BASC I TECNOLOGIA LINGÜÍSTICA

Itziar Aduriz*, Arantza Diaz de Ilarraza i Kepa Sarasola
IXA taldea (UPV-EHU)
* IXA taldea (UPV-EHU) i Linguistikako saila. Bartzelonako Unibertsitatea

1. INTRODUCCIÓ

Cada cop tenim accés a més informació en suport digital (vídeo, imatge, veu i text), per la qual cosa darrerament s'han creat eines i aplicacions informàtiques noves per processar-la. Ara bé, on se situa el basc en aquest món nou? Poden contribuir les eines esmentades a normalitzar-lo?

1.1. Més informació en format de text

El volum de texts disponibles en l'era digital és aclaparador: es calcula que a Internet hi ha més d'un bilió de paraules en anglès (un milió de milions de paraules!). La quantitat de text en basc present a la xarxa és, segons algunes estimacions (Alegría i Rodríguez, 2003), entre mil o deu mil vegades inferior a la d'anglès; per tant, actualment deu haver-hi uns mil milions de paraules. Per fer-nos una idea sobre la dimensió de les xifres esmentades, n'hi ha prou amb saber que un llibre de mida mitjana conté unes cent mil paraules, que una persona culta pot llegir unes deu mil paraules al dia, és a dir, 3,65 milions de paraules l'any i uns 300 milions de paraules al llarg de tota la vida. Així, doncs, si volguéssim llegir tots el texts existents en basc a la xarxa (disponibles en qüestió de segons), necessitaríem unes tres vides; si volguéssim llegir els texts existents en anglès, el nombre de vides necessàries ascendiria a tres mil.

1.2. Les eines per al tractament automàtic de la llengua són una realitat

Avui dia hi ha diverses aplicacions lingüístiques per al tractament de texts o de la parla: correctors ortogràfics, correctors d'estil, consultes de diccionaris en línia, traducció automàtica i traducció assistida, sistemes que converteixen la parla en text, sistemes capaços de llegir texts, sistemes per aprendre segones llengües, interfícies per utilitzar les aplicacions informàtiques en la nostra llengua, sistemes de cerca de respostes a preguntes (*question answering*), cercadors de documents (IR, *information retrieval*), extracció d'informació de documents (IE, *information extraction*), resum automàtic (*summarization*), classificadors de documents, encaminadors de documents (*routing*), agrupadors de documents (*clustering*), filtradors de documents (*filtering*) i generació automàtica de text.

Lamentablement no hi ha un sol lloc a Internet que reuneixi la informació sobre tots els productes relacionats amb el processament de la llengua. Ara bé, hi ha alguns productes —classificats per àrees o per aplicacions— a: Natural Language Software Registry (NLSR);[1] European Language Resources Association (ELRA);[2] Linguistic Data Consortium (LDC), especialitzat en productes USA (productes destinats a ser consumits als EUA),[3] i Association for Computational Linguistics

[1] http://registry.dfki.de
[2] http://catalog.elra.info
[3] http://www.ldc.upenn.edu/Catalog/catalogSearch.jsp

(ACL), que disposa d'informació dels recursos per a totes les llengües[4]. Per traduir automàticament es pot consultar Translation Directory[5] i Traduzione e computer[6].

1.3. Tanmateix, la majoria dels avenços beneficien sobretot les llengües majoritàries

Totes les aplicacions esmentades poden utilitzar-se en anglès, però no succeeix el mateix, ni qualitativament ni quantitativa, amb la resta de les llengües. La taula 1 ens mostra quants productes —i per a quantes llengües— contenen els tres llocs que ofereixen informació sobre tecnologies lingüístiques (ELRA, NLSR i LDC).

Taula 1. Nombre de productes de tecnologia lingüística i nombre de llengües per a les quals serveixen

	ELRA	NLSR	LDC
Alemany	428	106	14
Anglès	463	196	232
Àrab	49	0	57
Basc	4	61	0
Català	9	59	0
Danès	18	64	1
Espanyol	388	85	28
Francès	407	99	10
Holandès	46	69	3
Italià	349	76	2
Suec	29	67	2

La taula 2 ens mostra el nombre de sistemes de traducció que hi havia el 2005 per a les llengües oficials d'Europa. Aquí també és clar quines són les llengües que predominen.

	en	de	fr	es	it	pt	du	po	lt	gr	cs	hu	sw	fn	sl	rm	dk
Anglès		47	41	44	30	30	10	8	2	1	4	1		1	1		
Alemany	48		24	8	10	4	2	3	1	1	2	1	1	1		1	
Francès	40	23		11	13	8	4	1	1	1							
Castellà	41	7	11		9	8	1		1	1							
Italià	29	10	13	9		4	1		1	1							
Portuguès	29	5	7	8	4		1		1								
Holandès	10	2	4	1	1	1			1								
Polonès	7	2	1														
Lituà	2	1	1	1	1	1	1										
Grec	3		3														
Txec	1	1	1		1												
Hongarès	2	2															
Suec	2	1															
Suomi	2	1															
Eslovac																	
Romanès	1																
Danès		1															

Taula 2. Nombre de sistemes de traducció que hi havia el 2005 per a parells de llengües oficials d'Europa[7]

1.4. La tecnologia lingüística, objectiu i aliat per a la normalització del basc

El basc ha sofert un procés de recessió durant segles, les causes principals del qual, segons Amorrortu (2002), són les següents: d'una banda, el fet de no ser una llengua oficial i, de l'altra, haver romàs al marge del sistema d'ensenyament, dels mitjans de comunicació i del món laboral industrial. A més, cal tenir en compte el fet que hi hagi fins a vuit dialectes bastant distants i la falta de normalització consegüent, ja que han impedit històricament la difusió àmplia del basc escrit.

En les últimes dècades, no obstant això, s'ha avençat d'una manera qualitativa força important per canviar aquesta situació: el basc és llengua oficial en part del territori, es troba integrat en el sistema educatiu, hi ha mitjans de comunicació en basc i hi ha un estàndard majoritàriament acceptat. Així, el nombre de parlants de basc ascendeix, en l'actualitat, a 700.000, aproximadament un 25 % de la població.

Malgrat els avenços assenyalats, el futur del basc no està garantit. Per una banda, els passos que acabem de citar no són d'aplicació extensiva a tot el territori; per una altra banda, el basc continua al marge del

[4] http://aclweb.org/aclwiki/index.php?title=Resources_for_Basque

[5] http://www.translation-directory.com/machine.html

[6] http://www.federicozanettin.net/sslmit/cattools.htm#publications

[7] Segons el llibre de J. *Hutchins Compendium of Translation* i el projecte Euromatrix (http://www.euromatrix.net/).

món laboral industrial, inclòs el que té relació amb les tecnologies de la informació i la comunicació (TIC).

En aquest context i des de l'inici, el grup IXA ha investigat en tecnologia lingüística i, en la mesura que li ha estat possible, n'ha comercialitzat productes relacionats per impulsar l'ús del basc en les TIC.

En aquest article analitzarem com les tecnologies lingüístiques poden ajudar a fomentar i facilitar l'ús del basc: en el segon apartat es presenten diversos productes que poden contribuir als objectius esmentats; en el tercer, s'analitza la situació actual del basc en relació amb les tecnologies lingüístiques; en el quart, s'expliquen les raons que han determinat les prioritats en l'estratègia del grup IXA per desenvolupar les tecnologies lingüístiques; en el cinquè, se n'apunten les línies de futur, i, en el sisè, se'n presenten unes conclusions.

2. Productes que faciliten i fomenten l'ús del basc

Presentarem, en aquest apartat, diversos productes tecnològics desenvolupats pel grup IXA, que es troben a disposició del públic en general i que estan orientats a ajudar els parlants de basc a apropar-se al món digital.

2.1. Corrector ortogràfic

Escriure correctament comporta dificultats en qualsevol llengua, i els parlants de basc no en són una excepció, ja que a l'hora d'escriure els sorgeixen molts dubtes. Això és degut, d'una banda, que, de moment, el sistema educatiu actual no garanteix que tots els estudiants assoleixin correctament la capacitat d'escriure en basc, i de l'altra, que la gent que té una certa edat no ha pogut estudiar en basc. Com que la definició del basc estàndard és relativament recent i l'estandardització del lèxic no és completa, sorgeixen dubtes nous, ja que, de vegades, es produeixen canvis sobretot en les decisions referents al lèxic. D'altra banda, la gent que té una certa edat i sap basc no ha pogut estudiar-lo ni hi està alfabetitzada, per la qual cosa freqüentment sap com es diu una paraula, però no com s'escriu o quina n'és la forma estàndard.

En aquests casos el corrector ortogràfic XUXEN (Aduriz *et al.*, 1997) ofereix a l'usuari una ajuda inestimable per millorar la qualitat del text i perquè s'acostumi, a poc a poc, correcció després de correcció, a la forma estàndard. El programa XUXEN és, segons la nostra opinió, un aliat important del procés d'estandardització del basc, una eina que dóna confiança a l'usuari i que, per tant, serveix per promoure l'ús escrit. Es pot descarregar, gratis, de www.euskara.euskadi.net, cosa que ja han fet més de 20.000 usuaris, dada que mostra que l'ús del corrector es generalitza.

2.2. Consulta de diccionaris basada en la lematització i integrada a l'edició

Els productes creats en aquest tipus d'aplicació consisteixen en connectors per a l'editor de Word. L'usuari pot consultar fàcilment una paraula en un diccionari: n'hi ha prou amb fer clic sobre la paraula que vol consultar perquè aparegui en la pantalla, en una finestra dinàmica, la informació corresponent. La lematització és un instrument imprescindible per a llengües que, com el basc, tenen una morfologia rica, ja que, sense ella, no es podrien reconèixer les paraules sufixades. Suposem, per exemple, que volem saber l'equivalent en basc de la forma verbal *cabéssim*. Gràcies al connector, el programa sap que es tracta d'una forma del verb *cabr*e, per la qual cosa ens en mostra les equivalències en basc (*kabitu*, *sartu*, etc. –vegeu la figura 3-). Sens dubte, la lematització també és una gran ajuda per consultar paraules en basc. Per exemple, és capaç de trobar *kaikuenei* ('als més ximples', *kaiku* + *enei*) si busquem els sinònims de *tentelenei* ('als més ximples', *tentel* + *enei*). En aquest cas, el resultat de l'anàlisi morfològica no s'utilitza només per saber

quin és el lema, sinó també per oferir un sinònim amb el mateix sufix. Aquest sistema pot utilitzar-se amb quatre diccionaris: els de sinònims d'UZEI i d'Elhuyar, el castellà-basc d'Elhuyar i el francès-basc d'Elhuyar. D'aquí a poc temps podrà utilitzar-se també amb el diccionari anglès-basc d'Elhuyar.

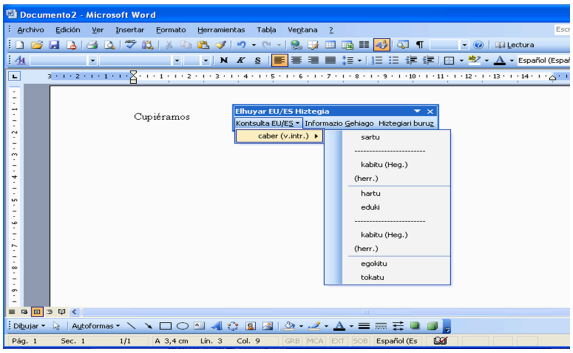

Figura 3. Consulta, amb lematització, d'un diccionari bilingüe

2.3. *Cerca de documents basada en la lematització*

El tipus d'aplicació denominada *recuperació de documents* (*information retrieval*) consisteix a triar, d'entre molts documents, un (o alguns) que contingui un concepte o una informació determinats. L'exemple més típic és el cercador d'Internet, com ara Google (www.google.com).

En els texts en basc, la cerca de paraules completes no dóna tan bons resultats com per al català, el castellà o l'anglès, ja que, sovint, les paraules contenen sufixos. Si busquéssim indicant només l'inici de la paraula, se'ns mostrarien també els resultats corresponents a altres paraules més llargues que comencen igual, amb la qual cosa ens apareixerien molts documents que no busquem. Per exemple, si volem buscar documents que continguin la paraula *ero* ('boig'), una solució possible seria detectar totes les paraules que comencen per *ero-* (buscar *ero**), però, de fer-ho així, apareixerien també documents que contenen *erosotasun* ('comoditat'), *erosi* ('comprar'), etc. i, és clar, no és això el que volem. Per tant, convé —sempre que sigui possible— fer cerques basades en la lematització per accedir a documents en basc.

ELEBILA és un cercador de documents especialitzat en la cerca de documents en basc.[8] Ara bé, com que crear tot un navegador web és una tasca fora del nostre abast, per poder buscar documents en basc, ELEBILA utilitza un cercador estàndard (MSN) que indexa i cerca paraules completes. La complexitat del programa, però, no es redueix a això. De manera inversa a MSN, Google, Yahoo i la resta de cercadors, ELEBILA distingeix les pàgines redactades en basc i, a més, si li demanem que busqui una paraula concreta, sol·licita al cercador estàndard que busqui totes les paraules que es poden obtenir afegint sufixos a la paraula sol·licitada.

2.4. *Altres aplicacions i eines públiques*

El grup IXA ha creat, dins de l'àmbit de la tecnologia lingüística, altres aplicacions informàtiques, eines i recursos lingüístics. Pel que fa a les aplicacions informàtiques, cal destacar el sistema de traducció Matxin (Alegría *et al.* 2007 o Mayor 2007) i el *Corpus de Ciencia y Tecnología* (Areta *et al.* 2007).

Als usuaris ens és fàcil entendre una llengua; a l'ordinador, en canvi, li costa molt. Quan, per exemple, llegim les paraules d'un text, no se'ns ocorren altres interpretacions curioses i extraordinàries, però sí a l'ordinador, ja que ha d'analitzar-les totes. El lematitzador del programa ajuda l'ordinador a triar la interpretació morfològica correcta entre totes interpretacions possibles, segons el context.

D'aquest manera, si li demanem que analitzi la frase *Itxura hori zuen gizonak ikusi du* ('L'ha vist l'home que tenia aquest aspecte'), l'analitzador morfològic analitza cada paraula sense tenir en compte el context (vegeu la figura 4) i ens dóna totes les interpretacions possibles per a cada paraula. Així, ens informa que la paraula *Itxura* ('aspecte') pot ser també un verb (*ADISIN*), a més d'un nom (*IZEARR*); que *hori* ('aquest'), a més d'un determinant

[8] http://www.elebila.eu

(*DETERK*), pot ser també verb i adjectiu (*ADJARR*), o que *ikusi* ('veure') pot ser nom i verb.

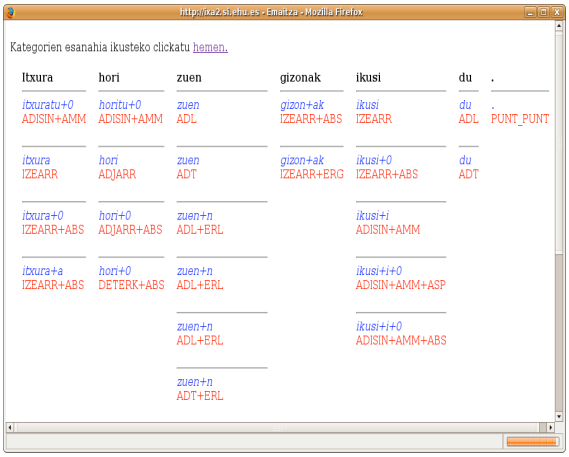

Figura 4. Resultat de l'analitzador morfològic

El lematitzador, en canvi, fa l'anàlisi morfològica tenint en compte el context de la paraula, de manera que selecciona una única anàlisi per a cada paraula (vegeu la figura 5).

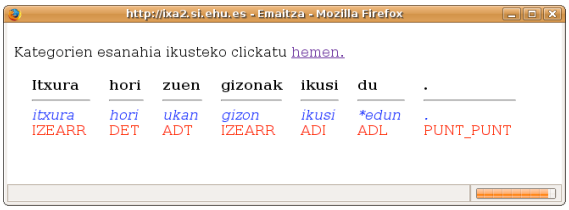

Figura 5. Resultat del lematitzador

L'analitzador morfològic MORFEUS fa una mitjana de 2,81 anàlisis diferents per a cada paraula en basc. Si tenim en compte només la categoria i la subcategoria sintàctiques, el nombre d'anàlisis per paraula es redueix a 1,5. El lematitzador, de manera contrària, selecciona únicament un lema i una categoria per a cada paraula, una vegada analitzat el context. S'equivoca, però el marge d'error no supera l'1 o el 2 %. Es tracta, per tant, d'una eina bàsica per a la tecnologia lingüística.

Per acabar, la pàgina de demostracions del grup IXA ens permet observar el funcionament d'altres eines com el reconeixedor d'entitats EIHERA, que detecta en el text noms de persones, llocs i entitats, i el divisor de sintagmes IZATI.

Pel que fa als recursos lingüístics, es poden consultar en aquesta mateixa pàgina el *Corpus de Ciencia y Tecnología* i la *Base de Datos Lexical del Euskera* (EDBL, la base lexical de tots els desenvolupaments). Cal esmentar, finalment, el recurs lexical EusWN (Agirre *et al.* 2002), el WordNet en basc. WordNet és una base de coneixement lèxic que estructura els significats de les paraules al voltant de les relacions lèxiques i semàntiques. El WordNet en basc segueix les especificacions de l'EuroWordNet; així, els significats de les paraules de llengües diferents s'interrelacionen per mitjà d'un índex interlingüístic (*Hizkuntza_Arteko_Indizea*). Es pot accedir de manera gratuïta a una interfície de WordNet en basc, català, castellà i anglès (ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl).

3. La situació actual del basc en l'àmbit de la tecnologia lingüística

En el catàleg del programari en basc SOFTKAT (www.ueu.org/softkat) hi ha 44 aplicacions relacionades amb el processament de la llengua: ajuts per a l'edició, tractament de la parla, mètodes d'aprenentatge del basc, lematitzador i eines per buscar informació, base de dades documental, corpus i vint recursos lexicals. El Govern Basc, al seu torn, està preparant l'«inventari de les TIC del basc», tot i que encara no l'ha completat.[9]

Segons el lloc Yourdictionary.com, al món hi ha unes 6.800 llengües, de les quals només 2.261 tenen expressió escrita i de les quals només 300 tenen diccionaris electrònics que es poden consultar a Internet. El basc és una d'aquestes 300

[9] www.euskara.euskadi.net/r59-734/eu

llengües, i si n'analitzem detalladament els productes relacionats amb la tecnologia lingüística, entraria, sense cap mena de dubte, en la llista de les cent i, fins i tot, de les cinquanta primeres. Això és fruit de la feina dels últims 25 anys, però potser no és suficient per fer front als reptes del futur immediat.

Si analitzem la situació del basc en l'àmbit de la informàtica amb una perspectiva global, i no ens atenem únicament a les aplicacions relacionades amb el processament de la llengua, la situació no és del tot dolenta. Si mirem de nou en el catàleg de programari en basc SOFTKAT, trobem les dades següents, classificades segons el tipus d'aplicació:

— 31 aplicacions ofimàtiques (processadors de text, comptabilitat, etc.).
— 30 de relacionades amb el lleure (música, jocs, etc.).
— 44 de relacionades amb la llengua (traductors, correctors, diccionaris, etc.).
— 56 per a Internet (navegadors, correu electrònic, etc.).
— 26 eines d'ús general (sistemes operatius, bases de dades d'Internet, cercadors, etc.).
— 74 de relacionades amb l'ensenyament o jocs pedagògics (matemàtiques, ciències, etc.).

4. Estratègia per al tractament de la tecnologia lingüística: prioritats

Tal com hem assenyalat anteriorment, el predomini de l'anglès en aquesta nova tecnologia és evident. L'anglès, de manera especial, i la resta de llengües principals, en un segon pla, han desenvolupat diversos productes i recursos de tecnologia lingüística. Què podem fer per no quedar-nos enrere? Com podem afrontar aquest desafiament?

Fa vint anys, el nostre primer projecte relacionat amb el basc tenia per objecte crear un sistema de traducció. En aquell temps, el grup comptava només amb quatre professors. De seguida ens vam adonar que no era el moment per desenvolupar un sistema de traducció, sinó que era preferible invertir totes les forces a crear eines de base per tractar la morfologia de el basc i establir uns fonaments sòlids per després desenvolupar-los. El sistema de traducció hauria estat molt limitat, amb un lèxic i gramàtiques «de nyigui-nyogui», no hauria pogut tractar la morfologia en profunditat i, per tant, res no hauria estat reutilitzable en altres aplicacions. Com que la morfologia del basc és tan diferent respecte de la de les llengües circumdants, ensopegàvem sempre amb dificultats serioses per adaptar-hi els productes d'altres llengües. Llavors, però, vam adonar-nos que el millor era començar immediatament l'anàlisi de la morfologia. Així, doncs, vam deixar per a una ocasió més bona la traducció i vam abordar en profunditat el lèxic i la morfologia. Més tard van arribar les aplicacions informàtiques relacionades amb la morfologia. Posteriorment, vam començar eines i aplicacions més complexes. La trajectòria de gairebé vint anys del nostre grup ha funcionat partint d'aquesta estratègia, que hem presentat i contrastat en fòrums internacionals (Alegría *et al.* 2001 o Sarasola 2007). Les idees centrals d'aquest projecte són les següents:

— Al principi, cal crear recursos bàsics i eines sòlides, per després començar les aplicacions de mercat.
— Cal utilitzar formats estàndard, tant en les dades inicials com en els resultats.[10]
— Sempre que sigui possible, s'ha d'utilitzar i crear programari lliure.

Sabem que aquests punts semblen «molt simples» i que són els que s'utilitzen per

[10] Els estàndards més adequats són XML i TEI.

desenvolupar qualsevol aplicació informàtica, però la nostra experiència ens mostra que en els projectes d'altres llengües amb pocs recursos no s'ha actuat així. Segons la nostra opinió, si el basc es troba entre les cent llengües principals en relació amb la tecnologia lingüística, es deu, en gran mesura, que s'ha adoptat aquesta estratègia.

Vet aquí les principals fites que marquen la nostra trajectòria:

— 1993: corrector ortogràfic XUXEN.
— 1996: EDBL, base de dades lexical de el basc.
— 1998: lematitzador.
— 2002: connector per consultar el diccionari Elhuyar Word.
— 2006: *Corpus de Ciencia y Tecnología* (ZT) i sistema de traducció Matxin.
— 2007: Euskal Wordnet, analitzador sintàctic superficial.
— 2008: traductor Matxin.

5. PROJECTES DE FUTUR

5.1. Cal crear recursos lingüístics per al basc: desenvolupament del corpus

En aquests moments, considerem d'importància estratègica per al desenvolupament de les tecnologies aplicades al basc la constitució de corpus tant monolingües com bilingües, anotats (en els àmbits morfològic, sintàctic o semàntic) i sense anotar (només text).

Pel que fa als corpus monolingües, se'n preveu recopilar un de cent milions de paraules i després anotar-les. En basc disposem ja d'alguns corpus anotats, com el *XX Mendeko Corpusa* ('el corpus del segle XX', amb 4,6 milions de paraules),[11] l'*Ereduzko Prosa Gaur* ('prosa exemplar actual', amb 9,6 milions de paraules)[12] i el *ZT* ('corpus de ciència i tecnologia', amb vuit milions de paraules)[13]; entre els corpus sense anotar destaquen *Susa,*[14] *Klasikoen Gordailua* ('clàssics'), *Ibinagabeitia Proiektua* i *Orotariko Euskal Hiztegia* ('diccionari general de basc').

Els corpus bilingües alineats són molt importants per investigar sobre traducció automàtica. Pretenem crear un corpus d'un mínim de trenta milions de paraules perquè les tècniques estadístiques ofereixin resultats mínimament acceptables. Doblar la mida del corpus pot suposar una millora de l'1 % en la qualitat de la traducció.[15] Mentre que en basc amb prou feines hi ha corpus bilingües alineats, amb pocs milions de paraules, en l'àmbit internacional hi ha molts corpus com l'Europarl,[16] que té trenta milions de paraules en onze llengües oficials.

5.2. Recopilar els continguts electrònics; biblioteca nacional

Un altre dels nostres objectius de futur és recopilar de manera sistemàtica tots els continguts en basc d'Internet (o, més ben dit, tot el que es publiqui en format digital). Aquesta recopilació de texts constituiria un recurs fonamental, tant per a la tecnologia lingüística com per a totes les ciències socials. Diverses biblioteques nacionals i altres institucions estan adoptant ja mesures en aquest sentit. El gener de 2002 es va fer una trobada internacional per tractar aquest tema.[17] A Europa hi ha actualment diverses convocatòries i projectes que tenen per objecte preservar el patrimoni històric. Entre d'altres, està molt avançada l'experiència de Dinamarca.[18] Altres referències interessants sobre aquest tema són l'European Digital

[11] http://www.euskaracorpusa.net
[12] http://www.ehu.es/euskara-orria/euskara/ereduzkoa
[13] http://www.ztcorpusa.net/cgi-bin/kontsulta.py
[14] http://www.susa-literatura.com
[15] Utilitzant la mètrica d'avaluació BLEU.
[16] http://www.statmt.org/europarl
[17] http://www.nla.gov.au/ntwkpubs/gw/56/p08a01.htm
[18] http://www.netarchive.dak

Library Project,[19] la Biblioteca Nacional de España[20] i DELOS, Network of Excellence on Digital Libraries[21].

5.3. Sintaxi, semàntica i altres aplicacions

Hem presentat al llarg d'aquest article les aplicacions desenvolupades en el camp de la morfologia, però també estan bastant avançades diverses eines per al tractament de la sintaxi i la semàntica. La combinació de mètodes simbòlics i estadístics permetrà millorar els resultats en aquests àmbits d'anàlisi i afrontar amb més perspectives d'èxit les aplicacions tecnològiques de les TIC: l'extracció i la recuperació d'informació, els sistemes d'ajuda a l'aprenentatge de llengües, els correctors ortogràfics i d'estil, els sistemes de cerca de respostes, el resum automàtic i els classificadors de documents.

6. CONCLUSIONS

La situació del basc ha millorat notablement gràcies als avenços registrats aquests últims trenta anys, però el futur no està garantit, ja que aquestes millores no són d'aplicació general i el basc continua estant al marge dels entorns industrials, incloent-hi els relacionats amb les tecnologies de la informació i la comunicació (TIC).

La nostra estratègia basada en l'estandardització, la reutilització i l'ús i la creació de programari lliure ha estat fructífera per al desenvolupament de productes. Seguint aquesta estratègia, el grup IXA ha fet diverses aportacions en investigació i en aplicacions pràctiques (corrector ortogràfic, consulta de diccionaris en línia, cercador de documents, sistema de traducció, etc.), fruit d'una anàlisi inicial profunda de la morfologia.

En l'àmbit internacional, la nostra experiència i eines poden ser molt útils en el processament d'altres llengües. Òbviament, el mercat de productes per a l'anglès és molt ampli en la indústria lingüística, però, segons la nostra opinió, aquests productes no s'han expandit adequadament cap a la resta de les llengües i, per tant, hi ha un espai ampli de treball que espera. Els bascos i, en general, els europeus estem acostumats a viure en el plurilingüisme. Per una part, el plurilingüisme ens suposa esforços, despeses i maldecaps, però, per una altra, ens situa en una posició molt avantatjosa per fer aportacions rellevants en l'àmbit internacional, ja que estem més capacitats per afrontar el multilingüisme. A més, el basc té unes característiques molt diferents en relació amb la tipologia, per la qual cosa pot convertir-se en un banc de proves perfecte per comprovar l'adaptabilitat dels productes lingüístics a altres llengües.

7. REFERÈNCIES BIBLIOGRÀFIQUES

ADURIZ, I., ALEGRÍA, I., ARTOLA, X., EZEIZA, N. i SARASOLA, K. (1997). «A spelling corrector for Basque based on morphology». *Literary & Linguistic Computing*, Oxford: Oxford University Press, 12, 1, 31-38.

AGIRRE, E., ANSA, O., ARREGI, X., ARRIOLA, J., DÍAZ DE ILARRAZA, A., POCIELLO, E. i URIA, L. (2002). «Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis». *Proceedings of First International WordNet Conference*. Mysore. Índia. 32-40.

ALEGRÍA, I., ARTOLA, X. i SARASOLA, K. (2001). «Hizkuntzaren tratamendu automatikoa: aplikazioak, tresnak, baliabideak eta oinarriak». Euskonews, http://suse00.su.ehu.es/euskonews/0110zbk/frgaia.htm.

[19] http://www.edlproject.eu
[20] http://www.bne.es/esp/bne/index.htm
[21] http://www.delos.info

ALEGRÍA, I. i RODRÍGUEZ, M. J. (2003). «Euskarararen presentzia Interneten neurtu nahian». *BAT soziolinguistika aldizkari,* 48, 89-100.

ALEGRÍA, I., DÍAZ DE ILARRAZA, A., LABAKA, G., LERSUNDI, M., MAYOR, A. i SARASOLA, K. (2007). «Transfer-based MT from Spanish into Basque: reusability, standardization and open source». *Cicling* 2007. LNCS 4394. 374-384

AMORRORTU, E. (2002). «Bilingual Education in the Basque Country: Achievements and Challenges after Four Decades of Acquisition Planning». *Journal of Iberian and Latin American Literary and Cultural Studies*, 2, 2.

ARETA, N., GURRUTXAGA, A., LETURIA, I., ALEGRÍA, I., ARTOLA, X., DÍAZ DE ILARRAZA A., EZEIZA N. i SOLOGAISTOA A. (2007). «ZT Corpus: Annotation and tools for Basque corpora». *Copus Linguistics.Birminghan.*

MAYOR, A. (2007). *Matxin: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz.* Facultad de Informática de Donostia (UPV).

SARASOLA, K. (2007). «Technology is an effective tool to promote use of Basque». *ICML Colloquium on "Language Revitalisation through Multimedia Technology". Pecs, Hungary* (en premsa).