Taligner: itzulpen corpus eleaniztunak sortzeko tresna

IBON URIBARRI ZENEKORTA Euskal Herriko Unibertsitatea (UPV/EHU)

Artikulu honetan, Taligner programa informatikoa aurkeztuko dut. Software horrek corpus paraleloak sortzeko, kudeatzeko eta analizatzeko balio du, corpus eleaniztunak eta itzulpen corpusak barne. Programaren jatorria eta ezaugarri nagusiak zehaztuko ditut, bereziki.

Jatorria

EHUko Itzulpen eta Interpretazioko jakintza-arloko ikerketa ekimenetan du jatorria Taligner programak. Bai TRACE ikerketa proiektuetan bai Gasteizen egin diren zenbait doktoretza tesitan, corpusak erabili dira zenbait itzulpen errealitate hobeto ezagutzeko. Hasiera batean, corpus horien kudeaketa oso oinarrizkoa izan zen: Word programarekin lerrokatzen ziren jatorrizko testua eta itzulpena(k). Egoera horrek ez gintuen asetzen, eta ordurako eskuragarri zeuden zenbait software probatu genituen. Jauzi nabarmena egin zen Elizabete Manterolaren tesiari esker. Batetik, itzulpen corpus paralelo eleaniztun zabala sortu zuen Manterolak (Atxagaren 12 testu oso edo lagin, eta haien itzulpenak 7 hizkuntzatara; milioi eta erdi hitz inguru guztira), eta, horretarako, Wordsmith Tools programa erabili zuen. Tresna hori zen une hartan aurkitu genuen onena; izan ere, corpusak analizatzeko tresneria garatuarekin uztartzen du corpus paraleloak sortzeko tresna. Hala ere, zenbait arazo edo oztopo topatu genituen: batetik, software ez-librea erabiltzen du; bestetik, artxibo mota propio ez-estandarrak baliatzen ditu, eta, batez ere, lerrokatzeko tresneria oso astuna du, hau da, jatorrizko testua eta itzulpena(k) esaldi mailan doitzeko prozesua oso neketsua da.



1. irudia: Atxagaren itzulpen corpusa Wordsmith Tools erabiliz.

Hala, ikusi genuen aurrera egin nahi bagenuen geure beharretara egokitutako tresna bat sortzea zela bidea. Une horretan, gainera, Hezkuntza Ministerioko TRACE proiektua lortu genuen, Leóngo Unibertsitateko talde batekin elkarlanean. Orduan, Leónen egin ziren Trace aligner 1.0. sortzeko lehen urratsak. Bitestuak lerrokatzeko oinarrizko programa bat sortu zen, baina oso erabilera mugatua zuen: lerrokatutako testuak (laburrak) html formatuan erakustera mugatzen zen.

Geroxeago, Gasteizen, Iñaki Albisua informatikaria kontratatzeko aukera izan genuen, TRACE proiektuen diru laguntzei esker hasieran eta TRALIMA/ITZULIK ikerketa taldeko baliabideei esker ondoren, eta horrek ekarri zuen TRACEaligner 2.0. garatzeko bultzada erabakigarria. Leónen egin zena oinarri hartuta, Albisuak programa garatu zuen, bi doktoretza tesiren eskakizunak betetzeko, Naroa Zubillagaren tesirako lehenik eta Zuriñe Sanzen tesirako ondoren: programaren aldaera berriak testu luzeekin lan egiteko aukera ematen zuen; tritestuak lerrokatzeko bidea ireki zuen (jatorrizko testu alemanez eta euskarazko itzulpenez gain tarteko gaztelaniazko testua ere behar baikenuen zeharkako itzulpenen kasuan); lerroketa doitzeko funtzioak gaineratu ziren; corpusak mysql formatuko datu base bihurtzen zituen, eta bilaketa motorra garatu zen sortutako corpusa arakatzeko. Une horretan genuen arazo nagusia datu basea sortzeko prozesu astuna zen (xampp programa erabili behar genuen sql datu baseak sortzeko). Programaren aldaera hori Aleuska egitasmoko zenbait lani esker eta, aldi berean, lan horien mesedetarako garatu zen. Alemanetik euskarara itzulitako testuen katalogoan oinarrituta (700 sarreratik gora), Naroa Zubillagak AleuskaHGL itzulpen corpus paraleloa sortu zuen (18 egileren 19 itzulpen zuzen eta 14 zeharkako; 1.275.000 hitz guztira); Zuriñe Sanzek, AleuskaPhraseo corpusa (30 egileren 34 itzulpen zuzen eta 14 zeharkako; 3.500.000 hitz guztira), eta Ibon Uribarrik, AleuskaFilo corpusa (13 egileren 33 itzulpen zuzen; 1.200.000 hitz guztira). Aleuska corpusek gaur-gaurkoz 5.500.000 hitz inguru biltzen dituzte osotara. Besteak beste, Touryren interferentziaren legea zehazteko balio izan du. Touryk zeharkako interferentzia aipatzen du (1995: 72):

Hypothetically identified relationships may also give rise to the assumptions that a target text drew on a text in a language other than the assumed one, or on more than one source text, in more than one language.

Baina, gero, ez du zehazten interferentziaren legeaz diharduenean. Izan ere, paratestuen mailan itzulpen zuzen moduan aurkezten diren testuetan, zeharkakotasun arrastoak nabari dira Aleuska corpusetan, zenbait mailatan. Gerta daiteke testua itzulpen zuzen moduan aurkeztea baina benetan gaztelaniatik itzulia izatea testua, ez alemanetik, hau da, ustez itzulpen zuzena zena (assumed direct translation) ezkutuko zeharkako itzulpena izatea (covert indirect translation). Testu interferentzia hori osoa izan daiteke, jatorrizko testua erabili ez denean, edo gertakari bakana izan daiteke, aldian-aldian jatorrizko testuaz gain bitarteko itzulpena(k) erabili d(ir)elako (itzulpen konpilatua litzateke hori). Zeharkako testu interferentzia horretaz gain, zeharkako interferentzia instrumentala ere identifikatu dugu, alemanetik euskarara itzultzerakoan zeharkako laguntza tresnak erabiltzen direlako (hiztegiak, bereziki). Eta, azkenik, zeharkako gogo interferentzia edo interferentzia kognitiboa ere aurkitu dugu, hau da, alemanetik zuzenean itzultzerakoan, euskal itzultzaileak zeharkako hizkuntza bat ere aktibatzen du (gaztelania, bereziki), eta, orduan, zeharkako itzulpena egiteko aukera sortzen da noizean behin, tarteka edo sarriago. Beraz, itzulpen zuzena/zeharkakoa dikotomia oso motz geratzen da itzulpen errealitate askoz ere konplexuago bat aztertzeko. Azaldutako itzulpen egoera horri zeharkako interferentzia diglosikoa deitu diogu. Ustez (guztiz) jatorrizkoak diren testu produkzioan gertatzen den fenomeno bera da, azken finean.

Azken TRACE proiektuari esker (EHU, Leóngo Unibertsitatea, Kantabriako Unibertsitatea, Elxeko Unibertsitatea), Taligner 3.0. garatzeko aukera ere izan dugu. Iñaki Albisuak hutsetik berridatzi du programa osoa. Oraintsu arte, adabakiz osatutako kode nahasi bat genuen, eta arazo teknikoak sortzen ziren, askotan. Orain, java hizkuntzan idatzitako programa multiplataforma bat dugu, hau da, java behar du ordenagailuak programa erabiltzeko, baina berdin dabil, printzipioz, Windows, Mac eta Linux plataformetan. Hauek dira, oraingoz, azkena den aldaera horren berritasun nagusiak: pluritestuz osatutako corpusak osatzeko aukera ematen du, hiru testuko muga gaindituz; testu dramatikoz osatutako corpusak sortzeko aukera berezitua ematen du; corpusaren testu analisia egiteko lehen tresnak biltzen ditu (testu hustuketa egiteko tresna); datu baseak sortzeko tresna integratu da (orain ez da xampp edo beste programa berezirik behar); hainbat corpus aldi berean kudeatzeko funtzioa gaineratu zaio. Aldaera hori Itzulpen Ikasketetako gradu amaierako lan batzuk egiteko baliatu da; Itzulpena eta Teknologiak graduondokoan ere erabili da, eta Garazi Arrulak 10 idazleren 10 testu eta haien autoitzulpenak biltzen dituen corpusa osatzeko baliatu du.

Programaren ezaugarri nagusiak

Hasiera batean poesia testuak ere lantzea pentsatu bazen ere, oraingoz, aukera hori ez da garatu, eta testu laua eta testu dramatikoa erabiltzeko dago prestatuta programa. Hasieran, testu laua bakarrik hartu genuen kontuan, eta, azken aldaeran, testu dramatikoa gaineratu diogu, hau da, elkarrizketa testuaren egitura konplexuagoa lantzeko aukera jaso dugu: pertsonaien izenak, pertsonaien solasa, iruzkinak.



2. irudia: Taligner 3.0. interfazea.

Orain, funtzio nagusiak aurkeztuko ditut, ezkerretik eskuinera. Lehenak *Limpiar* izena du (oraingoz, gaztelaniaz bakarrik dugu interfazea), eta testu garbiketan laguntzen duen eranskin bat dela esan daiteke. Hainbatetan, testuak eskaneatu egin behar izan ditugu, eta horrek testuak garbitzeko lan astuna ekartzen zuen gero. Testu garbiketa horretako eragiketa errepikakor batzuk kolpe batean egiteko tresna bat sortu zuen Albisuak, Access erabiliz, eta tresna autono-

moa zena programa nagusian integratu genuen, azkenean. Baina, zuzenean testu digitalarekin lan egiteko aukera izanez gero, funtzio hori ez da ezertarako behar.

Etiketazioa da bigarren funtzioa. Lerrokatzen hasi aurretik, testuak etiketatu egin behar dira. Abiapuntu gisa, txt formatuko testuak erabiliko ditugu beti, eta, etiketatze funtzio honekin, xml artxiboak lortuko ditugu. Testua aukeratu eta kargatu ondoren, *Etiquetar texto* botoia erabiliz, behin-behineko emaitza ikusiko dugu, eskuinean.



3. irudia: Testuak etiketatzen. Canettiren Die gerettete Zunge.

Gure kasuan, testu bat etiketatuta egoteak esan nahi du, alde batetik, paragrafo eta esaldiak etiketa (*tag*) bidez markatuta egongo direla, eta, bestetik, testuak metadatuz hornituko ditugula; gero bilaketak egiteko ezinbestekoa izango den informazioz, alegia. Etiketa horretan, testuaren izenburua, egilea, itzultzailea, eta itzulpen eta genero mota jasotzen dira, besteak beste.

Ondoren, sortuko den fitxategi berriari izen bat jarri, eta gorde egin beharko dugu (*Guar-dar texto*), eta, hala, xml fitxategi bat sortuko da; hor amaitzen da etiketatze prozesua. Eragiketa bera egin behar dugu corpusean jaso nahi dugun testu ororekin. Xml fitxategiak software jakin bati lotuak ez daudenez, testu kudeatzaile arrunt bat erabil dezakegu fitxategiaren barneinformazioa ikusteko.

<u> </u>) contractor	• 🕑 ••••		Configure Course
Ceperton and Garder Higherton	C TRACIANS			
	Ater			Daetah
		-		nd-Lengtud 1004
Rustachuit 1905-1911.	Take Education Alde		_	1#
Meine friheote Drinnerung . Meine friheote Drinnerung ist in Rut geteucht. Auf dem ärm einer Madrhene kumme ich zu einer Tür	Calecoles			34 Rastaclask 1905-1911.
Separaber von une, in seiber Nobe, officet sich Er tritt gene ombe an mich beran, bleibt steben	Pelle			34 Meine trüberte Erinnerung .
"Seag dae Junge"".	tabatar:	Ealers	-	All Makes Willow's Educations at in Roll principle.
lik strecke die Junge berwus, er greift in seine fr segt:	Made	Orgent	*	54 Auf dam Arm-sizer Midchenz instant ich zu einer Tür bernun, der Buden vor mir ist rei, und zur Linken geht eine
lob wege es alobe, die Dunge purickryslieben, er	Citilge			10 Geparather was use, in other Wate, offset sich eine Türe and ein lächelnder Mann tritt herson, der Brandlich auf mich
In jetzten äugenblick zieht er dus Mesoer rurick.	SubCerpusi General	Sedgesfice		THE Dr tritt gasz radie an mich herne, blabt atoben und sagt zu mic
Te himppt das Masser wieder zu und stecht as in a Zeden Borgen treien wir mas der Tür bereus auf de				in "Eng de Dauge".
Ech weld, was an angen wird und warte auf seland Ech weld, dab ar sie wir abschneiden wird und fo		Guerde Tento		14 ich strecke de Zange benous, er groß in seine Tasche, nie ein Taschennenser herver, offint in und führt die Känge gest
rer ing seption oment, sold es peschietet visie An Dit behalte es für mich und frage erst sehr visi :	1			100 Er magt
				11# "Joint schoolen wir ihm de Zange ab".

4. irudia: Metadatuak. Canettiren Die gerettete Zunge.



5. irudia: Xml dokumentua. Kafkaren Die Verwandlung.

Irudian, fitxategiaren zati bat besterik ez da ikusten. Garrantzitsuena bi zati handi daudela identifikatzea da: alde batetik, <header> etiketaren barnean biltzen dena, eta, bestetik, <text> barruan dagoena. Lehen zati horretan, metadatuak aurkitzen dira (izenburua, egilea, kodea, itzultzailea, hizkuntza, modua, generoa), eta bigarrenean, aldiz, testua bera dago. Testua, era berean, paragrafo eta esalditan banatuta dago. Paragrafo eta esaldi gisa markatzeko, berariazko etiketak erabiltzen dira: (eta <s>). Goiko irudian gorriz agertzen den *n*-a paragrafo eta esaldien atributu bat da, eta halakoak zenbatzeko balio du. Beheko eskeman ikus daiteke aipatu berri duguna, laburbilduta. Sarean erabiltzen ditugun corpusek ere egitura bera dute (TEI estandarra).



6. irudia: Xml testu baten egitura.

Testuak etiketatu ondoren, lerrokatu egin beharko ditugu, hirugarren funtzio nagusia erabiliz, *Alinear*. Bi, hiru edo testu gehiago lerrokatu daitezke aldi berean, aurretik sortu ditugun xml artxiboetatik abiatuz. Segundo gutxiren buruan lerrokatuko du programak testu bikote, hirukote, laukotea... (testuak oso luzeak baldin badira, denbora pixka bat har dezake kontuak). 7. irudian, behin testuak lerrokatuta programaren itxura ikus dezakegu.

Hala ere, gehienetan, lerrokatze-prozesu automatikoa ez da guztiz zehatza izango, are gutxiago literatur itzulpenak lerrokatu nahi baditugu, hau da, programak testu bakoitza esalditan banatzen du, eta testuak esaldiz esaldi uztartzen, baina uztarketa hori ez da *zuzena* izango beti; bereziki, gorabeherak izan badira itzulpen prozesuan puntuazioarekin. Orduan, guri egokituko zaigu, jatorrizko testuan oinarrituta, jatorrizko testuko esaldia xede testuko esaldiarekin bat datorrela ziurtatzea.

Addiese		
∾ _ ±-	- @ @	Corpus and corpus
👔 taya dasa 💽 tarka d		in television N
194 ⁰	Uni-Etz.ini-Langlud tor?	Unit-DE2.cml -Longitud: 1047
	1# URLO.	IN URLO.
	24 Famaren cian M.	20 Renerocion M
	3# Aditor	34 Adree
	ANT UNLO	or URLO.
nit.	5#	54
SUBRERAE ESIDIENAES OPERA.	OF COMEDIA LINICA IN THEIR ACTOR.	OF KOMERCHE OPER IN DREE ARCTER
	TH MÜBICA Y LETINA CON TRADUCCIÓN RITMICA.	74 Musik and riptlassicher Test.
	54	14 Is Bulkisch und Spanisch.
and a second	14 Did Prohiber Dr. ADDTE.	94 Von Predsyller Dr. ADD/R.
	IS DELOCADA.	104 In Verdung
anako Kondean.	11# A rubum anapi d Escena St. Coste Satera.	List Gewähnet seinem galen Frenade, dem Graden
	138 Propedul del Autor.	12# Eigentum des Komponisten.
	13# En bilhes 1911-1912.	13# Miles 1913-1912

7. irudia: R. M. Azkueren Urlo opera, euskaraz, gaztelaniaz eta alemanez.

• Tableger - Neveline					5.81	
谷	()	0		- Cardquere Cargos 🕑 540		
1						
🚰 Caparitanta 🔬 Capar abasatia 🏝 Sandar abasatia	(i) con 10 🖄		X			
Canation gold and - Longbuil 1940	Constitute and and	1420		Carettilungel2.col - Longitud SH1		
14 Postaclaik 1965-1911.	1# RUSCHUK - 1985-1911			IN MUSTBORIK (1985-1911).		
24 Mane Briteste Erinnerung .	24 My Earliest Mesory .			20 M records mix renots.		
34 Siltese Brideste Briaserung ist is Ret gelaucht.	3# 3/y submi memory a 4	lepel a.rel.		38 Mi recuerte mia resolo erta balado de rejo		
14 Auf dem Arm man Micklama komme ich m mar Tie beraus, der Boden vor mit ist net, und zur Linion geltt sine Treppe himselter, die	All 2 cause not of a door on in red, and to the left a share	the arm of a maid, the tare goes drown, reput	e floor in front of me dy cost.	10 Daigo por una puerta en loratos de se es rejo y a la impaiente deciende una en	a mulacia, are mi d noto njen godonete roja	
14 Cognitive ros sao, in other Hilts, affliet such sur Tare and en- lactation Man tell Sense, or Brandick auf such rapits.	14 Acres from us, at the same steps forth, walking it	Contrast Contrast on Contrast	enc, and a realing	10 Frente a accorticut, a la marma altura, un hombre constituite que nime amiguiés	or alore una puesta y aparece mente bacia mi	
14 Er teit gest ader as mich heras, Melht riefen und sag zu mar	of He stops right up close	Doartar Banco Ginela		10 De me aproxima macho, se delene, y me dice		
34 "Dag de Dauge".	14 'three as your tongot	Dev		14 "(Eurola la Impari".		
iet lich erweize die Jampe bernue, er greift in seine Turche, mehr ein Turchenzummer bervier, offliet en und führt die Uinge gant nater an	In 2 mick out my tempter, packloally, opener 6, and be-	Garder IM. Editor Milan Milan Instant	, pulle out a	10. To man kalenges, 4 paips on ra bai y scorrisodo la cochilacionie a mi lenges	db, eccer un camp, b der	
14 Er nagt "Jetat schonien wir dan die Sangr-de".	SH He says "New well of	Quiter columns		98 dox 'Mare is ortermoria imper	0	
100 leis wage er nelst, de lange meiskensisken, er kommt inner niker, gesch werd er or nat der Clauge berdaren	100 2 don't dare puil back o Made will truch our any or	Real a share in free op solgert, or coloris col.	cover and chore, the	100 No me store a retrie is lengts, di retaria ton is loga	et acetta cada vez mair barta	
118 In Inten Augustick unti er die Meuer narich, sogi	11# In the last moment, he	puls back the leafs.	repting	110 Es el últeur memorie relax la neve	ja y den	
Die "Nieste such sucht, margen".	128 'Blat today, tomorrow'			124 "Ney interta ne, malam".		
ZINDEL HACKOB	and the later in the features	-	Chipe Arts.		Case Magoria	

8. irudia: Lerrokatze funtzioaren doikuntza tresnak.

Itzulpena(k) jatorrizkoari egokitzeko, beharrezko aldaketak egin ahal izateko, goiko irudian ikus daitezkeen botoiak ditugu (saguaren eskuinaldean sakatuta ikusiko ditugu). *Combinar* botoiarekin, bi esaldi batu ditzakegu; laukiak gehitu nahi baditugu, *Insertar blanco*-n egingo dugu klik; hitzak berak dioen bezala, *Dividir*-ekin, laukiak zatitu daitezke; *Eliminar*-en klik eginez, nahi beste lauki kendu dezakegu; *Editar* funtzioarekin, lauki bateko testua aldatu/zuzendu dezakegu. Gero, gainera, goiko botoietan, *Deshacer* daukagu; aukera horrekin, azken agindua desegin dezakegu, hanka sartu baldin badugu.

Lerrokatze lana amaitu dugunean, edo lanerako denbora gehiago ez badugu, lanaren emaitza gordetzeko modu bat baino gehiago dago. Saguarekin, *Guardar xml* erabil dezakegu, eta, hala, lerrokatu ditugun testu guztiak gordeko ditugu, banan-banan. Lanarekin aurrera jarraitu nahi dugunean, testu bakoitzaren xml artxiboa igo beharko dugu berriz, banaka-banaka. Bestalde, goialdeko funtzioetan, *Guardar tmx* dugu; aukera horrekin, lerrokatutako testu guztiak artxibo bakarrean gordeko ditugu (OLI programek erabiltzen duten formatu berbera da).



9. irudia: Lerrokatutako tmx artxiboa.

Hala, gordetako artxibo bakar hori kargatu egin dezakegu gero, lanarekin jarraitzeko. Hala ere, tmx formatoak badu arazo bat: ez du metadaturik gordetzen. Horregatik, Taligner 3.0. aldaeran, taf formatua gaineratu dugu, zeina *Guardar alineación* funtzioarekin aktibatzen den. Kasu horretan, testu guztiak artxibo bakarrean gordetzen ditugu, metadatuak barne. Gero, artxibo hori berriro kargatu daiteke, lanean jarraitzeko. Lerrokatze lanak amaitu eta emaitza gorde badugu, datu basea elikatu behar dugu. Horretarako, testu bakoitza banan-banan igo behar dugu datu basera, saguaren funtzioen artean *Añadir a corpus* hautatuta. Aurreko prozedura hori hainbat aldiz errepikatu dezakegu corpusean sartu nahi ditugun testu bildumekin; hala, datu basea elikatzen jarraituko dugu.

Corpusa eratua dugunean, bilaketak egin ahal izango ditugu. Horretarako, goiko *Consultar corpus* atalean sartuko gara. Hor, bilaketa motorraren interfazea aurkituko dugu. Hala ere, bilaketak egin aurretik, corpusa modu egokian antolatu behar da, eta, horretarako, *Configurar corpus* atalean sartu, eta eragiketa batzuk egin beharko ditugu.

		-	•	141		isar-	
	r spull						
92	buote ceps and toote ceps						
-		,					
Ciere I	The	Adar	Talata	18ma	Gen	Airden	Description
	Atheismus im Christentum	Bloch		Deutsch		1	
3	Advismoe kristautesunean	Bloch	Pello Zabaleta Kortaberri				
4	Enletung in die Gesteswissenschaften	Diffey					
5	Izpiritu-zientzietarako samera	Dittey	Xabier Mendiguren Bere				
6	Die Freudsche psychoanalytische Meth	Freud					
7	Azterbide psikoanalitiko freudiarra	Freud	Lontxo Othertzabel Rez				
8	Phanomenologie des Gestes	regel					
9	diogoarwn fenomenologia	regel	Xabier Insausti, Ibon UK				
10	Bref über den Humanismus	Heidegger					
11	Gutune humanismoen buruz	Heidegger	Pello Zabaleta Kortaberri				
12	Abhandlung über den Ursprung der 6	Herder					
13	Hokuntzaren jatomari buruzko tratat	Herder	Antton Garkano Irureta				
2-6	über das Studium des Alterthums, un	Humboldt					
15	Antzinatearen eta berezki Greziamare	Humboldt	Pello Zabaleta Kortaberri				
16	Kritik der praktischen Vernunft	Kant					
	Americanan protein protein	KAVE	Bon Uribam Zenekorta				

10. irudia: Corpusa konfiguratzen. AleuskaFilo.

Hor, corpuseko testu bilduma osoa ikusi eta kudeatu ahal izango dugu. Bilaketa prozesuan arazorik ez izateko, testuen arteko harremanak zehaztu ahal izango ditugu, hau da, itzulpenen eta jatorrizkoen arteko harremanak, testuen gainean saguarekin *Editar metadatos y relaciones* aktibatuz. Bestalde, corpus osoan azpicorpusak zehazteko aukera ere badugu; hala, bilakete-tan corpus guztia erabili beharrean, azpicorpus bakoitzean egin dezakegu bilaketa. Gainera, aurretik osatutako corpusak inportatu edo sortu duguna esportatzeko aukerak ere badauzkagu hor. Corpusa ongi eratu ondoren, bilaketak txukun egin ahal izango ditugu *Consultar corpus* funtzioa baliatuz.

Constant respond Mone de biogram. Mone de biogram. Mone de biogram. Cardin de Card	± " 0			martin 14	. televete	- 25 -	
Create spuid Main is hispan. Nain is hispanite Adar Nain Nain Carel Equation Carel Equation Nain Nain Nain Nain Nain Nain Nain Nai			0	1	5		
Marce & Bridgen. Table W Made & Bridgen. Scallo W Made & Bridgen. Scallo W Made & Males Scallo W Made Table W Gred Expend Table W			_				
Men in tongan. Units in a Pada de Solana de Pada de Solana de Pada de Solana de Pada de Solana de Pada de					_	Consult	ar corpus
Ten bagan baran baran Baran baran b	a hour a name	0.00	-	Second Second	Contractor of		
Adar hile di Ruis tale di Crel Egenat tale di	e ten chara 2		_	Description of the local diversion of the loc	Contractor of the	Contractor Million	
nan bin di nah bin di Gel fipele	e Teste o han ar 2		_	Detercie a Tentos	Compare of	Compare and	
Nation Value (1) Cond Reported	e taken	Toolan .	-	-	10	×	6
Crel Special Lobs g	e Gres	Testine	M	Bacir Inductiones	U		
	er Artes y después:	(C.M)					DPOR

11. irudia: Bilatzailea.

Bilaketa interfazea aurrez aurre izanda, argiago ikus daiteke metadatuen funtzioa. Izan ere, testu bakoitza metadatuz hornituta dagoenez, posible da, esate baterako, bilaketak egile jakin baten testuetan bakarrik egitea. Eta gauza bera egin daiteke hizkuntza, izenburu, itzultzaile, itzulpen modu eta generoarekin. Horretaz gain, hitza(k) bilatzeko orduan, aukera bat baino gehiago dugu: *Exacto, Contiene, Empieza con, Termina con*, alegia. Hala, hitza bere horretan, hitz zati bat, hitzaren hasiera edo hitzaren amaiera bilatu dezakegu. Gainera, hitz bateko bilaketa sinplea egiteaz gain, hitz konbinazioak bilatzeko aukera ere badugu (hiru hitzeraino), hitz horien arteko distantzia zehaztuz. Gainera, emaitzak bistaratzeko unean, bilatu nahi dugun hitza edo esamoldea agertzen den esaldia bakarrik aukeratu dezakegu, edo aurreko eta ondorengo esaldi 1, 2 edo 3 ere ikus ditzakegu. Hona hemen adibide bat.

-	- 2	()	4	~ 0	Grade	- 0	Gradard	- 🔅	je tentare tita	corp	US (onlyane (organ	<u>ں</u>
Alter 0	elgen copa	Create copul										
										Co	onsultar corp	IS T
1	Norm in history.	German		Hole-in kingarde	East		Texto a	rbatier 2	2.eqe			12.50
	the post of the	E-ato	8	Terts of Justian 21			Detand	is a Tente Si	Culpers 1	1		- 65
,	tala de bisqueda	E-ato		Tests albusiar 3			Detund	ina foto	Culpins 1	Culpera	×	
	Mar.	Table		Tradutor:	1.44		The other		Toda			34.8
,	nde .	Cuelaire		Génera	Curre	an a	hour	Industries	8 .	diversity Pa.	- 19	5.4
4	Ored Expectation	hala	×	Artins - después:		×	Second		×			DPORTM.
Gade	Test				Cade	Text				Gde	fleet.	
Dange CE	Er tell gest to	be as much here	n, Nebi steb	m and stag to mar	SageEN 00	He steps right	up close to me, i	halts, and says:		Dept3	De me aproxima maci	ui, se Belieter, y me
DagCE (7)	16 "Zing de Zinge".				Dangellik (7)	'Show me your impor'.				DeepfER (7)	"/Earda k Impa/".	
DangeC.K. (D)	pCE lish stretche die Dange beruur, er greff in onne Tanche, nield ein Tuschenzenner bervor, offen ei und fichet die Könge gaaz sahe an meine Dange berun.			Dage(D) (0)	I stick out my tongoe, he reaches into his pocket, pulls out a jacklash, opens it, and brings the blade all the way to my tongoe.				2mpf2 (1)	To sace is bengue, 40 y y accertants is cuchili	nipa es en baixilo, a juste a en lengua	
tonget R	ngel E "Dag de Longe"			Congelite (7)	"Dow no you longu".				Dept3	"/facela 'a lenges?"		
Dangel H (F)	angellie lich strenzier der Longe kennun, er greff in onner Tauche, mehr ein () Tauchensnerser kennun, influet es und führt die Könge gest sahe an meine Longe kenn.			SingelD4 (R)	I stick out my tanger, he reaches into his pocket, puls out a picklesh, spece 8, and brings the blade all the way to my tangue.				Despt2 (1)	To saco la impas, di pajas m ra bolollo, y acorcando la cuchila parte a mi impas		
bertt	Er sam Viers	schenden wir de	en die Danes	41.	Dard)4	He serve Trippe	well-out-off his	I house and		Deedo	dice "Albom in contain	more in income.
					1	(* Coincie	lencias:	10				

12. irudia: Zunge hitzaren bilaketa Canettiren Die gerettete Zunge testuan.

Bilatu d(ir)en hitza(k) horiz markatuta agertzen d(ir)a, eta, gainera, esaldia identifikatuta dago lehen zutabean testuari eman zaion kodearekin eta esaldiaren zenbakiarekin. Gure kasuan, gainera, hitza agertzen den esaldiaren aurreko eta ondorengo esaldiak ere erakusten ditu bilaketak, testuinguru pixka bat gehiago izateko. Egiten diren bilaketak Excel fitxategi moduan gorde daitezke, gerora berriro kontsultatzeko.

Horretaz guztiaz gain, *Consultar ficheros* atala ere badugu. Hor, testuak analizatzeko aukerak nahi ditugu bildu, pixkanaka. Oraingoz, corpuseko testu bakoitzaren hustuketa terminologikoa egin daiteke, alfabetoaren arabera ordenatuta edo hitzen maiztasunaren arabera ordenatuta. Testu batean agertzen diren hitz guztiak eta euren maiztasuna ematen ditu horrek. Gainera, hitz kopuru gordina (*token*) eta hitz formen kopurua (*type*) kalkulatzen ditu.

Att Aligher - Nor Latina						- E	
🖗 🚥 🚫 🚥 🛓	‡~ Ø	Consider Corpus	🗿 (mata 10an	· 🔅	corpus	····· 🕑 ·	
pre cepe Cenda cepe Cenda Mero	20						
Courses X							
in obviously	, Analais						
Robusto .							
read a	Paldra	Cateled	Paldro	Calified			
dit dan matanin Termunifi, Kut		804 8	Car.	144 15			
		10	h	504			
	diruh	N	and .	100		Meeting machine	
	abbraha		P	1000		0	
	detaur		*	2542		Outer mineral	
	der	1047	rafe .	2146		8	
	distan)	m .:	2121		Hans & options	
	distant)	*	2004			
	dilate)		3/56 Total paintment	Total painteres: 1702		
	diteen)		3566	And the Party of the	Name de représent	
	dilitie	9	-	1509			
	dispe)	de la	(ALL) Perde apartoin	Meda apartoin per paldina 1	18	
	dillost)	ah .	340	Palatras per carpite films . Md		
	diplie)	plana .	3409	Node 10	. (7)	
	dipter	16	-	3067			
	dated)	der	3047		Constant	
	diploaft	9	**	1094		6.2	
	digitization)	da .	1010		100	
	diplote		~	1417			
	digitation	9	10.0	1044		Equitar	
	digite		-	3100			
	digites		piere .	34.00			
	dehile	2	in the second se	3442			
				and the second se			

13. irudia: Immanuel Kanten Kritik der reinen Vernuft.

Aurreko adibidean, 175.374 hitzez (*token*) osatutako testu bat dugu; tartean, 9.659 hitz ezberdin (*type*). Analisiaren emaitzak Excel taula batean gorde daitezke.

Amaitzeko, *Configurar Taligner* atala dugu. Hor, metadatuetan sar daitekeen informazioa egokitzeko tresna dugu, hala nola, hizkuntzak, itzulpen modua eta testuen generoa.

Etorkizunean

Etorkizunean programa garatzen jarraitzeko aukerarik balego, hauek lirateke egin beharreko hobekuntza batzuk: interfazea euskaratu eta ingeleseratu; bilatzailea testu lematizatuak erabiltzeko egokitu; testu analisirako tresneria zabaldu (Antconc programak eskaintzen dituen aukerak integratuz, besteak beste); eta abar.

BIBLIOGRAFIA

- Barambones, J. (2012). *Mapping the Dubbing Scene. Audiovisual Translation in Basque Television*. Bern: Peter Lang.
- Barambones Zubiria, J.; Manterola Agirrezabalaga, E.; Sanz Villar, Z.; Uribarri Zenekorta, I.; Zubillaga Gomez, N. (2015). «Itzulpen ikasketak eta euskara. Zenbait ekarpen Itzulpen Ikasketa Deskribatzaileei, eta haratago», in *Traducimos desde el sur: actas del VI Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación*, José Jorge Amigo Extremera (arg.). Las Palmas kanaria Handikoa: Las Palmas kanaria Handikoko Unibertsitatea.
- Manterola, E. (2012). Euskal literatura beste hizkuntza batzuetara itzulia. Bernardo Atxagaren lanen itzulpen moten arteko alderaketa, Bilbo: Euskal Herriko Unibertsitatea.
- Sanz Villar, Z. (2015). Unitate fraseologikoen itzulpena: alemana-euskara, UPV/EHUko tesia.
- Toury, G. (1995). *Descriptive Translation Studies and beyond*, Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Zubillaga, N. (2009). «Alemanetik euskaratutako haur eta gazte literatura edo amaigabe istorioa?», *Senez*, 36: 231-241.
- (2012). «A corpus based descriptive study of German children's literature translated into Basque: Preliminary results», in *Translating Fictional Dialogue for Children and Young People*, Fischer, M. eta Naro W. (arg.), Berlin: Frank&Timme.
- ——; Sanz, Z. eta Uribarri, I. (2015). «Building a trilingual parallel corpus to analyse literary translations from German into Basque», in *New directions in corpus-based translation studies*, Fantinuoli, C. eta Zanettin, F. (arg.), Berlin: Language Science Press.

Resumen

Taligner es un programa para la creación y análisis de corpus multilingües, desarrollado por un grupo de investigadores del Área de traducción de la EHU-UPV en colaboración con el informático Iñaki Albisua. La herramienta se ha utilizado ya en varios proyectos de investigación, lo que ha permitido, a su vez, desarrollar la aplicación. Taligner permite etiquetar y alinear textos narrativos y dramáticos, para poder analizar posteriormente los corpus creados, con ayuda de un buscador. Así mismo, permite gestionar diferentes corpus. En lo referente al análisis de los textos, por el momento la aplicación realiza el vaciado terminológico de los textos recogiendo la frecuencia de utilización de cada palabra.

Résumé

Le programme Taligner a été conçu par un groupe de chercheurs du département de traduction de l'Université du Pays Basque, avec l'aide de l'informaticien Iñaki Albisua, dans le but de créer et d'analyser des corpus multilingues. Il a été appliqué dans plusieurs projets de recherche et il a pu, en même temps, être développé grâce à ces projets. Il permet d'étiqueter et d'aligner les textes narratifs et dramatiques, pour analyser les corpus ainsi créés au moyen d'un chercheur. De plus, il permet également de gérer les différents corpus. Quant à l'analyse des textes, ce programme permet pour le moment de faire le dépouillement terminologique en comptabilisant la fréquence de chaque mot.

Abstract

The Taligner software was developed by a research team from the translation section at the University of the Basque Country in collaboration with IT expert Ińaki Albisua, with the aim of compiling and analyzing multilingual translation corpora. It has been used in several research projects and has been further developed thanks to those projects. It provides capabilities for tagging and aligning narrative and drama texts so that the compiled corpora may be analyzed using a search engine. Furthermore, the software makes it possible to manage different corpora simultaneously. With respect to text analysis, Taligner is currently able to produce term extractions of the texts, counting the frequency of each word.