Quasi-Parallel Corpora for Less-Resourced Languages: Parallelized Translations of Plato´s Faidon in Basque and Finnish.

Koldo J. Garai
University of Eastern Finland (UEF) & University of the Basque Country (EHU)
Digital Research Data and Human Sciences (DRDHum).

2024/12/10

OBJECTIVES

- To present a possible “easy” path to provide and enrich technologically less-resourced languages
  - Following the aims of the European Language Equality (ELE) program
  - It is a proposal to enlarge this investigation with more texts and non-Indo-European languages
- To present our outcomes to be tested regarding their efficiency and whether they meet the goals

What has been done

- Two independent translations of Plato's Faidon
  - Finnish Faidoni by Calamnius 1887
  - Basque Faidon by Zaitegi 1975
- They have been parallelized and automatically annotated in Universal Dependencies
    - Now we look at the terms "less-resourced language" and "Quasi-parallel"

Less-resourced language

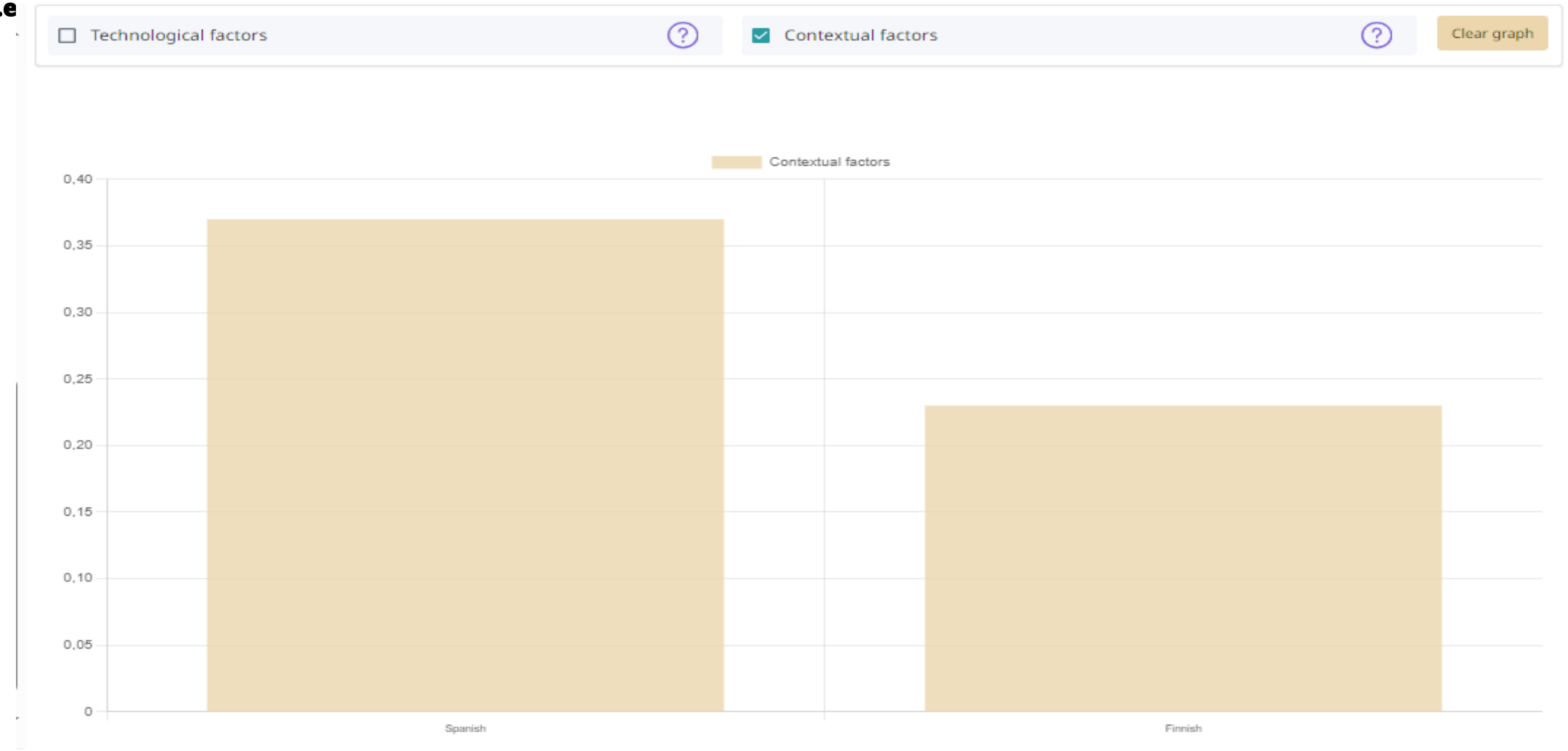
- **European Language Equality**
  - the mission of the ELE Programme is to achieve language equality in Europe by 2030, reducing the technology gap with English and addressing the lack of available language data.
  - “Digital Language Equality (DLE): all languages have the technological support and **situational context** necessary for them to continue to exist and to prosper as living languages in the digital age.” (Gaspari et al., 2023: 43).
  - DLE Metric is a measure that reflects the digital readiness of a language and its contribution to the state of technology-enabled multilingualism”.
    - It considers two-factor branches: technological and contextual

Technological & Contextual factors for metrics

- The technological factors (TFs) include language resources, tools, and services listed in the European Language Grid (ELG) catalog.
- The contextual factors (CFs) are based on societal, economic, and industrial conditions, and ecosystems of languages, to determine the potential for Language Resources and Technologies (LRTs), the extent to which a language has a **context that supports** the possibility of **evolving digitally** or not.

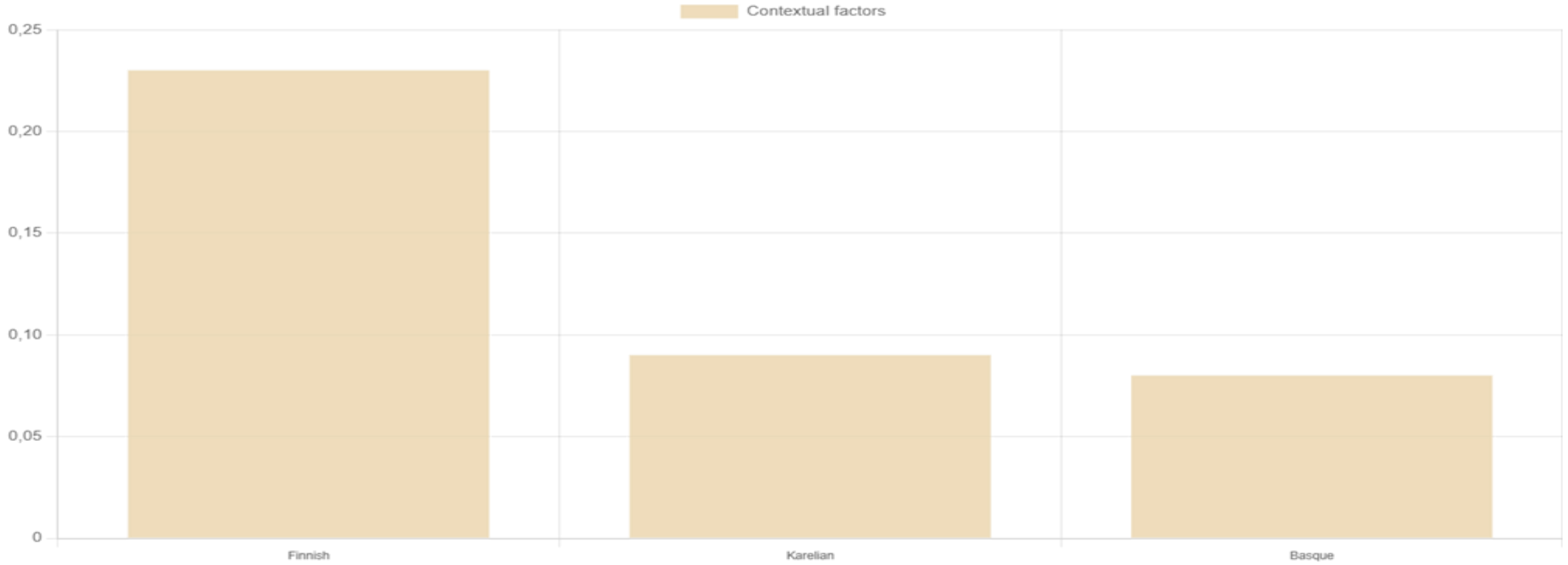
Contextual factors: Spanish vs Finnish

https://live.e

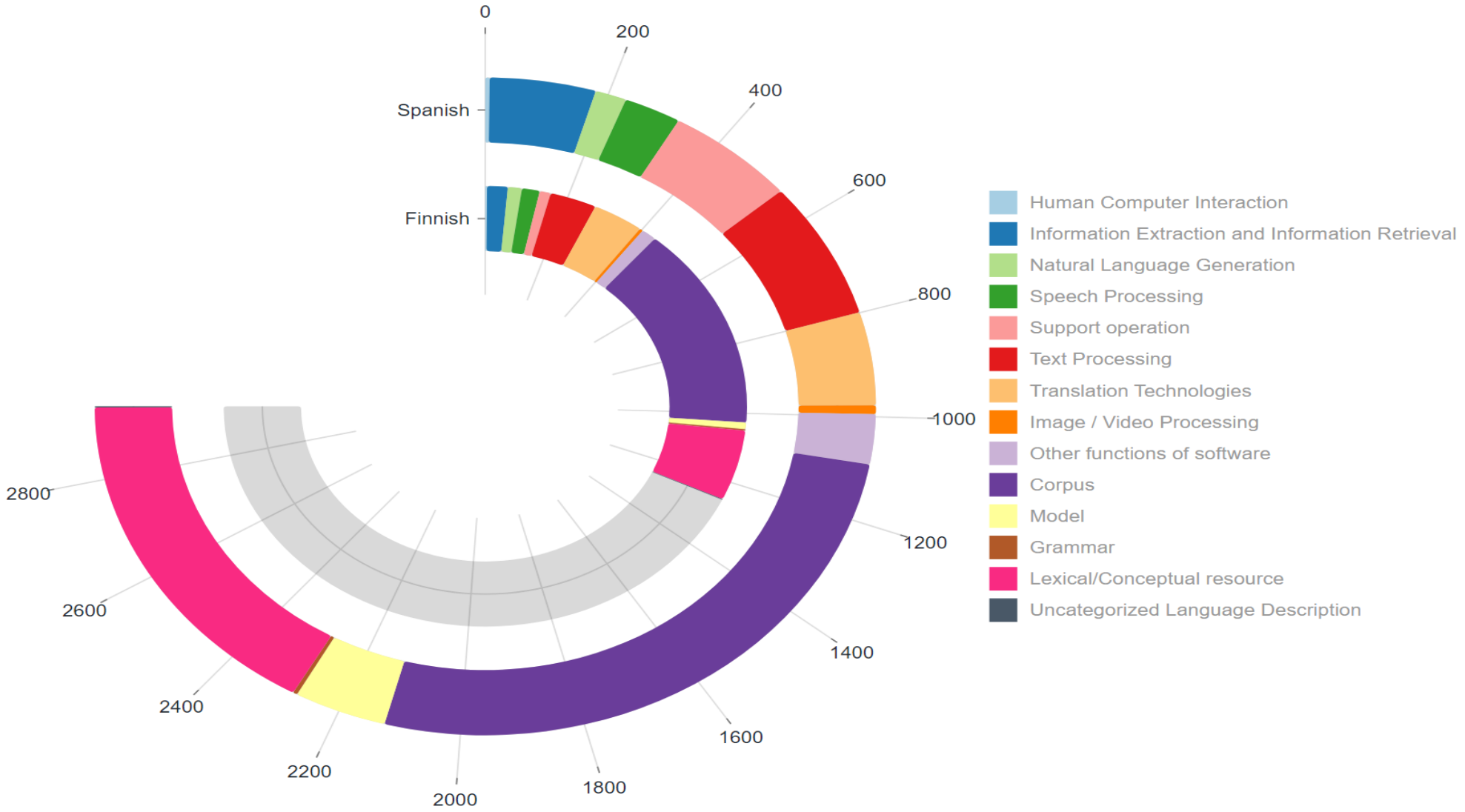


https://live.european-language-grid.eu/catalogue/dashboard

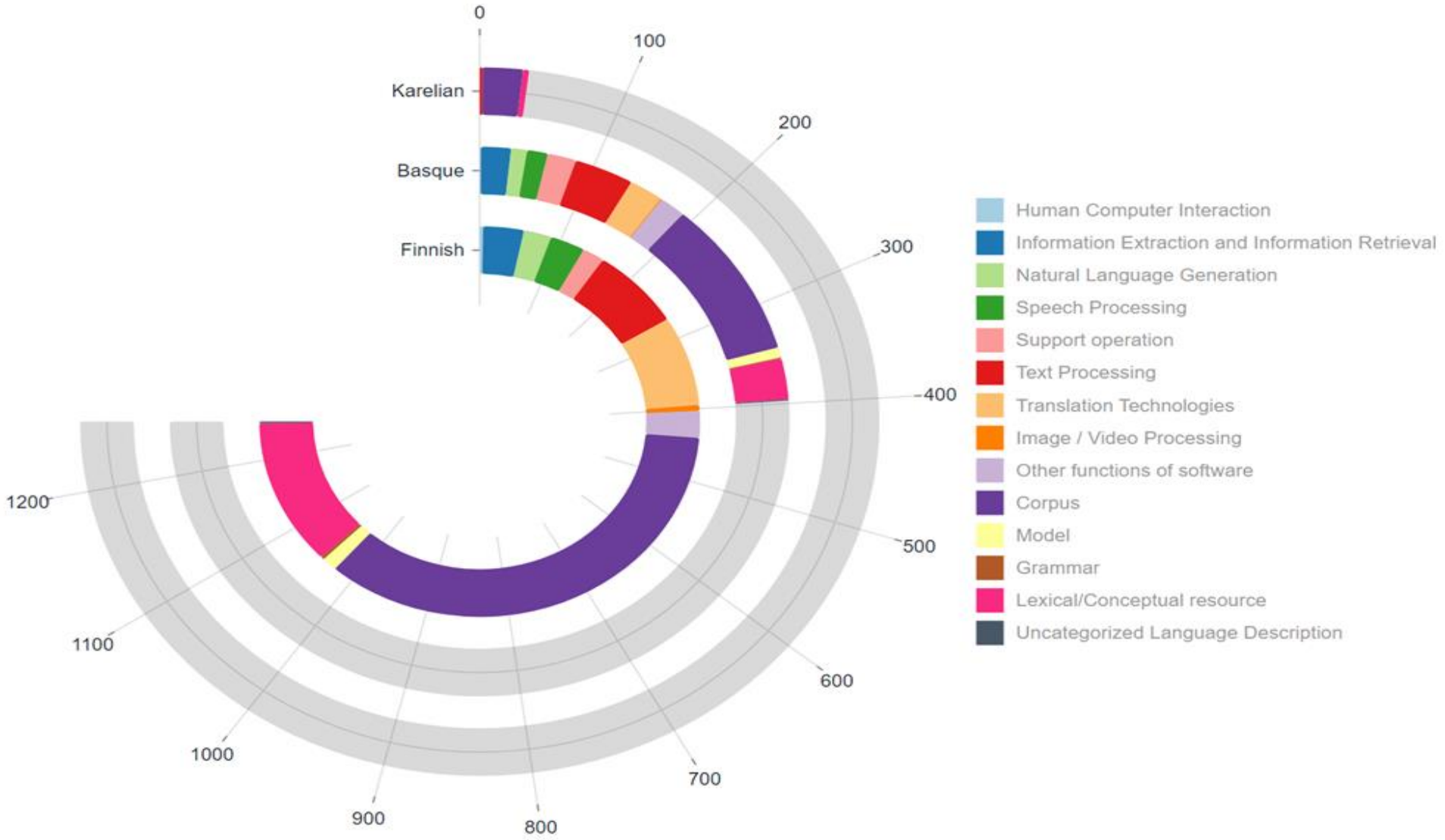
Contextual Factors: Finnish, Karelian, Basque



TFs: Spanish / Finnish



TFs: Finnish / Karelian / Basque



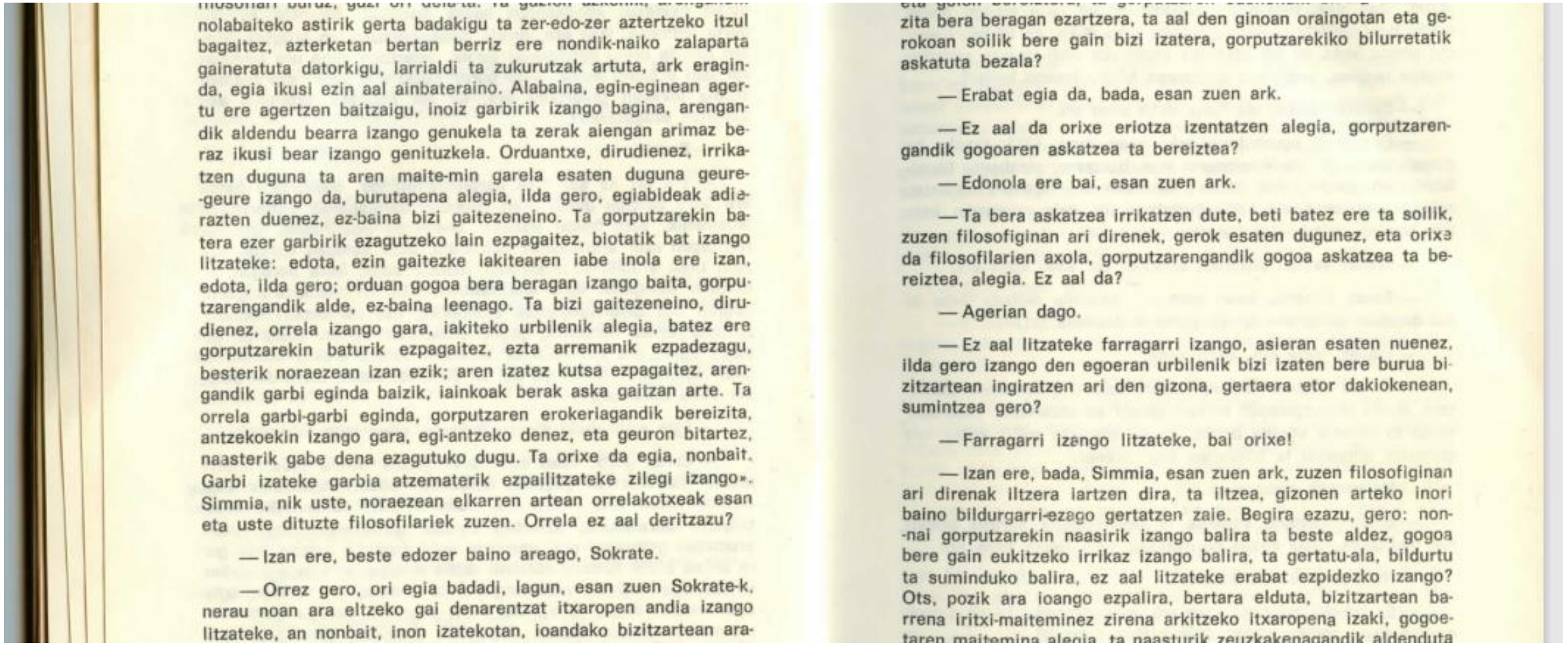
Parallel vs Comparable Corpora: Quasi-Parallel?

- Type A: Source texts plus translations (L1 > L2). This is the parallel corpora, strictly speaking.
- Type B: The comparability is given by the design of the same sampling frame. (L1 + L1'), similar balance and representativeness; the same proportions of the texts of the same genres, same sampling period, in the same domains, and not translations of each other.
- A combination of A and B . This also is included in the comparable category (McEnery and Xiao, 2018)

Process

- 1) Scanning the printed version and editing the OCR errors to have it in a machine-readable format of the Basque text,
- 2) Aligning both texts,
- 3) Standardizing the Basque version,
- 4) Annotating both texts, original Finnish and standardized Basque.

Digitizing: The text



nolabaiteko astirik gerta badakigu ta zer-edo-zer aztertzeko itzul bagaitez, azterketan bertan berriz ere nondik-naiko zalaparta gaineratuta datorkigu, larrialdi ta zukurutzak artuta, ark eraginda, egia ikusi ezin aal ainbateraino. Alabaina, egin-eginean agertu ere agertzen baitzaigu, inoiz garbirik izango bagina, arengandik aldendu bearra izango genukela ta zerak aiengan arimaz beraz ikusi bear izango genituzkela. Orduantxe, dirudienez, irrikatzen duguna ta aren maite-min garela esaten duguna geure-geure izango da, burutapena alegia, ilda gero, egiabideak adierazten duenez, ez-baina bizi gaitezeneino. Ta gorputzarekin batera ezer garbirik ezagutzeko lain ezpagaitez, biotatik bat izango litzateke: edota, ezin gaitezke iakitearen iabe inola ere izan, edota, ilda gero; orduan gogoa bera beragan izango baita, gorputzarengandik alde, ez-baina leenago. Ta bizi gaitezeneino, dirudienez, orrela izango gara, iakiteko urbilenik alegia, batez ere gorputzarekin baturik ezpagaitez, ezta arremanik ezpadezagu, besterik noraezean izan ezik; aren izatez kutsa ezpagaitez, arengandik garbi eginda baizik, iainkoak berak aska gaitzan arte. Ta orrela garbi-garbi eginda, gorputzaren erokeriagandik bereizita, antzekoekin izango gara, egi-antzeko denez, eta geuron bitartez, naasterik gabe dena ezagutuko dugu. Ta orixe da egia, nonbait. Garbi izateke garbia atzematerik ezpailitzateke zilegi izango». Simmia, nik uste, noraezean elkarren artean orrelakotxeak esan eta uste dituzte filosofilariek zuzen. Orrela ez aal deritzazu?

— Izan ere, beste edozer baino areago, Sokrate.

— Orrez gero, ori egia badadi, lagun, esan zuen Sokrate-k, nerau noan ara eltzeko gai denarentzat itxaropen andia izango litzateke, an nonbait, inon izatekotan, ioandako bizitzartean ara-

zita bera beragan ezartzera, ta aal den ginoan oraingotan eta gerokoan soilik bere gain bizi izatera, gorputzarekiko bilurretatik askatuta bezala?

— Erabat egia da, bada, esan zuen ark.

— Ez aal da orixe eriotza izentatzen alegia, gorputzarengandik gogoaren askatzea ta bereiztea?

— Edonola ere bai, esan zuen ark.

— Ta bera askatzea irrikatzen dute, beti batez ere ta soilik, zuzen filosofiginan ari direnek, gerok esaten dugunez, eta orixe da filosofilarien axola, gorputzarengandik gogoa askatzea ta bereiztea, alegia. Ez aal da?

— Agerian dago.

— Ez aal litzateke farragarri izango, asieran esaten nuenez, ilda gero izango den egoeran urbilenik bizi izaten bere burua bizitzartean ingiratzen ari den gizona, gertaera etor dakiokenean, sumintzea gero?

— Farragarri izango litzateke, bai orixe!

— Izan ere, bada, Simmia, esan zuen ark, zuzen filosofiginan ari direnak iltzera iartzen dira, ta iltzea, gizonen arteko inori baino bildurgarri-ezago gertatzen zaie. Begira ezazu, gero: nonnai gorputzarekin naasirik izango balira ta beste aldez, gogoa bere gain eukitzeko irrikaz izango balira, ta gertatu-ala, bildurtu ta suminduko balira, ez aal litzateke erabat ezpidezko izango? Ots, pozik ara ioango ezpalira, bertara elduta, bizitzartean barrena iritxi-maiteminez zirena arkitzeko itxaropena izaki, gogoetaren maitemina alegia, ta naasturik zeuzkakenagandik aldenduta

Digitizing: the process

- 1) Transform the **hard copy t**ext into txt: scanning and OCR reading
- 2) Process the txt: find **clusters** of the misread tokens (Word Smith Tools 8. Scott, Michael, 2021); **hyphens** are part of the token.
- 3) Extract the list of tokens, including the frequencies, to single out OCR mistakes from recurrent forms.
- 4) Organize the tokens **alphabetically**
  - The errors should be in the least frequent words,
  - the character sequences should cluster around the more frequent correctly read words.
- 5) Create the **concordances** per token and break the concordances alphabetically for each key token
  - We needed a wider context than single words to **disambiguate homo-form** cases.
  - Locate the reference in the .txt text,
  - Keep the scanned text also available to visually double-check the token.
- 7) Evaluate the speed of correction in a sample time, i.e., one hour of work.

Digitizing: the dictionaries

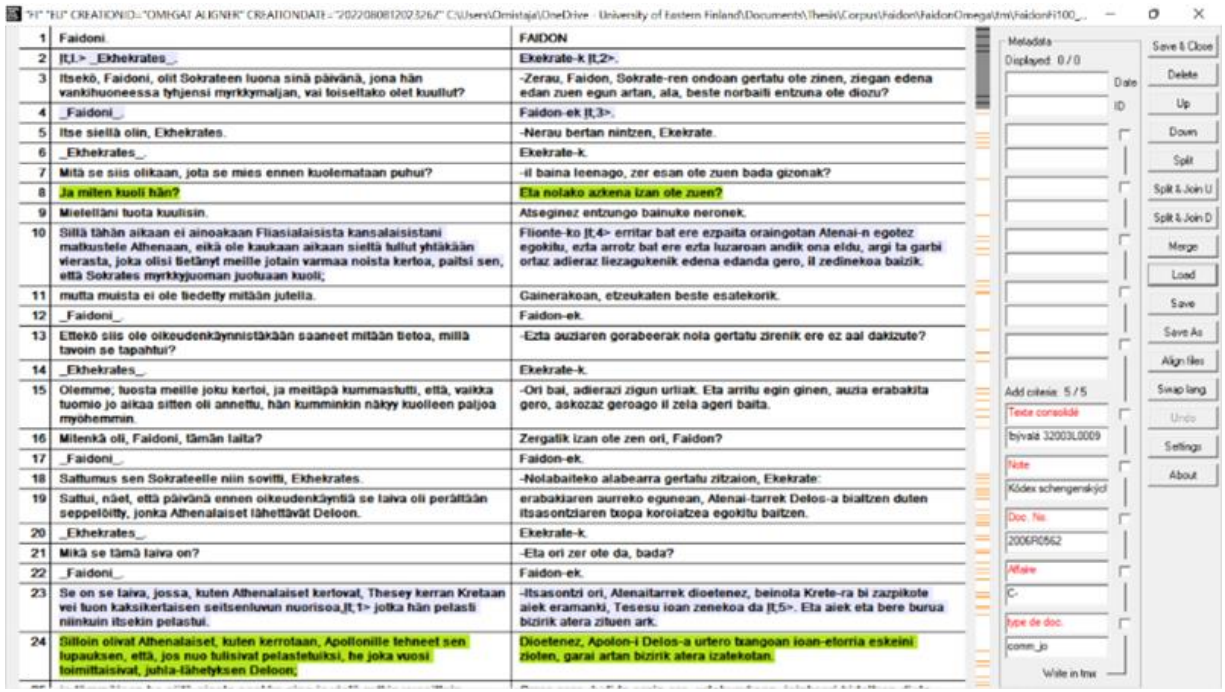
- Rationale:
  - Create a dictionary of **systematic** OCR **errors**
    - It could be used for same typeface with more texts.
- Outcome:
  - The **corrected original** text
  - A golden rule

AALDE	AAIDE
AALDEAK	AAIDEAK
AALDE-KLDE	AAIDE-KIDE
AALEGLN	AALEGIN
AALEGLNAK	AALEGINAK
AALEGLNKA	AALEGINKA
AAILK	AALIK
AALLK	AALIK
AAZTURLK	AAZTURIK
ADLERAZL-BEARREAN	ADIERAZI-BEARREAN
ALERBIDE	AGERBIDE
AGRTUKO	AGERTUKO
AGERTREN	AGERTZEN
ALN	AIN
ALNAKOA	AINAKOA
ALNBATEAN	AINBATEAN
ALNBESTE	AINBESTE

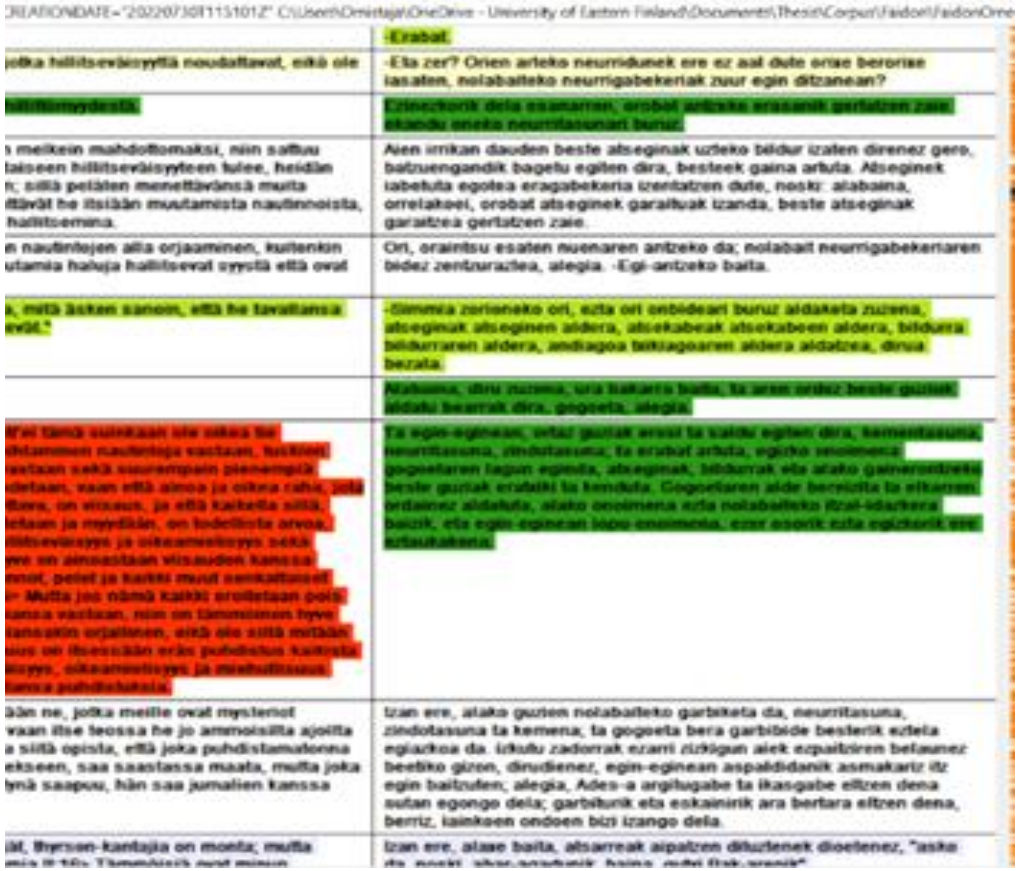
Alignment

- The leading **intuition**: to create ”fake” Translation Memories using CAT aligners (TM).
- The **problem**: the segmenting algorithms look at punctuation and string length
- The solution: manual aligning combining CAT tools, Aligners, and direct and inverse automatic translation
  - Content based aligning
- The question: is the product reusable in further texts?

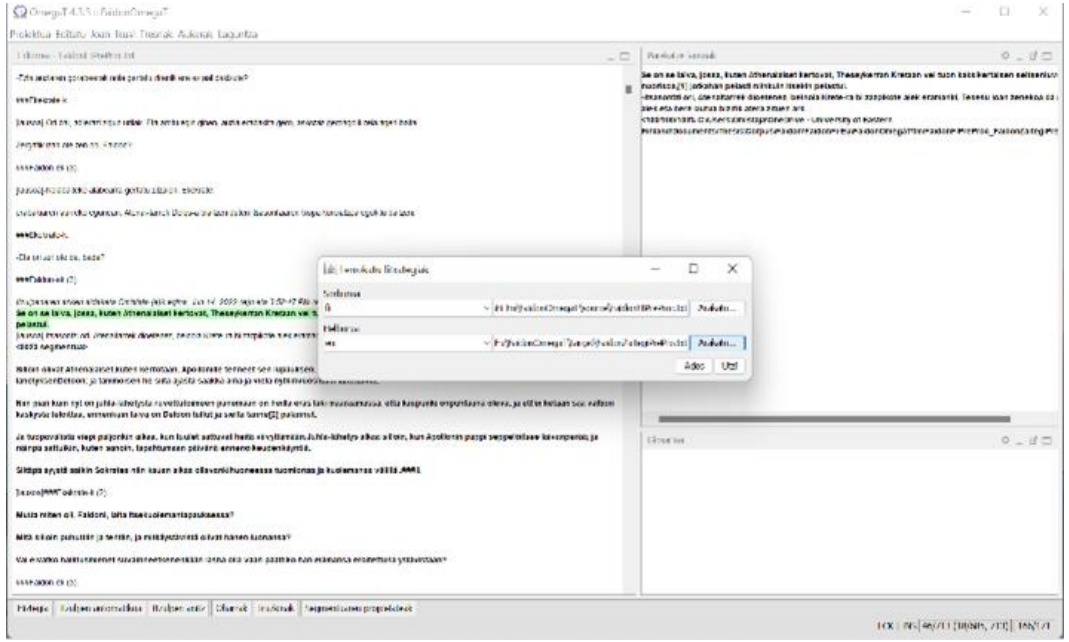
TMX Editor: the segmentation issue

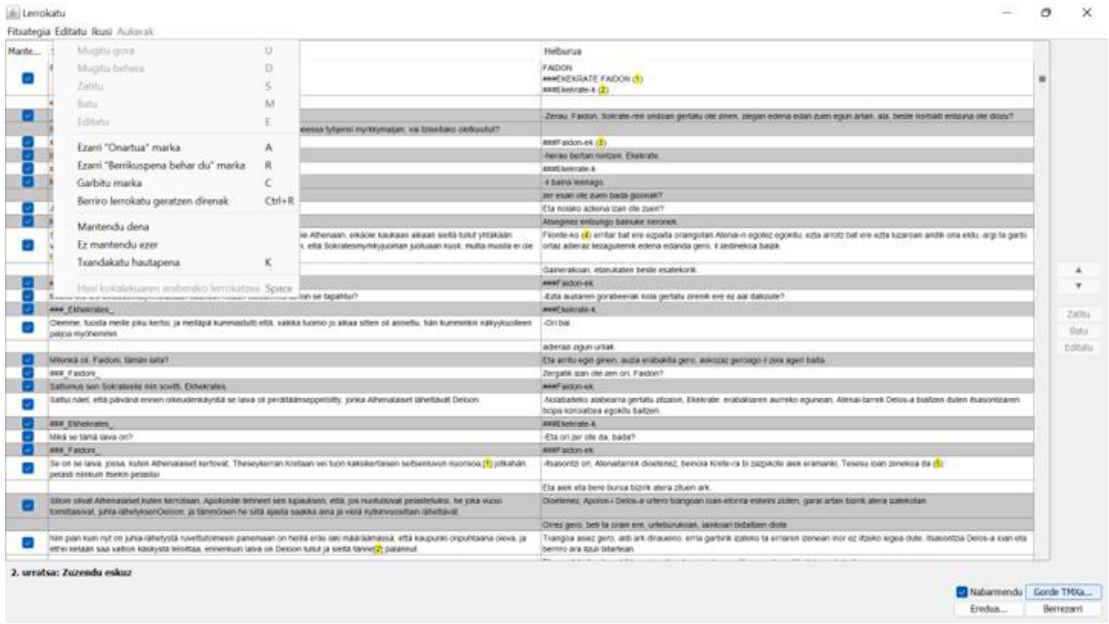


1	Faidoni.	FAIDON
2	‖t‖> _Ekhekrates_.	Ekekrate-k ‖t‖>.
3	Itsekō, Faidoni, olit Sokrateen luona sinä päivänä, jona hän vankihuoneessa tyhjensi myrkkymaljan, vai loiseltako olet kuullut?	-Zerau, Faidon, Sokrate-ren ondoan gertatu ote zinen, ziegan edena edan zuen egun artan, ala, beste norbaiti entzuna ote diozu?
4	_Faidoni_.	Faidon-ek ‖t‖>.
5	Itse siellä olin, Ekhekrates.	-Nerau bertan nintzen, Ekekrate.
6	_Ekhekrates_.	Ekekrate-k.
7	Mitä se siis olikaan, jota se mies ennen kuolemataan puhui?	-Il baina leenago, zer esan ote zuen bada gizonak?
8	Ja miten kuoli hän?	Eta nolako azkena izan ote zuen?
9	Mielelläni tuota kuulisin.	Atseginez entzungo bainuke neronek.
10	Sillä tähän aikaan ei ainoakaan Fliasialaisista kansalaisistani matkustele Athenaan, eikä ole kaukaan aikaan sieltä tullut yhtäkään vierasta, joka olisi tietänyt meille jotain varmaa noista kertoa, paitsi sen, että Sokrates myrkkyjuoman juotuaan kuoli;	Fliionte-ko ‖t.4> erritar bat ere ezpaita oraingotan Atenai-n egotez egokitu, ezta arrotz bat ere ezta luzaroan andik ona eldu, argi ta garbi ortaz adieraz liezagukenik edena edanda gero, il zedinekoa baizik.
11	mutta muista ei ole tiedetty mitään jutella.	Gainerakoan, etzeukaten beste esatekorik.
12	_Faidoni_.	Faidon-ek.
13	Ettekö siis ole oikeudenkäynnistäkään saaneet mitään tietoa, millä tavoin se tapahtui?	-Ezta auziaren gorabeerak nola gertatu zirenik ere ez aal dakizute?
14	_Ekhekrates_.	Ekekrate-k.
15	Olemme; tuosta meille joku kertoi, ja meitäpä kummastutti, että, vaikka tuomio jo aikaa sitten oli annettu, hän kumminkin näkyy kuolleen paljoa myöhemmin.	-Ori bai, adierazi zigun urliak. Eta arritu egin ginen, auzia erabakita gero, askozaz geroago il zela ageri baita.
16	Mitenkä oli, Faidoni, tämän laita?	Zergatik izan ote zen ori, Faidon?
17	_Faidoni_.	Faidon-ek.
18	Sattumus sen Sokrateelle niin sovitti, Ekhekrates.	-Nolabaiteko alabearra gertatu zitzaion, Ekekrate:
19	Sattui, näet, että päivänä ennen oikeudenkäyntiä se laiva oli perältään seppeliötty, jonka Athenalaiset lähettävät Deloon.	erabakiaren aurreko egunean, Atenai-tarrek Delos-a bialtzen duten itsasontziaren txopa koroiatzea egokitu baitzen.
20	_Ekhekrates_.	Ekekrate-k.
21	Mikä se tämä laiva on?	-Eta ori zer ote da, bada?
22	_Faidoni_.	Faidon-ek.
23	Se on se laiva, jossa, kuten Athenalaiset kertovat, Thesey kerran Kretaan vei tuon kaksikertaisen seitsenluvun nuorisoa ‖t.1> jotka hän pelasti niinkuin itsekin pelastui.	-Itsasontzi ori, Atenaitarrek dioetenez, beinola Krete-ra bi zazpikote aiek eramanki, Tesesu ioan zenekoa da ‖t.5>. Eta aiek eta bere burua bizirik atera zituen ark.
24	Silloin olivat Athenalaiset, kuten kerrotaan, Apollonille tehneet sen lupauksen, että, jos nuo tulisivat pelastetuiksi, he joka vuosi toimittaisivat, juhla-lähetyksen Deloon;	Dioetenez, Apolon-i Delos-a urtero txangoan ioan-etorria eskeini zioten, garai artan bizirik atera izatekotan.

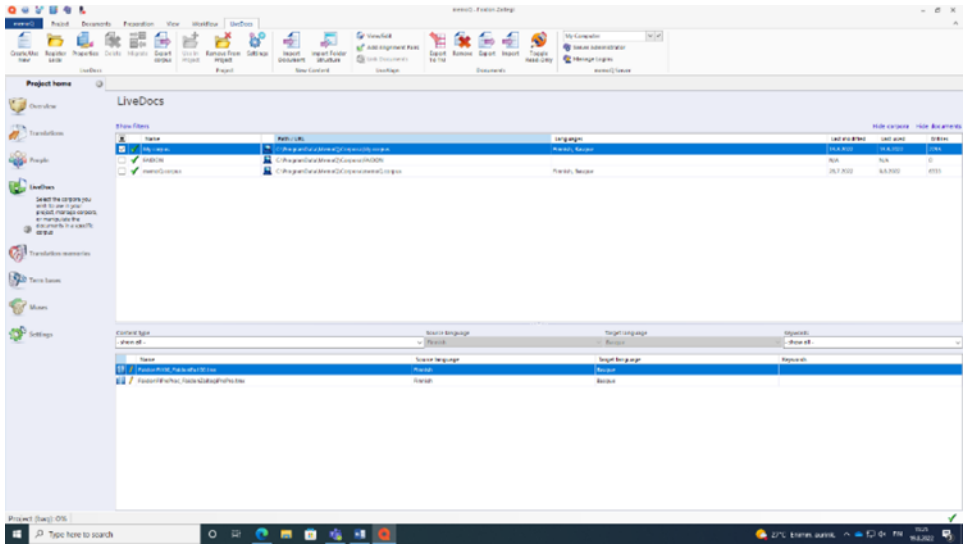


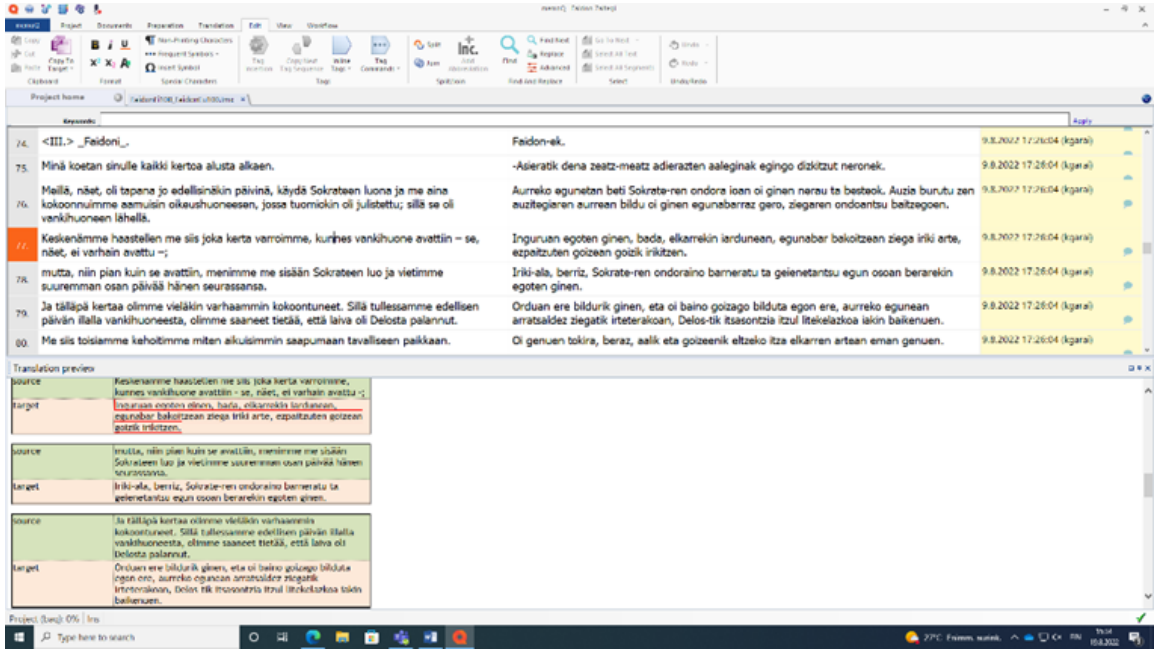
Fake Translation Memories with CAT systems: OmegaT



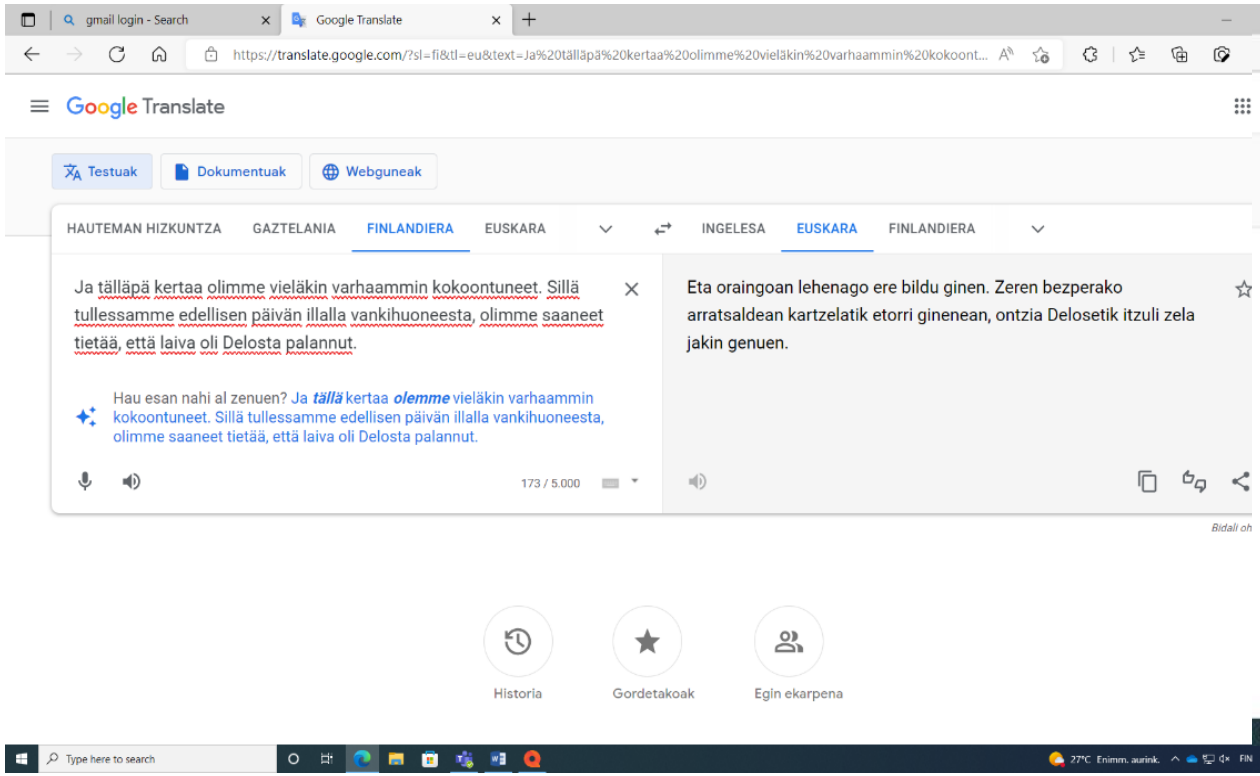


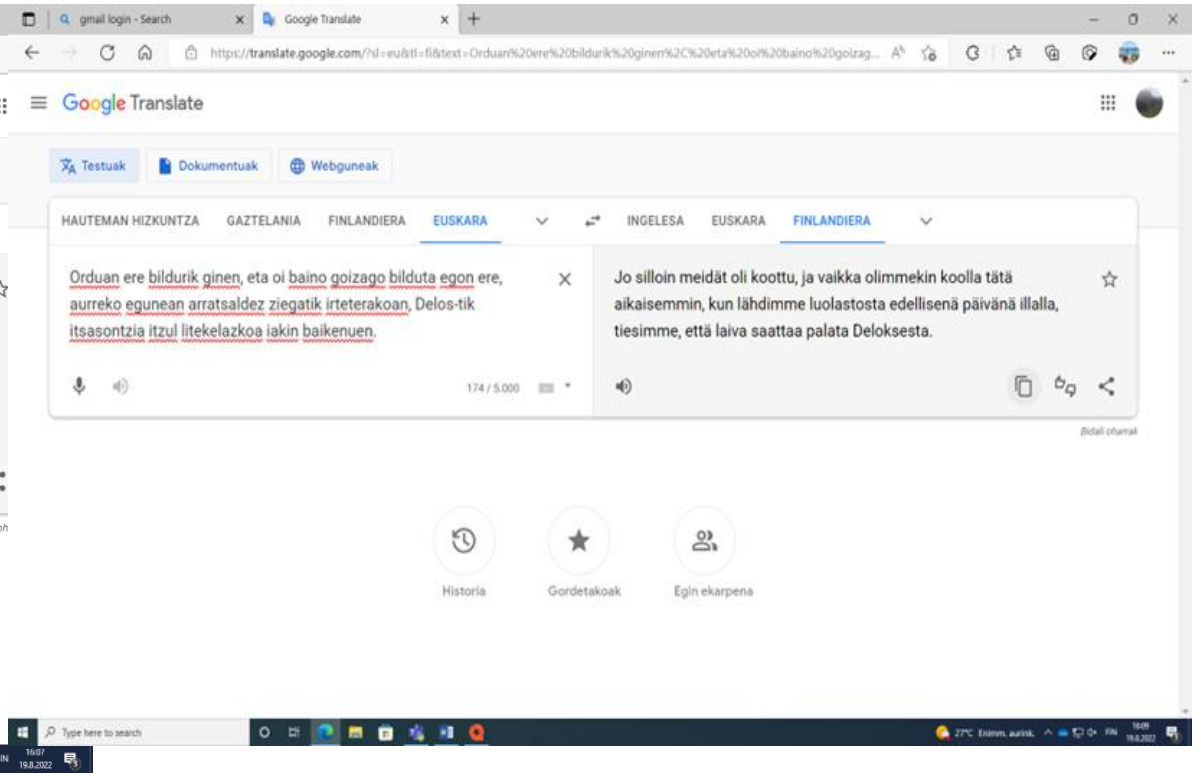
Recovery by Livedocs (memoQ)



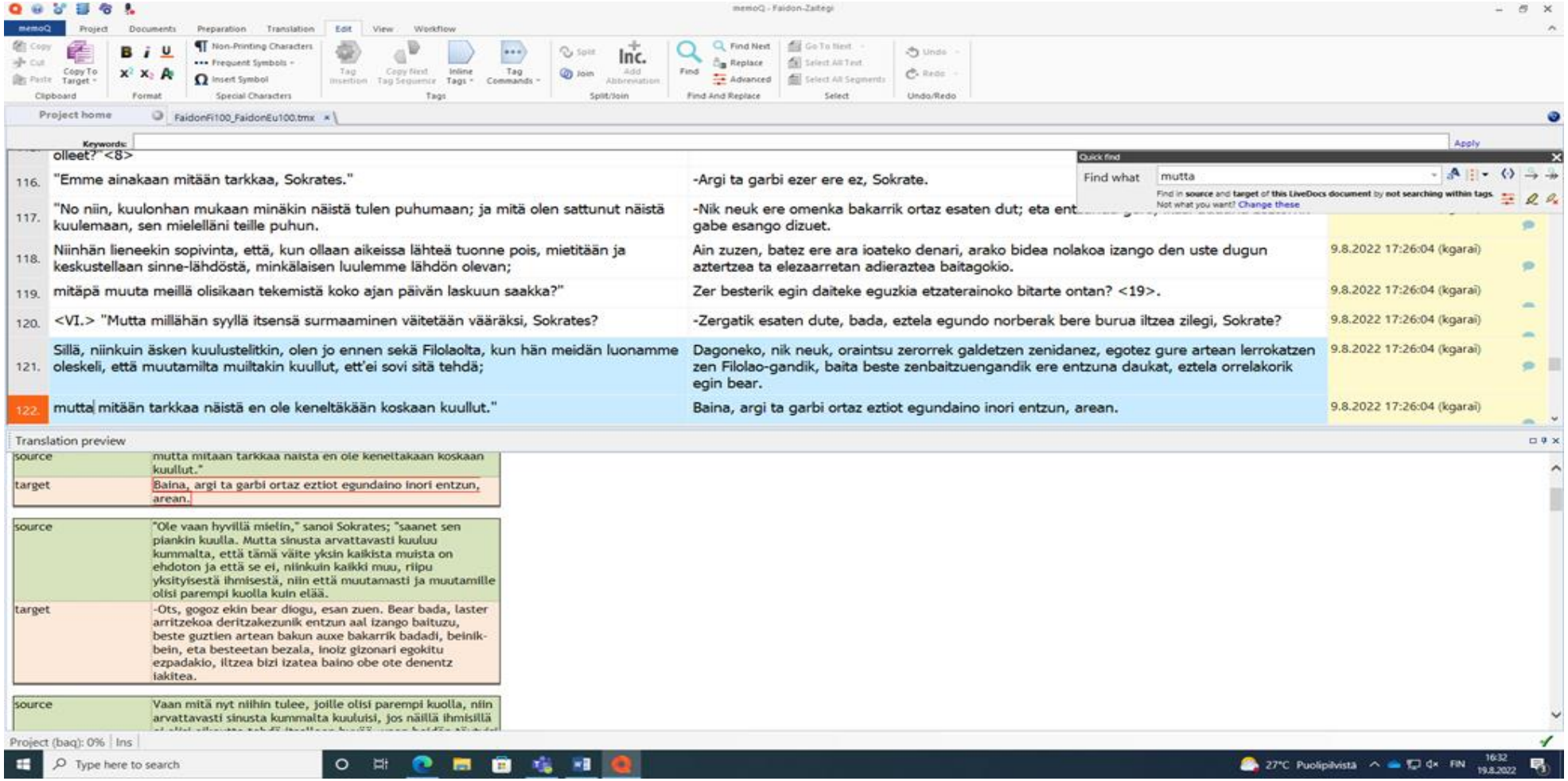


Content based manual alignment





Beyond punctuation and string length



Standardization

- The Basque text was produced before the adoption of a written standard.
- The NLPs work usually only with modern spelling
  - Using original (pre-standard) spelling brings annotation errors.
- The “standardizer” has two components:
  - 1) a dictionary of **non-lemmatized occurrences** built by hand
  - 2) the **language-independent program**: it substitutes the tokens in the input text by reading the dictionaries and creating a new output text.

Dictionaries (1)

- Problem: **how to discriminate and extract** those tokens that need to be standardized reliably and systematically
  - Are the spell-checkers enough?
- Procedure:
  - When the actualization is obvious, we add directly the standardized token in the chart, helped by the spelling corrector Xuxen (IXA group)
  - When there is a doubt check it in the Basque Language Academy dictionary (Euskaltzaindiaren Hiztegia, n.d.);
  - When the word does not appear in the Basque language academy, just “trust the text”.

Dictionaries

- The tokens are not lemmatized intentionally
  - For possible loses in grammatical information (multi word items, etc.)
  - What is **possible** vs. **probable** in language
    - Eventually the number of tokens will stabilize.
    - The substitution by rules can be prone to errors.

Word	WORD3
AAIDE	AHAIDE
AAIDEAK	AHAIDEAK
AAIDE-KIDE	AHAIDE-KIDE
AAL	AHAL
AALA	AHALA
AALBAITEKO	ALBAITEKO
AALEAN	AHALEAN
AALEGIN	AHALEGIN
AALEGINAK	AHALEGINAK
AALEGINEZ	AHALEGINEZ
AALEGINKA	AHALEGINKA
AALIK	AHALIK
AALTSUA	AHALTSUA
AALTSUAGORLK	AHALTSUAGORIK
AANTZ	AHANTZ
AANZTERAINO	AHANZTERAINO
AAZTEA	AHAZTEA
AAZTU	AHAZTU
AAZTURIK	AHAZTURIK
ABAR-AGADUNIK	ABAR-HAGADUNIK
ABERE-SORRETARA	ABERE SORRETARA
ADES	HADES
ADESA	HADESA
ADES-A	HADESA
ADES-AREN	HADESAREN
ADES-EKO	HADESEKO
ADES-EN	HADESEN
ADES-ERUNTZ	HADESERANTZ
ADES-KO	HADESKO
ADIERAZI-BEARREAN	ADIERAZI BEHARREAN
ADIMEN-GABEKERIA	ADIMEN GABEKERIA
ADIMEN-GABEKOA	ADIMEN GABEKOA
AGITZ	HAGITZ
AGORTU-EZINEKO	AGORTU EZINEKO
AIDEAK	AIREAK
AIEI	HAIEI
AIEK	HAIEK
AIEKIN	HAIEKIN
AIEN	HAIEN
AIENGAN	HAIENGAN
AIENGANDIK	HAIENGANDIK
AIGITO-N	EGIPTON

Sorting tokens by lenght

- To avoid errors:
  - Substitution could happen inside the token, if using rules

WORD2	**WORD3**	# occurs	Length
armoni-gabekoarekin	harmonia gabekoarekin	1	19
erakutsi-aaleginean	erakutsi ahaleginean	1	19
gorputz-inarkunetan	gorputz inarkunetan	1	19
biotz-pozgarritzat	bihotz pozgarritzat	1	18
eztirautenetaruntz	ez dirautenetarantz	1	18
iauretxe-lapurreta	jauretxe lapurreta	1	18
litzatekenearenera	litzatekeenearenera	1	18
adierazi-bearrean	adierazi beharrean	1	17
armoniagabearekin	harmonia gabearekin	1	17
bildurgarri-ezago	beldurgarri ezago	1	17
ta	eta	568	2
Ta	Eta	62	2
oi	ohi	23	2
il	**hil**	**22**	2
au	hau	21	2
an	**han**	**16**	2
io	jo	10	2
ar	har	9	2
An	**Han**	**3**	2
oe	ohe	2	2
as	has	1	2
eo	eho	1	2
Il	**Hil**	**1**	2
Io	Jo	1	2
Oi	Ohi	1	2

The ”inclusive standardizer”

- The program should perform these tasks:
  - 1- To sort the extracted list of tokens of the dictionary by decreasing number of characters, from the largest word (e.g. “zuzengabeagoarengandik”) to the shortest (e.g. “jo”).
  - 2-To find the token of the left column of the dictionary in the text and substitute it with the correspondent token of the column in the right.
  - 3- To repeat this process until no more instances of the left token are found in the plain text.

The program

- The steps the program should take are these:
  - Open the CSV file and read it into a dictionary where the keys are the original words, and the values are the standard words.
  - Open the text file and read it into a string.
  - Loop through the dictionary and use the replace method to replace all instances of the original words with the standard words.
  - Write the corrected text back to the file.

Universal Dependencies: practical analysis

- Askatze honen aurkakorik ez duela egin behar deritza benetan filosofo denaren gogoak.

- Aska-tze[verb nominalization] / hau-EN[gen] / aurka-ko[gen-locative]-rik[partitive] / ez / du[Aux]-eLa[completive-object] / egin / behar / d-eritz-a [3 argument defective synthetic verb] / benetan / filosofo / da[Aux]-EN[relative]-a[det]-rEN[gen] / gogo-a[det]-K[ergative].

- Freeing / this-of[gen] / against-of-(ellipsis)-“any”[part] / not / has-that / do / must / be-lieves / truly / philosopher / is-that[relative]-the-of[gen] / mind-the[erg]

- The mind of that who is a true philosopher believes that it must not do (anything) against this freeing.

Stanza

Part-of-Speech (XPOS):

XPOS is not available for this language at this time.

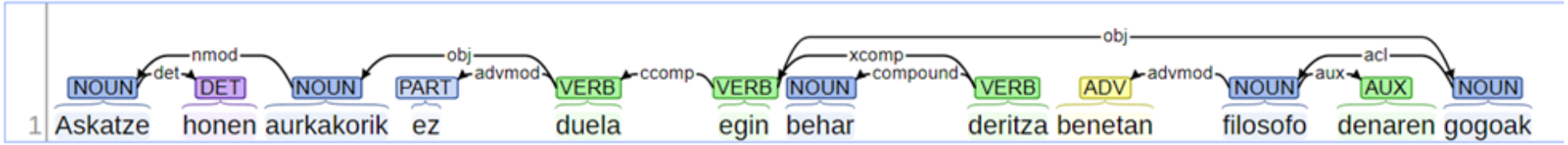
Lemmas:

	askatze	hau	aurka	ez	ukan	egin	behar	iritzi	benetan	filosofo	izan	gogo
1	Askatze	honen	aurkakorik	ez	duela	egin	behar	deritza	benetan	filosofo	denaren	gogoak

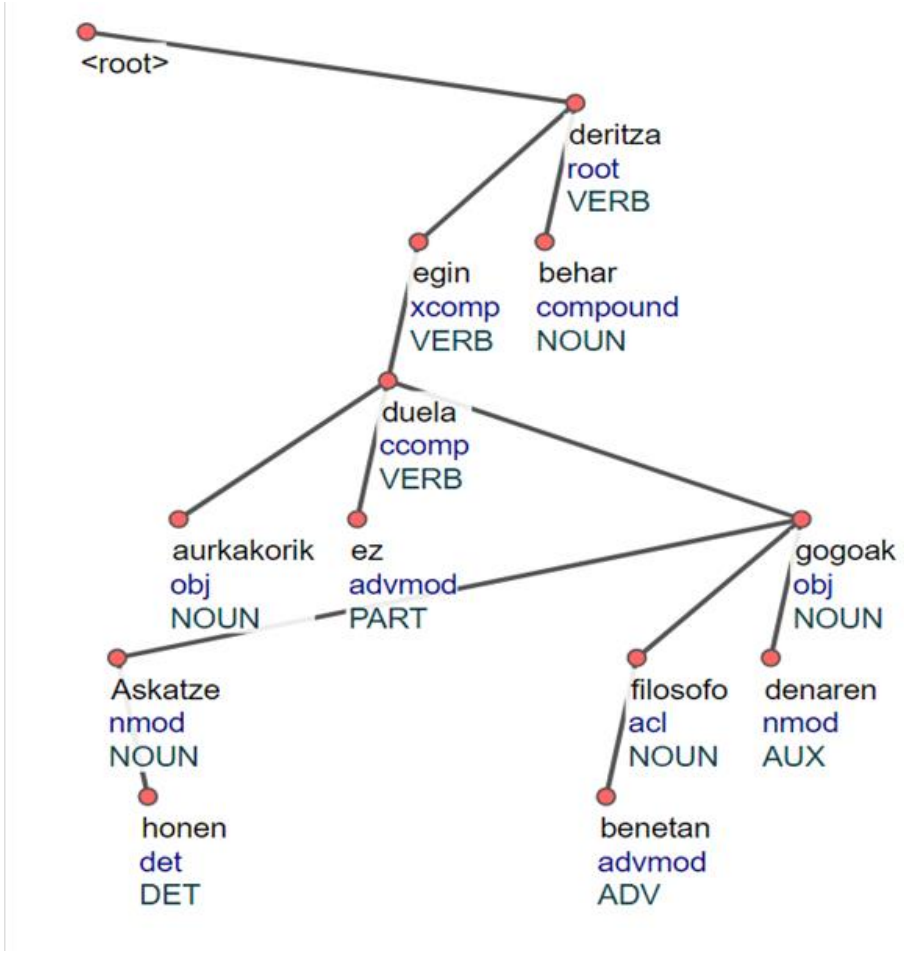
Named Entity Recognition:

NER is not available for this language at this time.

Universal Dependencies:



UDPipe Lindat



UDPipe/MaltIxa comparison. Lemmas and UPoS

Id	Form	Lemma1	Lemma2	UPosTag	Cate-gory	Subcate-gory
1	Askatze	askatze	askatu	NOUN	ADI	ADI_SIN
2	honen	hau	hau	DET	DET	DET_ERKARR
3	aurka-korik	aurkako	aurkako	NOUN	IZE	IZE_ARR
4	ez	ez	ez	PART	PRT	PRT
5	duela	ukan	ukan	VERB	ADT	ADT
6	egin	egin	egin	VERB	ADI	ADI_SIN
7	behar	behar	behar	NOUN	IZE	IZE_ARR
8	deritza	iritzi	iritzi	VERB	ADT	ADT
9	benetan	benetan	benetan	ADV	ADB	ADB_ARR
10	filosofo	filosofo	filosofo	NOUN	IZE	IZE_ARR
11	denaren	izan	dena	AUX	DET	DET_ORO
12	gogoak	gogo	gogo	NOUN	IZE	IZE_ARR
13	.	.	.	PUNCT	PUNT	PUNT_PUNT

UDPipes/MaltIxa Features comparison

Id	Form	Head1	Head2	DepRel1	Relation 2	Feats1	Feats2
1	Askatze	3	5	nmod	xmod	_	ADM:ADIZE
2	honen	1	3	det	ncmod	Case=Gen|Definite=Def|Number=Sing	KAS:GEN|NUM:S
3	aurkakorik	5	5	obj	ncobj	Case=Par|Definite=Ind	KAS:PAR
4	ez	5	5	advmod	ncmod	Polarity=Neg	_
5	duela	6	0	ccomp	ROOT	Aspect=Prog|Mood=Ind|Number[abs]=Sing|Number[erg]=Sing|Person[abs]=3|Person[erg]=3|VerbForm=Fin	ERL:KONPL|ASP:PNT
6	egin	8	7	xcomp	xmod	VerbForm=Part	ADM:PART
7	behar	8	5	compound	ncobj	Case=Abs|Definite=Ind	KAS:ABS
8	deritza	0	7	root	mw	Aspect=Prog|Mood=Ind|Number[abs]=Plur|Number[erg]=Sing|Person[abs]=3|Person[erg]=3|VerbForm=Fin	ASP:PNT
9	benetan	10	12	advmod	ncmod	_	_
10	filosofo	12	12	acl	ncmod	Animacy=Inan|Case=Abs|Definite=Ind	_
11	denaren	10	10	cop	detmod	Aspect=Prog|Case=Gen|Definite=Def|Mood=Ind|Number=Sing|Number[abs]=Sing|Person[abs]=3|VerbForm=Fin	KAS:GEN|NUM:S
12	gogoak	6	5	nsubj	ncobj	Animacy=Inan|Case=Abs|Definite=Def|Number=Plur	KAS:ABS|NUM:P
13	.	8	12	punct	PUNT		

Conclusions

- Outcomes
  - Dictionaries:
    - Systematized OCR errors with their corresponding corrections
    - Standardization dictionaries for that period
  - Original text in machine readable format.
  - Aligned texts Finnish-(original) Basque
  - Aligned texts Original vs Standard Basque
  - Automatically annotated texts, Finnish and Basque

.. And Questions:

- If hand-curation is expensive,
  - to which extend is this process “an easy path” to enrich less-resourced languages?
  - To which extend are the content based aligned segments reusable in other similar future texts of the same domain?

References

- Aranzabe, M. J., Atutxa, A., Bengoetxea, K., Diaz, A., de Ilarraza, Goenaga, I., Gojenola, K., & Uria, L. (2015). Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies. *14th International Workshop on Treebanks and Linguistic Theories*, 9.
- Bengoetxea, K., & Gojenola, K. (2010). Application of different techniques to dependency parsing of Basque. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, 31–39.
- Estarrona, A., Etxeberria, I., Etxepare, R., Padilla-Moyano, M., & Soraluze, A. (2020). Dealing with dialectal variation in the construction of the Basque historical corpus. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 79–89. https://aclanthology.org/2020.vardial-1.8
- Estarrona, A., Etxeberria, I., Soraluze, A., Etxepare, R., & Padilla-Moyano, M. (2021). The first annotated corpus of historical Basque. *Digital Scholarship in the Humanities*, fqab066. https://doi.org/10.1093/llc/fqab066
- *Euskaltzaindiaren Hiztegia—Bilaketa*. (n.d.). Retrieved May 21, 2023, from https://www.euskaltzaindia.eus/index.php?option=com_hiztegianbilatu&task=hasiera&Itemid=1693&lang=eu-ES
- *EUSTAGGER: Lemmatizer/tagger for Basque*. (2022, October 25). https://ixa2.si.ehu.eus/eustagger/
- Gaspari, F., Grützner-Zahn, A., Rehm, G., Gallagher, O., Giagkou, M., Piperidis, S., & Way, A. (2023). Digital Language Equality: Definition, Metric, Dashboard. In *European Language Equality* (pp. 39–73). Springer International Publishing. https://doi.org/10.1007/978-3-031-28819-7_3
- Hovy, E. (2022). Chapter 21. Corpus Annotation. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 508–517). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.013.011
- IXA Taldea. (2004). *Ixa taldeko etiketen eskuliburua*. https://garaterm.ehu.es/sites/default/files/dokumentuak/4083/Etiketen_eskuliburua_IXA.pdf
- Lapshinova-Koltunski, E., Popović, M., & Koponen, M. (2022). DiHuTra: A Parallel Corpus to Analyse Differences between Human Translations. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 337–338. https://aclanthology.org/2022.eamt-1.58

References

- *LF Aligner*. (2023, May 7). SourceForge. https://sourceforge.net/projects/aligner/
- Li, B., Gaussier, E., & Yang, D. (2018). Measuring bilingual corpus comparability. *Natural Language Engineering*, *24*(4), 523–549. https://doi.org/10.1017/s1351324917000481
- Lindén, K., & Dyster, W. (2023). Language Report Finnish. In G. Rehm & A. Way (Eds.), *European Language Equality: A Strategic Agenda for Digital Language Equality* (pp. 135–138). Springer International Publishing. https://doi.org/10.1007/978-3-031-28819-7_15
- *LiveDocs—memoQ*. (2024, November 15). Translation Software - memoQ. https://www.memoq.com/tools/livedocs
- *Maltixa | Ixa taldea*. (n.d.). Retrieved June 23, 2022, from https://ixa.ehu.eus/node/4457?language=en
- McEnery, T. (2022). Corpora. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 494–507). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.013.47_update_001
- McEnery, T., & Xiao, R. (2018). *Parallel and Comparable Corpora: What is Happening?* (pp. 18–31). Multilingual Matters. https://doi.org/10.21832/9781853599873-005
- memoQ (Director). (2019, September 18). *Hidden Treasures in memoQ* [Video recording]. https://www.youtube.com/watch?v=OjwGEppw92M
- *META-NET White Paper Series—META Multilingual Europe Technology Alliance*. (n.d.). Retrieved October 24, 2023, from http://www.meta-net.eu/whitepapers/overview
- Moseley, C., Nicolas, A., & Unesco. (2010). *Atlas of the world's languages in danger* (3rd ed., entirely rev., enl.updated, p. 1 atlas (154, 62 pages) : 29 color maps; 20 x 29 cm + 1 map (86 x 119 cm folded to 29 x 18 cm).) [Map]. Unesco Pub. Paris.

References

- Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). *Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection* (No. arXiv:2004.10643). arXiv. https://doi.org/10.48550/arXiv.2004.10643
- *OmegaT - The Free Translation Memory Tool—OmegaT*. (n.d.). OmegaT - The Free Translation Memory Tool. Retrieved August 19, 2022, from http://omegat.org/
- Pace-Sigge, M. T. L. (2013). The concept of Lexical Priming in the context of language use. *ICAME*, *37*, 149–173.
- Plato. (1978). *Platon. IV., Kriton eta Faidon* (I. Zaitegi Plazaola, Trans.). Euskaltzaindia.
- Plato, 427? BCE-347? BCE. (2006). *Faidoni Platonin keskustelma Sokrateen viimeisistä hetkistä jasielun kuolemattomuudesta* (J. W. (Johan W. Calamnius, Trans.). https://www.gutenberg.org/ebooks/19210
- *Project Gutenberg*. (n.d.). Project Gutenberg. Retrieved March 14, 2022, from https://www.gutenberg.org/
- Rehm, G., & Way, A. (Eds.). (2023a). *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer International Publishing. https://doi.org/10.1007/978-3-031-28819-7
- Rehm, G., & Way, A. (2023b). European Language Equality: Introduction. In G. Rehm & A. Way (Eds.), *European Language Equality: A Strategic Agenda for Digital Language Equality* (pp. 1–10). Springer International Publishing. https://doi.org/10.1007/978-3-031-28819-7_1
- Salminen, T. (2010). Europe and the Caucasus. In *Atlas of the world's languages in danger*. UNESCO. https://researchportal.helsinki.fi/en/publications/europe-and-the-caucasus
- Sarasola, K., Aldabe, I., Ilarraza, A. D. de, Estarrona, A., Farwell, A., Hernáez, I., & Navas, E. (2023). Language Report Basque. In G. Rehm & A. Way (Eds.), *European Language Equality: A Strategic Agenda for Digital Language Equality* (pp. 95–98). Springer International Publishing. https://doi.org/10.1007/978-3-031-28819-7_5

References

- Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education / Mike Scott and Christopher Tribble* (Issue v. 22). John Benjamins Publishing Co.
- Scott, Michael. (2021). *WordSmith Tools* (Version 8.0.0.61) [Windows].
- *TMX editor*. (2015, September 16). SourceForge. https://sourceforge.net/projects/tmxeditor/
- *UDPipe*. (2022). [C++]. ÚFAL. https://github.com/ufal/udpipe (Original work published 2016)
- *UNESCO WAL*. (n.d.). UNESCO WAL. Retrieved October 21, 2023, from https://en.wal.unesco.org/
- *Xuxen | Ixa taldea*. (n.d.). Retrieved May 30, 2022, from http://ixa.ehu.eus/node/4463?language=en
- Zaitegi, Iokin. (1974). Platon Euskeratzen. *Euskera: Euskaltzaindiaren lan eta agiriak = Trabajos y actas de la Real Academia de la Lengua Vasca = Travaux et actes de l'Academie de la Langue basque*, *XIX*(1), 358–361.
- Zaitegi Plazaola, I. 1906-1979., Lafitte, P. 1901-1983., & Zaitegi Plazaola, Iokin. (1962). *Platon'eneko atarian*. [s.n.].
- Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., & Hajič, J. (2014). HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, *48*(4), 601–637. https://doi.org/10.1007/s10579-014-9275-2
- Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., Ajede, C. K., Akkurt, S. F., Aleksandravičiūtė, G., Alfina, I., Algom, A., Alnajjar, K., Alzetta, C., Andersen, E., Antonsen, L., Aoyama, T., … Ziane, R. (2023). Universal Dependencies 2.12. *http://universaldependencies.org/*. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5150