eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Hiztegietan oinarritutako hizkuntza arteko dokumentuen berreskurapena

Xabier Saralegi
Tutoreak: Iñaki Alegria eta Eneko Agirre

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua

lortzeko bukaerako proiektua

2012eko iraila

Sailak: Lengoaia eta Sistema Informatikoak, Konputagailuen Arkitektura eta Teknologia, Konputazio Zientziak eta Adimen Artifiziala, Euskal Filologia**,** Elektronika eta Telekomunikazioak.

Laburpena

Kontsulten itzulpen-prozesuan gertatzen den anbiguotasun-problema ebazteko zenbait teknika azalduko dira. Euskara bezalako baliabide urriko hizkuntzetarako egokiak diren tekniketara mugatuko gara: Galdera egituratuak eta kookurrentzietan oinarritzen den beste teknika bat. Horrez gain, anbiguotasun-problema gainditzeko hain baliagarriak diren itzulpenen probabilitateak internetetik eskuratzeko metodo berri bat ere aurkeztuko dugu.

Abstract

This work addresses the specific problem of the translation selection. Dictionary-based approaches have been explored, because of the lack of sufficient parallel corpora for Basque. Specifically, we have concentrated on testing two methods: structured queries and cooccurrence-based methods. In addition, we introduce a method which exploits context similarity-based techniques in order to estimate word translation probabilities using the Internet as a bilingual comparable corpus.

Aurkibidea

1 Proiektuaren definizioa

Informazio eleaniztunaren tratamenduarekin lotutako ikerketa-gaien artean, C*ross-Language Information Retrieval* (CLIR) eta horren baitan dagoen *Cross-language Document Retrieval* (hizkuntza arteko dokumentuen berreskurapena) ditugu. Eleaniztasunaren eremura eramandako teknologia asko bezala, horiek ere elebakarrean egindako lanetik abiatzen dira. Elhuyar Fundazioak apustu estrategikotzat du gizarte eleaniztunean euskarak hizkuntza-estatus normalizatua izateko tresnak eta baliabideak garatzea. Elhuyar Fundazioak 2010-2012 urte-bitarterako onartu zuen Plan Estrategikoan, Hizkuntzen industria ezarri zen xede-sektore nagusietakotzat Hizkuntza Teknologien arloko Ikerkuntza eta Garapeneko jarduerarako. Aurkezten dugun lana egoki kokatzen da ildo estrategiko horretan, eta ezarritako helburua betetzen laguntzen du garatu nahi den teknologiak. CLIR teknologia Elhuyar I+G taldeak estrategikotzat duen Informazioaren berreskurapena eta erauzketa ikerketa-ildoan kokatzen da.

Aurrekarien inguruko atalean CLIR ikerketa-ildoa eta, batez ere, CLIR teknologiaren beharra justifikatu nahi dugu. Zoritxarrez, literaturan ez dago artikulurik CLIR teknologiaren beharra datu soziolinguistikoenetan oinarrituz azaltzen duenik. Hala ere, irizpide soziolinguistikoak oso garrantzitsuak dira edozein hizkuntza-produktu garatzerakoan[1]. Hori dela eta, kapitulu horretan mundu mailako datu estatistikoak erabiliz, CLIRen beharra justifikatzen saiatuko gara.

Dokumentuen berreskurapenerako sistema baten helburua erabiltzaileak egindako kontsultarekiko esanguratsuak diren dokumentuak edo dokumentu-zatiak automatikoki berreskuratzea da. Dokumentuak eta galdera hizkuntza ezberdinetakoak baldin badira, hizkuntza arteko sistema bati buruz ari gara. Kasu horretan sistemak gai izan behar du hizkuntza batean jasotako galderatik abiatuta zenbait hizkuntzatako dokumentuak berreskuratzeko. Hizkuntzaren mugari aurre egiteko itzulpena galderatik bildumara, bildumatik galderara edo tarteko hizkuntza batera burutu daiteke. Literaturan galdera itzultzeko hurbilpena aztertuena da, batik bat eskalagarriena delako.

Itzulpen-metodoari dagokionez hiru aukera nagusi daude; Itzulpen automatikoa, corpus paraleloetatik trebatutako itzulpen ereduak eta MRDetan (*Machine Readable Dictionary*) oinarritako itzulpenak. Lehenengo biekin lortzen dira emaitzarik onenak (itzulpen-hautapena ebazten dutelako), baina zoritxarrez soilik hizkuntza bikote gutxi batzuetarako daude

1 Miren Igone Zabala eta Mikel Lersundi. Hizkuntz produktuen diseinurako irizpide linguistiko eta soziolinguistikoak. HAP masterra.

eskuragarri. Domeinu berezituaz (adib. medikuntza, biologia...) ari bagara ere baliabide urriak dira. MRDak aldiz ugariagoak dira orokorrean. Hori dela eta, MRDetan oinarritako hurbilpena hartuko dugu lan honetan.

Hirugarren atalean kontsulten itzulpen-prozesuan gertatzen den anbiguotasun-problema ebazteko zenbait teknika azalduko ditugu. Aipatu beharra dago lan esperimental hauen garapenean Elhuyar I+Gko beste kide batek (Maddalen Lopez de Lacalle) ere parte hartu zuela. Atal esperimental hau ingelesez dago idatzita esperimentu horien inguruan bi artikulu zientifiko (Saralegi eta López de Lacalle, 2009; Saralegi eta López de Lacalle, 2010) dagoeneko argitaratuta dauzkadalako. Hortaz, itzulpen-lana aurrezte aldera erabaki hori hartu dut.

Gure lanean euskara bezalako baliabide urriko hizkuntzetarako egokiak diren tekniketara mugatuko gara. Lehenengo esperimentuan itzulpenen hautapenerako metodo desberdinak ebaluatuko ditugu. Bi teknika aztertuko ditugu nagusiki; galdera egituratuak eta kookurrentzietan oinarritzen den beste teknika bat, izan ere, badirudi azken metodo hauek egokienak direla corpus paraleloak eskuragarri ez daudenean. Bi metodoekin lortutako emaitzak konparatuko dira, metodo bakoitzaren ahuleziak zehaztuko dira eta azkenik, bi teknikak bateratzen dituen metodo hibrido bat proposatuko dugu. Bigarren esperimentuan hiztegi batetik lor daitezkeen itzulpenen-probabilitateak corpus paraleloak erabili gabe estimatzeko metodo bat landuko da. Hiztegi batean ordainen ordena ez denez beti erabilera-maiztasunarekin bat etortzen, hiztegia berrantolatzeko metodo bat diseinatuko dugu, horretarako internet corpus konparagarri bat balitz bezala erabiliz.

Azkenik, lan honetatik ateratako ondorioak plazaratuko ditugu. Ondorio horiek abiapuntutzat hartuta aurreikusten ditugun lanak ere zehaztuko ditugu.

2 Aurrekariak

2.1 CLIRen motibazioa

Atal honetan CLIR teknologiaren beharra edo motibazioa aztertu nahi dugu. Zehazki, aztertu nahi dugu ea ezagutzaren gizarte deitutakoan beharrezkoa den berreskurapen-atazak testuinguru eleaniztun batean garatzeko teknika berezituak (CLIR) inplementatzea. Azken finean ikusi nahi dugu ea proposatzen ditugun teknikak benetan aplikagarriak eta baliagarriak izan litezken eszenatokiren batean. Azterketa hori bi urratsetan aurkeztuko dugu. Lehenengo, IR sistemen erabiltzaileen eta bildumen profil soziolinguistikoak orokortzen saiatuko gara. Horrez gain, IR prozesuen eraginkortasuna eleaniztun testuinguruak tratatuz hobetzerik al dagoen aztertuko dugu. Adibidez, informazio esanguratsu osoa norberaren hizkuntzan baldin badago tratamendu eleaniztasunak ez luke hobetuko berreskurapen-prozesua erabiltzailearen ikuspuntutik. Aitzitik, erabiltzaile batek informazio esanguratsuaren parte nabarmen bat bere lehenengo hizkuntzan ez balu, orduan etekin handia aterako lioke hizkuntza arteko berreskurapen prozesuari. CLIRek eman dezakeen hobekuntza hori posible den aztertzeko ezinbestekoa da gaur egungo IR sistemen erabiltzaileen eta bildumen profil soziolinguistikoak zehaztea. Modu horretan jakin ahal izango dugu zein testuingurutan -hizkuntzak, gaia, kontsulta mota...- merezi duen hizkuntza arteko teknikak garatzeak.

2.1.1 Informazioren gizartea

Gaur egungo gizartean informazio-trukaketak garrantzi handia du jarduera eta sektore askotan. Joera hori 90. hamarkadatik dator, informazioa protagonismo handia hartzen hasi zenetik. Batzuek informazioren gizartea izenarekin bataiatu zuten gizarte-eredu berri hura. Informazioa ezinbesteko erregai bihurtu da fluxu komertzialak eta kulturalak dinamizatzeko. Teknologia berririk aipatu gabe informazioren eztanda hori ezin da ulertu. Beste alor batzuetan bezala informazioaren tratamenduan ere teknologia berriek ezinbesteko tresna bihurtu dira. Informazio eta komunikaziorako teknologiak (IKT) izena dute informazioaren tratamendurako zuzenduta dauden teknologiak. IKTek informazio-trukaketa errazten dute, espazio eta denboraren mugak gainditzen dituen komunikazio asinkronoa eta nonahikoa eskainiz.

Esan bezala, gaur egungo gizartean informazioa garrantzi handiko lehengaia da zenbait jarduera eta sektoretan. Adibidez, ezagutza-alor guztietan aritzen diren profesionalen erronka

batzuk egunean egotea, berritzaileak izatea eta batez ere ekarpenak egin eta ezagutza berria ekoiztea dira. Helburu horiek lortzea askoz errazagoa da gaur egungo IKT teknologiei esker, honako abantailak eskaintzen dituztelako:

- Informazio mota guztiak atzitzen erraztea
- Datuak prozesatzeko tresnak
- Berehalako komunikazio-kanalak
- Biltegiratze-ahalmena
- Atazak automatizatzea
- Interaktibotasuna

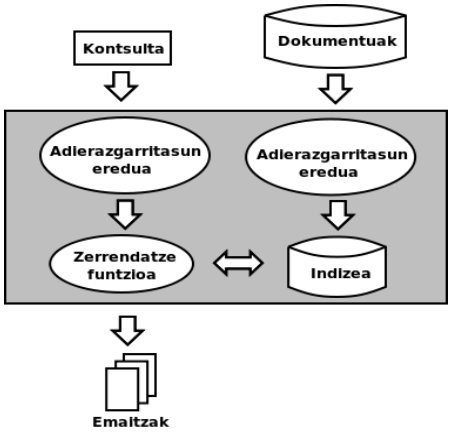
Gizarte berri honen erregaia informazioa bada ere, informazioak ez du zertan ezagutza inplikatu behar (Jacoby et al., 1974). Halere, ezagutza da benetan balio duena eta benetan zuzenean aplika dezaketena balio erantsi moduan ezagutza-alor guztietan aritzen diren profesionalek. Ildo horretan egile batzuek informazioaren gizartearen ordez ezagutzaren gizartea terminoa proposatu dute bereizketa hori nabarmendu nahian. Informazioa gertaerez osatuta dago, ezagutza, ordea urrunago doa, gertaerak testuinguru batean interpretatzean datza. Alegia, informazioa ez dela nahikoa gure jardueretan zuzenean aplikatzeko. Informazioak nolabaiteko sukalde-lanak behar ditu ezagutza bihurtzeko. Esate baterako, medikuek hartu beharreko erabaki askok diagnosia asmatu eta tratamendua aukeratzea hartzen dituzte. Mota honetako prozesu kognitiboetan (erabakiak hartzea) paradoxa bat gertatzen da: informazio gehiegi egoteak ulermena eta erabakiak hartzea oztopatu ditzake, alegia ezagutza erauztea eragoztea. Toffler-ek (1970) informazio-saturazioa[2] deitu zion fenomeno honi. Zenbait ikerketek fenomeno hau baieztatu zuten. Produktuen marken aukeraketaren inguruko azterketa baten arabera (Jacoby et al., 1974) produktuei buruzko informazio gehiago izateak erabaki okerragoak hartzea eragiten du. Beste egile batzuk ondorio berdinera iritsi ziren osasunaren alorrean egindako azterketa batean (Kim et al., 2007). Laburbilduz, honakoak izango lirateke informazioren saturazioa eragiten duten faktoreak:

- Informazio berriaren ekoizpenaren hazkundea
- Informazioren bikoizketa interneten
- Informazioa jasotzeko kanalen ugaritzea (telefonoa, sare sozialak, posta elektronikoa...)
- Zehaztasun falta eskura dagoen informazioan
- Informazioa konparatu eta prozesatzeko metodoen falta

2 http://en.wikipedia.org/wiki/Information_overload

- Informazio-zatien arteko lotura falta

Nielsen-ek (1995) hiru estrategia proposatzen ditu informazio-saturazioaren problemari aurre egiteko. Lehenengoa (kasu gehienetan eraginkorrena) erabiltzaile-interfaze onak diseinatu, eta informazioa ondo antolatuta erakustea da. Beste bi estrategiak informazioren berreskurapena (*Information Retrieval*) eta iragazketa (*Information Filtering*) dira (Belkin eta Crofy, 1992). Berreskurapenean erabiltzaileak modu aktiboan burutzen du bilaketa bat informazio zehatza behar izaten duenean. Iragazketa, ordea, etengabeko prozesu pasiboa da erabiltzailearen aldetik, non erabiltzaileak aldez aurretik gai-multzo bat adierazten duen. Adibidez, IBMko buruaren izena topatzea berreskurapen ataza bat izango litzateke. IBMk kaleratzen dituen inprimagailuen berri izatea, ordea, iragazketa-ataza bat izango litzateke.



1. irudia. IR sistema baten oinarrizko arkitektura

Informazioaren berreskurapen prozesu bat hasten da erabiltzaileak kontsulta bat sistemara bidaltzen duenean. Kontsultak erabiltzailearen informazio-beharrei dagozkien adierazpen formalak dira. Kasu batzuetan idatzitako esaldi bat, bestetan gako-hitz zerrenda bat. IR sistemak kontsulta hori hartu eta dagozkion informazio zati esanguratsuak (dokumentuak kasu gehienetan) aukeratzen ditu helburu-bilduma batetik (ikusi 1. irudia).

IR klasikoak (Baeza-Yates eta Ribeiro-Neto, 1999) informazio esanguratsua bilduman dagoela suposatzen du. Autore batzuk ez daude ados sinplifikazio horrekin eta bilatzeko prozesua askoz konplexuagoa dela defendatzen dute. *Information seeking*[3] (Informazioaren Bilaketa) paradigma proposatzen dute Informazioaren Berreskurapenaren ordez. Diziplina honek ez du suposatzen kontsulta baterako erantzun zuzenik existitzen denik. Paradigma honen arabera bilaketa prozesua konplexua da, eta erantzuna topatzeko modua *seeking* (bilaketa)

3 http://en.wikipedia.org/wiki/Information_seeking

prozesuaren beraren bidez ikasten da. Hortaz, esan daiteke *information seeking* (informazioaren bilaketa) erabiltzaileari zuzenduago dagoela, haren jarrera edo jarduera kontuan hartzen duelako.

Bereizketa hau egokia da guk nahi dugun azterketarako, eleaniztun egoera batek isla ezberdina izango baitu bi paradigmen arabera. Alde batetik, hizkuntza arteko bilaketa IRko problema teknikoa da. Bestetik, erabiltzaileak hizkuntzari dagokionean zer nolako bilaketa-estrategiak burutzen dituen Informazioaren Bilaketaren barruko problema da. Zehazki *seeking-behaviourr*ekin (bilaketa-portaera) zerikusi zuzena dauka. *Information Seeking Behaviour* (Informazioren bilaketan portaera) jendeak informazioa topatzeko eta erabiltzeko moduari buruzko diziplina da, eta honako gai nagusiak jorratzen ditu:

- Bilaketa-estrategiak
- Erabiltzaileen bilaketen patroiak
- Erabiltzaileen aurkikuntzen patroiak
- Kontsulta-terminoak eta interakzioak
- Informazio pertsonalaren kudeaketa
- Emaitzen bistaratzea eta bilaketa sozialak

Wilson-ek (2000) informazioaren bilaketan portaera honela definitu zuen: gizakiaren portaera informazioaren iturburu eta kanalekiko, bilaketa aktibo eta pasiboa, eta informazioaren erabilera kontuan hartuta. Informazioaren bilaketarako portaera mota guztietako informazio-sistemekin elkarreragiteko portaerari dagokion alderdia da. Literaturan zenbait eredu proposatu dira informazioaren bilaketa-portaera modelizatzeko (Foster, 2005; Kuhlthau 2006). Eredu guztiek orokorrean portaera hori dinamiko eta ez-lineala dela onartzen dute. Zentzu horretan erabiltzaileak informazio bilaketa prozesua ekintza, hausnarketa eta sentsazioen arteko interakzio-prozesu bat bezala hartzen du.

CLIR teknologiaren beharra erabiltzaileek bilaketetan erabilitako patroien araberakoa izan daiteke. Patroiak sailkatzeko Broder-ek (2002) taxonomia sinple bat proposatu zuen. Kontsultak hiru multzotan sailkatzen ditu erabiltzaileak duen helburuaren arabera: informazionalak informazioa lortzea helburu dutenak, nabigazionalak webgune topatzea helburu dutenak, eta azkenkik transakzionalak transakzioa helburu duten kontsultak. Eredu eta taxonomien hauen arabera eleaniztasun problema handiago edo txikiagoa izan daitekeela aurreikusi daiteke.

2.1.2 Eleaniztasuna eta informazioa

Testuinguru batzuetan (denda birtualak, posta elektronikoa, eztabaida-foroak, kazetaritzako dokumentazio-lanetan, interneteko kontsulta informazionalak...) bilatu nahi den informazio esanguratsua hainbat hizkuntzatan egon daiteke. IR sistema batean kontsultak hizkuntza batean egiten dira. Zentzua izango luke galderaren hizkuntzan ez dagoen bilduma batetik ere informazioa berreskuratzeak?

Erabiltzailea poliglota izango balitz edo bere hizkuntzara itzultzen duen MT sistema egoki bat eskuragarri izango balu emaitza aberatsagoak lor litzake. Benetan aberatsagoak izango lirateke baldin eta hizkuntza guztietatik jasotako informazio-zatiak osagarriak izango balira (adibidez, posta-elektronikoa, eztabaida-foroak, herrialde ezberdinetako egunkari digitalak...). Hortaz, ondoriozta daiteke CLIR sistemak beharrezkoa izan daitezen bi baldintza bete beharko lituzketela eszenatokiak: Alde batetik erabiltzaileak hizkuntza bat baino gehiago ulertzea, eta bestetik hizkuntza horien informazio esanguratsua osagarria izatea. Baldintza hauek IR sistema estandar bateko honako bi aldetan izango lukete eragina:

- Erabiltzailea/kontsulta:
  - Hizkuntza kopurua
  - Gaitasun-maila (edo MT sistemen kalitatea) hizkuntza bakoitzeko
- Informazioa/bilduma:
  - Hizkuntza kopurua
  - Hizkuntza arteko osagarritasuna

Testuingurua	**Erabiltzailea**		**Informazioa**	
	Hizkuntza kopurua	**Gaitasun-maila**	**hizkuntza-kopurua**	**osagarritasuna**
CLIRerako optimoa	Asko	Handia	Asko	Handia

1. taula. CLIR teknologiari etekin handiena aterako liokeen eszenatokia

Beraz, baldintza hauek betetzen dituzten eszenatokiek etekin handia aterako liokete CLIR teknologiari (Ikusi 1. taula). Mundu errealean bi alde (erabiltzailea eta bilduma) hauen izaerak eszenatokien araberakoak izaten dira. Adibidez, aisialdirako sare sozial baten erabiltzailearen profil soziolinguistikoa eta kazetari zientifiko batena ez dira berdinak izaten. Gauza bera bilatzen duten informazioren ezaugarriei dagokienez. Begi-bistakoa da bi alde

hauen ezaugarriak eszenatokiarekiko oso menpekoak direla. Eszenatokien artean ere bi alde hauen artean menpekotasun erlazio handia dago. Bataren ezaugarriek bestearenak baldintzatzen dituzte. Orokorrean, zenbat eta eduki digital gehiago hizkuntza batean egon orduan eta erabiltzaile gehiago lortuko dira hizkuntza horretako eta alderantziz[4]. Hortaz, esan daiteke bi baldintza hauek tentsio egoeran daudela, eta tentsio hori konpondu ahala batak besteak baino gehiago baldintza dezakeela. Horren ondorioz, erabiltzaile eta bildumen izaerak ez dira estatikoak denboraren zehar. Gandal-ek (2001), tentsio egoera dela eta, bi aukera aztertzen ditu etorkizuneko webeko edukien aniztasun linguistikoari begira:

a) Interneteko edukien proportzio handienak ingelesezkoa izaten segituko du. Horrez gain, denboraren poderioz egoera horrek jendea ingelesez ikastera bultzatuko du ingelesaren presentzia sendotuz. Txinerazko eta gaztelerazko hiztunak kopurua handiagoa izanda ere ingelesa nagusituko da.
b) Internet garatu ahala sareak mundu mailan erabilitako hizkuntzak hobeto islatuko ditu ingelesari pisua kenduz.

Quebecen egindako azterketaren arabera (Gandal, 2001) frantses hiztunek ingeleseko web-guneak bisitatzeko joera dute. Horrek esan nahi du emaitzak lehenengo aukeratik gertuago agertu zirela. Era berean, nahiz eta baieztatuta ez egon, hizkuntza nagusi bat[5] eta txiki bat[6] uztartzen diren inguruetako erabiltzaileen jarreran fenomeno bera gertatu daitekeela pentsa daiteke, Gandal-ek lortutako emaitzak beste hizkuntza batzuetara estrapolatuz gero. Halere, azterketa honek ez ditu kontuan hartzen joera horren kontra dauden indarrak. Euskal Herrian adibidez norabide horretako ekintza ugari daude. Erakunde publikoek nahiz internauten komunitateak euskararen presentzia handitzeko zenbait ekimen jarri dituzte martxan. Ildo horretan, ekintza hauen ahalmena neurtzea interesgarria izango litzateke.

Atalaren hasieran esan bezala, CLIR teknologikoa baldintza batzuk betetzen dituzten eszenatoki batzuetan baliagarria izan daitekeela ikusi dugu. Baldintza hauek IR sistema baten bi osagai diferentetan daukate eragina, erabiltzailearen partean eta edukien partean. Erabiltzaile poliglotek eskertuko lukete kontsulta eleaniztunak bere ama hizkuntzan egin ahal izatea. Bestetik, erabiltzaile poliglotek ere eskertuko lukete bigarren hizkuntzetan dagoen informazio esanguratsua jasotzea. Kasuistika ikaragarri handia denez, oso zaila da munduko eszenatoki guztiak zerrendatu eta parametro hauek estimatzea. Hori dela eta, hurrengo ataletan

4 United Nations (UN) report asserts, “Availability of content, in an appropriate language also affects the diffusion of the Internet. After all if you cannot find content in your language and you do not read other languages, how can you use the Internet?” ITU (1999), page 4, italics in original.

5 Hiztun askoko hizkuntza

6 Hiztun gutxiko hizkuntza

errealitatearen bi alde hauen argazki orokor bat ateratzen saiatuko gara. Erabiltzaile eta bildumen aldeen ezaugarriak independenteak eta globalak izango balira bezala aztertuko ditugu. Azterketa honek mundu mailako erabiltzaile eta informazioren ezaugarri soziolinguistikoak emango dizkigu. Modu horretan ikusiko dugu zenbateko potentziala daukan CLIR teknologiak mundu mailan. Azkenik, eszenatoki hautagai zehatz batzuk eta eszenatokien arabera CLIRen baliagarritasuna estimatzen duen neurri bat aipatuko ditugu.

2.1.2.1 Erabiltzaileen ezaugarriak

Munduko biztanle gehienak elebidunak edota eleaniztunak dira. Ama hizkuntzaz gain beste hizkuntza batzuk dakizkite. Fenomeno isolatua dela esaten bada ere, errealitatea oso bestelakoa da. Crystal-ek (1997) munduko umeen bi herenak ingurune elebidunetan hazten direla estimatzen du. Tabouret-Keller-ek (2004) Europako eleaniztasun-tasa populazioaren %50ean ipintzen du. Hortaz, esan liteke oso fenomeno naturala dela gaur egungo eta aurreko gizarteetan, bai herrialde garatuetan baita garapenean dauden herrialdetan ere.

Elebidunaren definizioari dagokionez literaturan proposamen asko topatu daitezke. Maximalisten korrontearen arabera pertsona elebidunak bi hizkuntzetan aritzeko gaitasun berdina izan behar du. Minimalistek ordea gaitasun minimo bati buruz hitz egiten dute. Beste teoriko batzuk (Garland, 2007) eleaniztasuna "*continuom*" bat bezala aztertzen hasi dira. Mutur batean bigarren hizkuntza batean gaitasun txikiena duten hiztunak kokatzen dituzte, bestean gaitasun handiena dutenak. Ringbom-ek (1985) ulermen (elebidun pasiboa) eta ekoizpenerako (elebidun aktiboa) gaitasunak bereizten ditu. IRren ikuspuntutik, sinplifikazio bat eginez, bereizketa hau interesgarria da pasiboek bigarren hizkuntza ulertzen dutelako (informazio esanguratsua ulertzeko), baina ez dira kapazak hitz egiteko (kontsultak sortzeko). Alegia, elebidun pasiboak ere CLIR sistemen erabiltzaile hautagaiak izan litezkeela. Bestetik, bigarren hizkuntzaren gaitasun mailak ere CLIR sistema baten eskakizunak edo ezaugarriak baldintzatu ditzake, adibidez informazio hori esangura-maila kalkulatzeko algoritmoetan irizpide moduan integratuz.

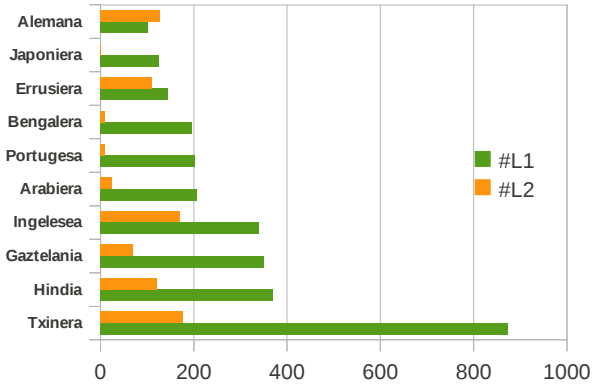
Pertsona eleaniztasunak esparru zehatzetan sortzen dira. Informazio hau interesgarria da jakiteko ze nolako esparruetan izan daitekeen aplikagarria CLIR teknologia. Elebidunen eremu nagusiak honakoak dira:

- Estatuko hizkuntza nagusiaz aparte tokiko hizkuntzaren hiztunak (Euskara-gaztelania, maltera-ingelesa...). Askotan diglosia egoerak[7] sortzen dira non hizkuntza ofizialak besteak baino prestigio handiagoa daukan.

7 http://en.wikipedia.org/wiki/List_of_diglossic_regions

- Arrazoi profesionalengatik edo aisialdiarengatik bigarren hizkuntza bat ikasten duten pertsonak (Gipuzkoako lantegi bateko gerenteak alemana, toki turistiko bateko ostalariak ingelesa...)
- Immigranteak eta haien ondorengoak (Alemaniako turkiarrak, Kanadako txinatarrak...)
- Hizkuntza diferenteak dituzten herrien arteko mugen inguruan bizi direnak

Munduan 6000 hizkuntza inguru existitzen dira (Grimes, 1992). Hauetako gehienak hiztun gutxiko hizkuntzak dira. Era horretan hizkuntzen hiztun kopuruak pareto banaketa[8] jarraitzen du. Horrek esan nahi du hizkuntza nagusi gutxi batzuk jakinda pertsona gehienekin komunikatu daitekeela. 60 hizkuntzek baino ez dauzkate 10 milioi hiztun baino gehiago. Horietako askori *lingua franca* izena ematen zaie. Nazioarteko esparru askotan adibidez ingelesa nagusitu da (komertziala, zientifikoa...) mundu mailako *lingua franca* bihurtuz. Guztira 510 milioi pertsonek dakite ingelesez. Hortaz, hizkuntza bakar bat -edo batzuk- nagusitzen ari denez CLIR teknologia ez dela beharrezko pentsa liteke. Baina, hiztunen kopuruak sakonago aztertzen baditugu ingelesezko hiztunak mundu osoko populazioaren %8ra ez direla iristen ikusten dugu. Alegia, populazioaren %92ak ingelesez ez daki. Horrez gain, 510 milioi horietatik 340 baino ez dira jatorrizkoak (Ikusi 2. irudia), eta ikerketa batzuen arabera jendea edozein ulermen edo adierazpen prozesutan seguruago dago ama hizkuntza erabiltzen duenean. IR sistemen erabileran ere hori betetzen dela adierazten duten ikerketak badaude (Rao eta Varma, 2010). Europako Batzordeko azterketa batek ere hori berresten du[9], eta interneteko erabiltzaileen %90ak bere hizkuntzan nabigatu nahiago duela dio. Gainera, erabiltzaileen %44ak uste du informazioa galtzen duela bere hizkuntzan bakarrik nabigatzen duenean.



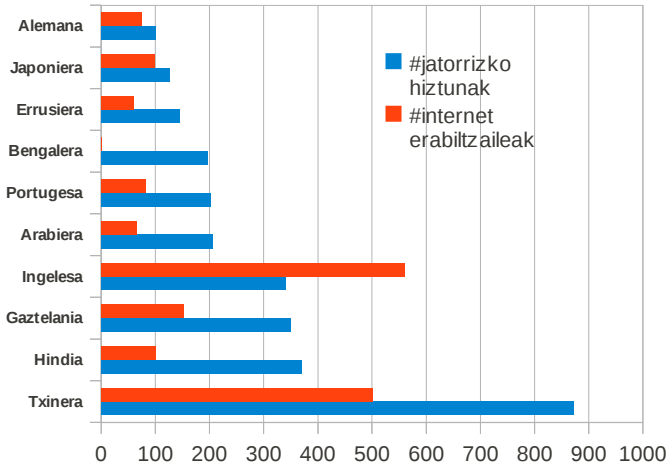
2. irudia. Elebidunen estatistikak

8 http://eu.wikipedia.org/wiki/Pareto_banakuntza

9 http://www.dw-world.de/dw/article/0,,15067034,00.html

Informazio digital gehiena hizkuntza erabilienetan eskuragarri dagoela kontuan hartuta, hizkuntza batetik hizkuntza nagusi horietara lan egiten duten CLIR sistemak beharrezkoak direla ondoriozta liteke. Are gehiago, CLIR sistema horien erabiltzaile hautagai kopuruak estimatu ere egin liteke. Erabiltzaile hautagaiak L2tzat hizkuntza nagusi bat dutenak izango lirateke. Esate baterako, hindia bigarren hizkuntzatzat dutenak (100 milioi). Berehalako erabiltzaile hautagaiak izango lirateke hindia bigarren hizkuntzatzat dutenak?

Hiztun hauen profil sozioekonomikoari erreparatuz gero ezetz erantzungo genuke, garapen ekonomiko eta teknologiko faltagatik erabiltzaile-hautagaiak ez direlako. Horrek esan nahi du CLIR sistemen erabiltzaile-hautagai kopuruak izaera eleaniztunaren arabera estimatzeko hizkuntzen egoera sozioekonomikoa adierazten duen faktore bat ere erabili behar dugula. Interneten barneratze indizeak (Ikusi figura 3.) ondo adierazten du hizkuntza baten egoera sozioekonomikoa.



3. irudia. Hiztunen estatistikak

Beraz, esan daiteke gaur egun L2tzat barneratze teknologiko handiko hizkuntza bat duten erabiltzaile gehienak CLIR sistemen berehalako erabiltzaile hautagaiak direla. Horietako batzuk L1tzat hizkuntza nagusia izango dute (adibidez, Quebeckeko frantses-ingeles hiztunak). Beste batzuk ordea hizkuntza txiki bat izango dute L1tzat (adibidez, euskara-gaztelania, katalana-gaztelania...). Europan egindako azterketa baten[10] arabera interneteko erabiltzaileen %55ek bigarren hizkuntza bat erabiltzen du edukiak irakurtzeko. Portzentaia hau %90-93 tartera

10 http://www.infotoday.eu/Articles/Editorial/Featured-Articles/Language-preferences-of-EU-internet-users-75474.aspx : http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf

igotzen da Txipre, Grezia, Luxenburgo, Malta eta Eslovenian. Nabarmena da, beraz, eremu urriko hizkuntzen hiztunek bigarren hizkuntza handiagoak (hiztun gehiago) erabiltzera joera dutela. Bestetik, aipatzekoa da zerrenda horretako herri gehienetan erabilitako L2a ez dagoela diglosia egoeran, euskara eta katalana dauden bezala. Interneteko ohiturak berdinak izango al dira egoera diglosikoetan?

Gandal-ek (2006) Quebeceko erabiltzaileen ohiturak aztertu zituen. Bere azterketaren arabera Quebeceko ingeles natiboak ingelesezko edukietara jotzen dute %87ko kasuetan. Frantses hiztunek, ordea, %64ko kasuetan irakurtzen zituzten ingelesezko edukia. Kasu asko dira frantsesak interneten presentzia handia duela kontuan hartuta. Katalunian bisitatutako web-orrien %82a gaztelerazkoak dira, %48.9a katalanez, eta %44,5a ingelesez[11]. Wei eta Kolkok (2005) Uzbekistango internet erabiltzaileen portaera aztertu zuten. Azterketan parte hartu zituzten erabiltzaileen %95ak errusieraren ezagutza bikaina aitortu zuen. %76ak uzbekeraren ezagutza bikaina zuten, eta %26ak ingelesaren ezagutza bikaina. Erabiltzaileen %66ak esan zuen errusiera zela interneten gehien erabiltzen zuen hizkuntza. Beste %34ak ingelesa erabiltzen zen gehien. Batek ere ez zuen uzbekera aukeratu. Euskal Autonomia Erkidegoko testuinguruan erabiltzaileen %23,4k ingelesez nabigatzen du. Euskaraz aldiz %22,8ak. Gaztelania %99,4rekin[12] nagusitzen da.

Orokorrean, interneteko erabiltzaileek hizkuntza handi bat -Nahiz eta haien L1a ez izan- erabiltzera jotzen dute. Portaera honen arrazoiak era askotakoak izan daitezke. Faktore nagusiak edukien bolumena edo erreferentzialtasuna izan daitezkeela pentsa daiteke. Lehen atalean aipatu dugun bezala lotura sendoa dago erabiltzaileen hizkuntzen erabilera eta edukien hizkuntzen artean. Erabiltzaile elebidunek CLIR sistema bat eskura izango balute zein hizkuntzatan egingo lituzkete kontsultak? Zeintzuk dira profil elebidun egokienak CLIR sistemak erabiltzeko?

2.1.2.2 Informazioren ezaugarriak

Informazio digitalaren bolumena esponentzialki hazten ari da azken urtetan. Honen froga da azkeneko 30 urte hauetan aurreko 5.000 urteetan (zibilizazioaren historia osoa) baino informazio gehiago ekoiztu izana. Aldi berean informazio digitala gordetzeko euskarriak ere ugaritu dira. Era horretan, gaur egun informazio digitala euskarri eta biltegi anitzetan gordeta dago informazio-fluxu nagusiak honakoak izanik (Craine, 2000):

- Posta elektronikoa: Gaur egun 225 miloi lagunek baino gehiagok jaso eta bidaltzen ditu mezuak.

11 http://www.uoc.edu/in3/wp/picwp1201/PICWP1201.pdf

12 http://www.eustat.es/elementos/ele0005100/not0005172_c.html#axzz1iO8feJC4

- WWW: 2,3 mila miloi dokumentu dauzka.
- E-merkataritza: Amazoneko kontu kopurua 2,2tik 8,4 milioira pasatu zen 1999. urtean.
- Informatika mugikorra (*Mobile Computing*): 400 milioi mugikor saldu ziren 2000. urtean.

Euskarri bakoitza testuinguru edo erabilera jakin batzuetara zuzenduta dago. Telefono mugikorretan adibidez bilaketa geolokalizatuak beste euskarri batzutan baino maizago izango da erabilgarri. Azken finean, testuinguru bakoitzak erabiltzaile eta informazio-behar jakin batzuk dauzka, non hizkuntzen aldagaia ere berezkoa den. Horren ondorioz, euskarri batzuk esparru soziolinguistiko jakin batzuetan kokatu daitezke CLIRen beharra neurri batean zehaztuz.

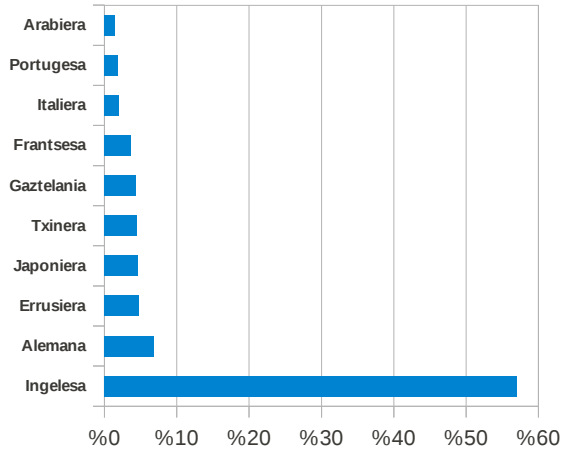
Euskarri edonolakoa dela ere informazio-beharrak eta erabiltzailearen profila nabarmen aldatu daitezke domeinu eta eremuaren arabera. Esate baterako, informazio zientifikoa edo lekuko gaien inguruko albisteak topatzea hizkuntza jakin batzutan edo bestetan bilatzera bultzatuko gaitu. Azken finean, erabiltzaile eleaniztuna informazio esanguratsua eskaintzen duen hizkuntzan bilatzen saiatuko da. Hortaz, nahikoa izango al da erabiltzaileak dakien hizkuntza nagusian bilatzea edo dakizkien beste hizkuntzetan ere informazio esanguratsua topatzea izango luke? Jarraian galdera horri erantzuten saiatuko gara informazio digitalaren ezaugarriak aztertuz.

Informazioaren bolumena

Biltegi digital eleaniztun esanguratsuena weba da. Weba mundu osoko biltegi elektronikotzat hartu daiteke eta atzigarri dago zenbait euskarritatik. Informazio digitalaren hizkuntzaren araberako estatistikak euskarri honetatik aterako ditugu interneten ondo islatuta baitago hizkuntza bakoitzean ekoiztutako informazio digitalaren bolumena.

Estatistika hauetan egin dezakegun lehen irakurketa hizkuntzen artean dagoen desoreka izango litzateke. Ildo horretan webeko edukien %57a ingelesez dagoela esanguratsua da (Ikusi 4. irudia). Erabiltzaileen hizkuntzarekin gertatzen den bezala ingelesa nagusitzen da gainerako hizkuntzetako edukiak askoz urriagoak direlarik. Izan ere, eduki gehien duen hurrengo hizkuntza %6,8ko kuota baino ez dauka. Distantzia horrek argi adierazten du ingelesaren edukien nagusitasuna. Alemanaren ondoren 5 hizkuntza baino ez daude %3ko kuota gainditzen dutenik. Bestetik, soilik 16 hizkuntzak gainditzen dute %0,5ko kuota. Horietako batzuk hiztun askoko hizkuntzak dira (adib. errusiera, txinera ...). Beste batzuk ordea ez (adib. Nederlandera).

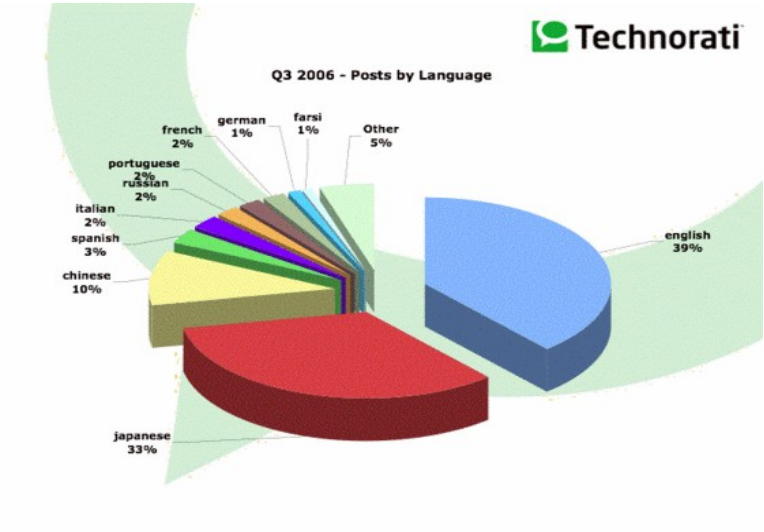
Eduki-bolumenetan ere erabiltzaile-kopuruekin gertatzen zen bezala garapenen ekonomiko handiarekin lotutako hizkuntzen presentzia handia da hiztun kopurua txikia izanda ere. Bestetik, aipatzekoa da hizkuntza bateko informazio-kopurua ez datorrela bat hiztun kopuruarekin. Adibidez, indonesiar batek eduki gutxi dauzka bere hizkuntzan (%0,3) nahiz eta mundu mailan hiztunen proportzioa askoz handiagoa den.



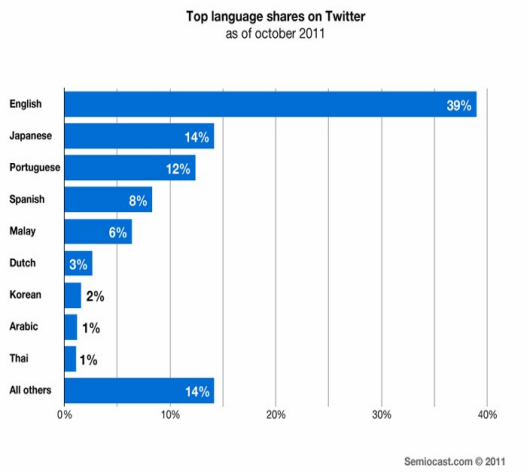
4. irudia. Hizkuntzen araberako web-edukiak[13]

Web osoaren estatistikak ikusitakoan honako ondorioa atera daiteke: Hizkuntza batzuetako hiztunak bilaketa-prozesu batzutan informazio esanguratsurik ez topatzerik gerta daiteke. Horren ondorioz, beste hizkuntza batzuetara jo beharko lukete. Jo behar horren maila hizkuntza bakoitzaren eduki kopuruaren araberakoa izango da. Problema hori nabarmena da bai hizkuntza txikietan baita garapen teknologiko gutxiko hizkuntzetan ere. Horrelako kasuetan ezinbestekoa da beste hizkuntza handiago batera jotzea. Egongo al da eremu edo testuingururik non hizkuntza batzuen nagusitasuna txikiagoa izango den?

13 http://w3techs.com/technologies/overview/content_language/all



5. irudia. Hizkuntzen erabilera blogetan[14]



6. irudia. Hizkuntzen erabilera Twitteren[15]

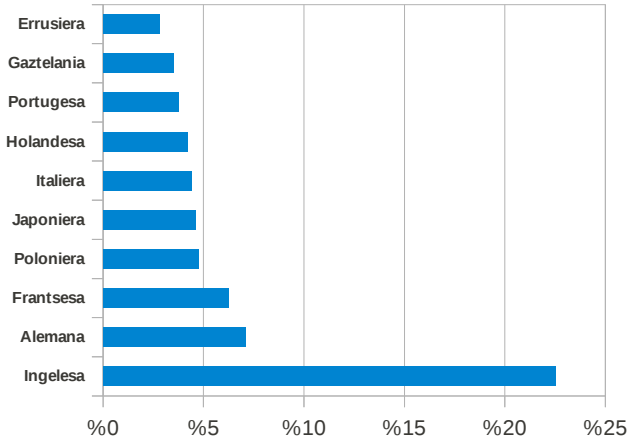
Web barruko sare sozialen eremura mugitzen bagara antzeko argazkia topatzen dugu, baina oraingoan ingelesaren ondoko hizkuntzen arteko atomizazioa txikiagoa delarik. Blog (Figura 5.) eta *microblogging* (Figura 6.) eremuetan japonierak eta beste hizkuntza batzuek informazioaren %10a gainditzen dute. Eduki entziklopedikoetara mugatzen bagara, fenomeno bera gertatzen da, ingelesak indarra galtzen du gainerako hizkuntzen aurrean (Figura 7.). Hori dela eta, hizkuntza horietan informazio esanguratsua topatzeko aukerak ugaritzen dira ingelesaren beharra arinduz. Dena dela, hizkuntza gehienen bolumenek oso txikiak izaten segitzen dute. Hortaz, badirudi webeko eremu gehienetan ingelesaren nagusitasuna -eta neurri

14 http://www.sifry.com/alerts/archives/000493.html

15 http://lsaweb.com/language/2011/11/new-study-reveals-top-languages-on-twitter/

batean beste hizkuntza gutxi batzuena- mantentzean dela. Hori dela eta, erabiltzaileak eduki gehiago lortuko du bere informazio-beharra asetzeko hizkuntza horietara jotzen badu.

Dena dela, askotan erabiltzaileak bere ama hizkuntzan lor dezakeen estaldura baino gehiago ez du behar. Alegia, bere bilaketak asetzeko masa kritikoa dagoela bere ama hizkuntzako edukietan. Gaztelerazko hiztun askok gazteleraz formulatzen dituzte bilaketa gehienak nahiz eta ingelesezko ezagutza handia izan. Portaera hori azaltzeko erosotasunaren faktoreaz aparte masa kritikoaren faktorea aipatu beharko litzateke. Masa kritikoaren falta edukien kuota txikia izatea baino arrazoi sendoagoa da beste hizkuntza nagusi batera jotzeko. Euskal hiztun askok adibidez askotan jo behar izaten dute beste hizkuntza batzuetara euskarazko edukietan informazio gutxi dutelako. Egoera horretan hizkuntza asko daude. %0.1ko kuota ezartzen badugu masa kritikoa lortzeko atalase gisa 36 hizkuntzek soilik gainditzen dute kuota hori. Gainerako hizkuntzak behetik geratzen dira. Nabarmentzekoa da hizkuntza horietako askok hiztun dezente dituztela (adib. hindia, urdua, tagaloa, swahilia...).



7. irudia. Wikipediako edukien banaketa hizkuntzen arabera[16]

Informazioaren osagarritasuna

Hizkuntza batean dagoen informazio-bolumena erabakigarria da erabiltzaile eleaniztun batek bilaketetarako hizkuntza aukeratzen duenean. Esan bezala, dakizkien hizkuntza guztietan gaitasun berdina badu beste hizkuntzetako edukiak baliagarriak izango dira baldin eta eduki horietan informazio esanguratsu berria aurkitzen badu. Beste era batera esanda, edukiek dibergenteak edo osagarriak izan behar dute hizkuntzen artean. Bestela informazio erredundante besterik ez du eskuratuko gainerako hizkuntzetatik. Hori dela eta, beste hizkuntza batera jotzea

16 http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

zentzuzkoa den aztertzeko osagarritasunaren faktorea ere kontuan hartu behar da. Osagarritasun falta hizkuntza bikote eta eszenatokien arabera aldatu daiteke. Webean bertan ere hizkuntza ezberdinen edukien osagarritasuna aldakorra da. Adibidez, dibulgazio zientifikoaren eremuan euskarazko edukiak eta ingelesezkoa asko teilakatzen dira norabide batean. Eztabaida-foroen kasuan hizkuntzen arteko dibergentzia, ordea, handia da. Gai edo domeinuaren araberako hizkuntzen arteko osagarritasun-mailen estimazioak 2. taulan erakusten ditugu. Esan beharra dago estimazio horiek gure intuizioan baino ez direla oinarritzen.

Gaia	**Dibergentzia**	**Gaia**	**Dibergentzia**
Artea	+	Osasuna	-
Jolasak	-	Komunikabideak	+
Umeak	+	Gizarte	+
Erreferentzia	+	Informatika	-
Erosketak	+	Etxea	+
Ekonomia	+	Aisialdia	+
Zientzia et Tek.	-	Sare sozialak	+
Kirolak	+		

2. taula. Gaiaren hizkuntza arteko osagarritasuna

Informazioren kalitatea

Beste puntu bat esangura-mailarekin zerikusia duena informazioaren kalitatea da. Bilaketa informazionaletan eta nabigazionaletan adibidez edukien sinesgarritasuna edo erredakzio maila bete beharreko faktore kritikoak dira erabiltzaileen ikuspuntutik. Kontsulta transakzionaletan interfazeen itxura ere kontuan hartu beharreko irizpidea da. Puntu edo faktore hauek okerragoak izan daitezke hizkuntza txiki edo garapen fasean dauden herrialdetako hizkuntzen kasuan hiztun askoko eta garapen handiko hizkuntzen kasuan baino. Azken finean kalitate-faktoreak baliabide ekonomikoen menpe daude neurri handi batean (adib. egunkariak, dendak, ...). Literaturan web-guneen kalitatea neurtzeko zenbait adierazle proposatu dira (Barnes eta Vidgen, 2000).

2.1.2.3 Eszenatoki hautagaiak

Aurreko ataletan interneteko eduki gehienak hiztun askoko eta garapen ekonomiko handiko hizkuntzetan daudela eskuragarri ikusi dugu. Hizkuntza mota horien artetik ingelesa da diferentzia handiz webean eduki gehien dituen hizkuntza. Aldi berean, ingelesa da txinerarekin batera L2 hizkuntza zabalduena. Horrek esan nahi du hizkuntza batetik ingeleserako norabideko kontsultak eskaintzen dituen CLIR sistemak izango liratekeela praktikoenak. Praktikotasun hori, arrazoi berdinengatik, hizkuntza batetik hiztun askoko eta garapen handiko hizkuntza baterako norabidera ere estrapola liteke. Praktikotasuna handia izan dadin helburu hizkuntzako edukiak ondo ulertzeaz gain jatorrizko hizkuntzan topatzen ez den informazio esanguratsu berria eskuratu beharko litzateke. Alegia, bigarren hizkuntzan dauden edukietan informazio osagarria eman behar dutela. Ildo horretan, aipatzekoa da Google bilatzailearen politika; erabiltzailearen hizkuntzan informazio esanguratsurik topatzen ez duenean bilatzaileak hizkuntza nagusietan topatu dituen emaitzak ekartzen dizkio erabiltzaileari[17].

Hortaz, badirudi ugaritasunaz $R(L_i)$ gain osagarritasuna $C(L_i,L_j)$ ere adierazle ona dela CLIR sistemen beharra islatzeko. Bestetik, bigarren hizkuntza batetik eskuratutako informazioa ulergarria izan dadin gutxieneko ulermen-gaitasuna bermatu behar da. Hizkuntzaren gaitasun-mailarekin G(L_i) erlazionatu daiteke trebetasun hori. Kasu horretan itzulpen automatikoaren kalitatea $Q\_TR(L_i,L_j)$ izango litzateke adierazlea. Itzulpen automatikoko sistemen erabilerak eragin izan dezake adierazle horretan. Azkenik, informazioaren kalitatea $Q(L_i)$ izango genuke hizkuntza bateko informazioaren baliagarritasuna neurtzeko adierazletzat.

Hainbat hizkuntza L_i eskaintzen dituen CLIR sistema baten beharra eszenatoki jakin batean neurtzeko aipatutako adierazleak barneratzen dituen honako neurria proposatzen dugu:

$$need_{CLIR}(\{L_i\}) = \sum_i G(L_i) + R(L_i) + Q(R(L_i)) + \sum_j C(L_i, L_j) \quad [18]$$

Neurria oso malgua da eta zehaztasun maila ezberdinetako eszenatokien gainean aplika daiteke. Ezaugarri hau oso garrantzitsua da literaturan baitago proposatuta eszenatokien taxonomia estandarrik. Hurrengo pausua neurri hau eszenatoki abstraktu eta errealen kontra egiaztatzea izango litzateke. Literaturan, ordea, ez da erabili horrelako neurrik CLIR sistemen baliagarritasuna kuantifikatzeko, eta aplikazio-adibideak, non CLIR lagungarria izan daitekeen, argudio sendorik gabe botatzen dira.

17 http://insidesearch.blogspot.com.es/2011/11/ten-recent-algorithm-changes.html

18 if (kal(tr(R(Li))>konp(Li)) konp(Li)=kal(tr(R(Li))

Oard-ek (1997) adibidez CLIR teknologiak online zerbitzu komertzialen merkatua zabaltzeko (adib. Dialog[19] eta Lexis/Nexis[20]) lagun dezakeela adierazten du. Beste ikerketa batek (Cleveland et al., 2007) dio hizkuntza muga nabarmena dela Dallas (EEBB) inguruneko komunitate txinatarrarentzat kalitatezko informazio medikoa (gehiena ingelesez) online topatzeko. Beste ikerlari batzuk (Chen eta Bao, 2009) aplikazio eremu gehiago aipatzen dituzte EEBBetan zentratuta:

- Ingelesez gaizki moldatzen diren immigranteek dokumentazio administratiboa bilatzeko zailtasunak dituzte.
- Atzerrian inbertsioak egin nahi dituzten inbertsoreak atzerriko erakunde eta enpresei buruzko informazioa topatu dezaten.
- Bigarren hizkuntza bat ikasten ari direnek ingelesezko kontsulta batetik bigarren hizkuntza horretan idatzitako informazio topatu dezaten.
- Botika edo tratamendu medikuei buruzko informazioa atzerrian bilatu nahi duten pazienteak.
- Atzerrira bidaiatzen dutenek bertako informazioa topatzea.

Gonzalok (2002) ere CLIR aplikagarria izan daiteken eszenatokiak zerrendatzen ditu. Kasu honetan eszenatokiak lehengoak baino orokorragoak dira.

- Galdera-erantzun sistemak.
- Irudien bilaketa.
- Bilaketa bibliografikoak.
- Web surfing.
- IR multimodala.

Dena dela, literaturan aipatutako erabilera-adibide horietatik ez dugu kuantifikatu CLIRen teknologia zenbatekoa den. Guk proposatutako neurria erabiliz, ordea, esan ahal izango dugu maila globalean eta eszenatoki zehatzagoetan CLIRen beharra edo praktikotasuna zenbatekoa den. Etorkizuneko lan bezala aurreikusten dugu hemen proposatuko neurriaren ebaluazioa.

19 http://www.krinfo.com

20 http://www.nexis.com

2.2 Artearen egoera

CLIR gero eta gai garrantzitsuagoa bihurtzen ari da hainbat arrazoirengatik. Alde batetik, informazio eleaniztunaren hazkundea etengabea da, eta bestetik munduko biztanle gehienak eleaniztunak dira. CLIR ohiko sistemek erabiltzaileei kontsultak bere ama hizkuntzan egitea eskaintzen diote, dokumentu esanguratsuak edozein hizkuntzatan idatzitako bilduma batetik hartuta.

Hizkuntzen arteko langa gainditzeko estrategia desberdinak proposatzen dira itzultzen den elementuaren arabera: kontsulta, dokumentua edo biak. Egile gehienek kontsultak itzultzeko estrategia lantzen dute, batez ere, estrategia hau oso arina delako memoria eta prozesu eskakizunei dagokienez (Hull and Grefenstette, 1998). Estrategia bakoitzarekin lortutako emaitzei erreparatuz gero dokumentuak itzuliz eraginkortasun onena lortzen da. Arrazoia honakoa da; dokumentuek kontsultek baino testuinguru zabalagoa dute. Hori dela eta, itzulpen-hautapena zuzena aukeratze errazagoa da. Horrez gain, dokumentu batean kontsulta batean baino esaldi edo adibide gehiago daude. Hitz bat gaizki itzuli daiteke esaldi batean, baina askoz zailagoa da esaldi guztietan gaizki itzultzea. Kontsultan, ordea, hitza ondo itzultzeko aukera bakarra dago. Oard-ek (1998) demostratu zuen itzulpenaren kalitatea eta berreskurapen-prozesuaren eraginkortasuna hobetzen direla bildumak itzultzen direnean. Beste ikerlari batzuek (McCarley, 1999; Chen and Gey, 2003) oraindik emaitza hobeak lortu zituzten kontsultak eta bildumak itzuliz. Ranking bana sortzen dituzte kontsultak eta bildumak abiapuntutzat hartuta. Ondoren, bi eratan lortutako rankingak konbinatzen dituzte.

Itzulpena burutzeko teknikak hiru multzo nagusitan banatu daitezke erabilitako ezagutza-iturriaren arabera: Itzulpen automatikoa, corpus paraleloak, eta hiztegi elebidunak. Azkeneko bi multzoetarako *framework* estatistiko diferenteak proposatu dira literaturan; hizkuntza arteko eredu probabilistikoak, eta hizkuntza arteko hizkuntza ereduak. Lehenengoa hiztegi elebidunekin konbinatzeko zuzenduta dago itzulpen anbiguoak tratatzeko eragile bereziak eskainiz. Bigarrenak corpus paralelo batetik lortutako itzulpen-probabilitateak *framework* formalago batean integratzen ditu (Hiemstra, 2000). Bi *framework* edo eredu hauen eraginkortasuna erabilitako baliabide motaren funtzioan badago ere hizkuntza arteko hizkuntza ereduekin emaitzak hobeak lortzen dira kasu gehienetan (Xu et al., 2001). Halere, lehen esan bezala, eredu honek corpus paraleloak eskatzen ditu, eta baliabide hau oso urria izaten da hizkuntza bikote gehienetarako. Hiztegi elebidunak, ordea, askoz ugariagoak dira baina zoritxarrez ez dituzte itzulpen-probabilitateak ematen. Hortaz, itzulpen anbiguoak tratatzeko metodoak inplementatu behar dira. Izan ere, itzulpen anbiguoak dira errore-iturri nagusia

hiztegietan oinarritutako hizkuntza arteko berreskurapen prozesuetan. 3. taulan adierazten den bezala (Saralegi eta Lopez de Lacalle, 2010). Artikulu horretan hiztegietan oinarritutako hizkuntza arteko berreskurapenean gertatzen diren problemen eragina aztertu zen. *Baseline* moduan hiztegiko lehenengo itzulpena hartzea finkatu zen. Hiztegik kanpoko hitzen (*Out-of-Vocabulary*) eta itzulpen anbiguoen tratamenduak eragiten zuten hobekuntza kalkulatu zen. Horretarako eskuz zuzendu ziren bi errore mota horiek. Emaitzetan ikusi zen hobekuntza handiena itzulpen anbiguoen tratamenduak eragiten zuela (0.27tik 0.34ra). Beraz, ondoriozta daiteke itzulpen anbiguoen tratamendua ezinbestekoa dela CLIR sistema bat garatzerakoan.

Pirkolak (1998) itzulpen-hautapena lantze aldera eredu probabilistikoetan kontsulta egituratuak integratzea proposatu zuen. Kontsulta egituratuek kontsulta bateko esanahi berdineko hitzak eragile baten bidez (*syn*) multzokatzeko aukera ematen dute. Modu horretan kontsultako hitz baten itzulpen guztiak hitz berdina izango balitz bezala tratatzen dira TF eta DF estatistikoak[21] kalkulatzen direnean. Halere, TF eta DF estatistikoen kalkulua kontserbadorea da. Jatorrizko hitz batek bi itzulpen hautagai baldin badauzka bi hautagai horien TF eta DF balioak batzen dira jatorrizko hautagaiarenak kalkulatzeko. Horrek esan nahi du hautagai bat oso hitz orokorra baldin bada, jatorrizko hitzari garrantzia kenduko diola kontsultako beste hitzen aldean. Hautagai orokor hori itzulpen zuzena baldin bada ez dago problemarik. Baina itzulpen okerra bada jatorrizko hitzak pisua galduko du erantzunen ranking kalkulatzen denean. Ildo horretan, autore batzuk (Darwish and Oard, 2003) galdera egituratu probabilistikoak proposatzen dituzte, non itzulpen-hautagaiei pisuak ematen zaizkien. Era horretan TF eta DF estatistikoak pisu horien arabera kalkulatzen dira modu orekatuago baten.

Galdera egituratuez gain kookurrentzietan oinarritutako metodoak proposatzen dira (Monz and Dorr, 2005; Ballesteros and Croft, 1998; Gao et al., 2001) itzulpen-hautapenari aurre egiteko. Metodo hauek helburu-bilduma hizkuntza eredu bat izango balitz bezala erabiltzen dute itzulpen hautapena bideratzeko. Literaturan proposatutako algoritmoek helburu-bilduman elkartze-maila handiena duten hautagaiak aukeratzen dituzte itzulpen zuzen moduan. Algoritmoek elkartze-maila orokor hori kalkulatzeko estrategia ezberdinak hartzen dituzte (Monz and Dorr, 2005; Gao et al., 2001; Gao et al., 2002; Liu et al., 2005).

Itzulpen anbiguotasunaz gain lehen aipatu bezala badaude beste problema batzuk kontsultaren itzulpen-prozesuan gertatzen direnak; Hiztegitik kanpoko hitzen presentzia eta hitz anitzeko unitateen itzulpena. Kognatuen detekzioa[22] da lehenengo errore mota hori konpontzeko

21 TF eta DF estatistikoek hitz batek dokumentu batean duen maiztasuna eta bera agertzen den dokumentu kopurua adierazten dute hurrenez hurren. Kontsulta bati dagozkion dokumentu esanguratsuen zerrenda kalkulatzeko erabiltzen dira estatistiko hauek.

22 Antzeko ortografia duten hitzak itzulpen-baliokideak izaten dira (adib. *sistema*<->*system*)

proposatzen den estrategia nagusia (Knight and Graehl, 1997). Hitz anitzeko unitate asko ezin dira hitzez hitz itzuli. Normalean hitz anitzeko unitateen zerrendak erabiltzen dira horien tratamendurako (Ballesteros and Croft, 1997).

Itzulpen-metodoa	**MAP**
1. itzulpena	0.27
OOV (orakuloa)	0.31
Itzulpen-ambiguoak (orakuloa)	0.34

3. taula. Itzulpen prozesuan gertatzen diren erroreen eragina

3 Metodologia

Atal honetan ebaluaziorako metodologia, eta esperimentuak burutzeko erabili ditugun datu-baliabideak eta algoritmoak azalduko ditugu. Lan honetan proposatzen diren teknikak hurrengo atalean deskribatuko dira.

IR sistemen ebaluazioa ikergai zabala da, informazioa bilatzeko erabilgarritasuna, eraginkortasuna eta behar diren baliabideak bere baitan hartzen dituena (Sanderson, 2010). Halere, ebaluazio gehienak sistemaren eraginkortasuna neurtzean zentratu dira. Eraginkortasuna neurtzeko oinarrizko modua erabiltzailearen informazio beharrarekiko IR sistemak itzulitako emaitzek duten esangura-maila zehaztean datza. Modu horretan, sistema bat eraginkorragoa da erabiltzailearen informazio beharrekiko emaitza esanguratsu gehiago itzultzen dituenean. IR ikergaiari buruzko argitalpen gehienetan eraginkortasuna neurtzeko oinarrizko modu hau dokumentu eta kontsulten bilduma batekin eta ebaluazio metrika batekin batera erabiltzen da. Hori dela eta, testerako bilduma eta kontsulta estandarrak finkatzea oso garrantzitsua da, ikerlarien komunitateak emaitza konparagarriak izan ditzan. Ildo horretan, horrelako bildumak sustatzen dituzten biltzarrak antolatzen hasi ziren. TREC[23], CLEF[24] eta NTCIR[25] dira helburu hori duten nazioarteko konferentzi nagusiak.

IR sistemak ebaluatzeko aipatutako metodologiak, Cranfield paradigma izenekoa (Cleverdon, 1962), honako abantailak dauzka; alde batetik dokumentu bildumen, kontsulta-zerrenden eta kontsultei dagozkien esangura-mailaren epaiketen eskuragarritasuna bermatzen da, eta bestetik IR sistema ezberdinen arteko konparagarritasuna eta esperimentuak errepikatzea ahalbidetzen du. Halere, benetako erabiltzaileen portaera mota honetako ebaluaketetan onartzen den portaeratik aldenduta dago. Erabiltzaileek IR sistemarekin elkarreragiteko joera dute sistemak itzulitako dokumentuak arakatuz edo kontsultak berriro bidaliz, beste modu batera formulatuta. Hortaz, IR sistemen ebaluazioan badago behar nabarmen bat erabiltzaile eta sistemaren arteko interazkioak kontuan hartzeko, 2.1.1 atalean aurkeztu bezala Informazioaren Bilaketan Portaera aipatu zenean. Ildo horretan zenbait TREC *track* proposatu ziren. TREC *Interactive Track*[26] atazaren helburua informazioaren berreskurapenean eragina duen interakzioa aztertzea da. Horretarako berreskurapen-prozesua bera eta emaitzak batera kontuan hartzen ditu. TREC *Session track*[27], ordea, kontsulten saio gaineko ebaluazioak lantzeko asmoarekin

23 http://trec.nist.gov/
24 http://www.clef-initiative.eu//
25 http://research.nii.ac.jp/ntcir/index-en.html
26 http://trec.nist.gov/data/interactive.html
27 http://ir.cis.udel.edu/sessions/

proposatu zen. Lan honetan, ebaluazio prozesua sinplifikatzeko eta emaitzak eta esperimentuak estandarizatze aldera Cranfield paradigman oinarritutako ebaluazioak burutu dira soilik. Hortaz, hurrengo lerroetan paradigma hau sakonago azalduko dugu.

Lehen esan bezala bilduma da Cranfield motako ebaluazio baten osagai nagusia. Ebaluazio bildumaren egitura klasikoak horrela dago antolatuta:

- Dokumentuen (*docid*) bilduma.
- Kontsulten (*qid*) zerrenda.
- Esangura-mailaren epaiketen zerrenda bat (*qrels* edo *query relevance set*), kontsulta bakoitzarekiko dokumentu bakoitzak duen esangura-maila adierazten duen *qid-docid* bikotez osatuta.

IR sistema baten ebaluazioak kontsulta zerrenda bat prozesatzea eskatzen du. Sistemak itzulitako *docid* zerrendari *run* izena ematen zaio. *Run* bateko edukia *qrels*-en arabera aztertzen da, sistemak bueltatutako dokumentuak esanguratsuak diren egiaztatzeko. Dokumentu esanguratsu kopurua eta dokumentuen posizioak kontuan hartuta sistemaren eraginkortasuna zehaztuko da. Eraginkortasun hori kuantifikatzeko ebaluazio-neurriak erabiltzen dira. Ebaluazio neurriak bilduma batekin batera erabiliz, eszenatoki zehatzetan sistemen bilaketa ahalmena simulatzen da. Modu horretan, ikerlariek sistema ezberdinen eraginkortasuna baldintza berdinetan aldera dezakete.

NIST erakundeak, TREC programaren bidez, lehenengo ebaluazio-bilduma publikoak argitaratu zituen. TREC kontsulten egitura kontsultaren atzean dagoen informazio-beharra adierazteko eredu formal bat ezartzeko asmoz diseinatu zen. TRECek proposatutako egitura honelakoa da:

- *Id* identifikadore bat kontsultarako
- Izenburu labur bat (*title*), erabiltzaileak botatzen dituen kontsulten modukoa
- Informazio-beharraren deskribapen labur bat (*desc*), gehienez esaldi bat
- Informazio-beharraren deskribapen sakonago bat (*narr*), zer nolako dokumentuak izango liratekeen esanguratsuak esanez

```
<top>
<num> Number: 200
<title> Topic: Impact of foreign textile imports on U.S. textile industry
<desc> Description: Document must report on how the importation of foreign textiles or textile products has
```

influenced or impacted on the U.S. Textile industry.
<narr> Narrative: The impact can be positive or negative or qualitative. It may include the expansion or shrinkage of markets or manufacturing volume or an influence on the methods or strategies of the U.S. textile industry. "Textile industry“ includes the production or purchase of raw materials; basic processing techniques such as dyeing, spinning, knitting, or weaving; the manufacture and marketing of finished goods; and also research in the textile field.
</top>

8. irudia. TREC egituraren adibidea.

TRECen arrakastaren ondoren antzeko ebaluazio-biltzarrak antolatzen hasi ziren munduko eremu ezberdinetan. CLEF (*Cross Language Evaluation Forum*) adibidez Europako hizkuntzak babesteko helburuarekin jaio zen. Lan honetan aurkezten diren esperimentuetan CLEFeko bildumak eta kontsultak erabili dira. Zehazki, *LA Times 94* eta *Glasgow Herald 95* bildumak. Bilduma hauek aipatutako egunkarietan argitaratutako artikuluak dituzte hurrenez hurren.

Bilduma	**# dok.**	**# token**
LA Times 94	110861	72M
Glasgow Herald 95	55892	27,7M

4. taula. Bildumen estatistikak

Sarreran esan bezala, esperimentuetan hiztegietan oinarritutako itzulpen metodoak aztertu ditugu. Bi baliabide lexikal erabili ziren hiztegi elebidun moduan: Alde batetik, Morris Euskara-Ingelesa hiztegia (77.864 ordain eta 28.874 sarrera) eta bestetik Euskalterm (72.184 ordain eta 56.745 sarrera). Demner-Fushman eta Oard-ek (2003) hiztegi elebidunaren estalduraren araberako berreskurapenaren eraginkortasuna aztertu zuten. Eraginkortasunaren hazkundea nabarmena zen 3.000 eta 20.000 sarreraren artean. Hortik aurrera hobekuntza oso txikia zen. Hortaz, esan dezakegu gure hiztegi elebidunaren estaldura egokia dela CLIR sistema batean integratzeko.

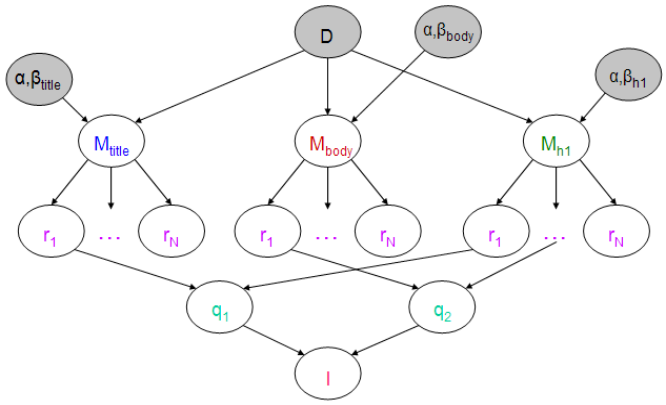
Lan honetan aurkeztutako esperimentu guztietan Indri *toolkit*-ak[28] eskaintzen duen *Indri* berreskurapen algoritmoa (Zhai and Lafferty, 2001) erabili zen. Algoritmoa kontsulta itzulitakoan aplikatzen da dokumentu esanguratsuen rankinga kalkulatzeko. *Indri* algoritmoa hizkuntza ereduen eta inferentzia-sareen berreskurapen-*frameworken* konbinazio bat da. Bi *framework* hauek, hizkuntza ereduak eta inferentzia sareak, oso erabiliak izan dira

28 http://www.lemurproject.org/indri/

berreskurapen ataza askotan. Hortaz, zentzuzkoa dirudi biak *framework* bakarrean konbinatzea. 9. irudian ereduaren ohiko instantzia bat erakusten da grafikoki. Grafoak bayersiar sare batek adierazten du, ausazko aldagaien bildumaren gaineko elkarrekiko probabilitatearen banaketa zehazten duena. Grafoaren nodo bakoitzak ausazko aldagai bat adierazten du. Grisez nabarmendutako nodoak ikusitako aldagaiak dira. Gainerakoak ezkutuko aldagaiak dira. Ertzek independentzia susmoak adierazten dituzte.

Sareak honako nodoak dauzka:

- Dokumentu-nodoa (*D*)
- Leuntze-parametroen (*Smoothing*) nodoak (*alpha*, *beta*)
- Eredu-nodoak (*M*)
- Kontzeptu-nodoak (*r*)
- *Belief* nodoak (*q*)
- Informazio-beharraren nodoa (*I*)



9. irudia. Indri ereduaren instantzia baten adibide grafikoa

Lehen esan bezala, bilduma eta kontsultez gain ebaluazio neurri bat beharrezkoa da IR sistema baten eraginkortasuna konputatzeko. Literaturan hainbat neurri proposatu dira helburu horrekin. MAP eta DCG dira IRko esperimentuetan erabilienak. MAP APen (*non-interpolated average precision*) dago oinarrituta:

$$AP = \frac{\sum_{rn=1}^{N} \left(P(rn) \times rel(rn) \right)}{R}$$

N sistemak itzulitako dokumentu kopurua da, *rn* ranking posizioa, *rel(rn)*, 1 edo 0 *rn.* posizioko dokumentuaren esangura-mailaren arabera; *P(rn) rn* posizioan neurtutako doitasuna da; eta *R* kontsultarekiko bilduman dauden dokumentu esanguratsuen kopurua. APk ranking posizio guztietako doitasuna kalkulatu eta batezbestekoa hartzen du. MAP kontsulta guztien APen batezbestekoa da.

MAP IRko ikerketa gehienetan neurrik erabiliena izan da duela gutxi arte. Azkeneko urteetan, ordea, ikerlari batzuk zalantzan jarri dituzte epaiketa boolearretan oinarritutako ebaluazioak eta neurriak. Testuinguru horretan Järvelin and Kekäläinenek (2000) *Cumulative Gain* (CG) eta horren aldaera batzuk proposatu zituzten. CG sistemak itzulitako lehenengo dokumentuei dagozkien esangura-mailen (*rel(i)*) batura da.

$$CG(n)=\sum_{i=1}^{N} rel(i)$$

CGk ez ditu kontuan hartzen dokumentuen posizioak. Hau problema bat da erabiltzailearen ikuspuntua ez duela ondo jasotzen. Sistema batek, dokumentu esanguratsuak lehenengo posizioetan itzultzen baditu, eraginkortasun hobea dauka. Hori dela eta *Discounted Cumulative Gain* (DCG) aldaera proposatzen da, non dokumentuen esangura-mailak zigortzen dira posizioak igo ahala:

$$DCG(n)=rel(1)+\sum_{i=2}^{N} \frac{rel(i)}{\log_b(i)}$$

Azkeneko urteetan DCG asko zabaltzen ari bada ere, lan honetan aurkezten diren esperimentuak landu zirenean neurria ez zegoen gaur bezala hain hedatuta. Hori dela eta, esperimentu guztiak ebaluatzeko MAP neurria erabili dugu.

Ebaluazio batek askotan sistema ezberdinen arteko konparaketa hartzen du. Sistema hauek *baseline* sistema baten aldaerak izan daitezke, edo haien artean zerikusirik ez duten sistemak ere izan daitezke. Konparaketa horren helburua zein sistemak ematen duen eraginkortasunik handiena ondorioztatzea da. Bi sistema ezberdinek itzulitako *run*ak alderatzeko estrategiarik erabiliena esangura-testak erabiltzea izaten da. IR esperimentuetan hipotesi nulua honako izaten da; bi sistemek ezaugarri berdinak dituztela, eta *run*etan egon daitezkeen diferentziak ausazkoak direla. Smucker eta besteek (2007) hainbat esangura-testa alderatu zuten bildumen gainean egindako esperimentuetan *randomization* testa onena zela ondorioztatuz. Hortaz, guk ere esangura-test hori erabili dugu proposatzen ditugun metodoak alderatzeko.

4 Gure hurbilpena

Lanaren muina atal honetan deskribatuko da. Kontsulten itzulpen-prozesuan gertatzen den anbiguotasun-problema ebazteko zenbait teknika azalduko dira. Euskara bezalako baliabide urriko hizkuntzetarako egokiak diren tekniketara mugatuko gara.

Lehenengo esperimentuan itzulpenen aukeraketarako metodo desberdinak ebaluatu ditugu. Bi teknika aztertu ditugu nagusiki, galdera egituratuak eta kookurrentzietan oinarritzen den beste teknika bat, izan ere, badirudi metodo egokienak horiexek direla corpus paraleloak eskuragarri ez daudenean. Bi metodoekin lortutako emaitzak konparatu dira, metodo bakoitzaren ahuleziak zehaztu dira eta azkenik, bi teknikak bateratzen dituen metodo hibrido bat proposatu da. Emaitzen arabera bi metodoen artean ez dago diferentzia handirik. Bestetik hibridazioak ez dakar hobekuntza esanguratsurik.

Bigarren esperimentuan hiztegi batetik lor daitezkeen itzulpenen probabilitateak corpus paraleloak erabili gabe estimatzeko metodo bat landu da. Hiztegi batean itzulpenen ordena ez denez erabilera maiztasunarekin beti bat etortzen, metodo bat diseinatu dugu hiztegia berrordenatzeko, internet corpus konparagarri bat bailitz bezala erabiliz. Metodo honekin, jatorrizko hitzaren testuinguruaren eta itzulpenen testuinguruaren arteko antzekotasunaren (*distributional similarity*) arabera berrordenatzen ditugu hiztegiko itzulpen-hautagaiak jatorrizko hitz bakoitzeko. Bai jatorrizko hitzaren testuingurua eta baita itzulpen-hautagaien testuinguruak web bilatzaileak eta *wac* tresnak erabiliz lortzen ditugu, WAC[29], Google newsArchive[30] edota GoogleBlog[31], besteak beste. Lortutako testuinguruak itzuli eta testuinguruak kosinua erabiliz konparatu ondoren, hiztegia berrordenatzen dugu lortutako antzekotasun balioaren arabera. Emaitzen arabera modu horretan berrordenatutako hiztegiek eragiten positiboa dute hizkuntza arteko berreskurapen prozesuaren eraginkortasunean.

4.1 Our approaches

we will now present our two experiments in turn. In Section 4.2 we will present the results of those experiments.

4.1.1 Comparing different approaches to treat Translation Ambiguity

29 http://webascorpus.org/searchwac.html
30 https://news.google.com/
31 http://www.google.com/blogsearch

Experiments introduced in this sections compare two alternatives proposed in the literature which do not require parallel corpora. The only resources used are a bilingual MRD and a corpus in the target language for the co-occurrence based method, which makes them suitable for less resourced languages like Basque. We have chosen a specific method for each approach: Pirkola's method (Pirkola, 1998), and a co-occurrence based method. Among all the co-occurrence based algorithms we have chosen the Monz and Dorr's algorithm (Monz and Dorr, 2005) assuming that being iterative yields better estimations, although we do not have any references that confirm this. In addition, we have designed an algorithm that combines both approaches. In this last case, we have used Darwish and Oard's probabilistic structured queries (Darwish and Oard, 2003) as a framework and Monz and Dorr's algorithm to estimate the weights of the translation candidates.

A. Dealing with ambiguous translations using Structured Queries

The #syn operator of structured queries is a suitable technique for dealing with ambiguous translations because among other things it is fast, offers good results and does not need external resources such as parallel corpora. The basic idea is to group together the translation candidates of a source word, thus making a set and treating them as if they were a single word in the target collection (Pirkola, 1998). Hence, when estimating the term frequency (*TF*) and document frequency (*DF*) statistics for query terms, the occurrences of all the words in the set are counted as occurrences of the same word. If we assume that s_i is a query term in a source language, D_k is a document term in a target language, d is a document and $T(s_i)$ is the set of translation candidate terms of s_i given by the MRD, *TF* and *DS* for source words are computed as follows:

$$TF_j(s_i) = \sum_{\left(k \mid D_k \in T(s_i)\right)} TF_j(D_k)$$

$$DF(s_i) = \left| U_{\left(k \mid D_k \in T(s_i)\right)} \left\{ d \mid D_k \in d \right\} \right|$$

where $TF_j(s_i)$ is the term frequency of s_i in document j, and $DF(s_i)$ is the number of documents that contain s_i.

If the translation candidates are correct or semantically related, the effect is an expansion of the query. The problem arises especially when wrong translations that are common

words occur, because *DF* of the #syn set can take high scores and the correct translation loses weight in the retrieval process. *TF* statistics can also be altered when wrong translations appear in the retrieved documents. But the probability that many wrong translations occur in retrieved documents is low. That is what we call retrieval time translation selection.

In order to test this method in development experiments, we have prepared a list of Basque topics translated from the English ones belonging to the CLEF 2001 edition (41-90), and the LA Times 94 collection and the corresponding relevance judgements, which are explained more fully in section 3. First, we have calculated the MAP for different numbers of translation candidates from the MRD (Figure 10), because a high coverage of translations and the precision level of the MRD affects the performance of this method (Larkey et al., 2002). Moreover, the translation equivalents of source words are usually ordered by frequency use in a MRD. Therefore, we can exploit that order to prune the least probable translations in the interests of query translation precision.

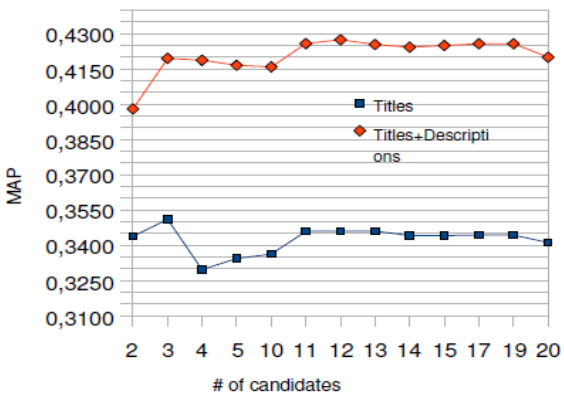


Figure 10. MAP values for different numbers of translation candidates

In the graph (Figure 10), we can see how the number of translation candidates from the MRD accepted for each source word affects the MAP. MAP curves are similar for both titles and titles+descriptions queries. They have local maximum in near points but the maximum global is reached by taking more candidates with the title+description set. The maximum MAP is achieved by taking the first three candidates for short queries, and the twelve first candidates for the long queries. This seems logical because there are more context words that can improve the retrieval time disambiguation.

B. Target co-occurrence based selection

As explained above, structured queries do not really do translation selection, and translations and statistics (*TF* and *DF*) can be wrong in some cases and decrease the retrieval performance. An alternative to executing the translation selection without using parallel corpora is to guide the selection by using statistics of the co-occurrence of the translation candidates in the target collection. The basic idea is to choose the ones that co-occur more frequently, assuming that the correct translation equivalents of query terms are more likely to appear together in target document collection than incorrect translation equivalents. The main problem of this idea is to compute that global correlation in an efficient way, because the maximization problem is *NP-hard*.

The algorithm we have used for the translation selection is the one introduced by (Monz and Dorr, 2005). Basically, it selects the translation candidates combination which maximizes the global coherence of the translated query by means of an *EM* (Expectation Maximization) type algorithm.

Initially, all the translation candidates are equally likely. Assuming that *t* is a translation candidate for a query term s_i given by the MRD, then:

Initialization step:

$$w_T^0(t|s_i) = \frac{1}{|tr(s_i)|}$$

In the iteration step, each translation candidate is iteratively updated using the weights of the rest of the candidates and the weight of the link connecting them.

Iteration step:

$$w_T^n(t|s_i) = w_T^{n-1}(t|s_i) + \sum_{t' \in inlink(t)} w_L(t,t') \cdot w_T(t'|s_i)$$

where *inlink(t)* is the set of translation candidates that are linked to *t*.

After recomputing each term weight they are normalized.

Normalization step:

$$w_L^n(t|s_i) = \frac{w_L^n(t|s_i)}{\sum_{m=1}^{|tr(s_i)|} w_L^n(t_{i,m}|s_i)}$$

The iteration stops when the variations of the term weights become smaller than a predefined threshold. There are different association measures to compute the association strength between two terms $w_L(t,t')$. We experimented with Mutual Information and Log Likelihood Ratio, and obtained the best results with the second one. That is the measure we use in the evaluation.

The question is whether by choosing the best translation of each query term we obtain a better MAP than grouping all the translation candidates by means of structured queries. As mentioned before, although in the structured queries some weights and translations can be wrong, an expansion that can benefit the MAP is also produced. For example, for the Basque query *"gene gaitz"*, when we select the best English translation *"gene disease"* and run it, we obtain an AP of 0.5046. However, when all the translation candidates given by the MRD are put in sets with the #syn operator, "*gene #syn(harm disease flaw ailment hurt malady defect difficult)*" , even if we incorporate incorrect translations, we get a greater AP value, 0.5548. So, in this example it is clear that the noise expanded translation gives a higher AP score than the best translation. Nevertheless, for the Basque query *"gose greba"* we construct a translated query like *#syn(hunger yearning desire famine urge ravenous craving famished hungry) #syn(#1(work stoppage) strike walkout)* obtaining an AP of 0.0741. Whereas if we choose the best translation manually, we get the query *"hunger strike"* and obtain an AP of 0.6743. Looking at this example, it seems that our co-occurrence method could provide a margin for improving the MAP compared with structured queries when query terms have many incorrect translation candidates. In order to estimate whether this case is general, a lexicographer manually disambiguated some Basque queries (built from 41-90 CLEF queries) translated into English by an MRD. We preprocessed the queries by keeping only the lemmas of content words and then translated them using the MRD. The work by the lexicographer was to select the best translation candidate for each source term of the queries (Example on Table 5.).

Translation method	**Query**
English query	Tainted-Blood Trial
Basque query	kutsatuko odolaren epaia
English query (content words)	Tainted-Blood Trial
Basque query (content words)	kutsatu odol epai
Structured translation into English	#syn(pollute impregnate infect) #syn(blood kinship) #syn(sentence crest judgment ridge notch scratch mark cut incision)

Best manual translation	infect blood sentence
Best manual translations	#syn(pollute ~~impregnate~~ infect) #syn(blood ~~kinship~~) #syn(sentence ~~crest judgment ridge notch scratch mark cut incision~~)

Table 5. Selecting the best translation of the structured query.

Then, we calculated the MAP by processing Basque queries (Table 6.) (titles and titles+description separately) for the different translation methods including the manual based one. The MAP results show the MAP obtained by manual disambiguation does not reach that obtained using structured queries. So it seems that there is no margin for improvement for the co-occurrences based method. However, the co-occurrences based method outperforms structured queries when we are dealing with short queries. It even outperforms the theoretical threshold marked by the manual disambiguation. It could be due to a better statistical selection of short queries, more adequate for relevances in that collection.

Translation method	**Titles**	**Titles+description**
English monolingual	0.4639	0.4912
Structured query (3 and 13 candidates[32])	0.3510	0.4274
Structured query (all candidates)	0.3352	0.4200
Best manual translation	0.3218	0.4127
Co-ocurrence-based	0.3564	0.3908
Best manual translations	0.3471	0.4308
Probabilistic structured query	0.3568	0.4268
Probabilistic structured query+threshold (0.8)	0.3594	0.4249

Table 6. MAP results for 41-90 topics

C. Combining structured queries and cooccurrence based algorithm

We think that we could take advantage of both techniques. Structured queries contribute to the translation less restrictiveness and query expansion in the retrieval phase, and the co-occurrence based method contributes translation selection and weighting capability. To do this, we propose that probabilistic structures queries (Darwish and Oard, 2003) be used, and the weights be estimated according to Monz and Dorr's algorithm. Thus, assuming $w_L(D_k|s_i)$ as the weight for the translation candidate D_k of a term s_i of a source query s we estimate *TF* and *DF* in this way:

$$TF_j(s_i) = \sum_{(k|D_k \in T(s_i))} TF_j(D_k) \cdot w_L(D_k|s_i)$$

32 See figure 10

$$DF(s_i) = \sum_{(k|D_k \in T(s_i))} DF_j(D_k) \cdot w_L(D_k|s_i)$$

As we did in subsection B, in order to estimate the possible improvement margin of this method, a lexicographer manually removed the wrong translations of the development queries, while maintaining only the correct ones (See Table 5.). We maintained all the possible candidates since this method is capable of selecting more than one candidate. Thus, for the Basque query "*gene gaitz*" ("*gene disease*" on English) we obtained a query "*gene #syn(disease ailment malady)*" achieving an AP of 0.5946. A higher score than the one achieved taking all candidates. However, contrary to what we expected, the MAP for 41-90 topics is not much higher than that achieved without doing any kind of selection (although pruning some translations of the MRD can be considered to be a general disambiguation method) for long queries, and for short queries it is even worse (Table 6). Therefore, better quality in the translations does not seem to imply a big improvement inn MAP. A further analysis will be conducted in the Section 4.2.1.

4.1.2 Mining Translation probabilities by using the web as a comparable corpus

Several techniques have been proposed for dealing with translation ambiguity for the query translation task on CLIR, such as structured query-based translation (also known as Pirkola's method) (Pirkola, 1998), word co-occurrence statistics (Ballesteros and Croft, 1998) and statistical translation models (Hiemstra and De Jong, 2001). Structured queries are adequate for less resourced languages, rare pairs of languages or certain domains where parallel corpora are scarce or even non-existent. The idea behind this method is to treat all the translation candidates of a source word as a single word (*syn* operator) when calculating *TF* and *DF* statistics. This produces an implicit translation selection during retrieval time. There are many works dealing with structured queries, and some variants are proposed. Darwish and Oard (2003) for example, proposes that weights or replacement probabilities be included in the translation candidates (*wsyn* operator). One drawback with this approach is that it needs parallel corpora in order to estimate the replacement probabilities.

Following this line of work, we propose a simple method based on the implicit translation probabilities of a dictionary, and also a more robust one which uses translation knowledge mined from the web. We have analyzed different ways of accessing web data: **Web As Corpus** tools, **News** search engines, and **Blog** search engines. Our aim is to examine how the characteristics of each access strategy influence the representation of the constructed

contexts, and also, how far these strategies are adequate for estimating translation probabilities by means of the cross-lingual context similarity paradigm. All experiments have been carried out taking Spanish as source language and English as target because no News search engines were available for Basque. In any case results obtained with the rest of the search-engines could be extrapolable to Basque.

4.1.2.1 Obtaining Translation Probabilities from a Dictionary

The first method proposed for estimating translation probabilities relies on the hypothesis that, in a bilingual MRD *D*, the position *pos* of the translation e_i among all the corresponding translation candidates $\{e_i\}$ for a source word *v* is inversely proportional to its translation probability $p(e_i|v)$. If we assume that it is an exponential decay relation, we can model the translation probability through this formula:

$$p(e_i|v) = \frac{1}{\left(\sum_{(v,e_i)\in D} \frac{1}{pos(D,v,e_i)}\right) \cdot pos(D,v,e_i)}$$

The principal problems of these assumptions are, firstly, that translations are not ordered in all MRD (partially or at all) by frequency of use, and secondly, that the proposed relation above does not fit all translation equivalents. So, we propose a method that is useful for ordering the translations of an MRD as well as for estimating more accurate translation probabilities, as presented in the following section.

4.1.2.2 Translation Probabilities by Context Similarity

The idea is to obtain translation probabilities by using the web as a bilingual comparable corpus. This strategy is based on estimating the translation probability of the translation candidates taken from the MRD in accordance with the context similarity of the translation pairs [5]. The hypothesis is that the more similar the contexts are, the more probable the translation will be. The computation of the context similarity requires a large amount of data (contexts of words), which has to be representative and from comparable sources. The Internet is a good source of large amounts of texts, and that is why, we propose that different search-engines be analyzed to obtain these contexts. These search engines have different features, such as domain, coverage and ranking, which affect both the degree of comparability and the representativeness of the contexts, as follows:

- **WebCorp**: This Web Concordancer is based on main search APIs. Therefore, navigational queries and popular ones are promoted. These criteria can reduce the representativeness of the contexts retrieved. Since we take a maximum number of snippets for each query, the selected contexts depend on the ranking algorithm. It guarantees good recall, but perhaps poor precision. Thus, the comparability degree between contexts in different languages can be affected negatively.
- **Google News Archive**: The content is only journalistic. It seems appropriate if we want to deal with journalism documents but not with other registers or more specialized domains. In short, it offers good precision, enough recall and a good degree of comparability.
- **Google Blog search**: The language used is more popular, and although the register is similar to that of journalism, the domain is more extensive. This could offer good recall but not very comparable contexts.

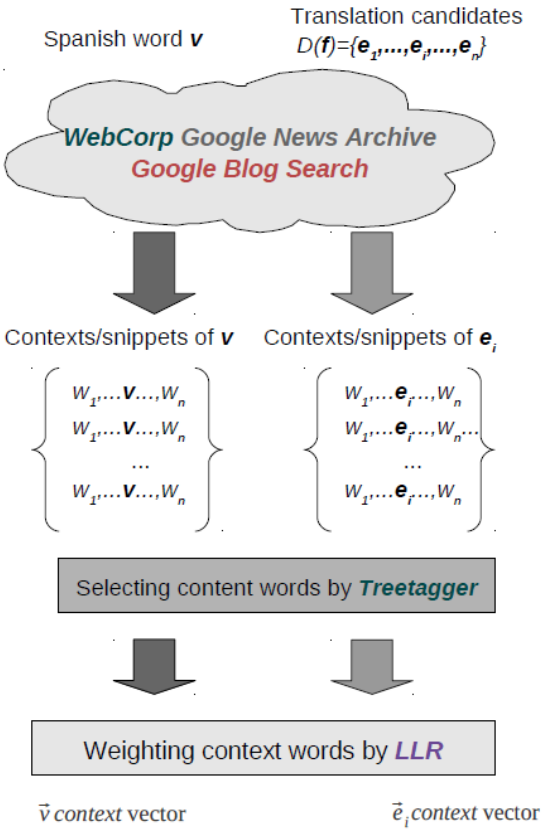


Figure 11. Extraction of context representations of bilingual equivalent words from the web

The method to estimate the translation probabilities between a source word *v* and its translations $\{e_i\}$ included in *D* starts by downloading, separately, the snippets of both words as returned by the search engines mentioned above (See figure 11.). Then, following the "bag-of-words" paradigm, the contexts of *v* and e_i are represented by weighted collections of words. We set up context vectors for the source $\vec{v}$ and the translation word $\vec{e}_i$ by taking keywordness (using loglikelihood ratio) of the content words (nouns, adjectives, verbs and adverbs selected by using *Treetagger*) belonging to all their snippets. Thus, the content words are weighted with regard to an open-domain corpus according to the log-likelihood ratio, and the context vectors of *v* and e_i are formed. Log-likelihood ratio provides the association degree of the context word with regard to the snippet by taking the open domain corpus as reference. The next step is to translate the Spanish context vector $\vec{v}$ into English $\overrightarrow{tr(v)}$. This is done by taking the first translation from a Spanish-English MRD *D* 34,167 entries. Once both vectors, $\vec{v}$ and $\{\vec{e}_i\}$) are put into the same space (same language) cross-lingual context similarity is calculated according to cosine measure which is transformed into translations probabilities:

$$p(e_i|v) = \frac{\cos\left(\overrightarrow{tr(v)}, \vec{e}_i\right)}{\sum_{(v,e_i)\in D} \cos\left(\overrightarrow{tr(v)}, \vec{e}_i\right)}$$

We analyze the differences between the translation rankings obtained with the different search engines and those in the original dictionary. We computed Pearson's correlation for the translation rankings obtained for the polysemous content words in all 300-350 Spanish CLEF topics. The correlation scores (cf. Table 7) show that the different characteristics of each search engine produce translation rankings which are quite different from those in the dictionary (**Dic.**) and also from each other.

	WebCorp	**News**	**Blog**
Dic.	0.42	0.31	0.40
WebCorp		0.44	0.54
News			0.49

Table 7. Mean of Pearson's correlation coefficients for translation rankings compared to each other

4.2 Results and discussion

4.2.1 Comparing different approaches to treat Translation Ambiguity

We evaluated the proposed translation methods using the collection from CLEF 2001 composed by LA Times 94 and Glasgow Herald 95. We translated from English to Basque two sets of topics: one for development (41-90) and the other one for test purposes (250-350). MAP values are calculated automatically with respect to existing human relevance judgments for queries and documents of the collections as explained in 3th section. The translation of the topics was carried out by professional translators and correctors of the Elhuyar foundation. The process was done in two steps: firstly, a translator translated the English topics into Basque, and then a corrector corrected the translations in order to minimize the possible bias and the possible lack of naturalness caused by the translation process.

We used the Indri as ranking model and the Porter Stemmer both for collections and translated topics. Before applying the proposed translation methods we removed words like *"documentuak...(documents)"* and selected the content words manually. Specifically, nouns, adjectives, verbs and adverbs. Postpositions like *"artean (between), buruz (about).. ."* were also removed. We used a Basque-English MRD which includes 34,167 entries. For the treatment of OOV (Out-Of-Vocabulary) words we looked for their cognates in the target collection. Transliteration rules (see table 8.) were applied and then LCSR (Longest Common Sequence Ratio) was computed. Those which reached a threshold (0.8) were taken as translation candidates in the translation phase.

Source word	**rule**	**Target word**
phase	ph → f	fase
action	tion → zio	akzio

Table 8. Example of transliteration rules

The runs were done by taking the titles as queries (short queries), and also by taking the titles and descriptions as queries (long queries) and carrying out Basque to English translation:

1. Monolingual: Titles and titles+descriptions of CLEF 250-350 English topics.
2. First translation: First translation from dictionary
3. Structured query: Group translation candidates from the dictionary in a #syn set using Pirkola's method.

4. Structured query (Optimized dictionary): first translation candidates of the dictionary grouped in a #syn set (three for titles and twelve for the titles+descriptions maximize MAP on development experiments) using Pirkola's method.
5. Co-occurrence-based translation: Best translation selected by Monz and Dorr's co-occurrence based algorithm.
6. Probabilistic structured query: all translation candidates of the dictionary grouped in a #wsyn set using Darwish and Oard's method, and weighted according Monz and Dorr's co-occurrence based algorithm.
7. Probabilistic structured query+threshold: Best translations selected according to a threshold and weighted by Monz and Dorr's co-occurrencebased algorithm and grouped by #wsyn set using Darwish and Oard's method.

The results are presented in Table 9 and Figures 12 and 13.

Run	**MAP**		**% of Mon.**		**Improvement Over First %**	
	Short	Long	Short	Long	Short	Long
English monolingual	0.3176	0.3778				
First	0.2118	0.2500	67	66		
Structured query	0.2342	0.2959	74	78	9.56*	15.51*
Structured query (optimized dictionary)	0.2359	0.2960	74	78	10.22*	15.54*
Co-occurrences based	0.2338	0.2725	74	72	9.41*	8.26*
Probabilistic structured queries+threshold	0.2404	0.2920	76	77	11.9*	14.38*
Probabilistic structured queries	0.2371	0.2941	75	78	10.67	14.99*

Table 9. MAP values for 250-350 topics

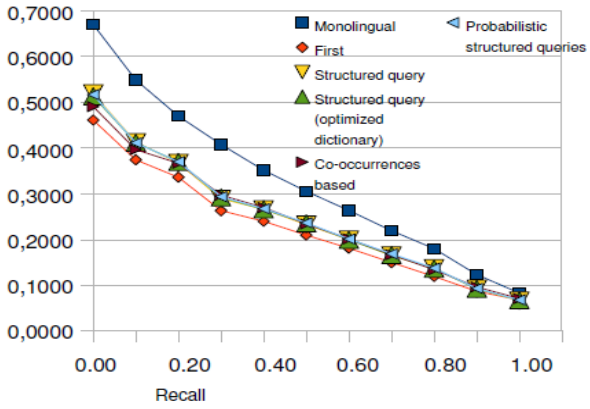


Figure 12. P-R curves (Titles)

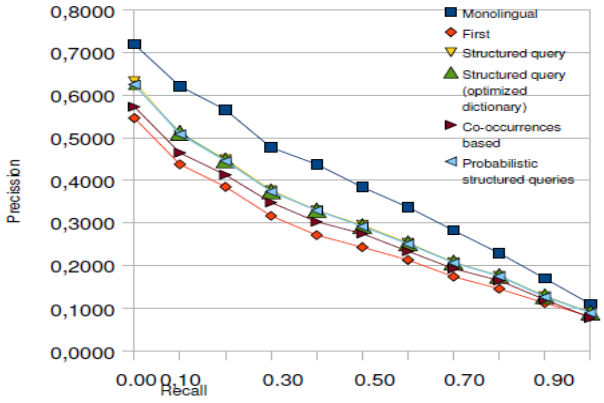


Figure 13. P-R curves (Titles + Description)

The achieved MAP is higher with long queries than with short queries in both cases, monolingual and cross-lingual. In the cross-lingual retrieval the translation methods proposed also offer greater improvement with long queries. This is logical because more context words help in the translation selection. Unlike the results in the development experiments, the methods do not show a different performance depending on the length of the queries. We have examined the queries translated by Monz and Dorr's method and the quality is quite adequate except for a few cases due to false associations. For example, the Basque query *"kutsatu odol **epai**"* is translated as *"infect blood **cut**"* by Monz and Dorr's method instead of *"infect blood **sentence** "*. We can assume that it happens due to the stronger relation between *epai* source word's translation candidate and *infect* and *blood* and *cut -epai* source word's translation candidate- than between *infect* and *blood* and sentence another translation candidate for *epai*. It seems to be because of the the limited representativity of the target collection where some words rarely co-occur. So this could be mitigated by using a bigger corpus. For short queries, too, the hybrid

method shows the best results, but statistically does not outperform Pirkolas's method significantly. Pirkolas's method achieves the best results when dealing with long queries. The optimized MRD improves the MAP but not significantly. All improvements that are statistically significant according to the Paired Randomization Test with alfa=0.05 are marked with an asterisk in table 9.

It seems that selecting and weighting translation candidates by means of Monz and Dorr's method in order to include them in structured queries do not imply a significant improvement in MAP terms with respect to Pirkola's method.

As in the earlier case, the queries translated by the hybrid method are adequate except for a few cases of false associations. In any case, as we have seen in subsection C, improving the quality of the translation does not always improve the MAP.

Translation phase	**query**	**AP**
English query (46)	**Embargo on Iraq**	
Basque query (46)	*Irakeko bahitura*	
Basque (content words)	Irak bahitura	
Structured translation	Iraq #syn(seizure mortgage kidnapping confiscation)	0.2989
Best translations	Iraq seizure	0.1302
English query (81)	**The reserve in the Antarctic in which hunting for whales is forbidden**	
Basque query (81)	*Baleak ehizatzea debekatuta dagoen Antarktikako erreserba*	
Basque (content words)	balea erreserba antarktika ehiza debekatu	
Structured translation	whale #syn(reservation reserve) Antarctica #syn(game hunting prey) prohibit	1
Best translations	whale #syn(~~reservation~~ reserve) Antarctica #syn(~~game~~ hunting ~~prey~~) prohibit	0.3333

Table 10. Selecting the best candidates from the structured query (Topics 46 and 81).

In our opinion, apart from the query expansion effect and retrieval time selection, another positive effect produced with structured queries is that the weight of some non relevant

terms are smoothed. It is a collateral effect that happens because non relevant words tend to be common words which inflate the *DF* statistic. We have examined the differences between AP values corresponding to 41-90 queries (when titles and descriptions are taken) translated by taking all translations of the MRD and by pruning the wrong ones manually. In theory, all the AP values corresponding to each query will be better with the pruned ones. However, there are 6 queries where AP is significantly higher when all translation candidates are taken, despite many of them being wrong (Fig 14).

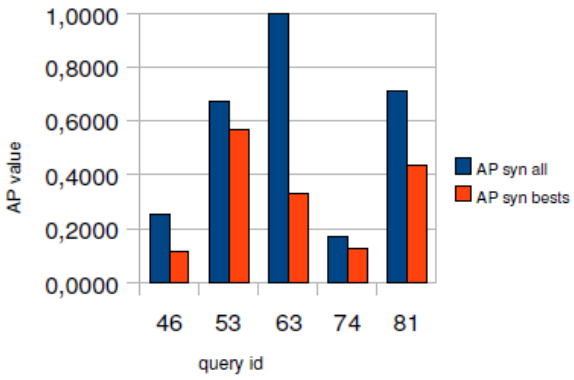


Figure 14. AP values for queries with significantly improved AP when taking all translations candidates

If we analyze these queries more deeply, we can detect two factors that explain this effect:

1. Wrong translations can turn out to be relevant terms: In the example (46) of Table 10. among all the translation candidates of the Basque source word *bahitura* only *kidnapping* appears in the relevant documents of the collection for that query.
2. Wrong translations can reduce non relevant or noise producer source term weight: in the example (81) of table 10. No of translations of *erreserba* and *ehiza* appear in the relevant documents. Thus, taking all candidates decreases the weight of these irrelevant sets, leading to a better AP score.

4.2.2 Mining Translation probabilities by using the web as a comparable corpus

We evaluated 50 queries (title+description) taken from 300-350 CLEF topics against collections from CLEF 2001 composed by LA Times 94 and Glasgow Herald 95 news. Previously, nouns,

adjectives, verbs and adverbs were selected manually both in Spanish and English topics. Indri was used as the retrieval model and the queries were translated using several methods: taking the first translation of the MRD (**First**); taking all the translations and grouping them by the *syn* operator (**All** or **Pirkola**); and weighting the translations by using the *wsyn* operator and the methods described in sections 2 (**Dic.**) and 3 (**Webcorp**, **News** and **Blog**). The results are shown in Table 11.

In the first column we show the MAP results obtained with each method, with the English monolingual results first. In the second column we show the percentage of the cross lingual MAP with respect to the monolingual result. We can see that using all translations with their replacement probability estimated according to the dictionary order produces better results than using only the first translation or using all translations, with a significant improvement (according to the Paired Randomization Test with α=0.05) over the **All** method. So, exploiting the translation knowledge latent in the position of the translations improves the MAP when provided by the dictionary.

Method	**MAP**	**% Monolingual**	**% Improv. over All**
Monolingual (en)	0.3651		
First	0.2462	67.43	
All	0.2892	79.21	
Dic.	0.2951	80.83	2.04
WebCorp	0.2943	80.55	1.76
News	0.2993	82.63	3.49
Blog	0.2960	81.07	2.35

Table 11. MAP for 300-350 topics

Otherwise, the web-based estimation techniques also improve significantly over the **First** and **All** strategies (α =0.05). However, there is no significant improvement over the **Dic.** method. It seems that context similarity calculated from **Blog** or **News** sources is more suited to estimating translation probabilities since they significantly outperform **WebCorp** in terms of MAP. Therefore, comparability between sources of both languages, domain precision and informational snippets seem to be important factors in order to obtain useful context for context-similarity, although deeper analyses must be carried out to determine the importance of each more precisely.

5 Ondorioak eta etorkizuneko lanak

Literaturan CLIRen inguruko lan esperimental dezente daude baina ikerketa-ildo hori justifikatzen duen azterketa soziolinguistikorik ez da existitzen. Lan honetan azterketa soziolinguistiko hori egiteko lehen gakoak aurkeztu ditugu. Aurreazterketa horren ondorioa da CLIR sistemak orokorrean baliagarriak direla, egoera eleaniztunak gaur egungo gizarteetan oraindik oso hedatuta baitaude. Hiztun gutxiko hizkuntzen kasuan CLIR sistemak oso beharrezkoak dira hiztunak kasu askotan informazio esanguratsua bere hizkuntzan topatzeko mugatuta egoten direlako.

Norabide horretan lan honetan euskarazko kontsultetatik abiatuta ingelesezko dokumentuak bilduma batetik berreskuratzeko teknikak landu dira. Zehazki kontsultak itzultzeko teknikak aztertu dira. Euskararen ezaugarriak kontuan hartuta baliabide gutxi eskatzen dituzten tekniketan zentratu gara; galdera egituratuak eta kookurrentzietan oinarritutako itzulpen teknikak hain zuzen ere. Bi teknika hauek itzulpen-ezagutza hiztegi batetik hartzen dute. Esperimentuetan lortutako emaitzen arabera teknika hauek egokiak dira euskara-ingelesa bikoterako. Horrek esan nahi du hiztegietan oinarritutako estrategia corpus paraleloetan oinarritutako estrategiaren aldean aukera ona izan daitekeela baliabide urriko hizkuntzez ari garenean.

Kookurrentzietan oinarritutako teknikak *baseline*arekin lortutako emaitzak gainditzen baditu ere galdera egituratuak eraginkorragoak agertzen dira. Izan ere, MAP terminoetan galdera egituratuek emandako eraginkortasuna berreskurapen elebakarretik gertu dago. Berreskurapen elebakarraren eraginkortasunaren %75a eta %78a lortzen dira kontsulta laburrak eta luzeak lantzen direnenak. Eraginkortasun balio hauek handiagoak izan litezke baldin eta erabilitako hiztegien estaldura handiagoa izango balitz edo hiztegian itzulpen probabilitateak sartuko bagenitu.

Itzulpen-probabilitateak eskuratzeko estrategia guztiak corpus paraleloez baliatzen dira. Guk corpus konparagarrietan oinarritutako estrategia bat proposatzen dugu. Era horretan lortutako itzulpen-probabilitateak baliagarriak dira hizkuntza arteko berreskurapen-prozesuaren eraginkortasuna hobetzeko. Hortaz, ondorioztatu daiteke etorkizun handiko metodoa dela, ez bakarrik CLIR-prozesuen eraginkortasuna hobetzeko, MT-sistemek ere etekina atera ahal diote metodo honi hizkuntza bikote edo domeinu baterako corpus paraleloak urriak direnean.

Bestalde, badago oraindik testuinguruko informazioa kontsulten itzulpen-prozesua hobetzeko baliagarria izan daitekeena. 2.1. atalean azaldu dugun bezala, benetako erabiltzaileen

portaera konplexua da. Erabiltzaileek IR sistemekin elkarreragiteko joera dute, galdera batentzat itzulitako dokumentuak arakatuz, hasierako galdera birformulatuz, hau da, hasierako informazio beharra beste hitz batzuekin adieraziz. Hasierako kontsultak eta birformulazioek saio bat osatzen dute. Saioko informazioa erabili nahi dugu kontsulten itzulpen kalitatea hobetzeko, eta bide batez, hobekuntza honek hizkuntza arteko informazioaren berreskurapen-prozesuaren eraginkortasunean ere eragina duela frogatu nahi dugu. Gure hipotesi nagusia honakoa da; saio berdineko kontsultek itzulpen hautapena modu egokian burutzen laguntzeko kalitatezko testuinguru erabilgarri bat osatzen dutela.

6 Bibliografia

Ricardo A. Baeza-Yates, Berthier A. Ribeiro-Neto. 1999. Modern Information Retrieval. ACM Press / Addison-Wesley.

Lisa Ballesteros and W. Bruce Croft. 1998. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '98). ACM, New York, NY, USA, 64-71.

Nicholas J. Belkin and W. Bruce Croft. 1992. Information filtering and information retrieval: two sides of the same coin?. *Commun. ACM* 35, 12 (December 1992), 29-38.

Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum* 36, 2 (September 2002), 3-10.

Aitao Chen and Frederich C. Gey. 2004. Combining query translation and document translation in cross-language retrieval. In *Proceedings CLEF-2003* , pp. 39.48. Trondheim.

C. W. Cleverdon, Report on the Testing and Analysis of an Investigation Into the Comparative Efficiency of Indexing Systems. ASLIB Cranfield Research Project. Cranfield, UK, 1962.

Kevin Craine. 2001. The Growth of Digital Information. in TDAN.com.

David Crystal. 1997. English as a Global Language. *Cambridge University Press*.

Kareem Darwish and Douglas W. Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (SIGIR '03). ACM, New York, NY, USA, 338-344.

Allen Foster. 2004. A nonlinear model of information-seeking behavior. *J. Am. Soc. Inf. Sci. Technol.* 55, 3 (February 2004), 228-237.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1* (COLING '98), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 414-420.

N. Gandal. 2006. Native Language and Internet Use. International Journal of the Sociology of Language, 182, 25 – 40.

Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, and Changning Huang. 2001. Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '01). ACM, New York, NY, USA, 96-104.

Jianfeng Gao, Ming Zhou, Jian-Yun Nie, Hongzhao He, and Weijun Chen. 2002. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '02). ACM, New York, NY, USA, 183-190.

Julio Gonzalo. 2002. Scenarios for interactive cross-language information retrieval systems. In: Proceedings of SIGIR 2002 Workshop on Cross-Language IR.

Djoerd Hiemstra. 2001. Using Language Models for Information Retrieval. Ph.D. Thesis, *Centre for Telematics and Information Technology*, University of Twente, January 2001.

Djoerd Hiemstra and Franciska De Jong. Statistical Language Models and Information Retrieval: natural language processing really meets retrieval. 2001. University of Twente, 2001.

David A. Hull and Gregory Grefenstette. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '96). ACM, New York, NY, USA, 49-57.

Jacob Jacoby, Donald E. Speller and Carol Kohn-Berning. 1974. Brand Choice Behavior As A Function of Information Load: Replication and Extension. Journal of Consumer Research, (June, 1974), 33-42.

Myung-Gil Jang, Sung Hyon Myaeng, and Se Young Park. 1999. Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (ACL '99). Association for Computational Linguistics, Stroudsburg, PA, USA, 223-229.

Kyunghye Kim, Mia Liza A. Lustria, Darrell Burke And Nahyun Kwon. 2007. Predictors of cancer information overload: findings from a national survey. *Information Research, 12*, paper 326.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Comput. Linguist.* 24, 4 (December 1998), 599-612.

Carol Collier Kuhlthau. 2006. Kuhlthau's Information Search Process, in Karen E. Fisher, Sandra Erdelez, and Lynne McKechnie (Eds.), Theories of Information Behavior (pp. 230–234), New Jersey: Information Today.

Leah S. Larkey, James Allan, Margaret E. Connell, Alvaro Bolivar and Courtney Wade. 2002. UMass at TREC 2002: Cross Language and Novelty Tracks. In Proceedings of TREC.

Yi Liu, Rong Jin, and Joyce Y. Chai. 2005. A maximum coherence model for dictionary-based cross-language information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '05). ACM, New York, NY, USA, 536-543.

J. Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval?. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (ACL '99). Association for Computational Linguistics, Stroudsburg, PA, USA, 208-214.

Christof Monz and Bonnie J. Dorr. 2005. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '05). ACM, New York, NY, USA, 520-527.

Jakob Nielsen. 1995. Multimedia and Hypertext: The Internet and Beyond, Morgan Kaufmann Publishers, 1995.

Douglas W. Oard. 1998. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup* (AMTA '98), David Farwell, Laurie Gerber, and Eduard H. Hovy (Eds.). Springer-Verlag, London, UK, 472-483.

Ari Pirkola. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM SIGIR*

conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 55-63.

Yan Qu, Alla N. Eilerman, Hongming Jin and David A. Evans. 2000. The effects of pseudo-relevance feedback on Mt-based. In *Proceedings of the Recherche d'Informations Assistee par Ordinateur (RIAO 2000)*, Paris, April 2000.

Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends® in Information Retrieval: Vol. 4: No 4, pp 247-375.

Vundavalli Srinivas Rao, and Vasudeva Varma. 2010. User Behavior in a Multilingual Information Access Task,, Indian Institute of Information Technology Allahabad, India. Report no: IIIT/TR/2010/30.

M. D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pp. 623–632, New York, NY, USA: ACM.

Tuomas Talvensaari. 2008. Comparable Corpora in Cross-Language Information Retrieval. Ph.D. Thesis, University of Tampere, Department of computer sciences, A-2008-7.

Alvin Toffler. 1970. Future Shock.

H. Turtle Strohman, D. Metzler and W.B. Croft. 2004. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2004.

Tom Wilson. 2000. Human Information Behaviour. *Informing Science* **3** (2): 49–55

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '01). ACM, New York, NY, USA, 334-342.