador temporal) passaria, neste caso em concreto, a ser integralmente reconhecida.

Com o alargamento do reconhecimento a toda a expressão temporal, a proposta do tempo levou também à reclassificação de certas EM que de outra forma seriam classificadas como VALOR QUANTIDADE. Considere-se o seguinte exemplo:

(8) Isso aconteceu há **3 anos** [`VALOR QUANTIDADE`]

No âmbito do primeiro HAREM, apenas a expressão *3 anos* seria identificada como EM, recebendo a classificação de VALOR QUANTIDADE (a qual não capta a referência temporal de toda a expressão). No âmbito da nova proposta, todo o complemento temporal (*há 3 anos*) seria reconhecido como EM, ao qual seria atribuído a subclassificação de `TEMPO_CALEND`.

(9) Isso aconteceu **há 3 anos** [`TEMPO TEMPO_CALEND DATA`]

Um outro aspecto que consideramos positivo no âmbito da nova especificação das EM temporais diz respeito ao tratamento conferido às expressões temporais que representam um intervalo específico de tempo. Considere-se o exemplo ilustrado em (10).

(10) A edição deste ano do Festival de Paredes de Coura decorre **entre 12 e 15 de Agosto**.

[5]De facto, no ReRelEM estava previsto o reconhecimento de relações entre entidades com a categoria TEMPO e entidades com as categorias OBRA, ACONTECIMENTO e PESSOA (ver Tabela 4.4 em Freitas et al. (2008b)

No Primeiro HAREM, seriam reconhecidas duas EM temporais independentes, *12* e *15 de Agosto*, o que levaria ao desmembramento da unidade sintáctico-semântica e consequente perda de informação veiculada por esta expressão. De facto, o significado individual de *12* e *15 de Agosto* não é idêntico ao significado de todo o constituinte, que também engloba, implicitamente, os dias 13 e 14.

Este tipo de expressão passou, pois, a ser tratado como uma única EM (do tipo INTERVALO) no âmbito da nova proposta de avaliação, tendo, aliás, motivado a decisão de considerar nas directivas do HAREM clássico também como entidades (com a categoria VALOR) os intervalos de valores, e respectivos especificadores (como ilustrado em (11) e (12)).

(11) Ele saltou **entre 7 a 10 metros** na sua fuga. [`VALOR QUANTIDADE`]

(12) O bilhete custa **mais de 5 euros** [`VALOR MOEDA`]

Ao nível da classificação, a nova proposta do TEMPO também é mais abrangente do que a anterior, ao considerar dois novos tipos de expressões temporais, os quais se encontram associados à noções de frequência e duração temporal (cf. (13) e (14), respectivamente).

(13) Costuma ir ao cinema **3 vezes por mês** [`TEMPO FREQUENCIA`]

(14) Demorou **3 anos** a escrever o livro [`TEMPO DURACAO`]

A proposta de análise das expressões temporais no Segundo HAREM introduziu a tarefa de normalização de expressões temporais. Embora consideremos que o novo desafio é importante para a análise temporal de um texto, não o iremos discutir por nos parecer uma tarefa que pode ser feita em paralelo ou posteriormente à do reconhecimento das entidades mencionadas, não fazendo, portanto, parte integrante do mesmo.

4 O tempo a regredir

A filosofia do Primeiro HAREM, e do HAREM em geral proposta pela Linguateca, assenta numa abordagem de baixo para cima ao reconhecimento de entidades mencionadas, partindo da análise de textos para determinar quais as entidades de interesse e propor uma classificação das mesmas com base no contexto semântico em que se encontram.

Na nova proposta do TEMPO, a noção de EM temporal é definida essencialmente com base num conjunto de critérios sintácticos e/ou formais, aproximando-se, em muitos casos, mais de uma unidade sintáctica do que lexical.

(15) Ficámos de nos encontrar **às três da tarde**

De facto, como ilustra o exemplo (15), a expressão destacada não corresponde a uma unidade lexical, mas a todo o complemento requerido pelo verbo *encontrar*. Porém, nem sempre assim é. Observa-se que, um número considerável de expressões que cabem na definição de EM temporal não corresponde a uma unidade linguística coesa, nem do ponto de vista lexical nem sintáctico-semântico. Um conjunto de restrições descritas ou ilustradas nas directivas leva, por exemplo, a que se considere que qualquer preposição que introduza um dado núcleo temporal faça parte integrante da expressão temporal, mesmo nos casos em que essa preposição nada tem a ver com o complemento temporal, servindo apenas de elemento de ligação entre esse complemento e o predicador (verbal, nominal ou adjectival) que a seleccionou, como acontece em (17), por oposição a (16) (e ao exemplo (15), ilustrado antes) .

(16) **À noite**, vou ao cinema

(17) Isso remonta **aos anos 80**

Os proponentes justificaram esta opção argumentando que os sistemas teriam dificuldade em distinguir os dois casos. Sendo uma simplificação, defenderam naturalmente a separação dos dois casos numa futura avaliação. Consideramos, no entanto, que este critério da simplificação leva a que se tenha uma colecção dourada com incorrecções do ponto de vista sintático e semântico, embora esteja conforme às directivas, o que não deveria acontecer.

No que respeita aos modificadores do núcleo de tempo (adjectivos, orações relativas ou sintagmas preposicionais), adoptou-se uma abordagem exactamente contrária à anteriormente mencionada. De facto, neste caso, a instrução geral das directivas é para não incluir os modificadores, mesmo que sejam modificadores obrigatórios do núcleo temporal (i.e., mesmo que sejam requeridos por esse nome), como exemplificado em (18). Um tratamento diferente é, no entanto, conferido aos modificadores adjectivais iniciados por maíusculas, como ilustrado em (19).

(18) Os poetas **do período** barroco

(19) Os poetas **do período Barroco**

Muito embora os restantes modificadores não possam, como referimos antes, ser identificados como fazendo parte da EM temporal, eles constituem um requisito para a identificação de certas expressões como entidades mencionadas. Observem-se, a título ilustrativo, os seguintes exemplos:

(20) **Nessa altura**, ele não estava consciente disso

(21) **Na altura do nascimento do filho**, ele não estava consciente disso

(22) **Na altura**, ele não estava consciente disso

De acordo com as novas directivas, apenas as expressões assinaladas em (20) e (21) devem ser consideradas EM, uma vez que o núcleo nominal *altura* é, respectivamente, especificado por um demonstrativo ou acompanhado de um modificador, neste caso, de tipo preposicional (*do nascimento do filho*), que, como dissemos, não fará, no entanto, parte da EM. No entanto, apesar de (20) e (22) serem construções quasi-equivalentes, a expressão temporal presente nesta última construção não deverá, de acordo com as directivas, ser reconhecida como EM, uma vez que não obedece a nenhuma das restrições anteriormente explicitadas.

A falta de sistematicidade no tratamento destes casos é, pois, algo que, na nossa perspectiva, deverá ser revisto numa futura edição de avaliação destas expressões temporais.

As questões anteriormente apontadas colocam-se fundamentalmente ao nível da identifição das EM. No que respeita à classificação propriamente dita das EM temporais, consideramos que, embora tenha havido uma maior especificação em relação à classificação de algumas EM, perderam-se também algumas distinções contempladas já no âmbito do Primeiro HAREM, que consideramos importantes. É, por exemplo, o caso da noção de PERIODO (anteriormente entendido como um período de tempo contínuo, não repetido), ilustrada nos exemplos (23) e (24) a seguir apresentados .

(23) Vou a Londres no próximo **Inverno** [`TEMPO PERIODO`]

(24) Nos **anos 80**, surgiram centenas de novas bandas musicais. [`TEMPO PERIODO`]

Actualmente, excepto nos casos em que esse período de tempo aparece expresso no texto através de dois limites temporais explícitos (que indicam, respectivamente, o início e o fim desse período), como acontece em (10), as restantes referências a períodos de tempo deixaram de ser classificadas como tal no âmbito desta avaliação. Os exemplos acima seriam, de acordo com as directivas do Segundo HAREM, anotados especificamente como datas.

(25) Vou a Londres **no próximo Inverno** [`TEMPO TEMPO_CALEND DATA`]

(26) **Nos anos 80**, surgiram centenas de novas bandas musicais [TEMPO TEMPO_CALEND DATA]

A solução adoptada, em que se valoriza a forma em detrimento da semântica das expressões, origina incongruência na análise de expressões, formalmente distintas mas semanticamente equivalentes, como, por exemplo, as ilustradas abaixo:

(27) **Entre 2006 e 2008** foram registados centenas de acidentes de viação [TEMPO TEMPO_CALEND INTERVALO]

(28) **Nos dois últimos anos** foram registados centenas de acidentes de viação [TEMPO TEMPO_CALEND DATA]

Em (27), como existem duas referências temporais explícitas, a entidade é marcada com o subtipo INTERVALO; a ausência das fronteiras explícitas desse intervalo de tempo, em (28), leva a que a EM em questão seja classificada com o subtipo DATA.

Por outro lado, o inverso também pode acontecer. Com as novas directivas do TEMPO, uma mesma expressão temporal que tenha sentidos diferentes, por se encontrar em contextos diferentes, pode ser anotada do mesmo modo. Por exemplo, a expressão *entre 12 e 15 de Agosto* no exemplo (10) e no exemplo (29) não referencia a mesma entidade temporal.

(29) O exame será realizado **entre 12 e 15 Agosto**.

Repare-se que em (10) o festival decorre em cada um dos dias expresso pelo intervalo, mas em (29) o exame só tem lugar num desses dias. Portanto neste último caso, a expressão está a ser usada para referir uma entidade que é uma data, embora formalmente não o seja.

A noção de tempo cíclico (abrangida pelo tipo CICLICO), que servia no Primeiro HAREM para representar períodos ou datas recorrentes / que se repetem no tempo, como é o caso das EM abaixo, também deixou de existir no novo modelo de classificação temporal.

(30) Costumo viajar na **Páscoa** [TEMPO CICLICO]

(31) A restauração da independência da República comemora-se a **1 de Dezembro** [TEMPO CICLICO]

No novo modelo de classificação, ambas as entidades *na Páscoa* e *a 1 de Dezembro* passariam a ser identificadas como datas. Na verdade, as expressões temporais abrangidas, no Primeiro HAREM, pelos tipos PERIODO ou CICLICO, são, na nova edição de avaliação, geralmente classificadas com o tipo DATA (como anteriormente ilustrado) ou ainda com o tipo GENERICO, como ilustrado em (32).

(32) Vários modelos inspirados no **século 18** [TEMPO PERIODO]

A classificação dessas expressões com os subtipos DATA ou GENERICO depende exclusivamente do critério geral de identificação e classificação das expressões temporais adoptado nas directivas do TEMPO no Segundo HAREM e que consiste na verificação de que a expressão temporal em questão, pode, ou não, responder em contexto à interrogativas "PREP quando?". Essa interrogativa é produtiva, por exemplo, em (30) e (31), o que leva a que as EM aí presentes sejam classificadas como TEMPO_CALEND DATA. Pelo contrário, essa interrogativa parece marginal ou mesmo inaceitável em (32), o que leva a que a EM presente nessa construção seja classificada como GENERICO.

Relativamente ao reconhecimento de expressões temporais com o tipo GENERICO, vale a pena ainda acrescentar que julgamos que muitas vezes estas expressões não representam/mencionam de facto entidades temporais. Compare-se, por exemplo, a classificação de *Natal* no Primeiro e Segundo HAREM na seguinte frase:

(33) Domingos Afonso, na maré do **Natal**, dava a todos os pobres um quilo de bacalhau [ABSTRACCAO ESCOLA]

(34) Domingos Afonso, na maré **do Natal**, dava a todos os pobres um quilo de bacalhau [TEMPO GENERICO]

De facto, *Natal* no contexto em causa, é uma referência ao *espírito natalício* e aos costumes da *época natalícia*, não sendo propriamente um locativo temporal: não é esta expressão que localiza temporalmente o evento referido na frase (a oferta do quilo de bacalhau), mas sim a ocorrência de *a noite de Consoada* no contexto anterior (não ilustrado)[6]. A classificação adequada desta expressão como ABSTRACCAO, em vez de TEMPO, fica mais clara no seguinte exemplo:

(35) Em Abril, Domingos Afonso, na maré do **Natal**, dava a todos os pobres um quilo de bacalhau [ABSTRACCAO ESCOLA]

[6]Destaque-se, aliás, que *Consoada* no Primeiro HAREM foi classificada como CICLICO (uma vez que está a representar várias noites de Consoada, em vez de uma Consoada em particular) e que no Segundo HAREM *a noite de Consoada* foi classificado como GENERICO (por, no contexto, não responder de forma adequada a uma pergunta com a estrutura "PREP quando?").

Finalmente, um aspecto que consideramos muito positivo no modelo de classificação do HAREM (Primeiro HAREM e HAREM clássico), a vagueza na classificação das entidades mencionadas, foi ignorado nas novas directivas do TEMPO[7].

Por exemplo, expressões que, num dado contexto, podem ser referências quer a um período (ou intervalo de tempo) quer a uma data, e cuja vagueza era possível de representar no antigo modelo de classificação, passaram a ter uma única classificação no âmbito da nova proposta.

(36) O mês de **Outubro** é marcado pela passagem para o euro de alguns produtos e serviços BPI.[TEMPO|TEMPO PERIODO|DATA]

Veja-se o caso de *Outubro*, ilustrado em (36), que tanto pode ser uma referência a um período, se a passagem ao euro foi sendo feita ao longo do mês, como a uma data, se a passagem ao euro se deu apenas num dia desse mês. No Primeiro HAREM, foi-lhe associados dois tipos (PERIODO e DATA). No Segundo HAREM, pelo contrário, seria necessário optar por anotar este caso (que incluiria *o mês de*), de uma de duas maneiras: (i) como GENERICO, porque que a expressão sintacticamente não corresponde a um complemento adverbial de tempo, e, consequentemente, não é possível formular com base na sintaxe da frase a pergunta "PREP quando" que a teria como resposta; (ii) como TEMPO_CALEND DATA, porque mesmo assim, com base na semântica da frase, é possível formular a pergunta *quando é que se deu a passagem para o euro de alguns produtos e serviços BPI?* e se trata formalmente de uma data.

5 *Troca de directivas entre as duas edições do HAREM*

De forma a tipificar e avaliar as diferenças no tratamento das expressões temporais em cada uma das edições do HAREM, procedemos a dois exercícios distintos, mas complementares, tomando como referência as directivas do TEMPO adoptadas no Primeiro e no Segundo HAREM.

O primeiro, cujos resultados descrevemos em 5.1, consistiu em verificar quais as entidades da CD do primeiro evento do Primeiro HAREM, classificadas como TEMPO de acordo com as directivas do Primeiro HAREM, que sofreriam alterações, tanto ao nível da segmentação como da classificação, se fossem analisadas de acordo com as directivas adoptadas no Segundo HAREM.

[7]Isso talvez seja uma consequência de a análise do contexto ser muito local, o que advém dos critérios de delimitação utilizados serem essencialmente lexico-sintácticos.

O segundo exercício, cujos resultados apresentamos em 5.2, consistiu no processo inverso ao anteriormente descrito. Em concreto, analisámos as entidades classificadas como TEMPO na CD do Segundo HAREM, por forma a verificar quais as alterações que essas EM sofreriam se tivéssemos utilizado, em vez das directivas do Segundo HAREM, as directivas do Primeiro HAREM.

Daqui resultaram novas colecções douradas que serão disponibilizadas no sítio da Linguateca dedicado ao HAREM (`http://www.linguateca.pt/HAREM/`).

As tabelas 3 e 5.2 sumarizam as principais modificações observadas, em cada um dos casos, respectivamente. Para simplificar a tabela usámos apenas o tipo ou subtipo mais específico, não mostrando a categoria nem o tipo (caso exista o subtipo). Relembramos então que:

Na classificação do Primeiro HAREM:

- DATA, HORA, CICLICO e PERIODO são tipos da categoria TEMPO
- EFEMERIDE e ORGANIZADO são tipos da categoria ACONTECIMENTO
- QUANTIDADE é um tipo da categoria VALOR
- PUBLICACAO é um tipo da categoria OBRA
- IDEIA é um tipo da categoria ABSTRACCAO

Na classificação do Segundo HAREM:

- GENERICO, DURACAO e FREQUENCIA são tipos da categoria TEMPO
- DATA, HORA e INTERVALO são subtipos do tipo TEMPO_CALEND, que é um tipo da categoria TEMPO
- EFEMERIDE é um tipo da categoria ACONTECIMENTO

Para efeitos de contabilização nas tabelas, considerámos que:

- a segmentação era diferente de uma edição para outra do HAREM quando de uma entidade se passasse a ter duas ou o contrário;
- a entidade seria alargada se passasse a incluir palavras em minúsculas, excluindo os casos em que a palavra é uma preposição;
- a entidade seria mais curta se deixasse de incluir palavras em minúsculas, excluindo os casos em que a palavra é uma preposição;
- a classificação TEMPO DATA do Primeiro HAREM e a classificação TEMPO TEMPO_CALEND DATA são a mesma classificação;

- a classificação TEMPO HORA do Primeiro HAREM e a classificação TEMPO TEMPO_CALEND HORA são a mesma classificação.

Falaremos dessas experiências com mais pormenor, em seguida.

5.1 Análise das entidades classificadas como TEMPO no Primeiro HAREM de acordo com as directivas do Segundo HAREM

Para efeitos deste exercício, tomámos como ponto de partida a CD usada no primeiro evento do Primeiro HAREM (Mota et al., 2008b; Mota et al., 2008a; Carvalho et al., 2008), a qual é constituída por 129 documentos (distribuídos por oito géneros: web, jornalístico, entrevista, expositivo, correio electrónico, literário, político e técnico), que compreendem, no seu conjunto, um total de 5.065 EM (as quais foram anotadas de acordo com as directivas estabelecidas no âmbito do Primeiro HAREM). Dessas EM, 441 (8.7%) encontram-se associadas à categoria TEMPO.

Ao analisarmos aquele subconjunto de EM[8] com as directivas do TEMPO do Segundo HAREM, verificámos que:

- 19% das EM (81 entidades) não sofreriam quaisquer alterações;
- 77% das EM (334 entidades) passariam a ser segmentadas de forma diferente, nomeadamente devido ou a) à inclusão da preposição que introduz a expressão temporal (223 entidades), como em (37), ou b) ao alargamento da própria noção de expressão temporal (32 entidades), como em (38), ou, finalmente, c) à observação de ambas as situações (79 entidades), como em (39).

  (37) Foi constituída por escritura pública **em Junho de 1992**

  (38) Realiza-se **dia 20 de Maio**

  (39) A Evolução Da Colônia portuguesa na América, **a partir da segunda metade do século XVII**, será profundamente marcada pelo novo rumo

- 18% das EM (79 entidades) passariam a receber uma nova classificação. Mais especificamente:

  – 45 entidades EM anteriormente classificadas como DATA ou PERIODO passariam a ser classificadas como INTERVALO (cf. (40) e (41), respectivamente);

  (40) O 17º congresso mundial em Gestão de Projectos decorrerá **entre os dias 4 a 6 do Junho**

  (41) foi o ditador alemão que comandou a Alemanha nazista na Segunda Guerra Mundial (**1939-1945**)

  – 21 entidades EM anteriormente classificadas como PERIODO ou CICLICO (cf. (42) e (43), respectivamente) passariam a ser classificadas como DATA;

  (42) Eles teriam vendido **em março** acima da média

  (43) Juntamo-nos sempre **pela Páscoa**

  – 8 entidades EM anteriormente classificadas como DATA, PERIODO ou CICLICO (cf. (44), (45) e (46), respectivamente) passariam a ser classificadas como GENERICO;

  (44) e dois meses depois veio **o 24 de Abril**

  (45) Comentários Jornadas da Juventude animam **mês de Abril**

  (46) Ainda me lembro do primeiro texto que eu estive a ler. Era sobre a **Primavera**

  – 2 entidades EM anteriormente classificada como DATA ou PERIODO passaria a ser classificada como DURACAO (ver, respectivamente, exemplos (47) e (48));

  (47) Dada a escassez de tempo e a abrangência da matéria, já se firmou um compromisso de **durante 1997**

  (48) encontraram a capivara no Brasil **durante o século XVI**

  – 1 entidade EM anteriormente classificada como CICLICO passaria a ser classificada como FREQUENCIA:

  (49) Desde aquele ano tem continuado este serviço na Igreja Metodista (**primeiro Domingo de cada ano**))

  – Apenas 2 entidades EM deixariam de ser reconhecidas como TEMPO (cf. (50) e (51))

  (50) os povoadores cristãos da **Reconquista**

  (51) parecia uma mina daquela época do **Velho Oeste**

[8] Para simplificar a análise, excluímos das 441, nove entidades por fazerem parte de uma estrutura `ALT`.

Mesma delimitação e classificação	81
Mesma delimitação, mas muda a classificação:	17
- PERIODO − > DATA	4
- DATA, CICLICO ou PERIODO − > GENERICO	4
- PERIODO − > INTERVALO	7
- DATA|EFEMERIDE − > EFEMERIDE	1
- PERIODO − > IDEIA	1
Segmentação diferente:	19
- Uma entidade PERIODO passa a duas entidades DATA	1
- Uma DATA passa a duas entidades (DATA e LOCAL)	1
- Duas entidades DATA formam INTERVALO	2
- Duas entidades ou mais formam INTERVALO:	31
– introduzido por preposição	24
– introduzido por preposição e incluindo modificadores	7
Entidade seria alargada:	26
- Para incluir modificadores	2
- Para incluir outros elementos (dia, ano, ...):	24
– Com mesma classificação	20
– PERIODO − > DATA	1
– DATA ou PERIODO − > GENERICO	3
Continua a ser TEMPO, mas com preposição inicial:	223
- Com mesma classificação	211
- CICLICO ou PERIODO − > DATA	7
- DATA − > GENERICO	1
- PERIODO − > INTERVALO	2
- DATA ou PERIODO − > DURACAO	2
Continua a ser TEMPO, mas com preposição inicial e outros elementos:	48
- Com mesma classificação	39
- PERIODO − > DATA	8
- PERIODO − > INTERVALO	1

Tabela 3: Análise das entidades classificadas com TEMPO no Primeiro HAREM com as directivas do Segundo HAREM

5.2 Análise das entidades classificadas como TEMPO no Segundo HAREM de acordo com as directivas do Primeiro HAREM

No segundo exercício, tomámos como ponto de partida a CD do Segundo HAREM (Carvalho et al., 2008; Mota et al., 2008b). Este corpo é constituído por 129 documentos distribuídos por 13 géneros: notícia, didáctico, blogue jornalístico, blogue pessoal, perguntas, ensaio, opinião, blogue humorístico, legislativo, promocional, entrevista, texto privado manuscrito e perguntas faq, que compreende um total de 7.836 EM, identificadas de acordo com as directivas em vigor no Segundo HAREM. Dessas EM, 1.195 (15%) encontram-se associadas à categoria TEMPO, quase três vezes mais do que as entidades temporais compreendidas na CD do primeiro evento do Primeiro HAREM.

Ao aplicarmos as directivas do Primeiro HAREM àquele subconjunto de EM, observámos que:

- 41% das entidades temporais (491 EM) deixariam de ser reconhecidas por não conterem nem dígitos nem maiúsculas, o que constituía, como já referimos antes, um requisito essencial no Primeiro HAREM. Isso mostra a importância de reconhecer também expressões em minúsculas, pois de outra forma uma grande parte do conteúdo temporal dos textos é perdida.

  Na sua maioria (62%), essas EM pertencem à nova categoria TEMPO_CALEND (291 são DATA, 11 são HORA e 3 são INTERVALO). Nos restantes casos, as expressões exprimem valores de frequência (70 EM) ou de duração (46 EM), não previstos nas directivas do Primeiro HAREM, ou, em vez disso, encontram-se associadas à categoria GENERICO (65 casos), que constitui igualmente uma novidade nas novas directivas do TEMPO. Tal como é referido nas directivas, estas últimas expressões, muito embora contenham uma unidade de tempo,

Não seria entidade no Primeiro HAREM	491
Mesma delimitação e classificação	53
Mesma delimitação, mas muda a classificação:	25
- GENERICO, DATA ou INTERVALO − > PERIODO	19
- DATA − > CICLICO	1
- DURACAO|QUANTIDADE ou FREQUENCIA − > QUANTIDADE	3
- EFEMERIDE|GENERICO − > EFEMERIDE	2
Segmentação diferente:	69
- Em duas entidades e mantém-se a classificação	1
- INTERVALO passa a duas entidades DATA	36
- INTERVALO passa a duas entidades HORA	2
- DURACAO ou QUANTIDADE|INTERVALO passa a duas entidades QUANTIDADE	2
- DATA faz parte de ORGANIZADO	19
- GENERICO ou DATA faz parte de EFEMERIDE	2
- DATA faz parte de PUBLICACAO	5
- INTERVALO passa a uma entidade PERIODO, e a outra não reconhecida	2
Entidade seria mais curta:	15
- Com mesma classificação	5
- GENERICO − > PERIODO	4
- DATA, DURACAO, GENERICO ou DURACAO|QUANTIDADE − > QUANTIDADE	5
- GENERICO − > EFEMERIDE	1
Continua a ser TEMPO, mas sem preposição inicial e sem outros elementos:	527
- Com mesma classificação	441
- DATA, DURACAO, GENERICO, ou INTERVALO − > PERIODO	81
- DATA ou GENERICO − > CICLICO	5
Deixa de ser TEMPO, sem preposição inicial e sem outros elementos:	15
- DATA, DURACAO, GENERICO, DATA|DURACAO ou DURACAO|QUANTIDADE − > QUANTIDADE	9
- DATA ou GENERICO − > EFEMERIDE	6

Tabela 4: Análise das entidades classificadas com TEMPO no Segundo HAREM com as directivas do Primeiro HAREM

não representam um tempo de calendário específico no contexto em questão. Estes casos encontram-se ilustrados, respectivamente, em (52), (53) e (54). Cinco casos estão ainda marcados como vagos entre DURACAO e GENERICO, DURACAO e DATA (3 deles), e DATA e GENERICO.

(52) **A maior parte das vezes** Mills fala de si mesmo como o originador da mensagem[FREQUENCIA]

(53) **Durante muitos anos** o fotógrafo trabalhou nas agências Sygma e Gamma [DURACAO]

(54) não há nada como **o outono** [GENERICO]

- 48% seriam reconhecidas como entidades temporais, mas seriam segmentadas de forma distinta (577 casos), quer por deixarem de fora da EM a preposição que as introduz (449 casos), quer por deixarem de incluir unidades lexicais - especificadores, modificadores, etc. (em minúsculas), que antes não seriam contempladas (9 casos), quer pela observação de ambas as situações (78 casos). Nos restantes 41 casos, a entidade seria segmentada em duas entidades temporais, deixando de fora outros elementos em mínúsculas (como a preposição inicial, por exemplo). Este último caso acontece sobretudo com entidades associadas à noção de intervalo (38 casos), que seriam fragmentadas e classificadas como duas EM independentes, pertencentes ao tipo DATA ou HORA. Cada um dos casos é ilustrado, respectivamente, em (55), (56), (57) e (58).

(55) O que nos é dito é que **em** [**Janeiro**] tudo vai mudar

(56) foi inaugurada **dia** [**9 de setembro de 2004**]

(57) deverá situar-se nos 407,4 euros (..) **a partir de** [**Janeiro**] **do próximo ano**

(58) Que acontenceu na Argélia **na noite de** [**17**] **para** [**18 de Agosto de 1994**]

De notar, no entanto, que a classificação dessas entidades não seria a mesma em 22,5% dos casos, tal como se ilustra de (59) a (62).

(59) assistimos **no** [**século XVI**] ao fermentar de um enorme debate [DATA -> PERIODO]

(60) evento que acontece sempre **dia** [**06 de janeiro**] [DATA -> CICLICO]

(61) Mas a festa vai continuar **ao longo do ano de** [**2008**] [DURACAO -> PERIODO]

(62) porque só funciona **das** [**08h00**] **às** [**09h00**] [INTERVALO -> HORA]

- Apenas 6% das entidades anotadas no Segundo HAREM como TEMPO (73 entidades) continuariam a ser TEMPO com a mesma delimitação, mas, dessas, 20 passariam a ser classificadas com um tipo diferente, como em (63) e (64). De referir, no entanto, que esta percentagem ascenderia aos 43% se tivermos igualmente em conta as EM que diferem destas simplesmente por causa da introdução da preposição. Isto sugere que sistemas, como o da Priberam, que por opção não incluíram a preposição, poderiam ter tido resultados significativamente melhores no reconhecimento das entidades temporais, uma vez que no Segundo HAREM entidades que não estivessem bem delimitadas não contribuíam para a pontuação final do sistema.

  (63) fecha **2.ª** [DATA -> CICLICO]

  (64) Afirmando esperar um ano de 2008 mais exigente do que **2007** [DATA -> PERIODO]

- Aproximadamente 4,5% das entidades temporais do Segundo HAREM (54 entidades) passariam a ser classificadas com uma categoria que não TEMPO, das quais apenas 5 casos manteriam a mesma delimitação. Na sua maioria, estas entidades fariam parte de uma entidade maior com a categoria ACONTECIMENTO ORGANIZACAO, como se ilustra em (65). Se tivermos em conta apenas as entidades que são reconhecidas de acordo com as directivas de ambas as edições do HAREM, então cerca de 8% não seriam classificadas como TEMPO, o que acentua as diferenças de interpretação do sentido das entidades no Primeiro e no Segundo HAREM.

  (65) O [Tour de França **de 2009**] vai começar no Mónaco [TEMPO -> ACONTECIMENTO ORGANIZADO]

Nos casos em que a entidade existe tanto no Segundo HAREM como no Primeiro HAREM, mesmo que com ajustes na delimitação, verificámos ainda que:

- as entidades com o tipo GENERICO continuariam a ser entidades temporais, mas com o tipo PERIODO (9 entidades) ou CICLICO (4 entidades), ou então passariam à categoria ACONTECIMENTO EFEMERIDE (8 entidades);
- as entidades com o tipo DURACAO continuariam a ser entidades temporais com o tipo PERIODO em 2 casos mas maioritariamente passariam a ser classificadas com VALOR QUANTIDADE (12 entidades);
- a única entidade com valor de frequência a manter-se, seria com a classificação de VALOR QUANTIDADE;
- as datas continuariam a ser datas na maioria dos casos, e passariam a ser classificadas como períodos em 82 casos e como datas cíclicas em 2 casos. Em 34 casos deixariam de ser tempo para passar a ser ou ser integradas em ACONTECIMENTO EFEMERIDE (5 casos), ACONTECIMENTO ORGANIZADO (19 casos), VALOR QUANTIDADE (5 casos) e OBRA PUBLICACAO (5 casos);
- Nenhuma entidade com o tipo HORA seria reclassificada;
- Entidades do tipo INTERVALO passariam a períodos, quando a entidade não é partida em duas (no Primeiro HAREM, datas separadas por '/' ou '-' correspondiam a uma única entidade), ou a datas e horas, quando a entidade é segmentada em duas que representam os limites do intervalo (como acontece, por exemplo, a *entre Novembro e Dezembro*).

5.3 Discussão

Estes dados mostram claramente que a noção de entidade temporal não é idêntica no Primeiro e no Segundo HAREM.

As principais diferenças observadas têm sobretudo a ver com a tarefa de identificação, uma vez que a percentagem de entidades temporais que seriam delimitadas do mesmo modo quer se usem umas directivas ou outras seria baixa (22% no caso da CD do Primeiro HAREM e 11% no caso da CD do Segundo HAREM, se não contarmos as entidades que não seriam identificadas por só conterem minúsculas - se tivermos em conta

também essas, então a percentagem seria ainda mais baixa: 6%).

No entanto, também se observam diferenças na classificação. Especificamente, 18% das entidades do Primeiro HAREM deixariamm de ter a mesma classificação, enquanto no Segundo HAREM 29% das entidades que seriam também reconhecidas no Primeiro HAREM (mesmo que com ajustes na delimitação) teriam uma classificação diferente. Isso é, naturalmente, uma consequência de as categorias serem diferentes em ambas as edições do HAREM, mas também se deve ao facto de os conceitos por elas representados serem diferentes. Por esse motivo, não é possível estabelecer um mapeamento entre as categorias das duas edições. Repare-se que mesmo as entidades DATA num dos HAREM podem não ser DATA no outro.

Ao nível da classificação, um facto interessante é que as directivas do Segundo HAREM tendem a confirmar que as entidades do Primeiro HAREM são entidades temporais, pois só duas entidades é que deixariam de serem reconhecidas como referências temporais. Porém, na situação inversa é mais provável que uma entidade temporal deixe de o ser ao aplicar-se as directivas do Primeiro HAREM, pois 8% das entidades que seriam reconhecidas por ambas as directivas deixariam de ser tempo no Segundo HAREM.

Se, por um lado, se pode argumentar que com o Segundo HAREM, as entidades temporais ficaram delimitadas com maior precisão e abrangendo um maior leque de entidades, por também permitir o reconhecimento de expressões só em minúsculas (relembre-se que 41% das entidades do Segundo HAREM não existiriam no Primeiro, e que 52% passaram a ser mais bem delimitadas), pelo outro, pensamos que o Segundo HAREM ficou a perder em termos da classificação das entidades referidas pelas expressões do texto (consideramos que em cerda de metade dos casos em que houve alteração da classificação, houve perda de significado).

De facto, no Primeiro HAREM, os critérios de reconhecimento de entidades temporais são essencialmente semânticos, obrigando muitas vezes a uma leitura que vai além da frase para poder determinar a classificação. No Segundo HAREM, pelo contrário, basta um contexto muito local para atribuir a classificação, pois os critérios são de natureza essencialmente formal.

É por essa razão que uma expressão que superficialmente é uma data, no Segundo HAREM é em geral classificada como DATA (ou será GENERICO quando não verifica o critério da pergunta-resposta), mas no Primeiro HAREM poderia ser classificada como DATA, PERIODO, ou CICLICO (ou mesmo, estar integrada em entidades não elementares que referenciam entidades não temporais), mostrando que uma mesma expressão pode designar referentes temporais com propriedades distintas. Por exemplo, *Páscoa* em (66) e (67) não designa o mesmo referente temporal. No primeiro caso, **Páscoa** designa um domingo de Páscoa concreto e único (DATA), mas, no segundo caso, designa vários domingos de Páscoa (CICLICO).

(66) Numa pista em bom estado, apesar da ameaça de chuva que pairou na região durante o domingo de **Páscoa**

(67) Juntamo-nos sempre pela **Páscoa**

6 Sugestões para o futuro do tempo no HAREM

De acordo com a análise que fizemos, e de modo a preservar quer o modelo semântico subjacente ao Primeiro HAREM, quer a maior precisão na identificação das entidades temporais do Segundo HAREM, julgamos que em futuras edições do HAREM se deve ter em conta os seguintes aspectos na avaliação do TEMPO:

- não constrangimento das entidades temporais a terem apenas maiúsculas ou números;
- inclusão de todos os modificadores que façam parte da entidade temporal de forma a delimitar todo o complemento adverbial de tempo;
- reconhecimento de intervalos de tempo cujos limites estejam expressos por meio de duas datas;
- reconhecimento de frequências e de durações;
- classificação semântica mais fina, nomeadamente voltar a considerar os tipos PERIODO e CICLICO.

Agradecimentos

Este trabalho foi desenvolvido no âmbito da Linguateca, co-financiada pelo governo português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, e também financiada pela UMIC e pela FCCN. O trabalho da segunda autora foi ainda financiado pela Fundação para a Ciência e a Tecnologia através de uma bolsa de pós-doutoramento com a referência SFRH/BPD/45416/2008. A primeira autora agradece ainda o apoio que o grupo Proteus da New York University lhe tem dado en-

quanto desenvolve o seu trabalho para a Linguateca.

Estamos igualmente gratas à Diana Santos pela motivação e revisão de versões anteriores do artigo, bem como ao José Carlos Medeiros e ao Pablo Gamallo pelo seu cuidadoso trabalho de revisão que nos forneceu valiosas sugestões de melhoria, que tentámos ter em conta tanto quanto o tempo permitiu.

A bibliografia foi construída com o apoio do SUPeRB (Cabral, Santos e Costa, 2008).

Referncias

Cabral, Luís Miguel, Diana Santos, e Luís Fernando Costa. 2008. SUPeRB: Building bibliographic resources on the computational processing of Portuguese. Em Daniela Braga, Miguel Sales Dias, e António Teixeira, editores, *Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies (full proceedings)*, September 10, 2008.

Cardoso, Nuno e Diana Santos. 2007. Directivas para a identificação e classificação semântica na colecção dourada do HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 211–238, 12 de Novembro, 2007. Documento original publicado no sítio do HAREM a 29 de Março de 2006. Republicado como Relatório técnico DI-FCUL TR-06-18 : Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Novembro de 2006.

Carvalho, Paula, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas, e Cristina Mota. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 11–31, 31 de Dezembro, 2008.

Freitas, Cláudia, Diana Santos, Paula Carvalho, e Hugo Gonçalo Oliveira. 2008a. Apêndice C: ReRelEM - Reconhecimento de Relações entre Entidades Mencionadas. Segundo HAREM: proposta de nova pista. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 31 de Dezembro, 2008.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota. 2008b. Relações semânticas do ReRelEM: além das entidades no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 77–96, 31 de Dezembro, 2008.

Grishman, Ralph e Beth Sundheim. 1996. Message understanding conference-6: a brief history. Em *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pp. 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede. 2008a. Apêndice B: Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 31 de Dezembro, 2008.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede. 2008b. Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 33–54, 31 de Dezembro, 2008.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede. 2010. Caracterização e processamento de expressões temporais em português. *Linguamática*, 2(1):63–76, Abril, 2010.

Mota, Cristina, Paula Carvalho, Cláudia Freitas, Hugo Gonçalo Oliveira, e Dia. 2008a. É tempo de avaliar o TEMPO. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 55–75, 31 de Dezembro, 2008.

Mota, Cristina, Diana Santos, Paula Carvalho, Cláudia Freitas, e Hugo Gonçalo Oliveira. 2008b. Apêndice H: Apresentação detalhada das colecções do Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 355–377, 31 de Dezembro, 2008.

Pustejovsky, James, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, e Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. Em *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, 15-17 de Janeiro, 2003.

Santos, Diana. 2007. O modelo semântico usado no Primeiro HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 43–57, 12 de Novembro, 2007.

Santos, Diana e Nuno Cardoso. 2007. Breve Introdução ao HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 1–16, 12 de Novembro, 2007.

Santos, Diana, Nuno Cardoso, e Nuno Seco. 2006. Avaliação no HAREM: Métodos e medidas. Relatório Técnico DI-FCUL TR-06-17, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Novembro, 2006. `http://www.linguateca.pt/Diana/download/SantosCardosoSecoMedidas2006.pdf`.

Santos, Diana e Paulo Rocha. 2003. AvalON: uma iniciativa de avaliação conjunta para o português. Em Amália Mendes e Tiago Freitas, editores, *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)*, pp. 693–704, Lisboa, 2-4 de Outubro de 2002, 2003. APL.

Seco, Nuno, 2007. *MUC vs HAREM: a contrastive perspective*, pp. 35–41. Linguateca, 12 de Novembro, 2007.

Apresentação de Projectos

Galnet: WordNet 3.0 do galego

Xavier Gómez Guinovart, Xosé María Gómez Clemente,
Andrea González Pereira, Verónica Taboada Lorenzo

Grupo TALG, Universidade de Vigo
sli@uvigo.es

Resumo

Neste artigo presentamos o proxecto Galnet do Grupo TALG da Universidade de Vigo, dirixido á construción da versión galega do WordNet 3.0. Trátase dun proxecto que se atopa na súa fase inicial de desenvolvemento, mais do que xa se obtiveron uns primeiros resultados que están dispoñibles para a consulta. Describiremos os trazos xerais do proxecto, a metodoloxía e as ferramentas utilizadas, algúns aspectos lingüísticos do traballo e os resultados obtidos nesta primeira etapa.

1. Introdución[1]

WordNet (Fellbaum, 1998; Miller et al., 1990) é unha base de coñecementos léxicos estruturada en forma de rede semántica. Nesta rede léxico semántica, cada nó é un concepto, e os fíos que conectan estes nós son as relacións semánticas (hiponimia, meronimia...) que se establecen entre os conceptos. Cada concepto na rede está representado polo grupo de lemas sinónimos que poden expresar ese concepto. Na terminoloxía asociada a WordNet, cada grupo de sinónimos é un *synset*, e cada sinónimo que forma parte dese grupo é unha *variant* (ou variante léxica dun mesmo concepto). WordNet inclúe, ao carón de cada synset, unha breve definición distintiva (ou *glosa*) do significado compartido por todas as variantes do synset e, en certos casos, exemplos de uso das variantes en contexto.

WordNet foi orixinalmente concibido para a lingua inglesa e, aínda que hoxe existen versións do WordNet en moitas linguas, o WordNet do inglés segue sendo arestora a máis desenvolvida e a de referencia. Os traballos do WordNet para esta lingua lévanse a cabo desde 1985 na Universidade de Princeton baixo a dirección do profesor George A. Miller. Na súa versión actual, o WordNet 3.0 do inglés contén 155.287 lemas (variantes) agrupadas en 117.659 grupos de sinónimos (synsets)[2].

WordNet constitúe, sen dúbida, o recurso de semántica léxica computacional máis importante na actualidade, especialmente, no ámbito do procesamento da linguaxe natural (PLN), onde é utilizada, por exemplo, en tarefas de desambiguación semántica automática (Agirre y Edmonds, 2006), de recuperación da información (Varelas et al., 2005), de clasificación automática de textos (Elberrichi, Rahmoun, e Bentaallah, 2008) ou de resumo automático (Barzilay y Elhadad, 1997).

Na actualidade existen versións do WordNet en diversas fases de desenvolvemento para moi diversas linguas[3], incluídas o hebreo (Ordan et al., 2007), o italiano[4] (Pianta, Bentivogli, e Girardi, 2002), o xaponés[5] (Isahara et al., 2008), o castelán (Fernández y Vázquez, 2010; Fernández-Montraveta, Vázquez, e Fellbaum, 2008), o catalán (Benítez et al., 1998) e o euskera[6] (Pociello, Agirre, e Aldezabal, 2011).

A maioria das versións en linguas distintas do inglés seguen o modelo de deseño de EuroWordNet (Vossen, 2002), na que os synsets que forman parte do WordNet da lingua propia están vinculados cos synsets do resto das linguas a través dun ILI (*interlingual index* ou índice interlingüístico) que é único para cada concepto e que principalmente está baseado nos synsets do WordNet inglés de referencia. Deste xeito, o conxunto de léxicos WordNet nos distintos idiomas permiten

[1]Este traballo foi financiado polo Ministerio de Ciencia e Innovación, dentro do proxecto *Multilingual Central Repository 2.0: TALG* do Subprograma de Acciones Complementarias, 2009 (ref. FFI2009-08317-E/FILO), concedido ao Grupo TALG da Universidade de Vigo.

[2]<http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>

[3]The Global WordNet Association mantén unha listaxe dos léxicos WordNet en distintas linguas na súa páxina web <http://www.globalwordnet.org/gwa/wordnet_table.htm>

[4]Estas dúas consultables en liña en <http://multiwordnet.fbk.eu/english/home.php>

[5]Consultable en liña en <http://nlpwww.nict.go.jp/wn-ja/index.en.html>

[6]Estas tres consultables en liña en <http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl>

a conexión entre os synsets de calquera par de linguas a través do ILI, constituíndo así un recurso de gran utilidade en aplicacións das tecnoloxías lingüísticas que precisan o procesamento pluringüe da linguaxe, como a tradución automática ou a recuperación interlingüística da información. Cómpre salientar tamén que os conceptos que forman parte do ILI están catalogados en xerarquías de dominios e ontoloxías, como a xerarquía de dominios IRST (Bentivogli et al., 2004) ou as ontoloxías SUMO (Pease, Niles, e Li, 2002) e Top Concept Ontology (Álvez et al., 2008), o que permite un mellor aproveitamento do recurso en diversas aplicacións.

Neste artigo presentamos o proxecto Galnet do Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo, dirixido á construción da versión galega do WordNet 3.0. Trátase dun proxecto que se atopa na súa fase inicial de desenvolvemento, mais do que xa se obtiveron uns primeiros resultados que están dispoñibles para a consulta. Nos seguintes apartados describiremos os trazos xerais do proxecto, a metodoloxía e as ferramentas utilizadas, algúns aspectos lingüísticos do traballo e os resultados obtidos nesta primeira etapa.

2. O proxecto Galnet

O obxectivo do proxecto Galnet non é outro que a construción dun WordNet para o galego (isto é, dun Galnet) aliñado co ILI xerado a partir do WordNet 3.0 do inglés. Este proxecto está incorporado nun proxecto máis amplo encamiñado á integración coordinada das versións castelá, catalá, galega e vasca do WordNet 3.0, no que participan os grupos de investigación do IXA (da Universidade do País Vasco), TALP (Universitat Politècnica de Catalunya), GRIAL (Universitat Autònoma de Barcelona, Universitat de Barcelona, Universitat de Lleida e Universitat Oberta de Catalunya), IULATERM (Universitat Pompeu Fabra) e TALG (Universidade de Vigo).

O marco de desenvolvemento no que se integra o Galnet é o do Multilingual Central Repository (MCR) (Atserias et al., 2004), unha plataforma de libre consulta[7] desenvolvida ao abeiro do proxecto europeo Meaning (IST-2001-34460) e dos proxectos de financiamento estatal KNOW (TIN2006-15049-C03) e KNOW2 (TIN2009-14715-C04-01). O MCR abrangue na actualidade os léxicos WordNet de cinco linguas (inglés, español, catalán, vasco e galego) enlazados interlingüisticamente polo ILI correspondente ao WordNet 3.0 e cos ILI categorizados na xerarquía de dominios IRST e nas ontoloxías SUMO e Top Concept Ontology.

A seguir describiremos a metodoloxía e as ferramentas empregadas na construción do Galnet na súa primeira etapa de desenvolvemento.

2.1. Metodoloxía e ferramentas

Os obxectivos desta primeira fase na construción do Galnet foron, en primeiro lugar, elaborar un conxunto de synsets básicos para a operatividade do recurso na lingua galega e, en segundo lugar, fornecer un conxunto suficiente de entradas que servise para ilustrar a utilidade do recurso e ampliar a súa operatividade.

A metodoloxía utilizada para levar a cabo o primeiro obxectivo consistiu na creación da versión galega dos synsets nominais e verbais pertencentes aos Basic Level Concepts (BLC). Para o segundo obxectivo, elaboramos as entradas galegas para os ficheiros lexicográficos do WordNet correspondentes aos nomes relacionados coas partes do corpo e coas substancias, e para unha parte dos correspondentes aos adxectivos de tipo xeral.

Os Basic Level Concepts (Izquierdo, Suárez, e Rigau, 2007) son un conxunto seleccionado de conceptos do WordNet que representan un compromiso entre dous principios de caracterización contraditorios: representar o maior número posible de conceptos (ser conceptos abstractos) e representar o maior número posible de trazos distintivos (ser conceptos concretos). Así, os BLC aparecen tipicamente na parte media das relacións semánticas xerárquicas de WordNet, sendo deste modo frecuentes e destacados, nin claramente xerais nin demasiado específicos. A primeira tarefa do proxecto Galnet consistiu en elaborar manualmente a versión galega dos BLC (649 synsets nominais e 616 synsets verbais) recollidos no apartado freqmin20/all da distribución oficial[8] dos BLC do WordNet 3.0, sen incluír na adaptación nin as glosas nin os exemplos incluídos nos synsets correspondentes da lingua inglesa.

Unha vez elaborado o núcleo inicial de synsets do Galnet, continuamos a ampliación do recurso a partir da tradución asistida dos ficheiros lexicográficos do WordNet para os nomes relacionados coas partes do corpo e coas substancias, e para unha parte dos adxectivos de tipo xeral. A ferramenta empregada nesta tarefa foi Google Translator Toolkit[9], unha ferramenta colaborativa en liña que nos permitiu a postedición asistida das propostas de tradución automática do tradu-

[7] <http://adimen.si.ehu.es/web/MCR/>

[8] <http://adimen.si.ehu.es/web/BLC/>

[9] <http://translate.google.com/toolkit/>

tor de Google.

A selección dos ficheiros lexicográficos relacionados coas partes do corpo e coas substancias veu motivada pola nosa vontade de aproveitar o material textual e terminolóxico elaborado en traballos previos do grupo e recollidos no Corpus Técnico do Galego[10] e na base de datos terminolóxica da Termoteca[11]. A incorporación dos adxectivos xustificouse en virtude dunha maior cobertura ingüística dos resultados nesta fase inicial do traballo. No apartado seguinte, describiremos algúns dos problemas lexicográficos máis relevantes con que nos atopamos nesta xeira.

3. Tratamento da microestrutura

A microestrutura das entradas de WordNet inclúen, para cada concepto ou nó da rede, o grupo de sinónimos ou synset que lexicalizan o concepto, e unha glosa ou definición que pode ir acompañada dun exemplo. En EuroWordNet, ademais, a cada concepto lle corresponde un ILI (índice interlingüístico), un identificador único do concepto que permite relacionar os conceptos entre os léxicos en formato WordNet das distintas linguas.

3.1. Synsets

O grupo de sinónimos ou synset pode conter unha única palabra ou varias. Igualmente, os sinónimos poden ser unidades monoléxicas ou pluriléxicas. A serie de sinónimos dun synset está composta por varias palabras que entran en relación de sinonimia parcial ou total, coa premisa básica de que nalgún contexto poden ser intercambiables. WordNet non diferencia os graos da relación de semellanza, nin ten en conta, polo de agora, a connotación do rexistro. Isto provoca que non se poida diferenciar cunha etiqueta de rexistro os usos coloquiais ou vulgares que se recollen nas series sinonímicas no WordNet orixinal inglés e que foron transladados ao Galnet en versión galega. Por exemplo, no synset correspondente ao concepto con código ILI 05514410-n, a serie sinonímica en inglés {female genitalia, female genitals, female genital organ, fanny} non inclúe na serie ningunha indicación de rexistro vulgar para o último dos sinónimos, o mesmo que sucede na súa adaptación galega {xenitais femininos, órgano xenital feminino, cona}.

A fraseoloxía no Wordnet 3.0 do inglés está tratada dun modo moi irregular, aparecendo entre as variantes nalgún synset illado. Porén, no Galnet desexaríamos que a inclusión de fraseoloxía fose algo común, como un tipo máis de expresións lexicalizadas que poden entrar en relación sinonímica con outras formas léxicas. Asemade, pensamos que Galnet pode contribuír á xeneralización do uso destas expresións pluriléxicas unidas claramente a aspectos socioculturais da nosa lingua. A inclusión de unidades fraseolóxicas no Galnet é, ata certo punto, independente da súa aparición no WordNet do inglés, como se pode apreciar nos seguintes catro exemplos (onde tampouco se marca na versión galega a connotación de rexistro coloquial que sería necesaria nalgúns casos):

01041061-v: {close up, clam up, dummy up, shut up, belt up, button up, be quiet, keep mum} >{calar, calar coma un peto, calar coma unha estoa, coser a boca, pechar o bico, non dar un chío, estar en silencio}

01823149-v: {care a hang, give a hoot, give a hang, give a damn} >{importar un pataco, importar un farrapo de gaita, pesarlle tanto a alguén un ombro coma outro, non lle importar a alguén algo unha chisca}

01825125-v: {begrudge, resent} >{dar de mala gana, dar a contragusto, dar torcendo o fociño, ter envexa}

01983597-v: {bristle, uprise, stand up} >{arrepiar as carnes, poñer os pelos de espeto, pór os pelos de punta, pór os pelos dereitos}

Canto ao tratamento da terminoloxía, no WordNet recóllense todos os termos asociados a un determinado concepto, tanto se pertencen a niveis de especialización altos como formas populares ou vulgarizadas, coa limitación xa indicada para o léxico xeral e fraseolóxico da imposibilidade de marcar as formas con etiquetas de rexistro:

14848642-n: {colutorio, enxaugadura} 'solución médica usada para facer gargarexos e para lavar a boca'

5573895-n: {fémur, óso da coxa} 'óso máis longo e groso do esqueleto humano. Esténdese dende a pelve ata o xeonllo'

Con todo, o problema principal que atopamos á hora de facer un tratamento axeitado da terminoloxía foi a aparición de erros na taxonomía da zooloxía incluída nos BLC nominais da WordNet do inglés:

01352574-n: {bacteria genus} >{xénero das bacterias} (Porén, as bacterias son un dominio, non un xénero zoolóxico)

01342529-n: {animal order} >{orde dos animais} (Animalia é un reino)

01862557-n: {mammal family} >{familia dos mamíferos} (Os mamíferos son unha clase)

[10] <http://sli.uvigo.es/CTG/>

[11] <http://sli.uvigo.es/termoteca/>

3.2. Glosas

Nas definicións de WordNet 3.0 non hai un patrón homoxéneo que sexa respectado na totalidade dos synsets, senón que existen diferentes esquemas entre os que se poden salientar os seguintes tipos:

a) Definición intensional: defínese o termo enumerando as características ou propiedades que fan que se sitúe dentro dun concepto determinado. Por exemplo:

14631295-n: {berilio, Be, glucinio, número atómico 4} 'elemento metálico, bivalente, tóxico, gris, crebadizo, duro e luminoso'

b) Definición negativa: defínese o termo polo seu contrario.

00023383-a: {impreciso, inexacto, inxusto} 'non exacto'. Ex: *unha tradución inexacta; o termómetro é impreciso*

c) Definición operacional: defínese o termo indicando a súa composición e outros elementos relativos á súa magnitude, como a masa, o tempo, a temperatura, etc.

15103911-n: {metal de Wood, aliaxe de Wood} 'aliaxe que contén o 50 % de bismuto máis chumbo, estaño e cadmio; fúndese a preto de 160 graos Fahrenheit'

d) Definición por xénero e diferenza: no inicio da definición clasifícase o termo polo xénero ao que pertence e despois engádense características propias que axudan a delimitar o concepto.

05445389-n: {mitocondria} 'orgánulo que contén as encimas responsables da produción de enerxía'

e) Definición teorética: defínese o termo seguindo unha determinada teoría ou disciplina.

14842703-n: {aire} 'un dos catro elementos que segundo Empédocles compoñen o universo'

Outro tipo de definicións máis problemáticas son as que recollemos a continuación:

f) Definicións circulares: definicións que non dan máis información da que xa se extrae do propio termo. Este tipo de definicións son moi abundantes no WordNet 3.0 do inglés. Por exemplo:

15075298-n: {toilet roll} 'a roll of toilet paper' >{rolo de papel hixiénico} 'rolo de papel fino e estreito que serve para usos hixiénicos' (Carballeira Anllo, 2009).

g) Definicións pouco específicas. Dentro deste tipo de definicións atopamos pouca especificación en diferentes elementos:

g1) Existen casos nos que a información que achega a definición non aclara o concepto, posto que non se dá información esencial (ou complementaria) que axude a identificar o termo:

15090742-n: {B-complex vitamin, B complex, vitamin B complex, vitamin B, B vitamin, B} 'originally thought to be a single vitamin but now separated into several B vitamins' >{complexo vitamínico B, complexo B, vitamina do complexo B, vitamina B, B} 'grupo de vitaminas hidrosolubles cada unha coa súa propia función metabólica diferente das demais' (Daviña Facal, 2000).

g2) Imprecisión na descrición dalgunhas características:

15056827-n: {spindrift, spoondrift} 'spray blown up from the surface of the sea' >{escuma de mar, espuma de mar} 'conxunto de burbullas que se forman na tona da auga do mar cando esta bate contra as rochas ou contra un corpo que se move no seu interior' (Carballeira Anllo, 2009).

g3) Nalgúns casos só se describen os usos ou a orixe do termo pero non as súas características esenciais:

15062468-n: {tallow} 'obtained from suet and used in making soap, candles and lubricants' >{sebo} 'graxa sólida dos animais, constituída principalmente por estearato e palmitato de glicerilo. Soluble en etanol e éter; emprégase na fabricación de xabóns e candeas, para o curtume do coiro e como produto intermedio' (Real Academia de Ciencias Exactas, Físicas y Naturales, 1996).

g4) En oposición ao punto anterior, existen tamén definicións que describen características pero non usos:

14673032-n: {crocolite} 'a rare lead chromite mineral that forms bright orange crystals' >{crocoíta} 'mineral raro de cromato de chumbo que forma cristais laranxas brillantes e que se emprega ás veces como pigmento' (definición propia).

h) Confusión conceptual da definición con respecto ao termo. Este tipo de problema apréciase no seguinte synset, no que se fala do lume como un combustible cando, en realidade, é o resultado dunha combustión:

14686186-n: {fire} 'fuel that is burning and is used as a means for cooking' Ex: *put the kettle on the fire; barbecue over an open fire* >{lume} 'chama ou labarada dunha substancia en combustión; materia en combustión' Ex: *no alto do monte había lume; no inverno acendemos o lume na lareira* (Carballeira Anllo, 2009).

i) Existen series de cohipónimos nos que non se establece un patrón de definición; por exemplo, nas veas e arterias debería aparecer sempre a orixe, o final e a función do vaso. Nos seguintes exemplos non se explicita a orixe do conduto sanguíneo:

05349906-n: {laryngeal artery, arteria laryngea} 'either of two arteries that supply blood to the larynx' >{arteria larínxea, laryngea arteria} 'unha das dúas arterias que fornecen sangue á larinxe'

05357366-n: {anastomotic vein, vena anastomotica} 'either of two communicating veins serving the brain' >{vea anastomótica, vena anastomotica} 'unha das dúas veas comunicantes que serve o cerebro'

3.3. Exemplos

Con respecto aos exemplos que poden complementar a glosa, non existe no WordNet 3.0 do inglés unha estrutura recorrente, isto é, non aparecen exemplos en todos os synsets e, no caso de apareceren, non hai unha pauta que unifique a cantidade dos mesmos. Nalgúns casos, como no dos adxectivos, os exemplos son, predominantemente, citas textuais.

Na versión galega, optamos por manter esta estrutura e respectar as citas para non alterar o paralelismo con outras linguas, agás nos casos nos que o exemplo identificaba o termo como unha realidade propia doutra cultura allea á galega. Nestes últimos tentamos que os exemplos se aproximasen, na medida do posible, á nosa cultura, como se pode observar nos seguintes exemplos:

00699651-a: {inflected} 'showing alteration in form (especially by the addition of affixes)' Ex: *'boys' and 'swam' are inflected English words; German is an inflected language* >{flexionado, flexivo} 'que mostra modificacións na forma (especialmente a través da adición de afixos' Ex: *'nenos' e 'nadou' son palabras flexionadas en galego; o alemán é unha lingua flexiva*

02297664-a {standard, received} 'conforming to the established language usage of educated native speakers' Ex: *standard English (American); received standard English is sometimes called the King's English (British)* >{estandarizado, estándar, dentro do estándaro} 'que concorda co uso da linguaxe establecida dos falantes cultos nativos' Ex: *galego estándar*

A continuación séguese unha breve descrición do tipo de exemplos que consideramos problemáticos no WordNet 3.0 do inglés e, no seu caso, das solucións achegadas para o GalNet.

a) Existen exemplos que se adaptan só a unha determinada zona, cultura ou lingua. O que intentamos facer, neste caso, foi modificar os exemplos para que fosen máis xerais ou, simplemente, axeitalos ao caso galego:

05514410-n: {female genitalia, female genitals, female genital organ, fanny} 'external female sex organs' Ex: *in England 'fanny' is vulgar slang for female genitals* >{xenitais femininos, órgano xenital feminino, cona} 'órganos sexuais femininos externos' Ex: *en Galicia 'cona' é un termo vulgar para referirse aos órganos xenitais femininos*

b) Algúns exemplos presentan características esenciais ou complementarias importantes para explicar o concepto, polo que serían elementos propios da definición. Un exemplo deste tipo atópase no seguinte synset:

14707361-n: {adenosine deaminase, ADA} 'an enzyme found in mammals that can catalyze the deamination of adenosine into inosine and ammonia' Ex: *ADA deficiency can lead to one form of severe combined immunodeficiency disease; the gene encoding ADA was one of the earlier human genes to be isolated and cloned for study* >{adenosina deaminase, ADA} 'encima que se atopa nos mamíferos que poden catalizar a desaminación da adenosina en inosina e amoníaco' Ex: *a discapacidade da ADA pode levar a unha forma de enfermidade da inmunodeficiencia combinada severa; o xene que codifica a ADA foi un dos xenes humanos que antes de puido illar e clonar para estudalo*

Neste caso o primeiro dos exemplos é unha característica que delimita o concepto, polo que o normal sería que aparecese na definición.

c) Hai exemplos que non contribúen esencialmente a aclarar o concepto e que transmiten outro tipo de informacións, algunhas delas de marcado carácter ideolóxico. Na adaptación do inglés ao galego das entradas para nomes de partes do corpo e de substancias, este tipo de exemplos foron modificados, na medida do posible, para facelos máis neutrais:

14699752-n: {gem, gemstone, stone} 'a crystalline rock that can be cut and polished for jewelry' Ex: *he had the gem set in a ring for his wife; she had jewels made of all the rarest stones* >{xema, pedra preciosa} 'pedra cristalina, que pode ser cortada e pulida para xoiaría' Ex: *os rubís son xemas; tiña xoias feitas de todas as pedras preciosas máis raras*

00512261-a: {incompetent, unqualified} 'legally not qualified or sufficient' Ex: *a wife is usually considered unqualified to testify against her husband; incompetent witnesses* >{incapacitado, legalmente incapacitado, legalmente non capacitado} 'privado de capacidade xurídica pola concorrencia de determinadas causas legais, legalmente non cualificado' Ex: *en xeral non se aceptan nun xuízo as testemuñas legalmente incapacitadas*

	WN30		Galnet	
	Vars	Syns	Vars	Syns
N	117798	82115	9183	5646
V	11529	13767	1414	616
Adx	21479	18156	4864	3114
Adv	4481	3621	0	0
TOTAL	155287	117659	15461	9376

Táboa 1: Resultados iniciais.

4. Resultados

Na Táboa 1 preséntanse, agrupados en categorías (nomes, verbos, adxectivos e adverbios) e diferenciando entre synsets e variantes, os resultados acadados desde un punto de vista cuantitativo no estado actual de desenvolvemento do proxecto Galnet[12]. Estes resultados corresponden a 649 sysnets (1.333 variantes léxicas) dos BLC de categoría nominal, 616 synsets (1.414 variantes) dos BLC de categoría verbal, 2.014 synsets (3.550 variantes) do ficheiro lexicográfico de nomes relacionados coas partes do corpo, 2.983 synsets (4.300 variantes) do ficheiro lexicográfico de nomes de substancias, e 3.114 synsets (4.864 variantes) do conxunto de adxectivos de tipo xeral incluídos en WordNet 3.0. Tendo en conta os resultados obtidos en todas as categorías, podemos concluír que o crecemento lexicográfico do Galnet nesta primeira fase do proxecto ofrece unha cobertura semántica moi próxima ao 10 % con relación a cobertura de referencia da WordNet 3.0 en lingua inglesa.

De momento, os resultados iniciais do proxecto Galnet poden ser consultados en liña a través da interface WEI (Web EuroWordNet Interface)[13] do Multilingual Central Repository e forman parte tamén da plataforma RILG de Recursos Integrados da Lingua Galega[14]. Porén, está prevista a distribución libre dos recursos léxicos xerados no Galnet para toda a comunidade científica con licenza GPL (GNU General Public License da Free Software Foundation)[15].

5. Conclusións

Neste artigo presentamos o proxecto Galnet do Grupo TALG da Universidade de Vigo, dirixido á construción da versión galega do WordNet 3.0. Aínda que se trata dun proxecto en curso, algúns dos seus resultados xa se poden consultar libremente en Internet mediante a interface web de consulta do WEI (Web EuroWordNet Interface) accesible en <`http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl`> ou a través da plataforma RILG de Recursos Integrados da Lingua Galega dispoñible en <`http://sli.uvigo.es/RILG/`>. Así mesmo, está prevista a dispoñibilización libre en Internet con licenza GPL das sucesivas versións do Galnet xeradas ao longo do proxecto.

[12] Os datos cuantitativos completos sobre o WordNet 3.0 do inglés poden consultarse en <`http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`>

[13] <`http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl`>

[14] <`http://sli.uvigo.es/RILG/`>

[15] <`http://www.gnu.org/licenses/gpl.html`>

Bibliografía

Agirre, Eneko e Philip Edmonds. 2006. *Word Sense Disambiguation*. Springer, Berlin.

Álvez, J., J. Atserias, J. Carrera, S. Climent, A. Oliver, e G. Rigau. 2008. Consistent Annotation of EuroWordNet with the Top Concept Ontology. En *Proceedings of the 4th Global WordNet Conference (GWC'08)*.

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, e P. Vossen. 2004. The MEANING Multilingual Central Repository. En *In Proceedings of the Second International WordNet Conference*, páxinas 80–210.

Barzilay, Regina e Michael Elhadad. 1997. Using lexical chains for text summarization. En *Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization*, páxinas 10–17.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini, e Emanuele Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. En *Proceedings of the Workshop on Multilingual Linguistic Ressources*, páxinas 101–108.

Benítez, Laura, Sergi Cervell, Gerard Escudero, Mònica López, German Rigau, e Mariona Taulé. 1998. Methods and Tools for Building the Catalan WordNet. En *In Proceedings of ELRA Workshop on Language Resources for European Minority Languages*.

Carballeira Anllo, Xosé María, editor. 2009. *Gran dicionario xerais da lingua*. Edicións Xerais da Lingua, Vigo.

Daviña Facal, Luís. 2000. *Diccionario das ciencias da natureza e da saúde*. Deputación Provincial da Coruña, A Coruña.

Elberrichi, Zakaria, Abdellatif Rahmoun, e Mohamed Amine Bentaallah. 2008. Using wordnet for text categorization. *Int. Arab J. Inf. Technol.*, 5(1):16–24.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

Fernández, A. e G. Vázquez. 2010. La construcción del wordnet 3.0 en español. En M. A. Castillo e J. M. García, editores, *La lexicografía en su dimensión teórica*, páxinas 201–220, Málaga. Universidad de Málaga.

Fernández-Montraveta, A., G. Vázquez, e C. Fellbaum. 2008. The Spanish Version of WordNet 3.0. En A. Storrer, editor, *Text Resources and Lexical Knowledge*, páxinas 175–182, Berlín. Mouton de Gruyter.

Isahara, Hitoshi, Fransis Bond, Kiyotaka Uchimoto, Masao Utiyama, e Kyoko Kanzaki. 2008. Development of the Japanese WordNet. En *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

Izquierdo, R., A. Suárez, e G. Rigau. 2007. Exploring the automatic selection of basic level concepts. En *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'07)*.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, e Katherine J. Miller. 1990. Introduction to WordNet: an On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Ordan, Noam, Bar Ilan, Noam Ordan, e Shuly Wintner. 2007. Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19:39–58.

Pease, Adam, Ian Niles, e John Li. 2002. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. En *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*.

Pianta, Emanuele, Luisa Bentivogli, e Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. En *Proceedings of the First International Conference on Global WordNet*.

Pociello, Elisabete, Eneko Agirre, e Izaskun Aldezabal. 2011. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45:121–142.

Real Academia de Ciencias Exactas, Físicas y Naturales. 1996. *Vocabulario científico y técnico*. Espasa, Madrid.

Varelas, Giannis, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides G.M. Petrakis, e Evangelos E. Milios. 2005. Semantic similarity methods in wordnet and their application to information retrieval on the web. En *Proceedings of the 7th annual ACM international workshop on Web information and data management*, páxinas 10–16.

Vossen, Piek. 2002. WordNet, EuroWordNet and Global WordNet. *Revue française de linguistique appliquée*, 7:27–38.

Bancos de Fala para o Português Brasileiro

Vanessa Marquiafavel Serrani
Universidade Estadual Paulista (IBILCE)
marquiafavel@gmail.com

Luis Felipe Uebel
ASR Labs
luis.uebel@asrlabs.com

Resumo

Reconhecimento e síntese de fala demandam grandes quantidades de dados para aplicações comerciais. Em inglês americano, existem grandes quantidades de bancos de fala para a produção de modelos acústicos. Isto não é realidade para muitas línguas, incluindo o Português Brasileiro. Este trabalho apresenta o desenvolvimento de dois bancos de fala para reconhecimento com 248 locutores (224 sentenças) e outro com 1.226 locutores (550 já gravados e 665 sentenças), e um banco para síntese com 1.220 sentenças. Também é mostrado um sistema para selecionar as sentenças gravadas, dicionário fonético para suportar esta seleção, e modos de gravar e validar um grande banco de fala.

1. Introdução

O reconhecimento de fala permite que um computador possa traduzir um comando vocálico produzido por um humano em comando para abrir uma página na Internet, comandar um robô, ditar uma carta, ou transcrever uma conversa em vídeo ou via telefone. O processo inverso da pronúncia de um texto é chamado de síntese de fala.

O reconhecimento e a síntese de fala necessitam de amostras de vozes humanas para poderem trabalhar com eficiência. No reconhecimento de fala, a métrica mais importante é o nível de reconhecimento. Caso o sistema não traduza corretamente para um texto o que o usuário pronunciou, de nada vale o mesmo. Outros fatores como o consumo de memória ou de processamento deixam de ser relevantes caso o fator mais importante, o nível de reconhecimento, não for elevado. Com o objetivo de alcançar altos níveis de reconhecimento, são necessárias amostras de todos os fonemas usados na aplicação, sendo pronunciados por uma quantidade muito elevada de locutores com todos os sotaques presentes naquela língua. Isto equivale a dizer que uma grande quantidade de locutores necessita gravar uma quantidade elevada de sentenças.

Na síntese de fala, é necessária uma quantidade elevada de sentenças sendo pronunciadas por um único locutor. Existem técnicas de produção de vozes sintetizadas que necessitam de pouca quantidade de sentenças, ou a voz pode ser produzida por uma quantidade grande de locutores usando poucas sentenças cada. Para alcançar o nível de naturalidade e inteligibilidade a nível comercial é necessário que um único locutor grave centenas de sentenças com a mesma entonação, volume, velocidade e expressão, o que usualmente é chamado de "*persona*" da voz (Cohen 2004), ou seja, associar voz personalizada a uma marca ou empresa.

De um lado existe a necessidade de uma grande quantidade de fala para a formação do banco (reconhecimento e síntese) e de outro lado existem aspectos financeiros e de tempo para a construção de um banco desta magnitude.

Dependendo da aplicação, o banco pode ser constituído simplesmente dos comandos usados nesta aplicação. Comandos como "siga", "pare", "para direita", "para a esquerda", podem ser gravados quando a aplicação for simplesmente comandar algumas funções de um robô. Outro aspecto interessante é determinar qual o público alvo da aplicação. Caso sejam jovens entre 18 e 20 anos, bastar encontrar uma grande quantidade de jovens nesta faixa etária dispostos a gravar esses quatro comandos.

Quando a aplicação é mais complexa, como ditar uma carta ou reconhecimento de fala espontânea presentes em vídeos postados na Internet, a quantidade de pessoas necessárias e a quantidade de sentenças aumentam. Neste caso é preciso achar um conjunto de sentenças que possuam todos os fonemas presentes no Português Brasileiro.

Quanto maior o conjunto de sentenças a serem gravadas, maiores são os custos e o tempo das gravações, e maiores são os custos da validação dos dados.

O trabalho descreve o desenvolvimento de três bancos de fala para o Português Brasileiro (dois para reconhecimento e um para a síntese de fala), dicionário fonético das palavras coletadas com múltiplas transcrições regionais e sentenças gravadas.

2. Seleção das Frases compostas do Banco de Fala

O primeiro banco gravado, denominado ASR-DB1, é constituído de 200 sentenças definidas no projeto NURC (Alcaim 1992), sendo o mesmo utilizado em outros sistemas de reconhecimento de fala (Ynoguti 1999). Com o objetivo de aumentar a abrangência do conjunto de sentenças, 24 outras foram acrescentadas. Este conjunto adicional cobre números cardinais e ordinais, direções, comandos, meses do ano e nomes do zodíaco. Alguns números estavam contidos nas duzentas frases do projeto NURC, mas foram repetidos para melhorar o treinamento dos modelos acústicos. As outras palavras representam aplicações usuais de reconhecimento de fala. As mais de 700 palavras únicas contidas no conjunto de frases são também as mais usualmente verificadas nos *links* de páginas da Internet brasileira, objetivo principal do banco.

Com o objetivo de aumentar a cobertura fonética, foi desenvolvido um algoritmo de seleção das sentenças que constituem o banco, de modo a ter a maior cobertura possível com a menor quantidade de sentenças. Em (Rebollo et al 2005) foi desenvolvido um sistema de seleção de sentenças que divide o corpus CETENFolha (Corpus de Extractos de Textos Electrônicos NILC/Folha de São Paulo) em 400 conjuntos de 1000 sentenças cada e o conjunto selecionado é o que possui a maior quantidade de fonemas distintos. O algoritmo desenvolvido neste trabalho seleciona a sentença do corpus que possui a maior quantidade de fonemas distintos que ainda não foram encontrados. Portanto, este conjunto de fonemas que ainda não foram selecionados varia cada vez que uma nova frase é selecionada. Aí reside a complexidade do algoritmo. O processo de seleção das sentenças consiste em:

- Coleta de um conjunto de sentenças corretamente transcritas;
- Separação dos textos em sentenças;
- Produção de um dicionário fonético com todas as palavras contidas no
- conjunto de sentenças usadas no sistema;
- Expansão fonética das frases em "*cross-word*" para determinar o
- conjunto de fonemas presentes no conjunto de sentenças;
- Seleção das sentenças até que o sistema cubra todos os fonemas.

Bons linguistas podem demorar até um ano para obter um conjunto de sentenças que cubra os fonemas de uma língua, mas não têm como determinar qual o conjunto de fonemas mais usualmente pronunciados nesta língua. Com o objetivo de determinar primeiramente qual é este conjunto, um banco com sete milhões de sentenças foram coletadas. Cada palavra dessas sentenças foi checada frente a um dicionário ortográfico e 3,9 milhões de sentenças, corretamente transcritas, foram selecionadas. O sistema leva 55 minutos para processar este conjunto de sentenças e selecionar qual o menor conjunto de sentenças que descreve todos os fonemas com um determinado número mínimo de ocorrências no conjunto de sentenças processadas. O número mínimo de fonemas presente no conjunto de sentenças é determinado pelo usuário.

O algoritmo foi adaptado para selecionar sentenças a serem usadas em aplicações de adaptação ao locutor usando *lattice based MLLR* e *discriminative linear transforms* (Uebel 2001, 2002). As duas técnicas aumentam o nível de reconhecimento em condições difíceis (ruído, descasamento entre áudio do treinamento e teste). Neste caso, os fonemas são agrupados e contados como se fossem um único.

2.1 Algoritmo de Seleção de Sentenças

Primeiramente, o algoritmo procura os fonemas com o mínimo de ocorrência definido pelo usuário. O segundo passo é descobrir qual sentença possui a maior quantidade de fonemas procurados. Os fonemas encontrados são retirados da lista de fonemas a serem procurados e uma nova sentença que contém a maior quantidade de fonemas procurados é selecionada. A pesquisa é concluída quando não existem mais fonemas a serem procurados. O algoritmo acaba selecionando fonemas não procurados, uma vez que uma sentença irá ser constituída de fonemas procurados e não procurados.

Com a velocidade que o algoritmo é executado, o linguista pode redefinir uma

seleção de sentenças baseadas na quantidade de fonemas a serem cobertos pela quantidade de sentenças, e avaliar o tempo de gravação e os custos envolvidos.

2.2. Sentenças usadas na Seleção

Foram 3.946.744 sentenças extraídas de jornais, revistas, notícias e livros encontrados na Internet. O conjunto de sentenças é constituído de 44.518.978 palavras e 213.654 palavras únicas (sem repetição). As palavras mais freqüentes no banco são: "de" (4,24%), "o" (3,27%), "a" (3,25%), "e" (2,54%), "do" (2,15%), "que" (2,10%), "da" (1,81%), "em" (1,16%), "para" (1,12%) e "um" (0,93%).

Da expansão fonética do tipo "*word internal*", resultaram em 211.927.983 fonemas (trifones, bifones e monofones) e 23.107 fonemas únicos, e na expansão do tipo "*cross word*" foram 36.373 fonemas únicos. Considerando-se que podem ocorrer erros na geração do dicionário fonético, um corte no número mínimo de ocorrências de um fonema é necessário. Outro parâmetro importante é o número de fonemas por sentença. Sentenças muito longas são difíceis de serem lidas e resultam em muitos erros de leitura e, consequentemente, aumento no tempo e custos de validação.

O número mínimo de ocorrências de um fonema influencia no número de sentenças selecionadas, já que uma quantidade maior de fonemas deverá ser coberta nas sentenças selecionadas. A figura abaixo apresenta a cobertura fonética em função da quantidade máxima de fonemas em uma sentença selecionada dado pelo número de corte (CORTE). Na figura são apresentados cortes entre 50 fonemas e 150 fonemas por sentença no máximo e 28 fonemas no mínimo. Foram 781 simulações de aproximadamente 55 minutos cada (716 horas no total).

2.3. Considerações sobre as Sentenças Selecionadas

O processo de seleção de um determinado conjunto de sentenças para gravar um banco de fala depende, não somente de características técnicas de uma determinada língua (fonemas mais usualmente pronunciados), mas também de aspectos mais subjetivos como a facilidade em pronunciar uma determinada sentença, menor quantidade de palavras pronunciadas, e menor quantidade de palavras de origem estrangeira. O processo de seleção do conjunto de sentenças levou em consideração o seguinte:

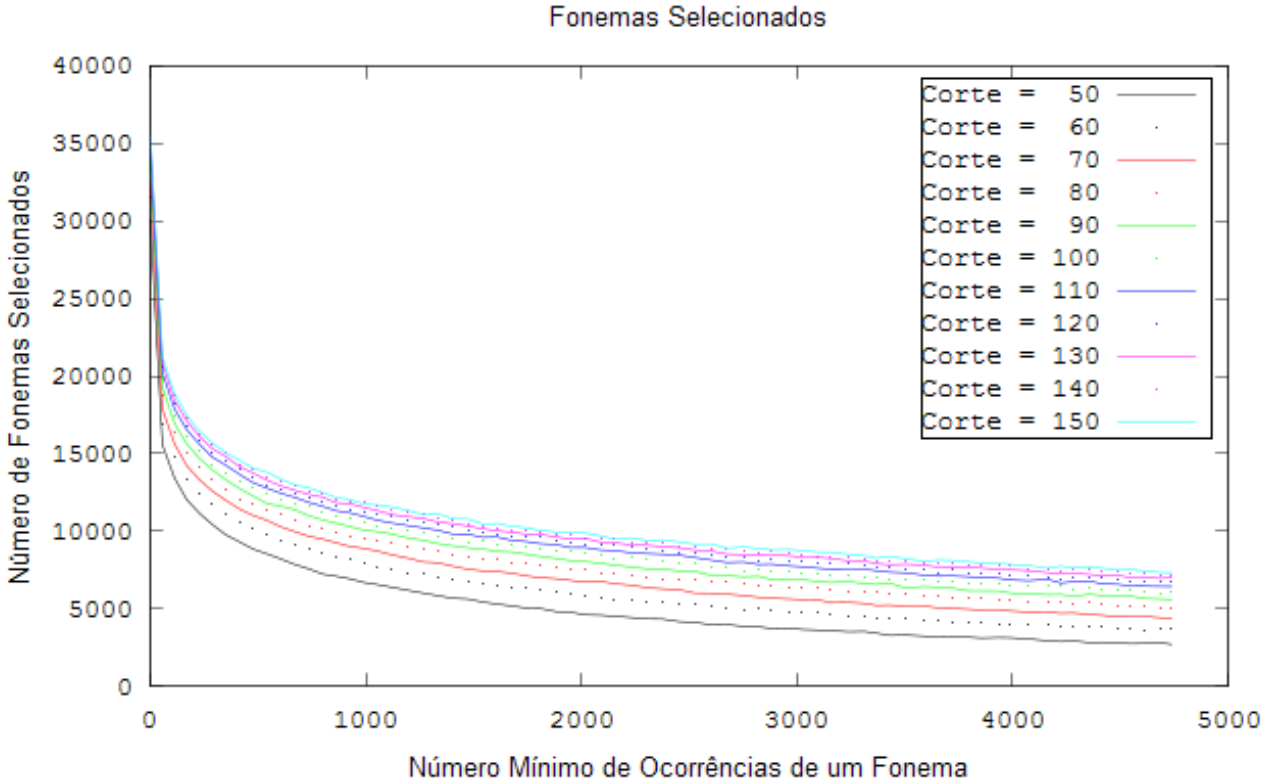


Figura 1. Número de Fonemas nas Sentenças Selecionadas.

1. Cobertura fonética de no mínimo 93,5 % de fonemas presentes na língua Portuguesa, sendo que esta cobertura foi verificada em função da distribuição dos fonemas presentes no banco de fala (3.946.744 de sentenças, 211.927.983 fonemas e 36.373 distintos);
2. O conjunto de sentenças selecionadas possui, no mínimo, os sete mil fonemas mais frequentemente encontrados na língua Portuguesa. O conjunto selecionado possui 353 sentenças, 4.175 palavras, e 7.216 fonemas distintos, que representam 93,8 % do total de fonemas presentes no banco de dados.

Portanto, em função da cobertura fonética, o conjunto de sentenças pode ser utilizado para o treinamento de modelos acústicos de excelente qualidade, com uma cobertura fonética excepcional e que podem ser utilizados no reconhecimento de fala contínua e espontânea, uma vez que as sentenças foram selecionadas utilizando-se uma expansão fonética do tipo "*cross word*".

Outras sentenças e palavras avulsas foram acrescentadas ao conjunto de sentenças para melhorar o nível de reconhecimento em situações difíceis, como é o caso dos números, comandos usados para comandar o Navegador Internet e outras tarefas de comando de um modo geral. Os números acima mencionados não levam em consideração estas sentenças e palavras avulsas.

As palavras que constam das sentenças gravadas no banco de fala representam 65,48% de todas as palavras do banco de dados, ou seja, as palavras que foram gravadas representam mais de 2/3 de todas as palavras que constam do banco de dados. Este é mais um dado que mostra quão acertada foi a seleção do conjunto de sentenças.

2.4. Sentenças do ASR-DB2

As primeiras cinco sentenças do conjunto de 353 encontram-se na Tabela 1. Além das 353 sentenças, foram acrescentadas 23 sentenças curtas para aumentar a cobertura fonética, 6 sentenças para adaptação ao locutor (aproximadamente 30 segundos de fala), 19 sentenças de domínio específico (similares as usadas no ASR-DB1), e 49 comandos de controle de um Navegador Internet.

Tabela 1. Lista de Sentenças Foneticamente Balanceadas

	Texto
1	O Corinthians passou para a segunda fase ao desclassificar o Vitória da Bahia.
2	Quem comprou telefones em noventa e três acabou o ano com uma boa surpresa.
3	Esse consumidor está em busca de um produto cada vez melhor e mais barato.
4	Os freios a disco nas quatro rodas podem ter ABS opcional, não travam.
5	Depoimento garante retorno ao gabinete civil, da sucursal de Brasília.

Como algumas aplicações de reconhecimento de fala demandam o reconhecimento de palavras isoladas, foram acrescentadas 45 palavras avulsas relacionadas a números, 28 de letras do alfabeto, 46 de direção e comando, 18 de partes do corpo humano, 36 de utensílios domésticos, e 42 de sensores, cores e estados emocionais. O total de sentenças é de 665, que agrupam a maioria das aplicações em reconhecimento de fala (fala espontânea, ditado, comandos, algarismos e robótica).

2.5. Sentenças para a Síntese de Fala

Diversos mecanismos podem ser utilizados para a elaboração da síntese de fala, como, por exemplo, a concatenação de sentenças ou palavras. O conjunto de sentenças deve possuir mais de uma mostra do mesmo fonema para melhor modelamento acústico.

Desta forma os fonemas possuem, no mínimo, duas ocorrências no conjunto de sentenças selecionadas e, pelo menos, 1.266 ocorrências no banco. O resultado são 632 sentenças foneticamente balanceadas para o Português Brasileiro, 8.046 fonemas (94,6% dos fonemas existentes no banco) e 7.529 palavras. Foram acrescentadas as sentenças para o reconhecimento de fala, 30 nomes próprios mais comumente encontrados no Brasil e outras palavras avulsas. O total de sentenças avulsas é de 558, sendo 47 sentenças para adaptação ao locutor (transformação da voz), e o total de sentenças para o banco de fala em síntese de fala (TTS-DB1) é de 1.220. Comparando a outros bancos, o CMU Artic (Kominek 2004), conjunto de sentenças para o

inglês americano, possui 1.132 sentenças, 10.003 palavras, sendo 2.970 palavras únicas, o que mostra uma grande repetição de palavras nas sentenças. O SynthFemale01 possui 4,392 sentenças em coreano e o projeto NEMLAR possui 2.032 sentenças (42 mil palavras) em árabe.

Comparado a outros bancos citados, o conjunto aqui selecionado possui menos sentenças, facilitando a leitura, maior cobertura fonética e menor repetição de palavras, resultando em 16 vozes sintetizadas usando o Flite+HTS (Black 2007).

3. Dicionário Fonético

O dicionário fonético é constituído pela transcrição ao nível fonético das palavras contidas nos bancos de fala e outras palavras que se queira reconhecer. Os fonemas são a menor unidade de som de uma língua e permitem que seja possível reconhecer palavras não contidas nestes bancos. De um modo simplificado, o "ão" de "tubarão" pode ser utilizado no treinamento de modelos que irão reconhecer "coração" ou "leão". O mesmo processo pode ser usado na síntese de fala.

3.1. Corpus e Transcrição Fonética

Todas as palavras que constituem as sentenças dos bancos de fala foram transcritas foneticamente segundo as variantes padrão (norma oficial) e não-padrão (em decorrência das diferentes regiões geográficas, classes sociais, faixas etárias, etc.) do Português Brasileiro, perfazendo um total de 13 grandes variações dialetais descritas: DF, GO, ES, AM, RJ, SP, interior de SP, MT, MG, PA, BA, Sul e Nordeste em geral.

Segundo pesquisas realizadas (Callou 2003), não podemos tomar como modelo apenas a pronúncia de uma pessoa, de uma classe social ou de uma região. Por isso, apresentamos em nosso dicionário múltiplas transcrições fonéticas das variantes e de vícios de linguagem (alterações sonoras como assimilação, epêntese, etc.), uma vez que a língua não é usada da mesma forma pelas pessoas em todos os momentos. Foram transcritas todas as palavras contidas nos três bancos de fala (ASR-DB1, ASR-DB2 e TTS-DB1) e o sistema de notação adotado para tal foi o SAMPA, alfabeto fonético computável, constituído pelo mapeamento dos símbolos do sistema IPA (Alfabeto Fonético Internacional) para códigos ASCII.

A correta escolha do conjunto de sentenças utilizado para a gravação de um banco de fala é de vital importância para o projeto. Tal importância não se restringe unicamente à cobertura fonética representativa dos sons da língua, mas também à facilidade com que os locutores as leem. Outro item de suma importância deste projeto é a construção de um dicionário fonético de qualidade, pilar para que se obtenham bons resultados nas tarefas relacionadas ao reconhecimento e síntese de fala.

3.2. Coleta do Banco de Fala

A coleta do banco de fala com vários locutores lendo as mesmas sentenças previamente escolhidas em função de sua contribuição para a cobertura fonética e para possíveis aplicações possui vantagens como: facilidade na coleta dos dados, uma vez que o conjunto de sentenças não é modificado; rapidez na validação do banco de fala, pois pode ser usada uma validação vertical, ou seja, validar cada sentença para todos os locutores ao invés de todas as sentenças de um locutor, o que possibilita ganhos de produtividade em torno de cinco vezes; e menor taxa de erros por parte do locutor e do validador. Como descrito anteriormente, este tipo de banco de dados permite definir, no menor conjunto de sentenças, todos os trifones mais encontrados na língua.

O perfil dos locutores escolhidos varia entre 13 e 59 anos, devido às aplicações vislumbradas pelos bancos de fala. As aplicações são de reconhecimento de fala espontânea, ditado, comandos de controle de robôs móveis e controle de objetos.

As gravações do banco ASR-DB1 foram realizadas no estado de São Paulo, e o ASR-DB2 no estado de São Paulo e outros estados da Federação. Muitas pessoas gravadas no estado de São Paulo são oriundas de outras regiões, possibilitando uma cobertura de sotaques das 13 regiões dialéticas do Brasil. O tempo médio das gravações do ASR-DB1 variava entre 20 minutos e 2 horas (média de 25 minutos) e

entre 1:10h e 1:30h para o ASR-DB2. Algumas pessoas podem levar até 3 horas, ou menos, como 55 minutos. O tempo varia conforma a velocidade de leitura do locutor, facilidade de leitura, nervosismo, timidez, e problemas de coordenação motora ao utilizar os equipamentos (computador e *mouse*). Algumas pessoas ficam nervosas por acharem que estão sendo testadas. Neste caso o técnico deve orientar o locutor e acalmá-lo para que a gravação tenha prosseguimento. Ruídos causados pelos lábios ao serem abertos no início das gravações e o ruído causado pela respiração resultam em picos nos respectivos espectrogramas e problemas na segmentação dos dados no treinamento acústico. Neste caso o técnico pedia ao locutor para regravar a sentença novamente. O locutor deixava um silêncio antes e depois de pronunciar cada sentença para que a pronúncia da mesma não fosse cortada.

As frases mais problemáticas de serem lidas são as que possuem palavras estrangeiras e excertos literários, por possuírem palavras desconhecidas pelo locutor, este não entende o sentido da frase e acaba cometendo erros de pronúncia e entonação.

3.3. Validação dos Dados

O processo de coleta dos dados é seguido de sua validação, ou seja, verificar se o locutor realmente leu o que deveria ler. O processo é iniciado na coleta dos dados. O técnico fica atento ao que o locutor está pronunciando e, caso não esteja correto, pede que o mesmo repita a leitura da sentença de uma forma correta. Este procedimento tem por finalidade diminuir o tempo e custos do processo de validação do banco de fala.

A validação consiste em escutar todos os arquivos de áudio e transcrever fielmente o que o locutor realmente pronunciou. A transcrição fiel dos arquivos sonoros deve ser a mais verossímil possível. Caso o locutor tenha pronunciado "desta" em vez de "dessa" ou pronunciou "pobema" em vez de "problema", estes fatos devem constar nos arquivos de texto validados. Um dos aspectos que facilitaram e aceleraram o processo de validação dos dados foi utilizar uma validação vertical dos dados, ou seja, frase por frase, em oposição ao método tradicional que valida locutor por locutor. O treinamento do técnico responsável pelas gravações reduziu significativamente o tempo de validação dos arquivos e o processo de validação vertical diminuir os erros na validação uma vez que o validador vai escutar centenas de vezes a mesma sentença e detectar de uma forma mais rápida e precisa pequenos erros como a falta da pronúncia de um simples "s" em uma palavra ou os erros descritos acima.

A validação é peça fundamental no treinamento de modelos acústicos no reconhecimento e síntese de fala, principalmente quando técnicas mais avançadas de modelamento acústico são empregadas, como treinamento discriminativo. Uma validação automática foi empregada para detectar arquivos corretamente pronunciados, facilitando (acelerando) o processo de validação e diminuindo seus custos.

4. Conclusões

O trabalho apresenta o desenvolvimento de três bancos de fala para o Português Brasileiro, dois especificamente para reconhecimento e um para síntese. O ASR-DB1 possui 248 locutores, 224 sentenças, microfone de 100 a 10 kHz, 16 *bits* e 48 kHz de taxa de amostragem. O ASR-DB2 irá gravar 1.226 locutores (550 já gravados), 665 sentenças, microfone de alta fidelidade do tipo *headset* (banda passante de 80 Hz a 15 kHz) e placa de som com 24 *bits* e 96 kHz de taxa de amostragem. O TTS-DB1 possui 1.220 sentenças, microfone de 20 Hz a 20 kHz, 24 *bits* e 96 kHz de amostragem. O trabalho também apresentou um algoritmo de seleção de sentenças que permite determinar com exatidão os fonemas mais usuais do Português Brasileiro usando um conjunto de 3,9 milhões de sentenças ortograficamente corrigidas.

5. Agradecimentos

Agradecemos o apoio da FAPESP (PIPE – 2005/59953-0), Finep (Subvenção Econômica à Inovação – 4717/06) e CNPq (RHAE Pesquisador na Empresa – 558052/2008-8) pelos recursos destinados ao projeto.

Agradecemos o pesquisador Mauro Miazaki pela implementação da primeira versão do sistema de seleção de sentenças, Henrique Ferraz, Vinicius Mittitier, Franclin Barros e

Roseli Oliveira Silva pelas gravações do banco ASR-DB2.

6. Referências

Alcaim, Abraham, Solemicz, José Alberto e Morais, João Antônio de. 1992. Frequência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas para o Português Falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações*, Vol. 7.

Black, Alan, Zen, Heiga e Tokuda, Keiichi. 2007. Statistical Parametric Speech Synthesis. *ICASSP 2007*, Honolulu, Havaí, EUA: 1229-1232.

Callou, Dinah & Leite, Ione. 2003. *Iniciação à Fonética e à Fonologia*, 5 ed., Rio de Janeiro: Zahar.

Rebollo Couto, Leticia; Moraes, J.A. de, Resende Jr, F.G.V.; Cirigliano, R.J. R.; Barbosa, F.L.F.; Vianna, C.M. 2005. Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro Obtido Utilizando a Abordagem de Algoritmos Genéticos, *Anais do XXII Simposio Brasileiro de Telecomunicacoes (SbrT 2005)*, Campinas, Brazil, pp. 544-549.

Cohen, Michael H.; Giangola, James. P.; Balogh, Jennifer. 2004. *Voice User Interface Design.* Addison-Wesley Professional, 368 p.

Kominek, John & Black, Alan. W. 2004. The CMU Artic Speech Databases. Proceedings of *5th ISCA Speech Synthesis Workshop*, Pittsburgh, EUA, pp. 223-224.

Uebel, Luis Felipe & Woodland, P. C. 2001. Improvements in Linear Transform based Speaker Adaptation. *In: IEEE – International Conference on Acoustics Speech and Signal Processing*, 7-11 May 2011, Salt Lake City, Utah, EUA, Vol. 1, pp. 49-52.

Uebel, Luis Felipe. 2002. *Speaker Normalisation and Adaptation in Large Vocabulary Speech Recognition*. Tese de Doutorado. University of Cambridge. England.

Ynoguti, Carlos Alberto. 1999. Reconhecimento de Fala Contínuo usando Modelos Ocultos de Markov. Tese de Doutorado, UNICAMP, Campinas-SP.

Chamada de Artigos

A revista Linguamática pretende colmatar uma lacuna na comunidade de processamento de linguagem natural para as línguas ibéricas. Deste modo, serão publicados artigos que visem o processamento de alguma destas línguas.

A Linguamática é uma revista completamente aberta. Os artigos serão publicados de forma electrónica e disponibilizados abertamente para toda a comunidade científica sob licença *Creative Commons*.

Tópicos de interesse:

- Morfologia, sintaxe e semântica computacional
- Tradução automática e ferramentas de auxílio à tradução
- Terminologia e lexicografia computacional
- Síntese e reconhecimento de fala
- Recolha de informação
- Resposta automática a perguntas
- Linguística com corpora
- Bibliotecas digitais
- Avaliação de sistemas de processamento de linguagem natural
- Ferramentas e recursos públicos ou partilháveis
- Serviços linguísticos na rede
- Ontologias e representação do conhecimento
- Métodos estatísticos aplicados à língua
- Ferramentas de apoio ao ensino das línguas

Os artigos devem ser enviados em PDF através do sistema electrónico da revista. Embora o número de páginas dos artigos seja flexível sugere-se que não excedam 20 páginas. Os artigos devem ser devidamente identificados. Do mesmo modo, os comentários dos membros do comité científico serão devidamente assinados.

Em relação à língua usada para a escrita do artigo, sugere-se o uso de português, galego, castelhano, basco ou catalão.

Os artigos devem seguir o formato gráfico da revista. Existem modelos LaTeX, Microsoft Word e OpenOffice.org na página da Linguamática.

Datas Importantes

- Envio de artigos até: 31 de outubro de 2011
- Resultados da selecção até: 30 de novembro de 2011
- Versão final até: 15 de dezembro de 2011
- Publicação da revista: dezembro de 2011

Qualquer questão deve ser endereçada a: `editores@linguamatica.com`

Petición de Artigos

A revista Linguamática pretende cubrir unha lagoa na comunidade de procesamento de linguaxe natural para as linguas ibéricas. Deste xeito, han ser publicados artigos que traten o procesamento de calquera destas linguas.

Linguamática é unha revista completamente aberta. Os artigos publicaranse de forma electrónica e estarán ao libre dispor de toda a comunidade científica con licenza *Creative Commons*.

Temas de interese:

- Morfoloxía, sintaxe e semántica computacional
- Tradución automática e ferramentas de axuda á tradución
- Terminoloxía e lexicografía computacional
- Síntese e recoñecemento de fala
- Extracción de información
- Resposta automática a preguntas
- Lingüística de corpus
- Bibliotecas dixitais
- Avaliación de sistemas de procesamento de linguaxe natural
- Ferramentas e recursos públicos ou cooperativos
- Servizos lingüísticos na rede
- Ontoloxías e representación do coñecemento
- Métodos estatísticos aplicados á lingua
- Ferramentas de apoio ao ensino das linguas

Os artigos deben de enviarse en PDF mediante o sistema electrónico da revista. Aínda que o número de páxinas dos artigos sexa flexíbel suxírese que non excedan as 20 páxinas. Os artigos teñen que identificarse debidamente. Do mesmo modo, os comentarios dos membros do comité científico serán debidamente asinados.

En relación á lingua usada para a escrita do artigo, suxírese o uso de portugués, galego, castelán, éuscaro ou catalán.

Os artigos teñen que seguir o formato gráfico da revista. Existen modelos LaTeX, Microsoft Word e OpenOffice.org na páxina de Linguamática.

Datas Importantes

- Envío de artigos até: 31 de outubro de 2011
- Resultados da selección: 30 de novembro de 2011
- Versión final: 15 de decembro de 2011
- Publicación da revista: decembro de 2011

Para calquera cuestión, pode dirixirse a: `editores@linguamatica.com`

Petición de Artículos

La revista Linguamática pretende cubrir una laguna en la comunidad de procesamiento del lenguaje natural para las lenguas ibéricas. Con este fin, se publicarán artículos que traten el procesamiento de cualquiera de estas lenguas.

Linguamática es una revista completamente abierta. Los artículos se publicarán de forma electrónica y se pondrán a libre disposición de toda la comunidad científica con licencia *Creative Commons*.

Temas de interés:

- Morfología, sintaxis y semántica computacional
- Traducción automática y herramientas de ayuda a la traducción
- Terminología y lexicografía computacional
- Síntesis y reconocimiento del habla
- Extracción de información
- Respuesta automática a preguntas
- Lingüística de corpus
- Bibliotecas digitales
- Evaluación de sistemas de procesamiento del linguage natural
- Herramientas y recursos públicos o cooperativos
- Servicios lingüísticos en la red
- Ontologías y representación del conocimiento
- Métodos estadísticos aplicados a la lengua
- Herramientas de apoyo para la enseñanza de lenguas

Los artículos tienen que enviarse en PDF mediante el sistema electrónico de la revista. Aunque el número de páginas de los artículos sea flexible, se sugiere que no excedan las 20 páginas. Los artículos tienen que identificarse debidamente. Del mismo modo, los comentarios de los miembros del comité científico serán debidamente firmados.

En relación a la lengua usada para la escritura del artículo, se sugiere el uso del portugués, gallego, castellano, vasco o catalán.

Los artículos tienen que seguir el formato gráfico de la revista. Existen modelos LaTeX, Microsoft Word y OpenOffice.org en la página de Linguamática.

Fechas Importantes

- Envío de artículos hasta: 31 de octubre de 2011
- Resultados de la selección: 30 de noviembre de 2011
- Versión final: 15 de diciembre de 2011
- Publicación de la revista: diciembre de 2011

Para cualquier cuestión, puede dirigirse a: `editores@linguamatica.com`

Petició d'articles

La revista Linguamática pretén cobrir una llacuna en la comunitat del processament de llenguatge natural per a les llengües ibèriques. Així, es publicaran articles que tractin el processament de qualsevol d'aquestes llengües.

Linguamática és una revista completament oberta. Els articles es publicaran de forma electrònica i es distribuiran lliurement per a tota la comunitat científica amb llicència *Creative Commons*.

Temes d'interès:

- Morfologia, sintaxi i semàntica computacional
- Traducció automàtica i eines d'ajuda a la traducció
- Terminologia i lexicografia computacional
- Síntesi i reconeixement de parla
- Extracció d'informació
- Resposta automàtica a preguntes
- Lingüística de corpus
- Biblioteques digitals
- Evaluació de sistemes de processament del llenguatge natural
- Eines i recursos lingüístics públics o cooperatius
- Serveis lingüístics en xarxa
- Ontologies i representació del coneixement
- Mètodes estadístics aplicats a la llengua
- Eines d'ajut per a l'ensenyament de llengües

Els articles s'han d'enviar en PDF mitjançant el sistema electrònic de la revista. Tot i que el nombre de pàgines dels articles sigui flexible es suggereix que no ultrapassin les 20 pàgines. Els articles s'han d'identificar degudament. Igualement, els comentaris dels membres del comitè científic seràn degudament signats.

En relació a la llengua usada per l'escriptura de l'article, es suggereix l'ús del portuguès, gallec, castellà, basc o català.

Els articles han de seguir el format gràfic de la revista. Es poden trobar models LaTeX, Microsoft Word i OpenOffice.org a la pàgina de Linguamática.

Dades Importants

- Enviament d'articles fins a: 31 d'octubre de 2011
- Resultats de la selecció: 30 de novembre de 2011
- Versió final: 15 de desembre de 2011
- Publicació de la revista: desembre de 2011

Per a qualsevol qüestió, pot adreçar-se a: `editores@linguamatica.com`

Artilulu eskaera

Iberiar penintsulako hizkuntzei dagokienean, hizkuntza naturalen prozedura komunitatean dagoen hutsunea betetzea litzateke Linguamática izeneko aldizkariaren helburu nagusiena. Helburu nagusi hau buru, aurretik aipaturiko edozein hizkuntzen prozedura landuko duten artikuluak argitaratuko dira.

Linguamática aldizkaria irekia da oso. Artikuluak elektronikoki argitaratuko dira, eta komunitate zientefikoaren eskura egongo dira honako lizentziarekin; *Creative Commons*.

Gai interesgarriak:

- Morfologia, sintaxia eta semantika konputazionala.
- Itzulpen automatikoa eta itzulpengintzarako lagungarriak diren tresnak.
- Terminologia eta lexikologia konputazionala.
- Mintzamenaren sintesia eta ikuskapena.
- Informazio ateratzea.
- Galderen erantzun automatikoa.
- Corpus-aren linguistika.
- Liburutegi digitalak.
- Hizkuntza naturalaren prozedura sistemaren ebaluaketa.
- Tresna eta baliabide publikoak edo kooperatiboak.
- Zerbitzu linguistikoak sarean.
- Ezagutzaren ontologia eta adierazpideak.
- Hizkuntzean oinarrituriko metodo estatistikoak.
- Hizkuntzen irakaskuntzarako laguntza tresnak.

Arikuluak PDF formatoan eta aldizkariaren sitema elektronikoaren bidez bidali behar dira. Orri kopurua malgua den arren, 20 orri baino gehiago ez idaztea komeni da. Artikuluak behar bezala identifikatu behar dira. Era berean, zientzi batzordeko kideen iruzkinak ere sinaturik egon beharko dira.

Artikulua idazterako garaian, erabilitako hizkuntzari dagokionean, honako kizkuntza hauek erabili daiztezke; portugesa, galiziera, gaztelania, euskara, eta katalana.

Artikuluek, aldizkariaren formato grafikoa jarraitu behar dute. "Linguamática" orrian LaTeX, Microsoft Word eta OpenOffice.org ereduak aurki ditzakegu.

Data garratzitsuak:

- Artikuluak bidali ahal izateko epea: 2011ko urriaren 31.
- Hautapen-prozesuaren jakinarazpena: 2011ko azaroaren 30a.
- Azken bertsioaren bidalketa: 2011ko abenduaren 15a.
- Argitarapena aldizkarian: 2011ko abendua.

Edozein zalantza argitzeko, hona hemen helbide hau: `editores@linguamatica.com`.

Volume 3, Número 1- Junho 2011

linguaMATICA

ISSN: 1647-0818

Dossier

Artigos de Investigação

Apresentação de Projectos