RETOUR D'EXPÉRIENCE SUR LE PROJET DORECO (LANGUAGE DOCUMENTATION REFERENCE CORPUS)

Matthew STAVE[1], Ludger PASCHEN[2], François DELAFONTAINE[3], Frank SEIFART[2] & François PELLEGRINO[1]

(1) Laboratoire Dynamique Du Langage, UMR5596 – CNRS, Université de Lyon, France
(2) Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), Berlin, Allemagne
(3) Université de Fribourg, Suisse

glose2023 : June 28, 2023 – Paris, France
Corpus Glosés: de la construction à l'exploitation automatique

Overview

- The DoReCo project in a nutshell
  - ✓ Why, Who, Where, What?
  - ✓ Key figures and illustrations

- Focus on Glosses/Annotations
  - ✓ The alignment / reinjection process
  - ✓ Consistency issues
  - ✓ A bird’s eye view across languages

DoReCo in a Nutshell: Why

- To describe human language…
  - ✓ Necessary to study naturalistic language data from a wide sample of languages
  - ✓ Not just the WEIRD ones (Blasi, *et al.*, 2022; Henrich, Heine & Norenzayan, 2010)
- Language documentation projects have accumulated highly valuable data for decades: let's gather them in a common framework as FAIR as possible

  → Corpora created by experts (incl. Martine ☺)
  who'd worked on the languages in collaboration with language communities
- Language selection aimed at providing a diverse sample from all continents
- Same spirit as Multi-CAST: eight languages in both (Haig & Schnell, 2022)

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D. & Majid, A. (2022). Over-reliance on English hinders cognitive science. *TiCS*.

Haig, G. & Schnell, S. (eds.). (2022). *Multi-CAST: Multilingual corpus of annotated spoken texts*. Bamberg: University of Bamberg https://multicast.aspra.uni-bamberg.de/

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). "Most people are not WEIRD". *Nature*.

Wilkinson, M. D., et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data*.

https://doreco.huma-num.fr/

DoReCo in a Nutshell: Who

- Who
  - ✓ PIs: Frank Seifart (ZAS, Berlin) & François Pellegrino (DDL, Lyon)
  - ✓ Postdocs
    - Ludger Paschen (+ Florian Schiel, BAS Munich)
    - Matt Stave (+ François Delafontaine, then-DDL, now University of Fribourg)
  - ✓ 20 research assistants and interns
  - ✓ ~100 corpus creators
    - Data collection field work, expert analyses and annotation
    - Answered questions during DoReCo data processing
    - Made data available in open access
  - ✓ Sébastien Flavier (DDL, CNRS): Publication on Huma-Num

- Subsidies
  - ✓ Main funding: ANR-DFG, 2019-2022
  - ✓ Additional funds: LabEx ASLAN + synergy with F. Seifart's other projects

https://doreco.huma-num.fr/

DoReCo in a Nutshell: Where

- Available since summer of 2022
  - ✓ Creative Commons CC-BY license ➔ “as open as possible, as closed as necessary”
  - ✓ Additional restrictions (NC, SA, ND) may apply to comply with the ethical aspects agreed on with the speakers community (decided by the corpus creator(s))

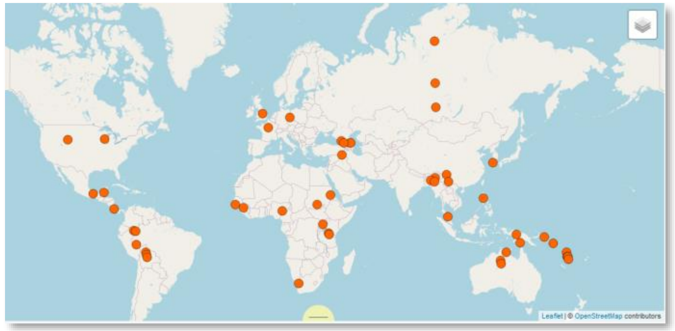
- Hosted on Huma-Num (French public infrastructure for data in H&SS)
  - ✓ All annotations files hosted on Nakala and accessible through the website
  - ✓ Audio files on Nakala for most languages (external repositories for 6 languages)
  - ✓ Each dataset identified by its unique DOI
  - ✓ Dataset = Publication authored by the corpus creator(s)
    - ➔ We insist that the corpus creators’ authorship is recognized by including full bibliographical citations for each DoReCo dataset

  - ✓ Additionally, several tools available on GitHub (https://github.com/DoReCo)

https://doreco.huma-num.fr/

ON S'ÉNERVE PAS!
ON Y VA! ON Y VA!
LET'S GO!
ON SE CALME!
PAS DE PANIQUE!
Gotlib

DoReCo in a Nutshell: What



- Coverage
  - ✓ Natural speech (mostly narrative)
  - ✓ 51 languages from 32 linguistic families/isolates
  - ✓ Mostly fieldwork-based documentation (small/endangered languages)

https://doreco.huma-num.fr/

DoReCo in a Nutshell: What

Leaflet | © OpenStreetMap contributors

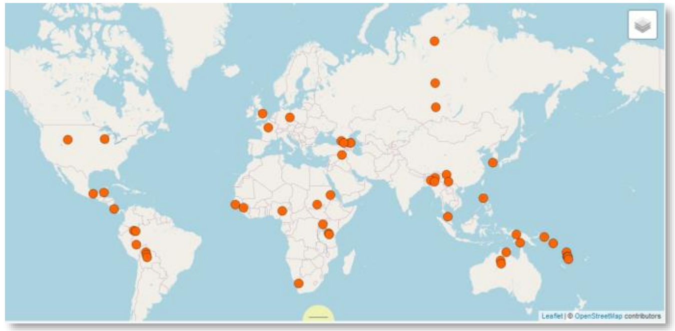
- Coverage
  - ✓ Natural speech (mostly narrative)
  - ✓ 51 languages from 32 linguistic families/isolates
  - ✓ Mostly fieldwork-based documentation (small/endangered languages)
- Time-alignment (two-pass)
  - ✓ Manually corrected phonemic time-alignment (Berlin)
  - ✓ MAUS alignment tool (Munich)
- 30 (+8 partially) languages with morphological annotation
  - ✓ Morpheme breaks, glosses, and often part-of-speech tags
  - ✓ Standardization, documentation, and re-alignment (Lyon)

https://doreco.huma-num.fr/

DoReCo in a Nutshell: What

- Coverage
  - ✓ Natural speech (mostly narrative)
  - ✓ 51 languages from 32 linguistic families/isolates
  - ✓ Mostly fieldwork-based documentation (small/endangered languages)
- Time-alignment (two-pass)
  - ✓ Manually corrected phonemic time-alignment (Berlin)
  - ✓ MAUS alignment tool (Munich)
- 30 (+8 partially) languages with morphological annotation
  - ✓ Morpheme breaks, glosses, and often part-of-speech tags
  - ✓ Standardization, documentation, and re-alignment (Lyon)

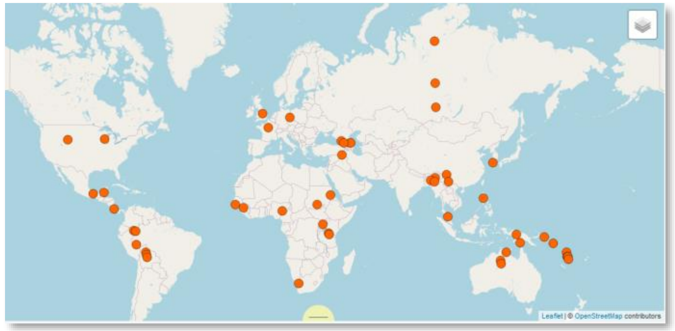
https://doreco.huma-num.fr/



Credits: Giovanni Handal, wikimedia

DoReCo in a Nutshell: What

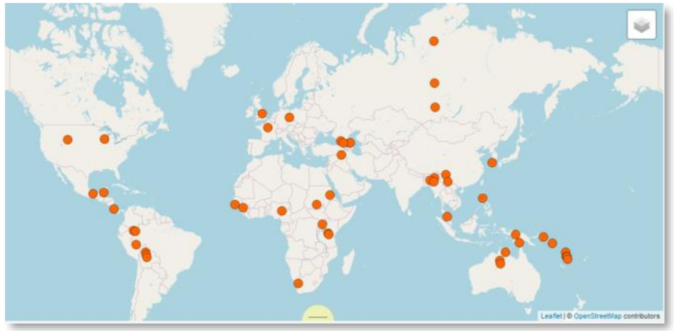
- Coverage
  - ✓ Natural speech (mostly narrative)
  - ✓ 51 languages from 32 linguistic families/isolates
  - ✓ Mostly fieldwork-based documentation (small/endangered languages)
- Time-alignment (two-pass)
  - ✓ Manually corrected phonemic time-alignment (Berlin)
  - ✓ MAUS alignment tool (Munich)
- 30 (+8 partially) languages with morphological annotation
  - ✓ Morpheme breaks, glosses, and often part-of-speech tags
  - ✓ Standardization, documentation, and re-alignment (Lyon)
- Each dataset includes
  - ✓ Speech files (or link to speech files if externally archived)
  - ✓ Elan, Praat, and xml annotations files
  - ✓ Two csv recap files (phoneme and word levels)
  - ✓ Metadata



Credits: Giovanni Handal, wikimedia

https://doreco.huma-num.fr/

Main Figures



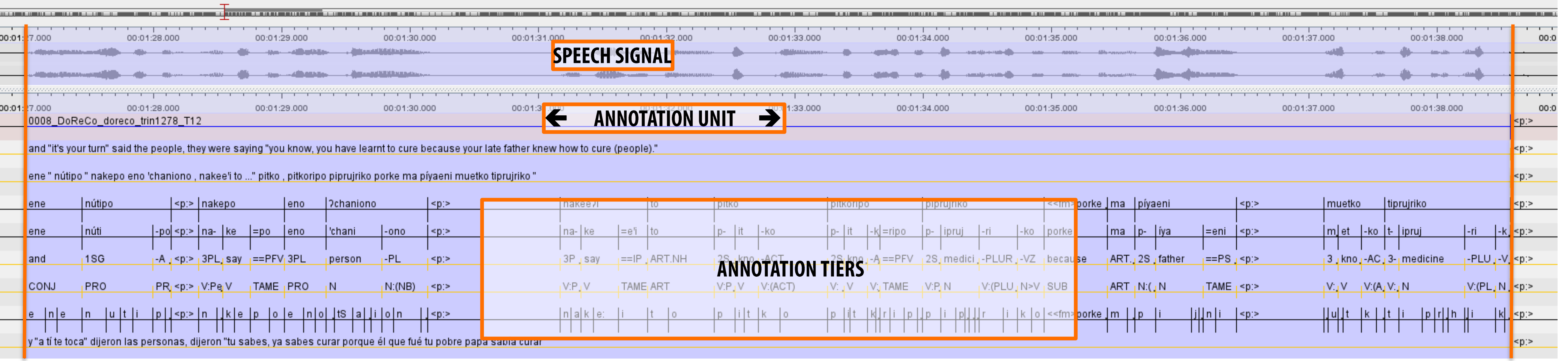
- CORE set (aka time-aligned dataset)
  - ✓ ~112 hours of recordings (~96 hours of actual speech) in 51 languages
  - ✓ 1.9 M syllables; 969,000 “words” in 51 languages (approximately and arguably)
  - ✓ 1.0 M morphs in 38 languages

- EXTENDED set (same language and source but without time alignment)
  - ✓ ~770,000 words
  - ✓ Useful for linguistic analyses based on transcription and analysis
  - ✓ Available for NLP development (fine-tuning, etc.)

LG_CODE	Morph Count	Syllable Count	Word Count	Raw Duration	Speech Duration
anal1239	-	49,650	27,234	173	139
apah1238	20,088	34,768	16,215	80	78
arap1274	18,596	36,080	9,588	169	127
bain1259	40,694	54,524	22,904	159	131
beja1238	62,214	66,428	30,664	218	218
bora1263	38,091	57,323	17,584	196	147
cabe1245	27,526	45,304	22,688	133	112
cash1254	30,626	56,388	20,756	175	137
dolg1241	32,788	42,920	18,102	152	127
even1259	41,582	52,222	18,884	230	166
goem1240	20,672	23,504	18,812	107	74
goro1270	29,372	42,250	21,598	109	99
hoch1243	24,982	46,958	15,662	205	133
jeha1242	11,404	23,356	14,418	105	88
jeju1234	25,444	31,686	14,478	97	85
kaka1265	22,042	27,714	19,452	106	75
kama1351	12,206	16,678	7,568	87	71
kark1256	-	46,628	18,338	135	105
komn1238	23,880	45,506	20,576	137	134
ligh1234	-	20,264	17,888	116	102
lowe1385	-	33,512	21,032	155	129
movi1243	29,488	46,220	20,834	160	143
ngal1292	8,354	19,076	7,038	68	64
nisv1234	29,646	42,456	21,576	118	114
nngg1234	22,328	27,246	20,080	96	88
nort2641	28,702	35,944	19,654	105	87
nort2641	28,702	35,944	19,654	105	87
nort2875	22,772	34,720	17,552	125	115
orko1234	17,736	32,986	20,546	107	99
pnar1238	22,190	27,208	17,742	108	78
port1286	22,692	36,210	22,952	116	100
resi1247	-	38,638	13,672	184	113
ruul1235	38,698	51,284	17,942	138	127
sadu1234	-	28,540	23,544	95	88
sanz1248	14,840	24,871	10,908	99	78
savo1255	23,954	39,562	18,546	120	89
sout2856	20,854	30,194	15,326	129	81
sout3282	19,610	-	18,116	91	75
stan1290	-	-	26,690	110	101
sumi1235	30,838	36,924	21,700	87	84
svan1243	-	42,266	20,024	170	138
taba1259	19,118	21,820	10,556	78	64
teop1238	28,714	46,968	23,838	122	111
texi1237	35,876	38,460	21,898	133	113
trin1278	33,140	44,290	16,018	186	145
tsim1256	-	-	9,511	123	62
urum1249	42,200	59,510	23,296	235	175
vera1241	30,336	38,366	25,016	119	99
warl1254	-	43,730	13,984	154	122
yong1270	-	16,936	9,474	82	68
yuca1254	-	34,828	21,304	119	106
yura1255	-	122,436	45,106	356	341

LG_CODE	Morph Count	Syllable Count	Word Count	Raw Duration	Speech Duration
anal1239	-	49,650	27,234	173	139
apah1238	20,088	34,768	16,215	80	78
arap1274	18,596	36,080	9,588	169	127
bain1259	40,694	54,524	22,904	159	131
beja1238	62,214	66,428	30,664	218	218
bora1263	38,091	57,323	17,584	196	147
cabe1245	27,526	45,304	22,688	133	112
cash1254	30,626	56,388	20,756	175	137
dolg1241	32,788	42,920	18,102	152	127
even1259	41,582	52,222	18,884	230	166
goem1240	20,672	23,504	18,812	107	74
goro1270	29,372	42,250	21,598	109	99
hoch1243	24,982	46,958	15,662	205	133
jeha1242	11,404	23,356	14,418	105	88
jeju1234	25,444	31,686	14,478	97	85
kaka1265	22,042	27,714	19,452	106	75
kama1351	12,206	16,678	7,568	87	71
kark1256	-	46,628	18,338	135	105
komn1238	23,880	45,506	20,576	137	134
ligh1234	-	20,264	17,888	116	102
lowe1385	-	33,512	21,032	155	129
movi1243	29,488	46,220	20,834	160	143
ngal1292	8,354	19,076	7,038	68	64
nisv1234	29,646	42,456	21,576	118	114
nngg1234	22,328	27,246	20,080	96	88
nort2641	28,702	35,944	19,654	105	87

LG_CODE	Morph Count	Syllable Count	Word Count	Raw Duration	Speech Duration
nort2641	28,702	35,944	19,654	105	87
nort2875	22,772	34,720	17,552	125	115
orko1234	17,736	32,986	20,546	107	99
pnar1238	22,190	27,208	17,742	108	78
port1286	22,692	36,210	22,952	116	100
resi1247	-	38,638	13,672	184	113
ruul1235	38,698	51,284	17,942	138	127
sadu1234	-	28,540	23,544	95	88
sanz1248	14,840	24,871	10,908	99	78
savo1255	23,954	39,562	18,546	120	89
sout2856	20,854	30,194	15,326	129	81
sout3282	19,610	-	18,116	91	75
stan1290	-	-	26,690	110	101
sumi1235	30,838	36,924	21,700	87	84
svan1243	-	42,266	20,024	170	138
taba1259	19,118	21,820	10,556	78	64
teop1238	28,714	46,968	23,838	122	111
texi1237	35,876	38,460	21,898	133	113
trin1278	33,140	44,290	16,018	186	145
tsim1256	-	-	9,511	123	62
urum1249	42,200	59,510	23,296	235	175
vera1241	30,336	38,366	25,016	119	99
warl1254	-	43,730	13,984	154	122
yong1270	-	16,936	9,474	82	68
yuca1254	-	34,828	21,304	119	106
yura1255	-	122,436	45,106	356	341



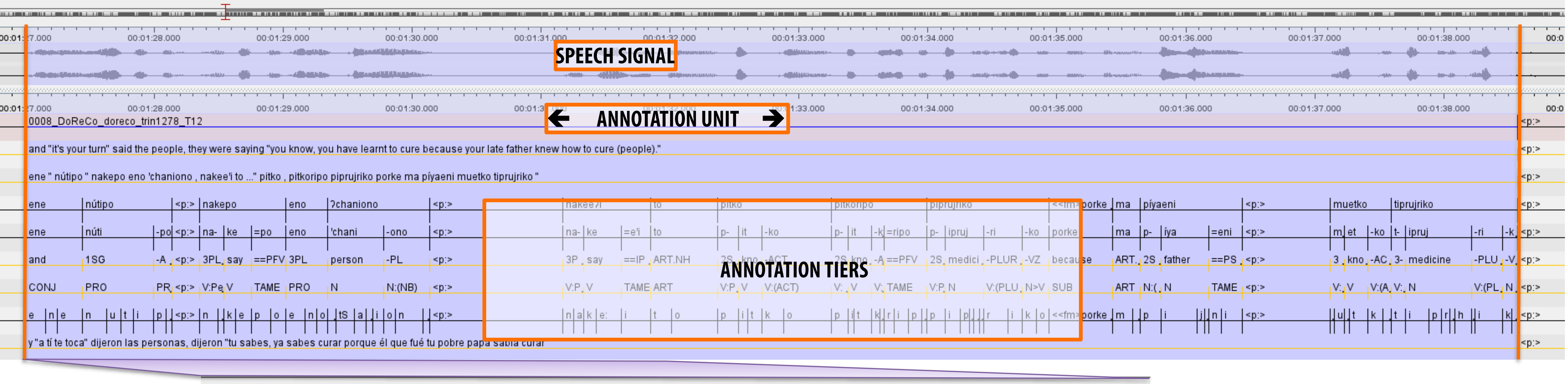
0008_DoReCo_doreco_trin1278_T12

and "it's your turn" said the people, they were saying "you know, you have learnt to cure because your late father knew how to cure (people)."

ene " nútipo " nakepo eno 'chaniono , nakee'i to ..." pitko , pitkoripo piprujriko porke ma píyaeni muetko tiprujriko "

y "a tí te toca" dijeron las personas, dijeron "tu sabes, ya sabes curar porque él que fué tu pobre papa sabia curar

Rose, F. (2022). Mojeño Trinitario DoReCo dataset

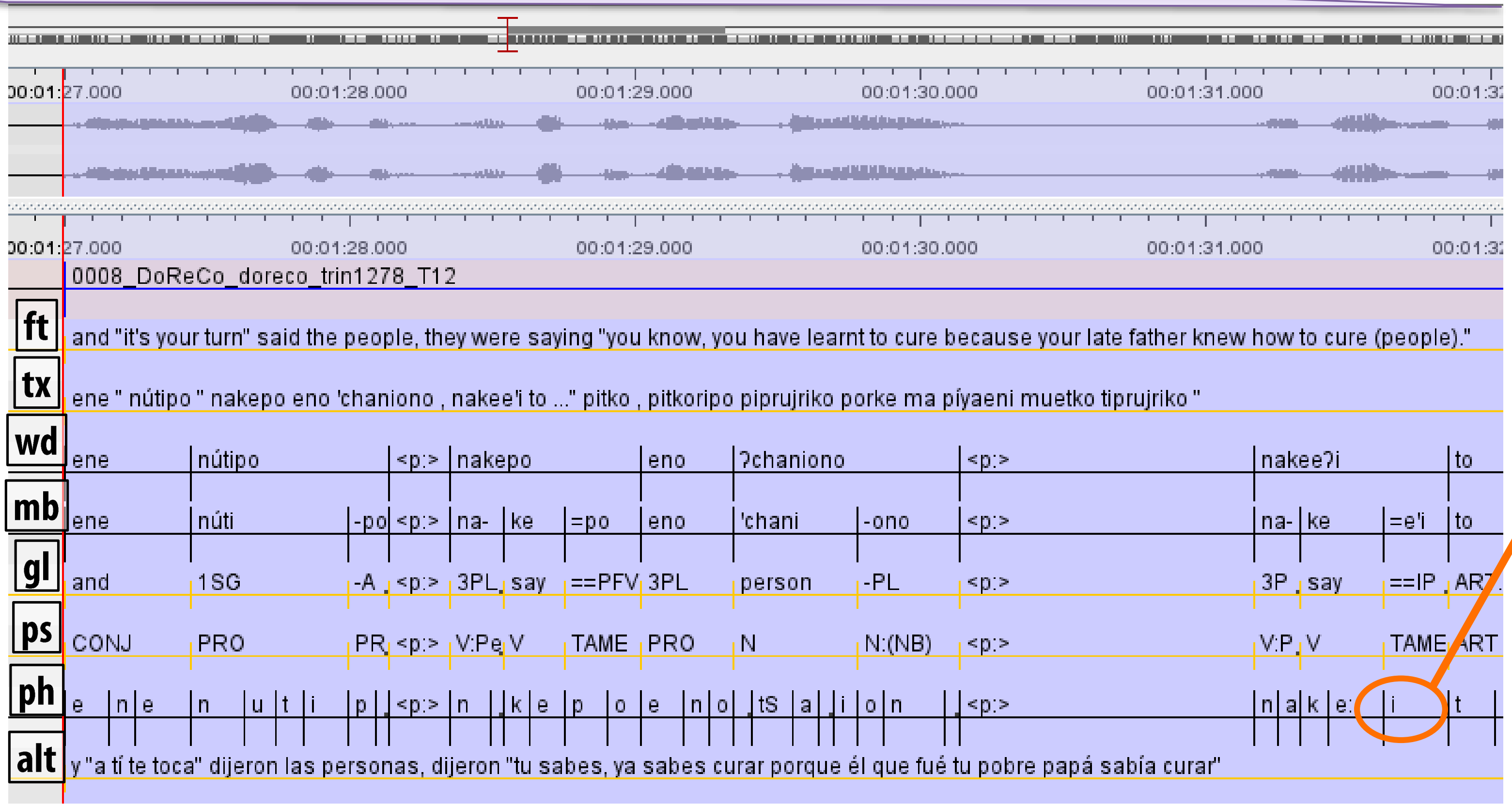


0008_DoReCo_doreco_trin1278_T12

and "it's your turn" said the people, they were saying "you know, you have learnt to cure because your late father knew how to cure (people)."

ene " nútipo " nakepo eno 'chaniono , nakee'i to ..." pitko , pitkoripo piprujriko porke ma píyaeni muetko tiprujriko "

y "a tí te toca" dijeron las personas, dijeron "tu sabes, ya sabes curar porque él que fué tu pobre papá sabía curar"



0008_DoReCo_doreco_trin1278_T12

and "it's your turn" said the people, they were saying "you know, you have learnt to cure because your late father knew how to cure (people)."

ene " nútipo " nakepo eno 'chaniono , nakee'i to ..." pitko , pitkoripo piprujriko porke ma píyaeni muetko tiprujriko "

ene nútipo <p:> nakepo eno ?chaniono <p:> nakee?i to

ene núti -po <p:> na- ke =po eno 'chani -ono <p:> na- ke =e'i to

and 1SG -A <p:> 3PL say ==PFV 3PL person -PL <p:> 3P say ==IP ART

CONJ PRO PR <p:> V:Pe V TAME PRO N N:(NB) <p:> V:P V TAME ART

e n e n u t i p <p:> n k e p o e n o tS a i o n <p:> n a k e: i t

y "a tí te toca" dijeron las personas, dijeron "tu sabes, ya sabes curar porque él que fué tu pobre papá sabía curar"

free translation
transcription
words
morphs
gloss
pos
phones
(2nd translation)

LG_CODE	trin1278
FILE	doreco_trin1278_T12
SPK_ID	ANM
ph_ID	p007796
ph	i
start	91.625
end	91.857
Annotation Unit	0008_DoReCo_doreco_trin1278_T12
tx	ene…
ft	and …
wd_ID	w004383
wd	nakee?i
mb_ID	m007290
mb	=e?i
ps	TAME
gl	==IPFV

Rose, F. (2022). Mojeño Trinitario DoReCo dataset

DoReCo in a nutshell

- A unique *and* accessible resource for NLP and linguistics
- High scientific potential, especially for linguistic comparative studies

DoReCo in a nutshell

- A unique *and* accessible resource for NLP and linguistics
- High scientific potential, especially for linguistic comparative studies
- But are the datasets comparable?

DoReCo in a nutshell

- A unique *and* accessible resource for NLP and linguistics
- High scientific potential, especially for linguistic comparative studies
- But are the datasets comparable?

Credits: Gotlib/Dargaud

- Differences potentially due to
  - A Language & Speaker
  - B Documentation context & Corpus creator

ILLUSTRATION #1
DURATION OF THE ANNOTATION UNITS

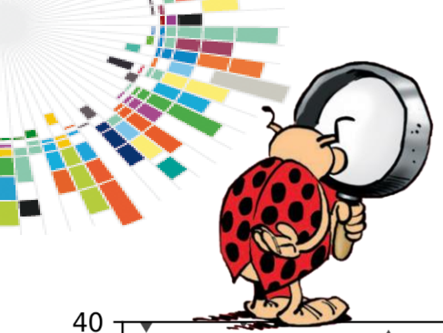


ILLUSTRATION #1
DURATION OF THE ANNOTATION UNITS

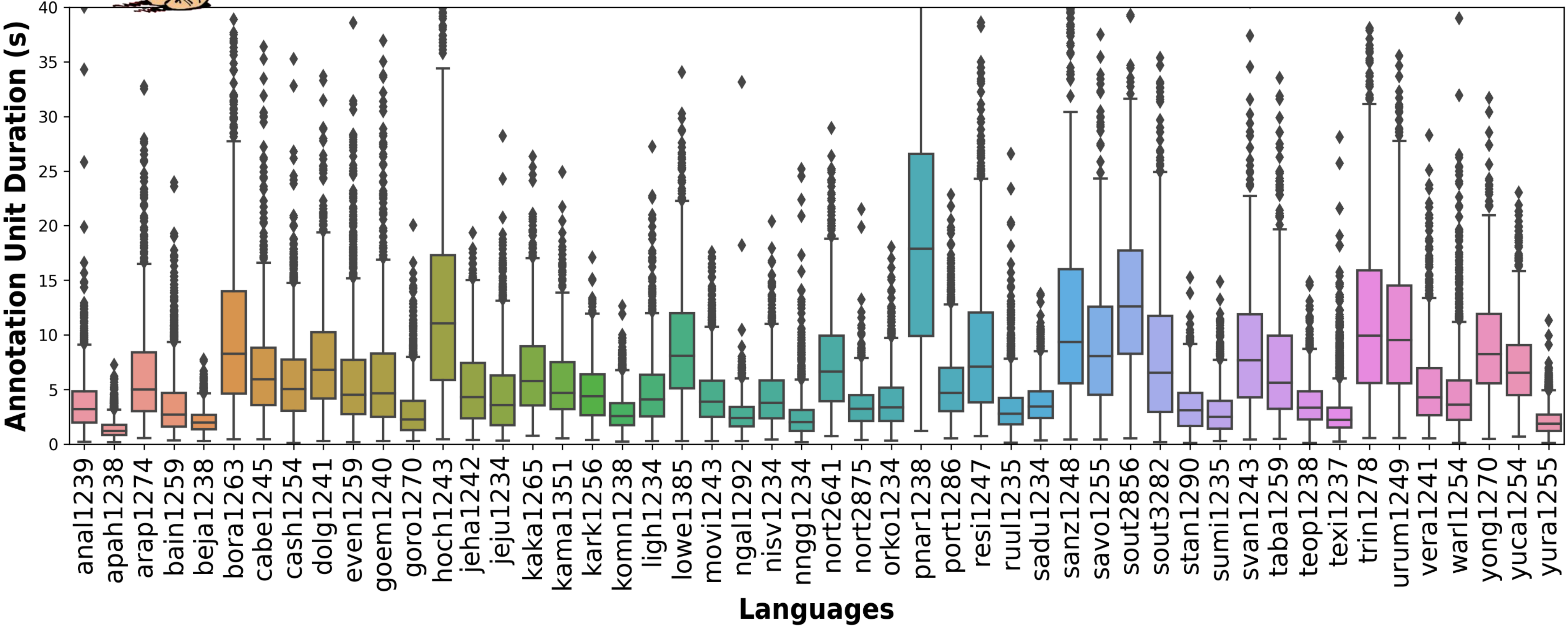
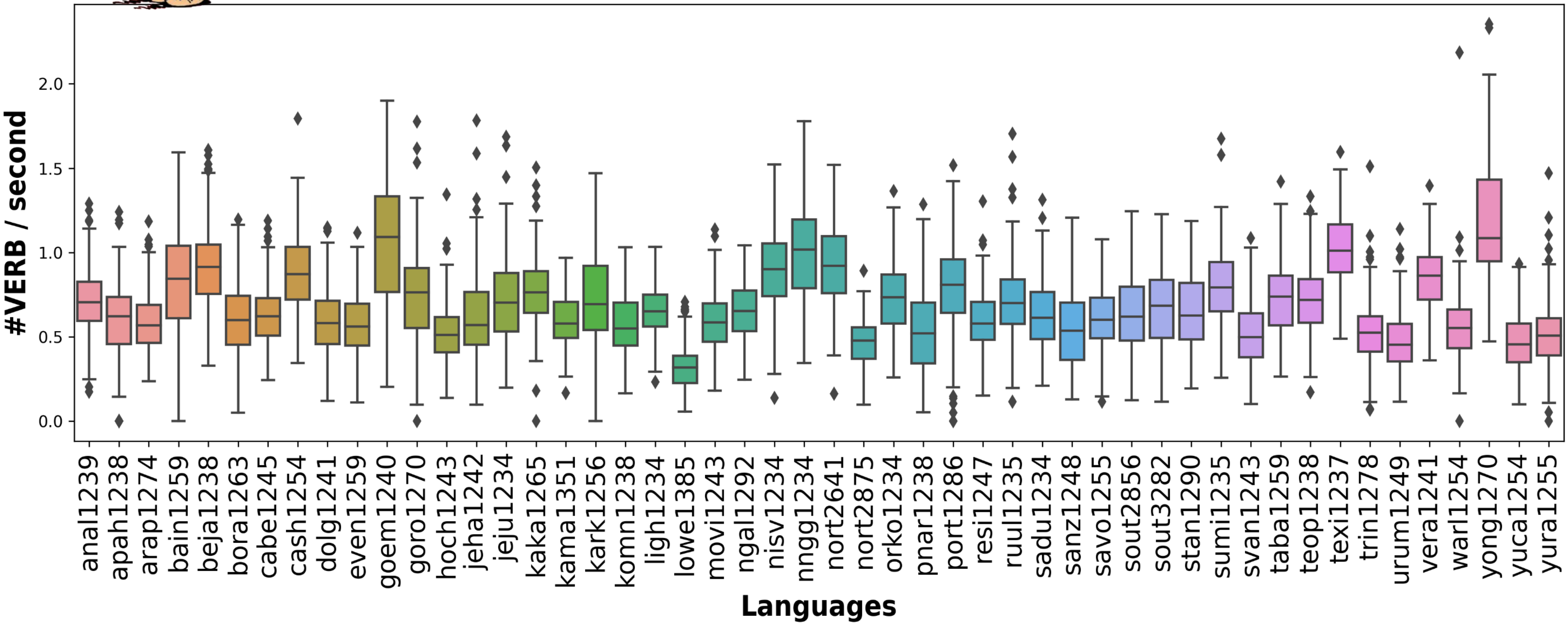
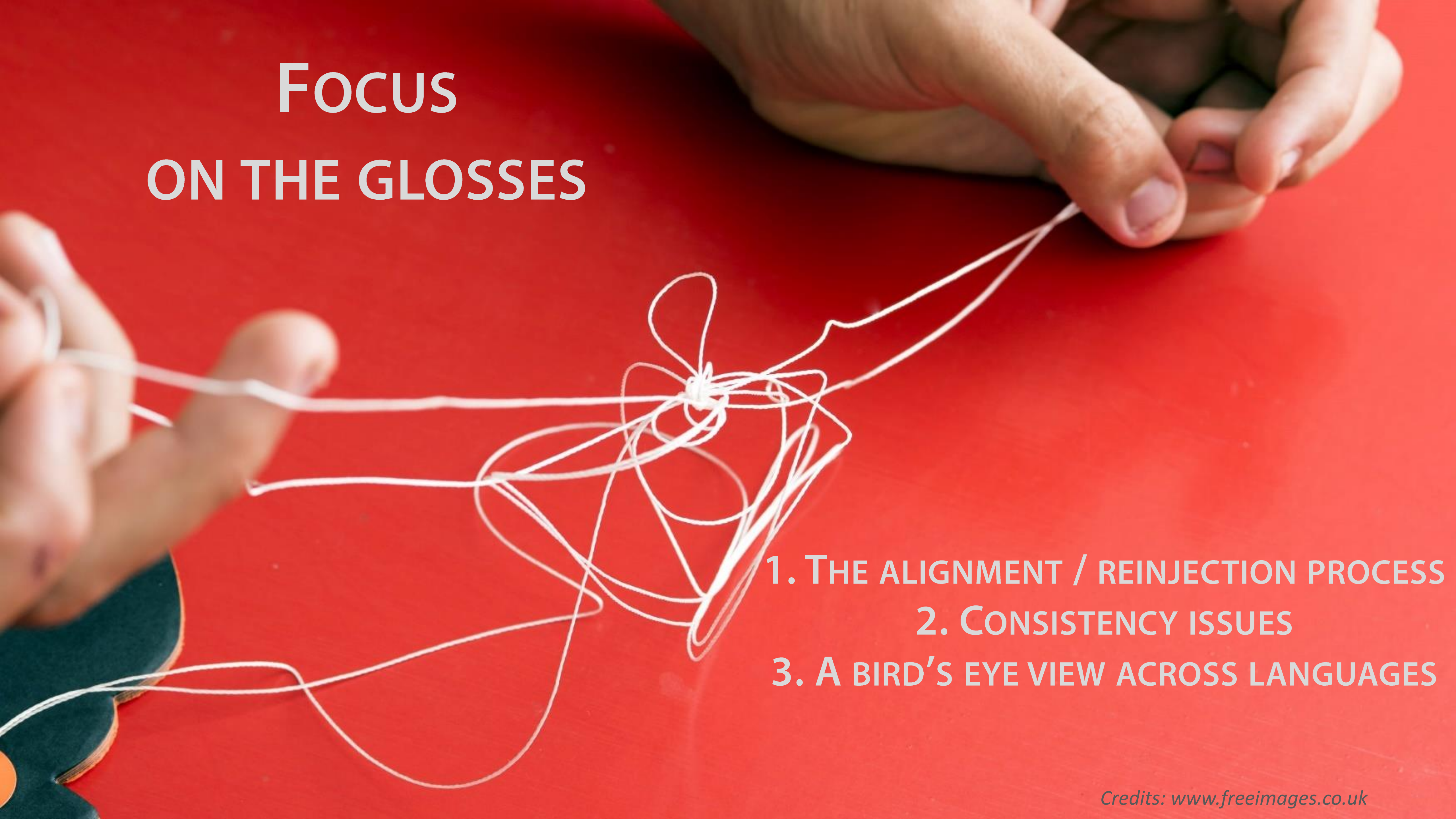


ILLUSTRATION #2
NUMBER OF VERBS PER SECOND*

* Estimated by a POS tagger applied to *the translation*





Focus
on the glosses

1. The alignment / reinjection process
2. Consistency issues
3. A bird’s eye view across languages

Credits: www.freeimages.co.uk

DATA PROCESSING PIPELINE

- Receiving language documentation data
- Selection of DoReCo-compatible datasets
- Automatic time-alignment using MAUS I
- Manual correction and labelling
- Automatic time-alignment using MAUS II
- Creating consistent and uniform morphological annotations
- Re-injection of morphological annotation into time-aligned transcription
- Creation of annotation files in various formats: TextGrid, EAF, TEI XML and CSV
- Making audio and annotation files available for download

1. ALIGNMENT / REINJECTION WORKFLOW: PRINCIPLES

- Original ELAN files contain many levels of annotation
  - ✓ Reference tier, morphological glosses, POS tags, other annotations

- Newly created TextGrid files contain time-aligned words and phones

- Reinjection
  - ✓ First ELAN words must be aligned with TextGrid words
  - ✓ Then ELAN morphological annotations must be aligned with TextGrid phones

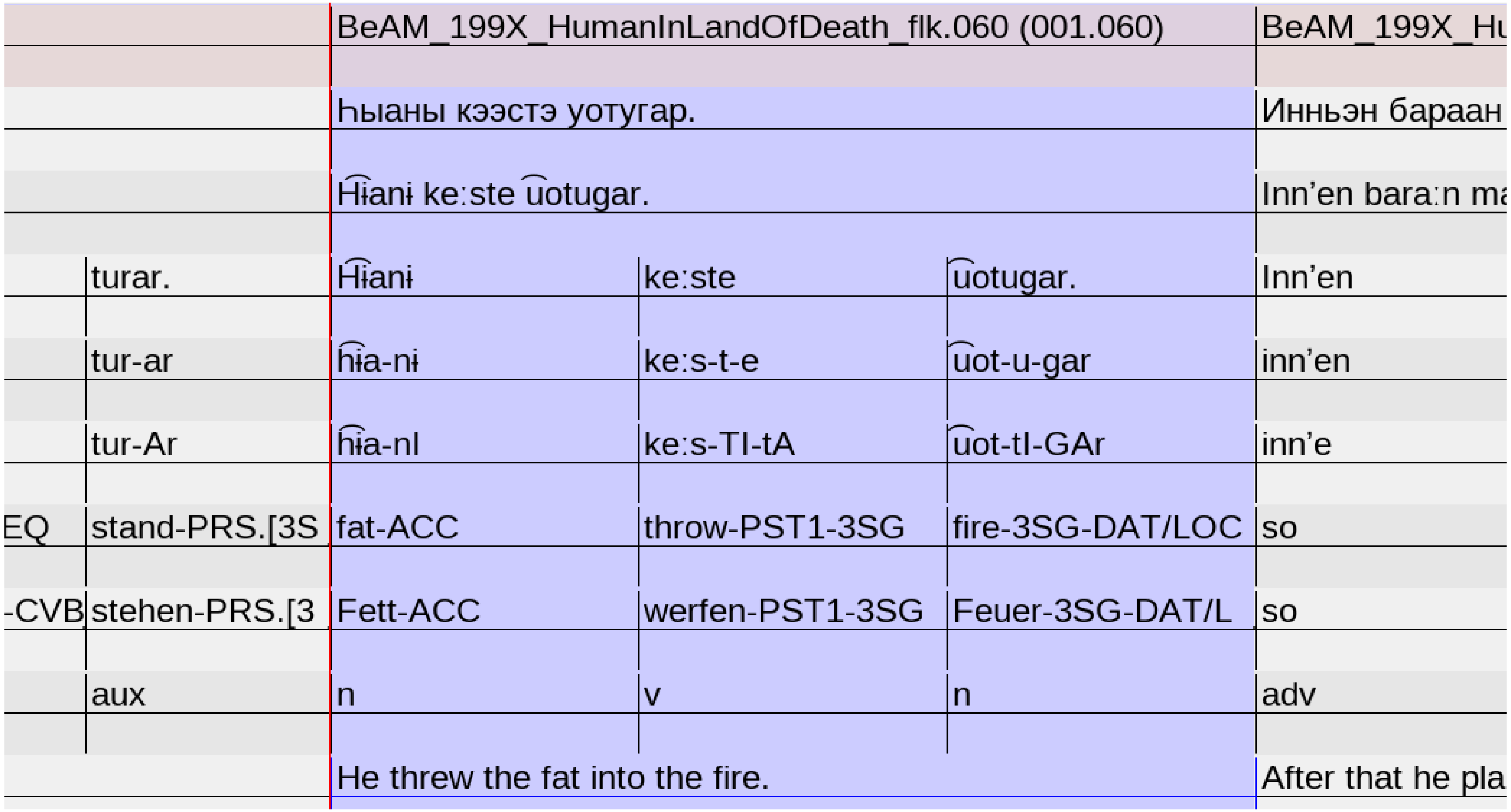
1. Alignment / reinjection workflow: principles

- Original ELAN files contain many levels of annotation
  - ✓ Reference tier, morphological glosses, POS tags, other annotations
- Newly created TextGrid files contain time-aligned words and phones
- Reinjection
  - ✓ First ELAN words must be aligned with TextGrid words
  - ✓ Then ELAN morphological annotations must be aligned with TextGrid phones
- Neither of these alignments are trivial
  - ✓ Words: During time-alignment, words may be added, removed, or changed, to match the acoustic signal
  - ✓ Morphs: Morphs are typically in their canonical forms, which do not perfectly match the time-aligned phones in the TextGrid
- Prior to any of this, however, files must be standardized
  - ✓ Classify tier types, rename tiers, standardize EAF/XML structure, inject time-aligned tiers, much more

1. Alignment / reinjection: process

- How to align strings of words with gaps, additions, and changes?
- And how to align strings of morphs with phones that don’t match?
- Needleman-Wunsch algorithm
  - ✓ Dynamic programming algorithm widely used in bioinformatics (optimal alignment of DNA sequences)
  - ✓ Also useful for aligning natural language sequences
- First stage
  - ✓ ELAN words with MAUSed words, to adjust words and utterances
- Second stage
  - ✓ ELAN morphs with MAUS phones, to adjust morphs, glosses, and POS tags

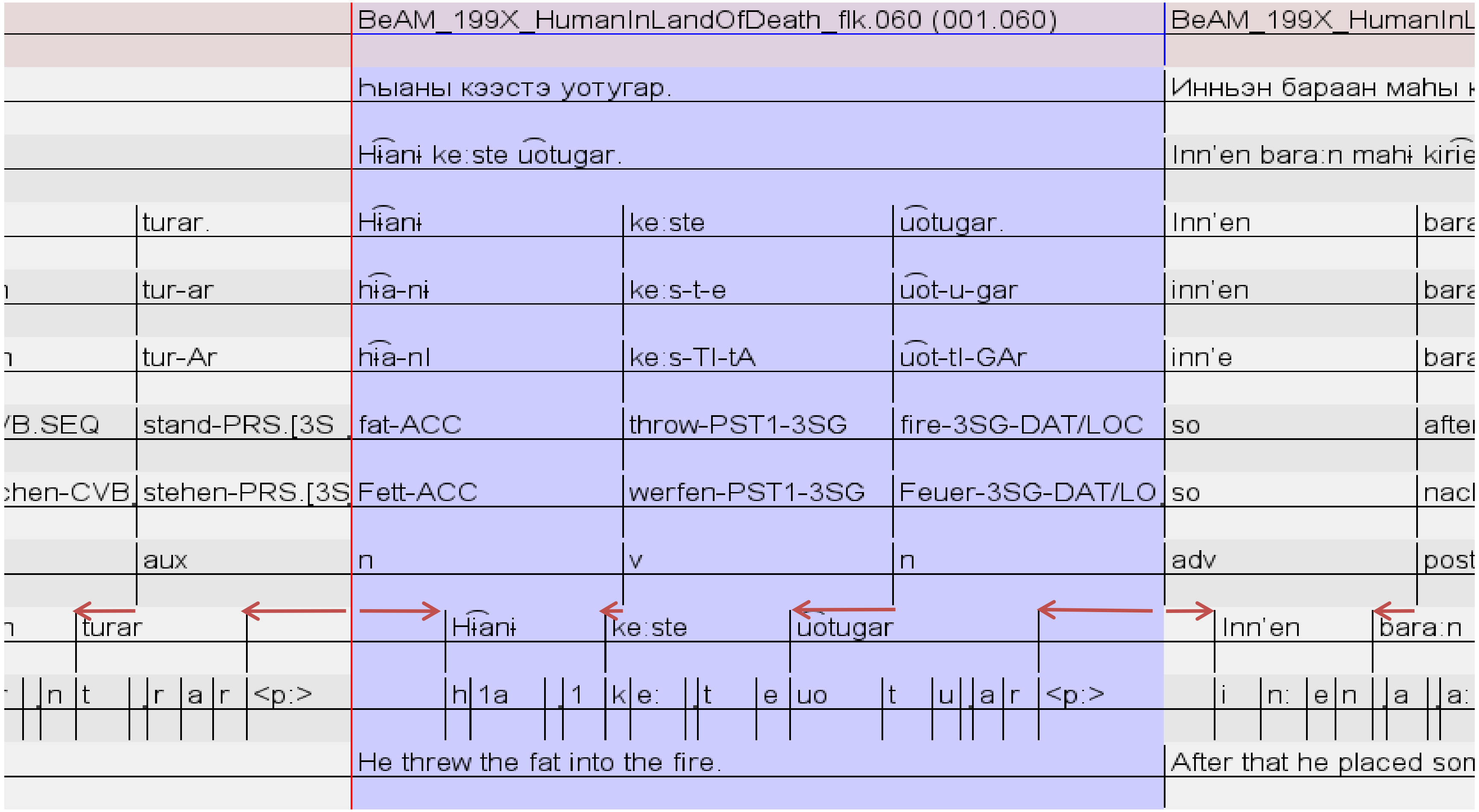
1. Alignment / reinjection: Illustration (SOURCE)



		BeAM_199X_HumanInLandOfDeath_flk.060 (001.060)			BeAM_199X_Hu
		Һыаны кээстэ уотугар.			Инньэн бараан
		H͡ɨanɨ ke:ste u͡otugar.			Inn’en bara:n ma
	turar.	H͡ɨanɨ	ke:ste	u͡otugar.	Inn’en
	tur-ar	h͡ɨa-nɨ	ke:s-t-e	u͡ot-u-gar	inn’en
	tur-Ar	h͡ɨa-nI	ke:s-TI-tA	u͡ot-tI-GAr	inn’e
EQ	stand-PRS.[3S	fat-ACC	throw-PST1-3SG	fire-3SG-DAT/LOC	so
-CVB	stehen-PRS.[3	Fett-ACC	werfen-PST1-3SG	Feuer-3SG-DAT/L	so
	aux	n	v	n	adv
		He threw the fat into the fire.			After that he pla

Däbritz, Chris Lasse; Kudryakova, Nina; Stapert, Eugénie. 2019. "INEL Dolgan Corpus." Version 1.0. Publication date 2019-08-31. http://hdl.handle.net/11022/0000-0007-CAE7-1. Archived in Hamburger Zentrum für Sprachkorpora. In: Wagner-Nagy, Beáta; Arkhipov, Alexandre; Ferger, Anne; Jettka, Daniel; Lehmberg, Timm (eds.). *The INEL corpora of indigenous Northern Eurasian languages.*

1. ALIGNMENT / REINJECTION: ILLUSTRATION (PROCESS)



		BeAM_199X_HumanInLandOfDeath_flk.060 (001.060)			BeAM_199X_HumanInL	
		һыаны кээстэ уотугар.			Инньэн бараан маһы к	
		Hɨ͡anɨ ke:ste u͡otugar.			Inn'en bara:n mahɨ kiri͡e	
	turar.	Hɨ͡anɨ	ke:ste	u͡otugar.	Inn'en	bara
ı	tur-ar	hɨ͡a-nɨ	ke:s-t-e	u͡ot-u-gar	inn'en	bara
ı	tur-Ar	hɨ͡a-nI	ke:s-TI-tA	u͡ot-tI-GAr	inn'e	bara
/B.SEQ	stand-PRS.[3S	fat-ACC	throw-PST1-3SG	fire-3SG-DAT/LOC	so	after
:hen-CVB	stehen-PRS.[3S	Fett-ACC	werfen-PST1-3SG	Feuer-3SG-DAT/LO	so	nach
	aux	n	v	n	adv	post
ı	turar	Hɨ͡anɨ	ke:ste	u͡otugar	Inn'en	bara:n

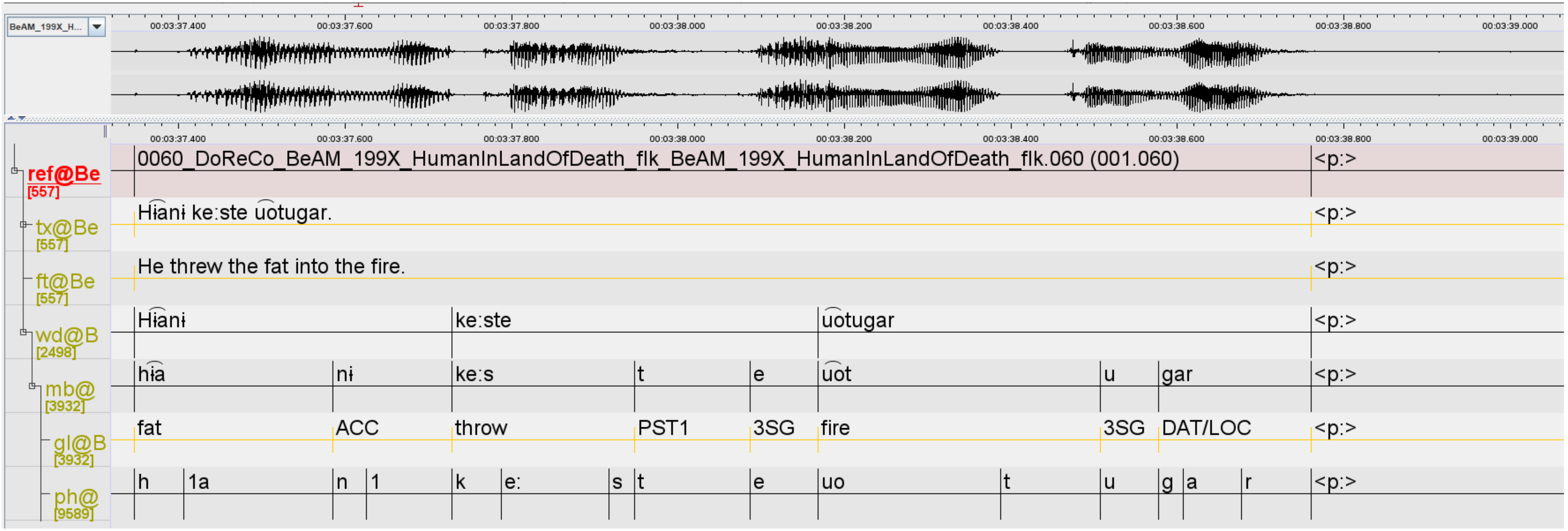
' n t r a r <p:> h 1a 1 k e: t e uo t u a r <p:> i n: e n a a:

He threw the fat into the fire. After that he placed son

Original ELAN annotations

Time-aligned tiers

1. ALIGNMENT / REINJECTION: ILLUSTRATION (OUTPUT)



ref@Be [557]	0060_DoReCo_BeAM_199X_HumanInLandOfDeath_flk_BeAM_199X_HumanInLandOfDeath_flk.060 (001.060)														<p:>
tx@Be [557]	Hɨ͡ani ke:ste u͡otugar.														<p:>
ft@Be [557]	He threw the fat into the fire.														<p:>
wd@B [2498]	Hɨ͡ani				ke:ste						u͡otugar				<p:>
mb@ [3932]	hɨ͡a		nɨ		ke:s			t	e	u͡ot		u	gar		<p:>
gl@B [3932]	fat		ACC		throw			PST1	3SG	fire		3SG	DAT/LOC		<p:>
ph@ [9589]	h	1a	n	1	k	e:	s	t	e	uo	t	u	g	a	r	<p:>

Däbritz, Chris Lasse, Nina Kudryakova, Eugénie Stapert and Alexandre Arkhipov. 2022. Dolgan DoReCo dataset. In Seifart, Frank, Ludger Paschen and Matthew Stave (eds.). Language Documentation Reference Corpus (DoReCo) 1.0. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). (Accessed on 29/07/2022).

2. Consistency

- Consistency: Unified coding scheme & Missingness

- Determines whether two units are considered as part of the same category
  - ✓ Obviously this can have big effects of frequency analysis (e.g. if you use Type-Token Ratio. Free hint: ***Don't***. see Oh & Pellegrino, 2022).

- Unified coding scheme
  - ✓ Internally-consistent coding (word, POS, gloss, gesture, etc.)
  - ✓ Sources of error: spelling errors, format changes, updated analyses

- Not trivial: Multiple coders, updates to coding scheme over many years
  - ✓ No blame on the corpus creators here!

- Within-corpus pitfall
  - ✓ Automated analyses treat spelling/coding variants as separate categories

- Across-corpus pitfall
  - ✓ Idiosyncratic coding variants complicate cross-linguistic comparison

Oh, Y. M., & Pellegrino, F. (2022). Towards robust complexity indices in linguistic typology: A corpus-based assessment. *Studies in Language*.

2. CONSISTENCY: ILLUSTRATIONS

- Unified coding scheme: morphemes

- What is the best way to define a morpheme within a corpus?
  - ✓ If only using the morph form, homophony slips in

    -s -> PL, 3Sg.PRS, POSS
  - ✓ If only using the gloss, allomorphy slips in

    -PL -> /s/, /z/, /əz/

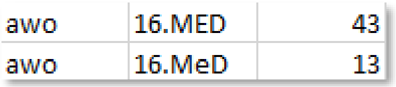
- Maybe better to use the combination of morph and gloss
  - ✓ But this can be upset by a lack of unified coding scheme

2. CONSISTENCY: ILLUSTRATIONS

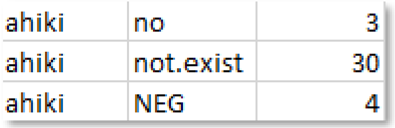
- Unified coding scheme: morphemes
- What is the best way to define a morpheme within a corpus?
  - ✓ If only using the morph form, homophony slips in

    -s -> PL, 3Sg.PRS, POSS
  - ✓ If only using the gloss, allomorphy slips in

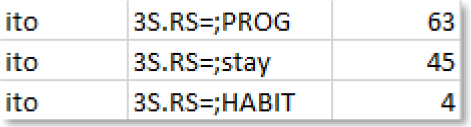
    -PL -> /s/, /z/, /əz/
- Maybe better to use the combination of morph and gloss
  - ✓ But this can be upset by a lack of unified coding scheme
- Some examples of a morph form, its glosses, and their frequencies



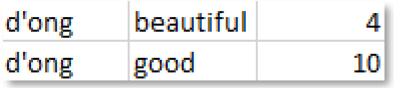
awo	16.MED	43
awo	16.MeD	13



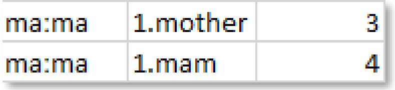
ahiki	no	3
ahiki	not.exist	30
ahiki	NEG	4



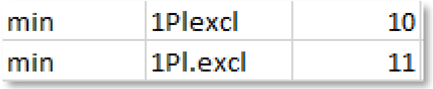
ito	3S.RS=;PROG	63
ito	3S.RS=;stay	45
ito	3S.RS=;HABIT	4



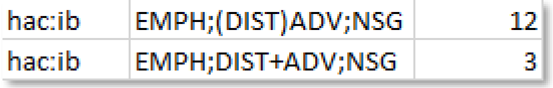
d'ong	beautiful	4
d'ong	good	10



ma:ma	1.mother	3
ma:ma	1.mam	4



min	1Plexcl	10
min	1Pl.excl	11



hac:ib	EMPH;(DIST)ADV;NSG	12
hac:ib	EMPH;DIST+ADV;NSG	3

2. Consistency: illustrations (cont'd)

2. Consistency: illustrations (cont'd)

- Equally problematic to define a morpheme across multiple corpora
  - ✓ E.g. how to identify first person singular pronominals

First person pronominal forms
1/2-
1SG
1.SG
1S.
1sgS-
1s.poss
1sg>3sg
PRO.1sg
PS1SG
1S/3S

2. CONSISTENCY: ILLUSTRATIONS (CONT'D)

- Equally problematic to define a morpheme across multiple corpora
  - ✓ E.g. how to identify first person singular pronominals

First person pronominal forms
1/2-
1SG
1.SG
1S.
1sgS-
1s.poss
1sg>3sg
PRO.1sg
PS1SG
1S/3S

- Or to separate glossed elements within a morpheme
  - ✓ Using separator symbols like "."

"."-separated glossed elements
sing to/for s.o.
sweet.potato
NOM.SG
yes(Ar.)
3.P(ATTR)
like.this
17.LOC
say\PFV.3PL
qué.cosa

2. CONSISTENCY: ILLUSTRATIONS (CONT’D)

- Equally problematic to define a morpheme across multiple corpora
  - ✓ E.g. how to identify first person singular pronominals

First person pronominal forms
1/2-
1SG
1.SG
1S.
1sgS-
1s.poss
1sg>3sg
PRO.1sg
PS1SG
1S/3S

- Or to separate glossed elements within a morpheme
  - ✓ Using separator symbols like “.”

“.”-separated glossed elements
sing to/for s.o.
sweet.potato
NOM.SG
yes(Ar.)
3.P(ATTR)
like.this
17.LOC
say\PFV.3PL
qué.cosa

- Other glossing symbols must be dealt with as well

Glossing symbols to look out for
DIR<wōl>
<1E.U>-
find\IPFV.[3SG.M]
put.PRS.3SG[IMP]
be_there\AOR
Neg:Fut
RED:eat
SG:SBJ>3SG.MASC:OBJ:PST:IPFV/call

2. Consistency: Identifying roots, affixes and clitics

- Ideally, every morph should be formatted such that its morph type is immediately obvious
  - ✓ Affixes: pref-, -suf, -inf-
  - ✓ Clitics: procl=, =encl
  - ✓ Reduplication: red~, ~red

- Inconsistencies
  - ✓ Affixes: Roots in nominal compounds often separated by "-"; infixes somehow indicated but never tokenized (e.g. "<inf>stem")
  - ✓ Clitics: Sometimes marked as "-" but tagged as "clitic" on legacy tier
  - ✓ Reduplication: Annotation by template (e.g. "CVdup") breaks re-injection alignment

- One of the work packages in the **AIRAL** project (Acoustic Insights into the Root-Affix Asymmetry across Languages, Ludger Paschen, 2022-2025, ZAS Berlin)

2. Consistency and beyond (thank you Captain Obvious)

- Missingness: empty cells in the data
- Unintelligible speech, unknown meaning, ran out of time
- Can introduce biases if missingness is high and/or systematic
- No magical recipe here, but have to keep this in mind

3. Beyond… A bird's eye view across languages

- Even if within-language consistency is improved
  (manually or semi-automatically),
  across-language heterogeneity remains the rule rather than the exception

3. Beyond… A bird's eye view across languages

- Even if within-language consistency is improved (manually or semi-automatically), across-language heterogeneity remains the rule rather than the exception

- Let's look at the metadata: List of abbreviations used in glosses

Beja1238 (excerpt)	
Gloss	**Meaning**
1	first person
2	second person
3	third person
ABL	ablative
ACC	accusative
ACMP	(unclear)
ADJVZ	adjectivizer
ADRE	addressee
ADRF	form of address
AOR	aorist
CAUS	causative
EMPH	emphatic
NMLZ	nominalizer
SMLT	simultaneity
VN	verbonominal

ngal1292 (excerpt)	
Gloss	**Meaning**
BEN	benefactive
CAUS	causative
CSTVZR	causativizer
DYAD	dyadic suffix
EMPH	emphasizer
ERG	ergative
h	high(er) on scale of animacy
INTERJ	interjection
NEG	negation
PCUST	customary past
PI	past imperfective
POSS	possessive
RR	reflexive/reciprocal
SEQ	sequential
VBLZR	verbalizer

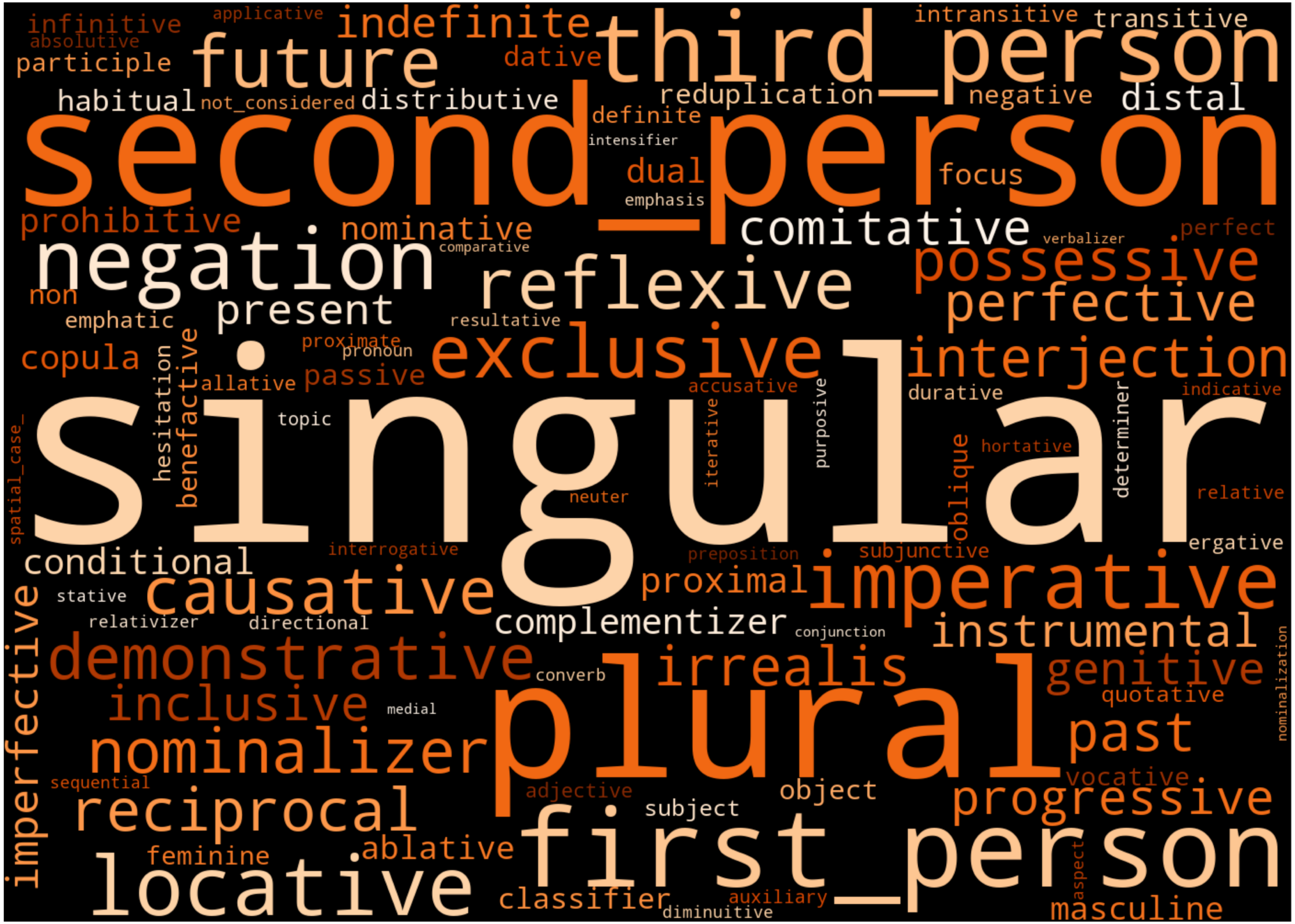
Ponsonnet, Maïa. 2022. Dalabon DoReCo dataset. In Seifart, Frank, Ludger Paschen and Matthew Stave (eds.). Language Documentation Reference Corpus (DoReCo) 1.2. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). https://doreco.huma-num.fr/languages/ngal1292 (Accessed on 20/06/2023). DOI:10.34847/nkl.fae299ug

Vanhove, Martine. 2022. Beja DoReCo dataset. In Seifart, Frank, Ludger Paschen and Matthew Stave (eds.). Language Documentation Reference Corpus (DoReCo) 1.2. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). https://doreco.huma-num.fr/languages/beja1238 (Accessed on 20/06/2023). DOI:10.34847/nkl.edd011t1

Meanings across languages

- ➔ The larger the font, the more pervasive the *Meaning* across the language descriptions

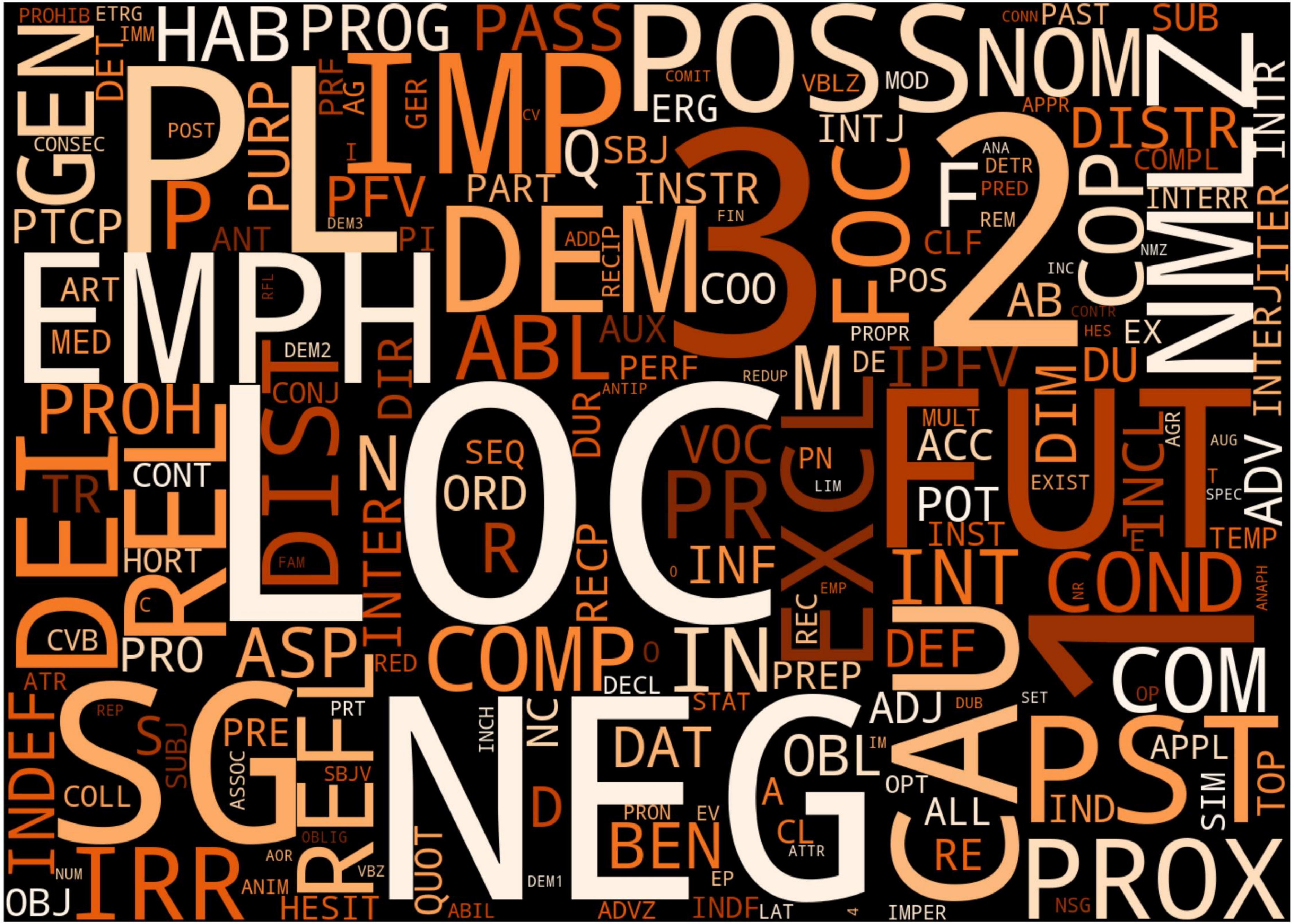
- Not an index of the *Meaning* token frequency within each language



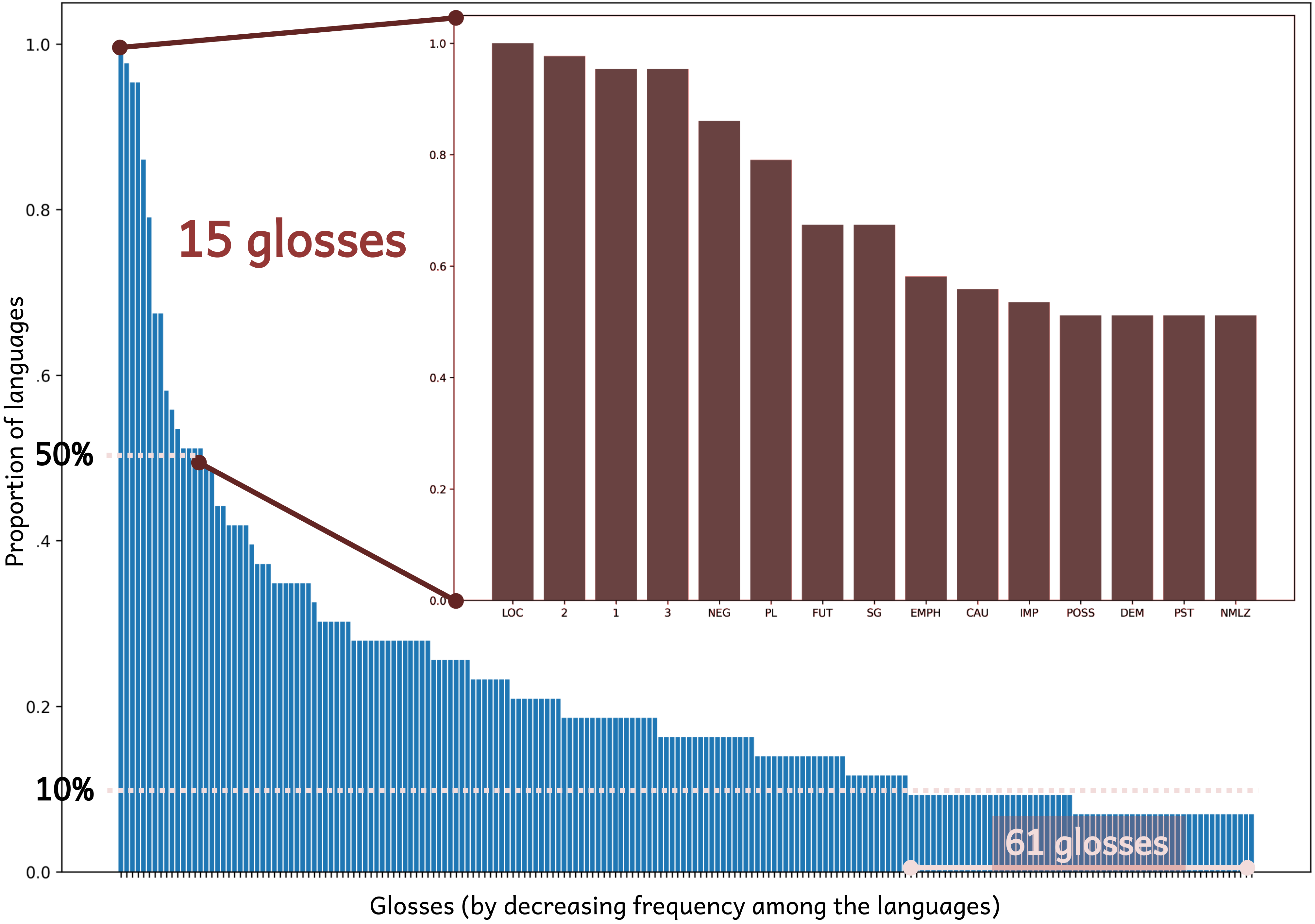
Glosses across languages

➔ The larger the font, the more pervasive the *Gloss* across the language descriptions

⚠ Not an index of the *Gloss* token frequency within each language

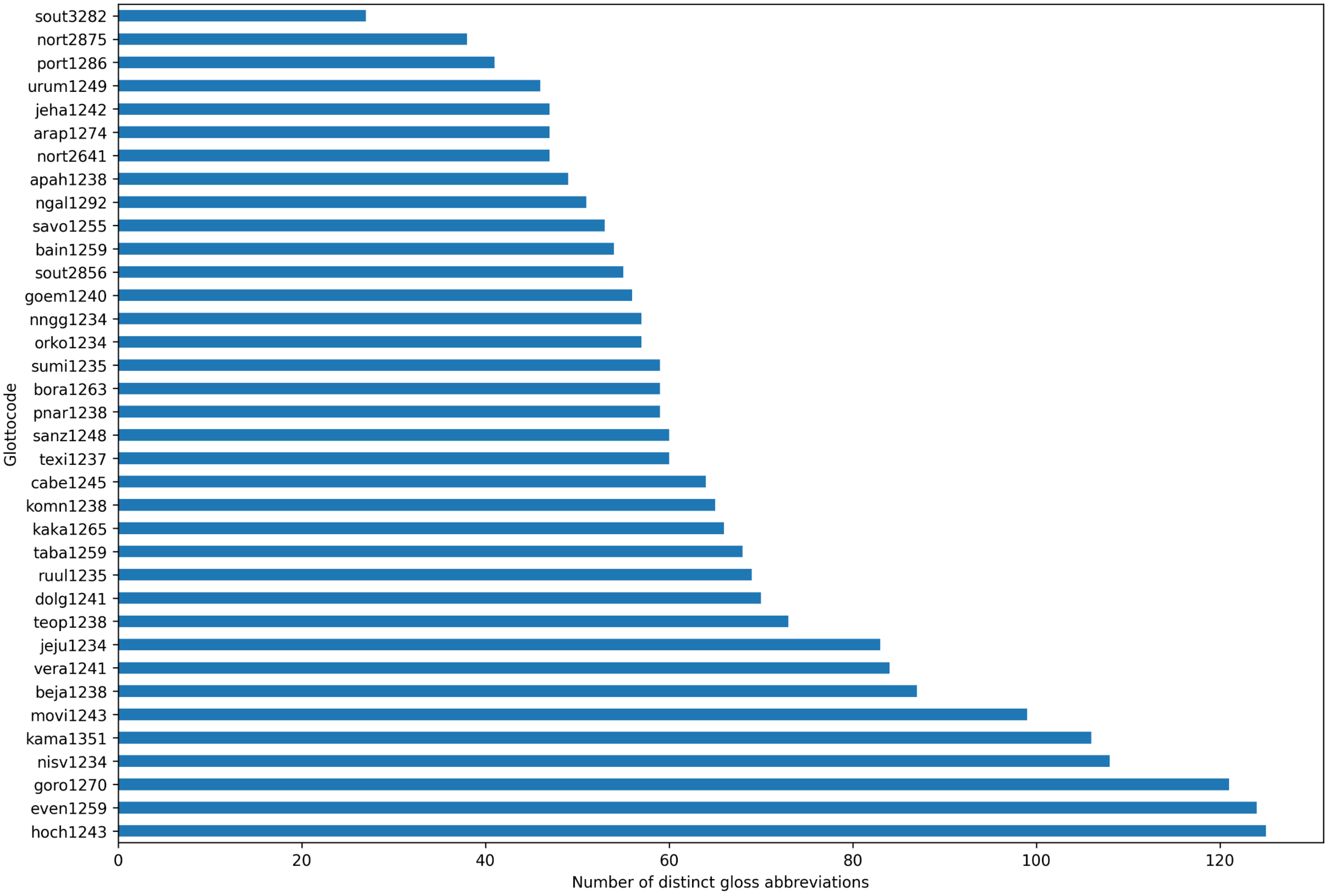


Glosses across languages

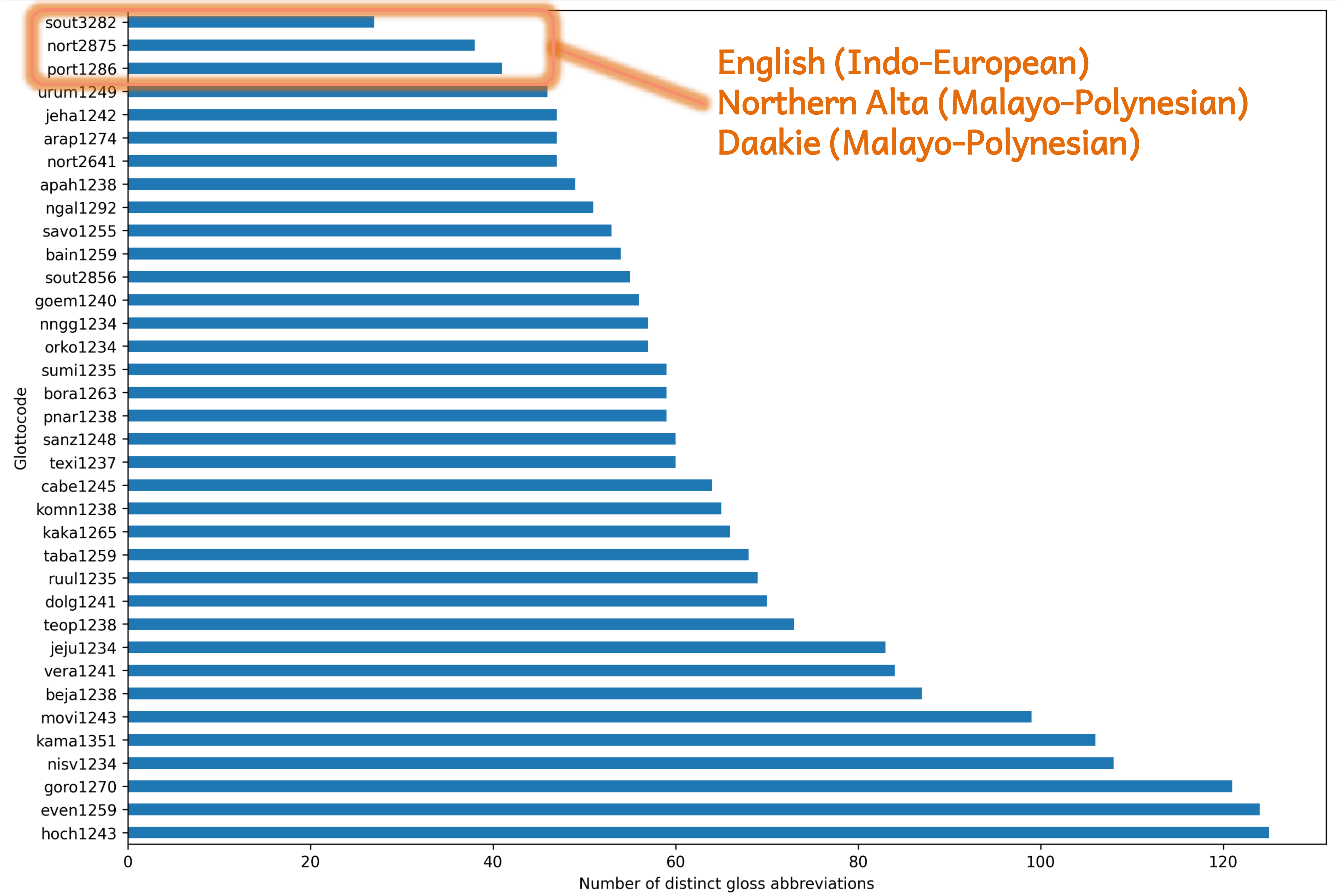


- 15 glosses present in 50+ % of languages
- 61 glosses in less than 10%

Number of Distinct glosses per language

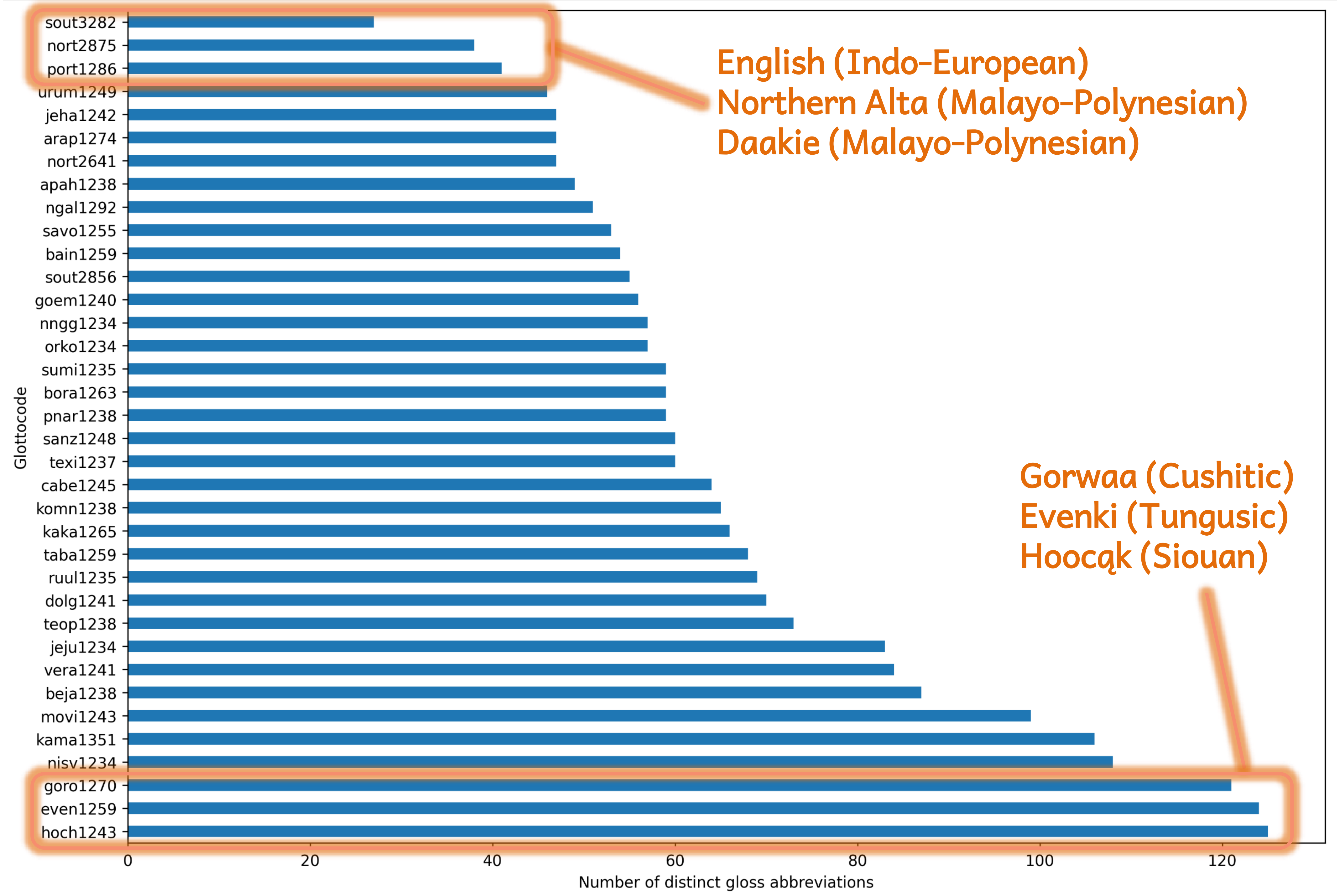


Number of Distinct glosses per language



English (Indo-European)
Northern Alta (Malayo-Polynesian)
Daakie (Malayo-Polynesian)

Number of Distinct glosses per language



English (Indo-European)
Northern Alta (Malayo-Polynesian)
Daakie (Malayo-Polynesian)

Gorwaa (Cushitic)
Evenki (Tungusic)
Hoocąk (Siouan)

CONCLUSION

- 51 languages with time-aligned words and phonemes
  - ✓ Including 38 languages with time-aligned interlinear glosses

- All initial goals achieved despite a heavily time-consuming procedure

➔ An unrivaled resource to study the temporal aspects of language in a typological perspective

Conclusion

- Trade-off between:

  across-language conventionalization

  and

  faithfulness to the source analysis

✓ DoReCo leans on the “faithfulness” side: “If it’s not broken, don’t fix it!”

➔ Beyond typology, a testbed for improvements

✓ Lange & Aznar (2022); von Prince & Nordhoff (2020)
✓ CLD 2025 ANR-DFG project; Autogramm ANR project, CREAM ANR project
✓ And more generally for resourcing under-resourced languages

Lange, H., & Aznar, J. (2022). RefCo and its Checker: Improving Language Documentation Corpora's Reusability Through a Semi-Automatic Review Process. In *Proc. 13th LREC*.
von Prince, K., & Nordhoff, S. (2020). An empirical evaluation of annotation practices in corpora from language documentation. In *Proc. of 12th LREC.*
Autogramm: https://autogramm.github.io/
CREAM: https://sites.google.com/view/creamproject/home

THANK YOU!

DoReCo corpus references

Aznar, J. (2022). Nisvai DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.2801565f

Bogomolova, N., Ganenkov, D., & Schiborr, N. N. (2022). Tabasaran DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.ad7f97xr

Burenhult, N. (2022). Jahai DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.6a71xp0p

Cobbinah, A. Y. (2022). Baïnounk Gubëeher DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.a332abw8

Cowell, A. (2022). Arapaho DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.36f5r1b6

Däbritz, C. L., Kudryakova, N., Stapert, E., & Arkhipov, A. (2022). Dolgan DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.f09eikq3

Döhler, C. (2022). Komnzo DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.c5e6dudv

Forker, D., & Schiborr, N. N. (2022). Sanzhi Dargwa DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.81934177

Franjieh, M. (2022). Fanbyak DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.02084446

Garcia-Laguia, A. (2022). Northern Alta DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.efea0b36

Güldemann, T., Ernszt, M., Siegmund, S., & Witzlack-Makarevich, A. (2022). N‖ng DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.f6c37fi0

Gusev, V., Klooster, T., Wagner-Nagy, B., & Arkhipov, A. (2022). Kamas DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.cdd8177b

Haig, G., Vollmer, M., & Thiele, H. (2022). Northern Kurdish (Kurmanji) DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.ca10ez5t

Hartmann, I. (2022). Hoocąk DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.b57f5065

Harvey, A. (2022). Gorwaa DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.a4b4ijj2

Haude, K. (2022). Movima DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.da42xf67

Hellwig, B. (2022). Goemai DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.b93664ml

Kazakevich, O., & Klyachko, E. (2022). Evenki DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.5e0d27cu

Kim, S.-U. (2022). Jejuan DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.06ebrk38

Mosel, U. (2022). Teop DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.9322sdf2

Ponsonnet, M. (2022). Dalabon DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.fae299ug

Quesada, J. D., Skopeteas, S., Pasamonik, C., Brokmann, C., & Fischer, F. (2022). Cabécar DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.ebc4ra22

Reiter, S. (2022). Cashinahua DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.a8f9q2f1

Riesberg, S. (2022). Yali (Apahapsili) DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.9d91nkq2

Ring, H. (2022). Pnar DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.5ba1062k

Rose, F. (2022). Mojeño Trinitario DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.cbc3b4xr

Schiborr, N. N. (2022). English (Southern England) DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.9c271u5g

Schnell, S. (2022). Vera'a DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.3e2cu8c4

Seifart, F. (2022a). Bora DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.6eaf5laq

Seifart, F., Paschen, L., & Stave, M. (Éds.). (2022). *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.7cbfq779

Skopeteas, S., Moisidi, V., Tsetereli, N., Lorenz, J., & Schröter, S. (2022). Urum DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.ac166n10

Teo, A. (2022). Sümi DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.5ad4t01p

Thieberger, N. (2022). Nafsan (South Efate) DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.ba4f760l

Vanhove, M. (2022). Beja DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.edd011t1

Vydrina, A. (2022). Kakabe DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.d5aeu9t6

Wegener, C. (2022). Savosavo DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.b74d1b33

Wichmann, S. (2022). Texistepec Popoluca DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.c50ck58f

M. Witzlack-Makarevich, A., Namyalo, S., Kiriggwajjo, A., & Molochieva, Z. (2022). Ruuli DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Éds.), *Language Documentation Reference Corpus (DoReCo) 1.2.* ZAS & DDL. https://doi.org/10.34847/nkl.fde4pp1u

ACKNOWLEDGMENTS

DoReCo
Language Documentation Reference Corpora

- Corpus creators
- François Delafontaine (Lyon)
- Ludger Paschen (ZAS Berlin)
- Frank Seifart (ZAS Berlin)
- Matthew Stave (DDL)

- Research assistants & interns
  - ✓ Webb Abernethy, Celia Birle, Frederic Blum, Alejandra Camelo Cruz, Laura Günther, Indira Hajnács, Nora Hofmann, Francie Höhler, Hannah Ida Hullmeine, Johanna Kimmerl, Cheslie Klein, Elena Lazarenko, Runzhi Lou, Stephan Lünser, Magdalena Nischik, Emma Ritz, Laura Schleicher, Jianqi Sun, Michelle Elizabeth Throssell Balagué, and Christin Walch.

- Funding Organizations

anr agence nationale de la recherche

DFG Deutsche Forschungsgemeinschaft German Research Foundation

LABEX ASLAN UNIVERSITÉ DE LYON

APPENDICES

ADDITIONAL FILES

- README with general information on DoReCo
- CONVENTIONS: labels, tier names
- More specific dataset information
- Metadata table
- Tier name changes
- Transcription (g2p) mappings
  - http://clarin.phonetik.uni-muenchen.de/BASWebServices/
  - services/runMAUSGetInventar?LANGUAGE=sampa
- List of abbreviations used in glosses (for some datasets)

TRANSCRIPTION CONVENTIONS

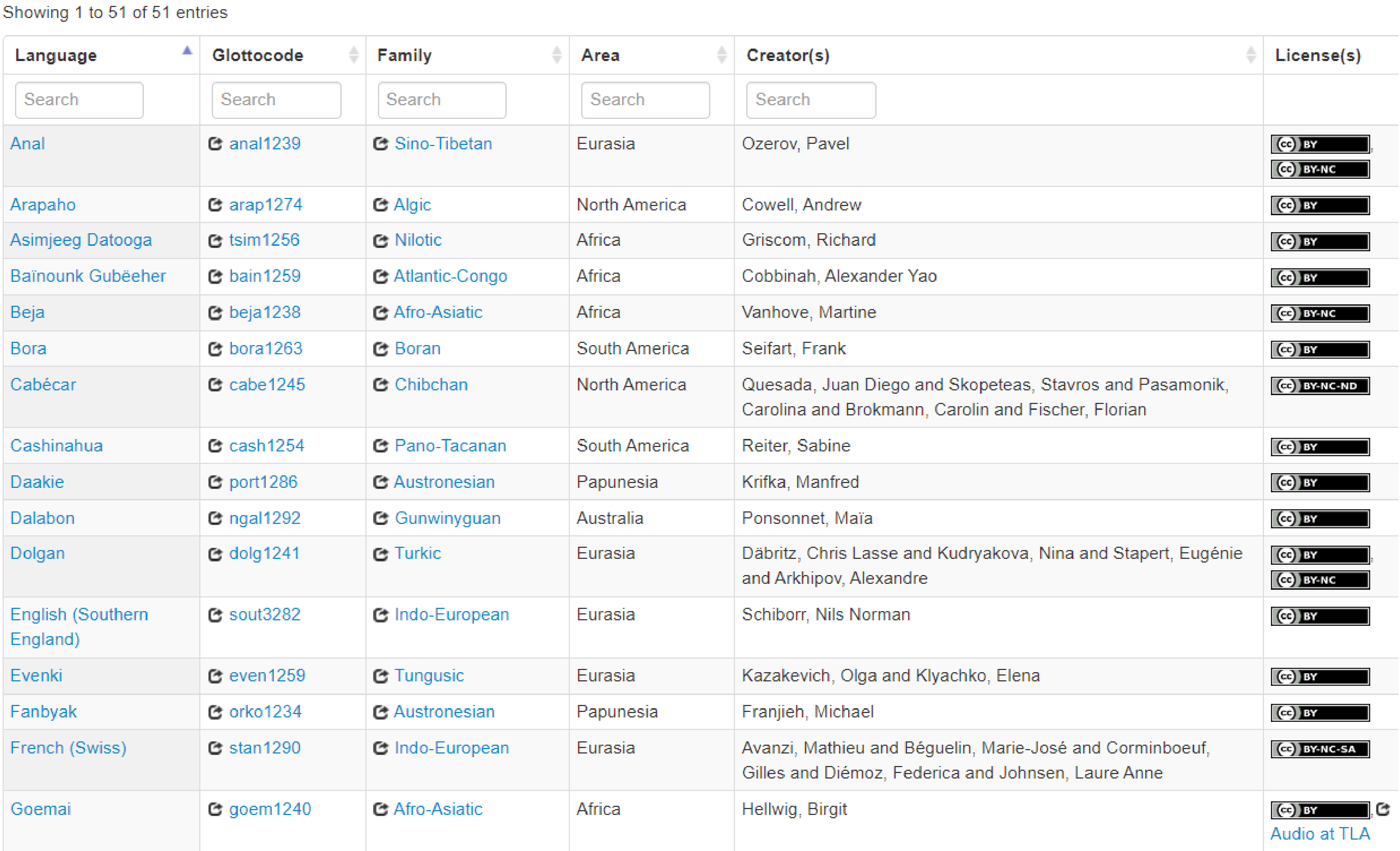
- Grapheme-to-phoneme mapping table used for creating forced alignments
- Mostly phonemic, but also includes frequent allophones if their distribution is well enough understood
- Using the language-independent model of MAUS and the X-SAMPA format for machine readability*
- Full list of symbols available at:

http://clarin.phonetik.uni-muenchen.de/BASWebServices/services/runMAUSGetInventar?LANGUAGE=sampa

Conventions: Labels

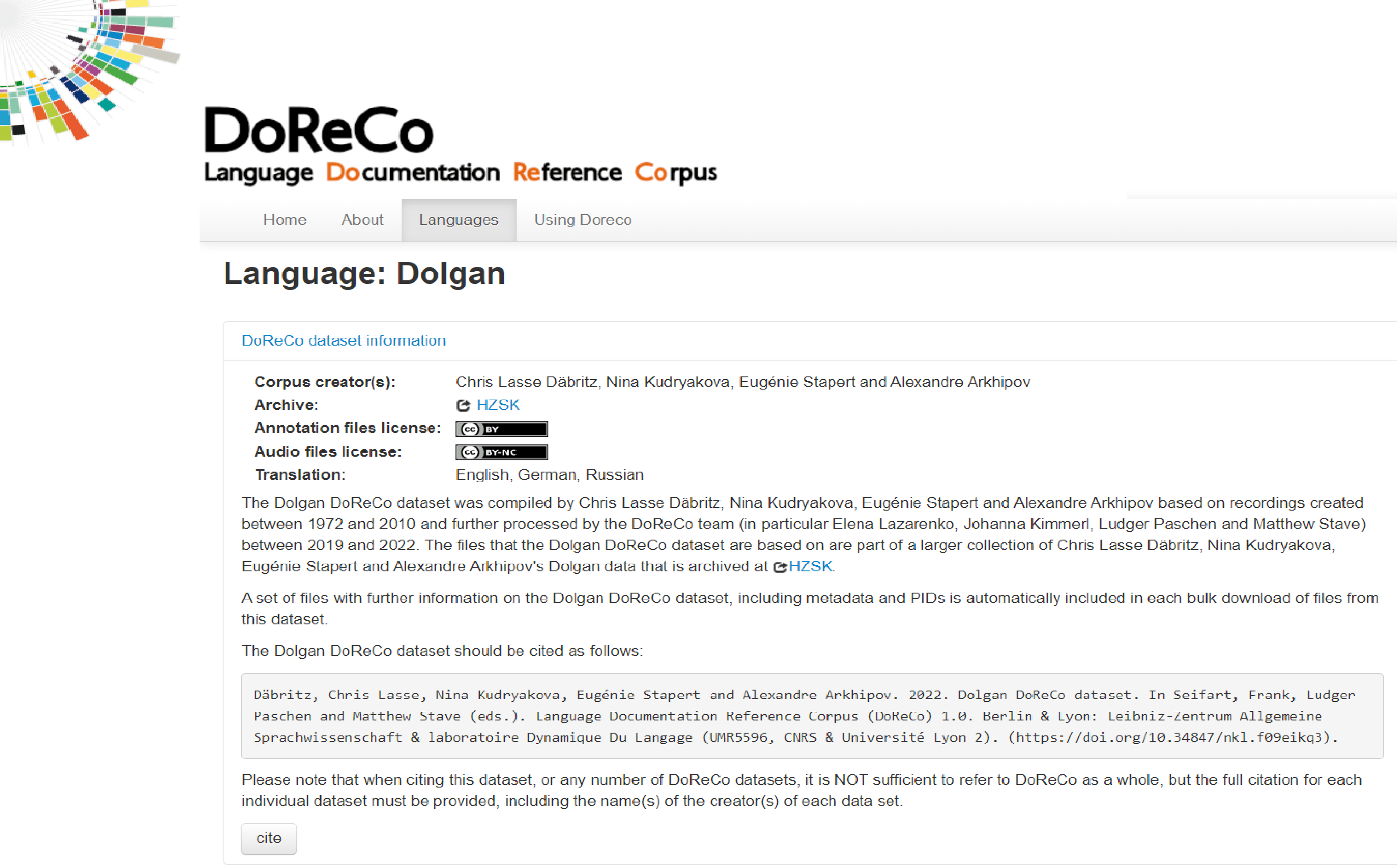
Filled pause	<<fp>>
False start	<<fs>>
Prolongation	<<pr>>
Singing	<<sg>>
Backchannel	<<bc>>
Ideophone	<<id>>
Onomatopoeic	<<on>>
Word–internal pause	<<wip>>
Unidentifiable	<<ui>>
Silent pause	<p:>

NAVIGATING TO A DORECO DATASET



Showing 1 to 51 of 51 entries

Language	Glottocode	Family	Area	Creator(s)	License(s)
Search	Search	Search	Search	Search	
Anal	anal1239	Sino-Tibetan	Eurasia	Ozerov, Pavel	CC BY, CC BY-NC
Arapaho	arap1274	Algic	North America	Cowell, Andrew	CC BY
Asimjeeg Datooga	tsim1256	Nilotic	Africa	Griscom, Richard	CC BY
Baïnounk Gubëeher	bain1259	Atlantic-Congo	Africa	Cobbinah, Alexander Yao	CC BY
Beja	beja1238	Afro-Asiatic	Africa	Vanhove, Martine	CC BY-NC
Bora	bora1263	Boran	South America	Seifart, Frank	CC BY
Cabécar	cabe1245	Chibchan	North America	Quesada, Juan Diego and Skopeteas, Stavros and Pasamonik, Carolina and Brokmann, Carolin and Fischer, Florian	CC BY-NC-ND
Cashinahua	cash1254	Pano-Tacanan	South America	Reiter, Sabine	CC BY
Daakie	port1286	Austronesian	Papunesia	Krifka, Manfred	CC BY
Dalabon	ngal1292	Gunwinyguan	Australia	Ponsonnet, Maïa	CC BY
Dolgan	dolg1241	Turkic	Eurasia	Däbritz, Chris Lasse and Kudryakova, Nina and Stapert, Eugénie and Arkhipov, Alexandre	CC BY, CC BY-NC
English (Southern England)	sout3282	Indo-European	Eurasia	Schiborr, Nils Norman	CC BY
Evenki	even1259	Tungusic	Eurasia	Kazakevich, Olga and Klyachko, Elena	CC BY
Fanbyak	orko1234	Austronesian	Papunesia	Franjieh, Michael	CC BY
French (Swiss)	stan1290	Indo-European	Eurasia	Avanzi, Mathieu and Béguelin, Marie-José and Corminboeuf, Gilles and Diémoz, Federica and Johnsen, Laure Anne	CC BY-NC-SA
Goemai	goem1240	Afro-Asiatic	Africa	Hellwig, Birgit	CC BY, Audio at TLA


DoReCo
Language Documentation Reference Corpus
Home
About
Languages
Using Doreco
DoReCo dataset information
cite

Language: Dolgan

Corpus creator(s):	Chris Lasse Däbritz, Nina Kudryakova, Eugénie Stapert and Alexandre Arkhipov
Archive:	HZSK
Annotation files license:	cc BY
Audio files license:	cc BY-NC
Translation:	English, German, Russian

The Dolgan DoReCo dataset was compiled by Chris Lasse Däbritz, Nina Kudryakova, Eugénie Stapert and Alexandre Arkhipov based on recordings created between 1972 and 2010 and further processed by the DoReCo team (in particular Elena Lazarenko, Johanna Kimmerl, Ludger Paschen and Matthew Stave) between 2019 and 2022. The files that the Dolgan DoReCo dataset are based on are part of a larger collection of Chris Lasse Däbritz, Nina Kudryakova, Eugénie Stapert and Alexandre Arkhipov's Dolgan data that is archived at HZSK.

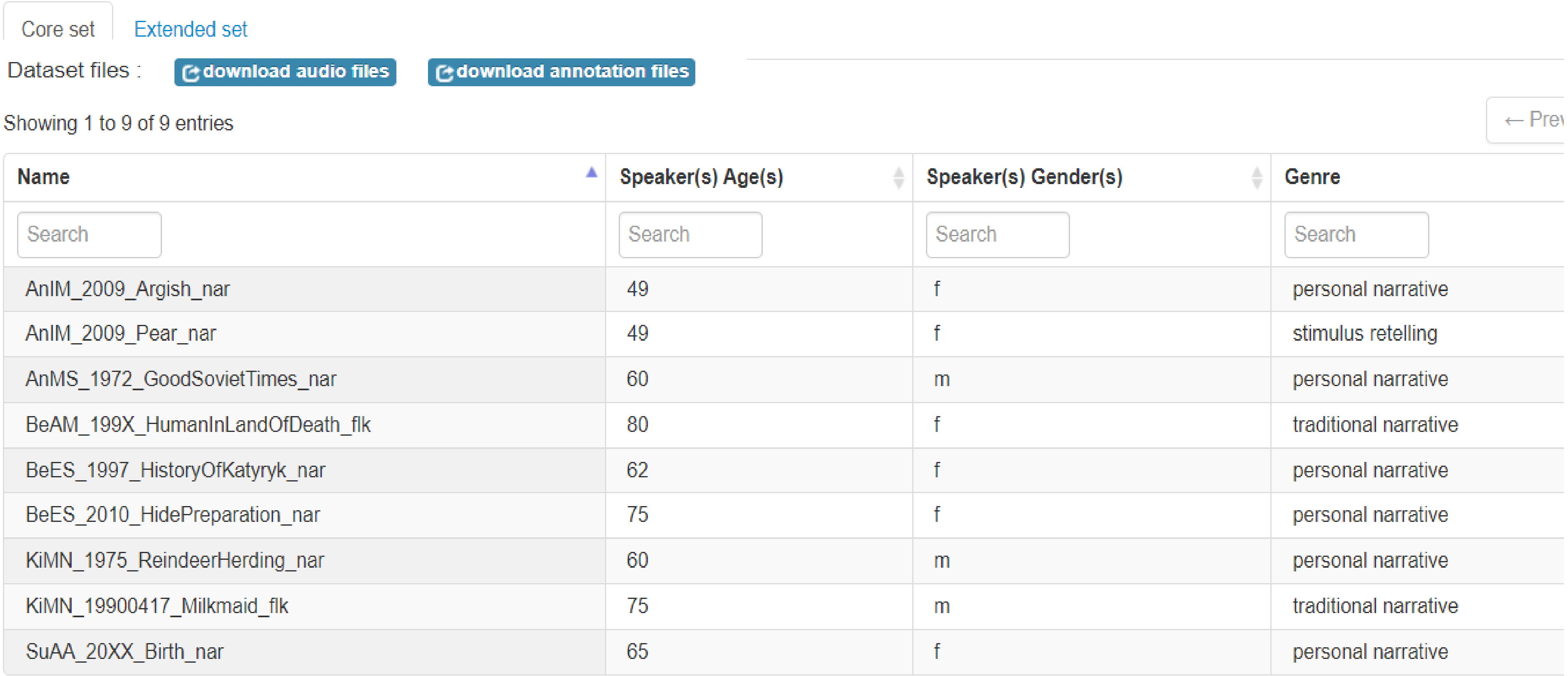
A set of files with further information on the Dolgan DoReCo dataset, including metadata and PIDs is automatically included in each bulk download of files from this dataset.

The Dolgan DoReCo dataset should be cited as follows:

```
Däbritz, Chris Lasse, Nina Kudryakova, Eugénie Stapert and Alexandre Arkhipov. 2022. Dolgan DoReCo dataset. In Seifart, Frank, Ludger
Paschen and Matthew Stave (eds.). Language Documentation Reference Corpus (DoReCo) 1.0. Berlin & Lyon: Leibniz-Zentrum Allgemeine
Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). (https://doi.org/10.34847/nkl.f09eikq3).
```

Please note that when citing this dataset, or any number of DoReCo datasets, it is NOT sufficient to refer to DoReCo as a whole, but the full citation for each individual dataset must be provided, including the name(s) of the creator(s) of each data set.

NAVIGATING TO A DORECO DATASET

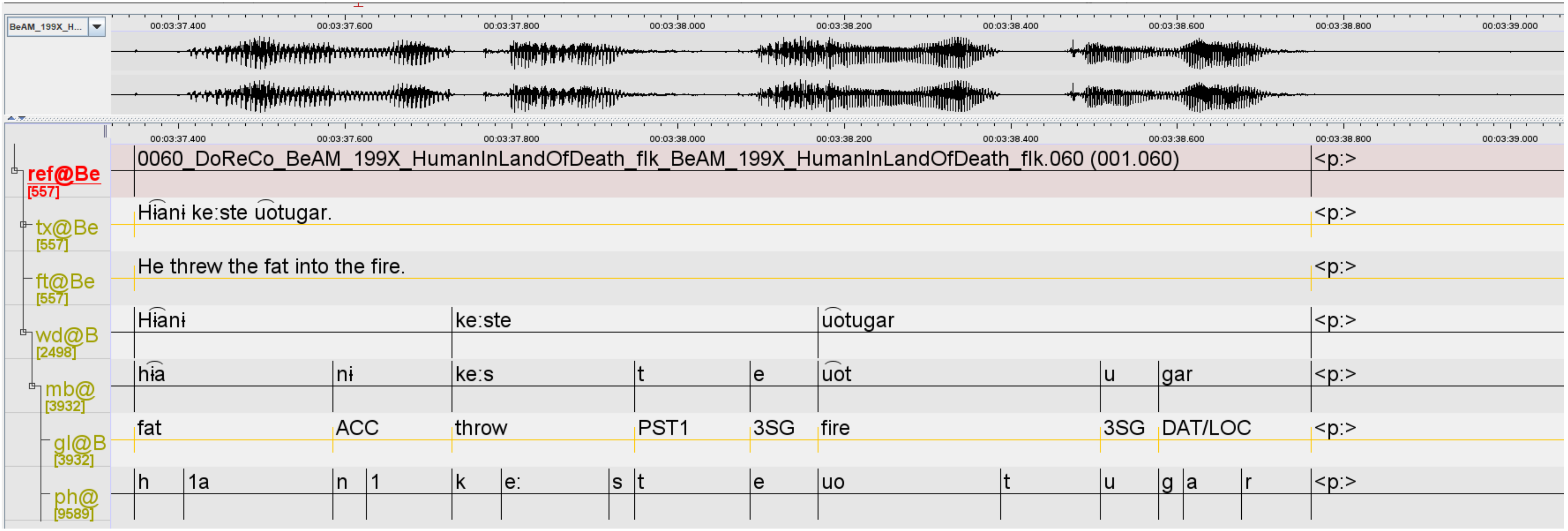


Dataset files :

Showing 1 to 9 of 9 entries

Name	Speaker(s) Age(s)	Speaker(s) Gender(s)	Genre
Search	Search	Search	Search
AnIM_2009_Argish_nar	49	f	personal narrative
AnIM_2009_Pear_nar	49	f	stimulus retelling
AnMS_1972_GoodSovietTimes_nar	60	m	personal narrative
BeAM_199X_HumanInLandOfDeath_flk	80	f	traditional narrative
BeES_1997_HistoryOfKatyryk_nar	62	f	personal narrative
BeES_2010_HidePreparation_nar	75	f	personal narrative
KiMN_1975_ReindeerHerding_nar	60	m	personal narrative
KiMN_19900417_Milkmaid_flk	75	m	traditional narrative
SuAA_20XX_Birth_nar	65	f	personal narrative

ANNOTATION FILES (1/4): ELAN EAF



ref@Be [557] 0060_DoReCo_BeAM_199X_HumanInLandOfDeath_flk_BeAM_199X_HumanInLandOfDeath_flk.060 (001.060) <p:>

tx@Be [557] Hɨ͡anɨ ke:ste u͡otugar. <p:>

ft@Be [557] He threw the fat into the fire. <p:>

wd@B [2498] Hɨ͡anɨ ke:ste u͡otugar <p:>

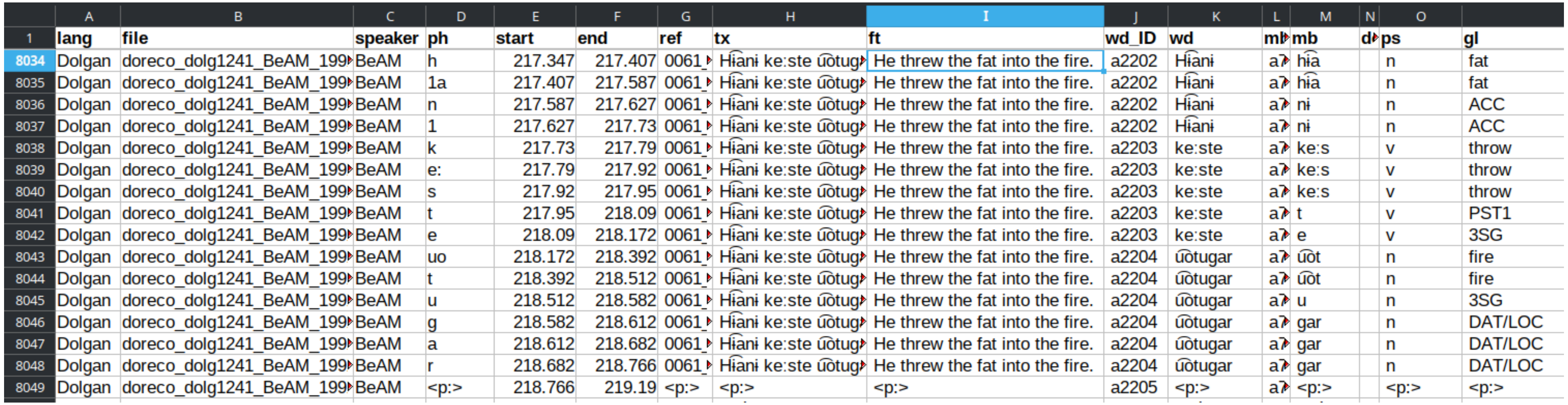
mb@ [3932] hɨ͡a nɨ ke:s t e u͡ot u gar <p:>

gl@B [3932] fat ACC throw PST1 3SG fire 3SG DAT/LOC <p:>

ph@ [9589] h 1a n 1 k e: s t e uo t u g a r <p:>

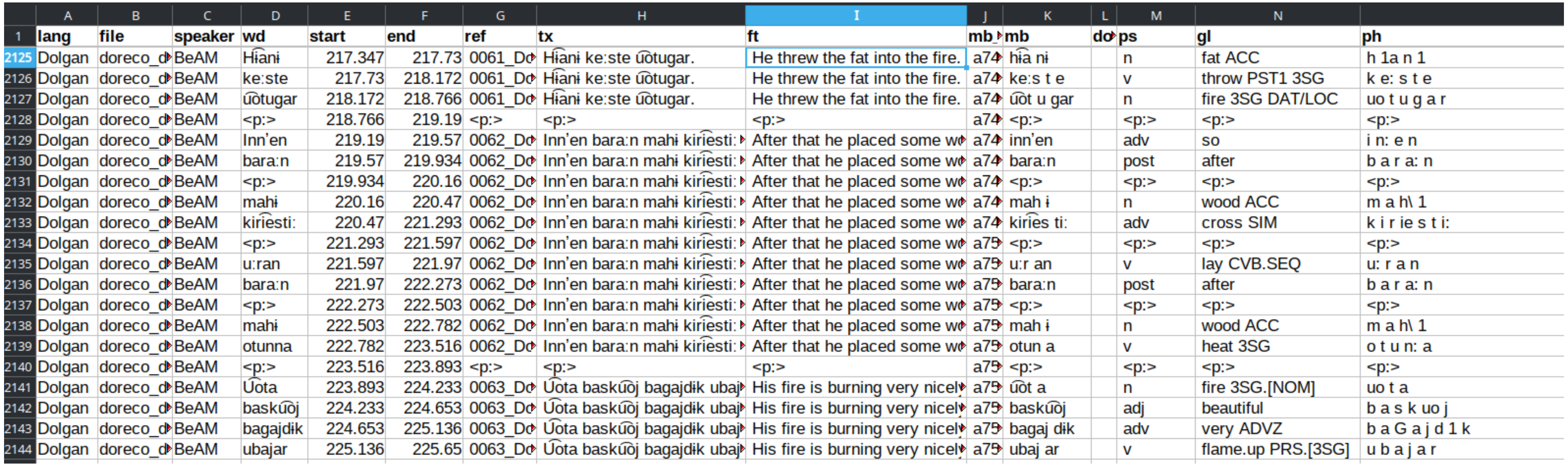
Däbritz, Chris Lasse, Nina Kudryakova, Eugénie Stapert and Alexandre Arkhipov. 2022. Dolgan DoReCo dataset. In Seifart, Frank, Ludger Paschen and Matthew Stave (eds.). *Language Documentation Reference Corpus (DoReCo) 1.0*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). (Accessed on 29/07/2022).

Annotation files (2/4): csv (ph level)



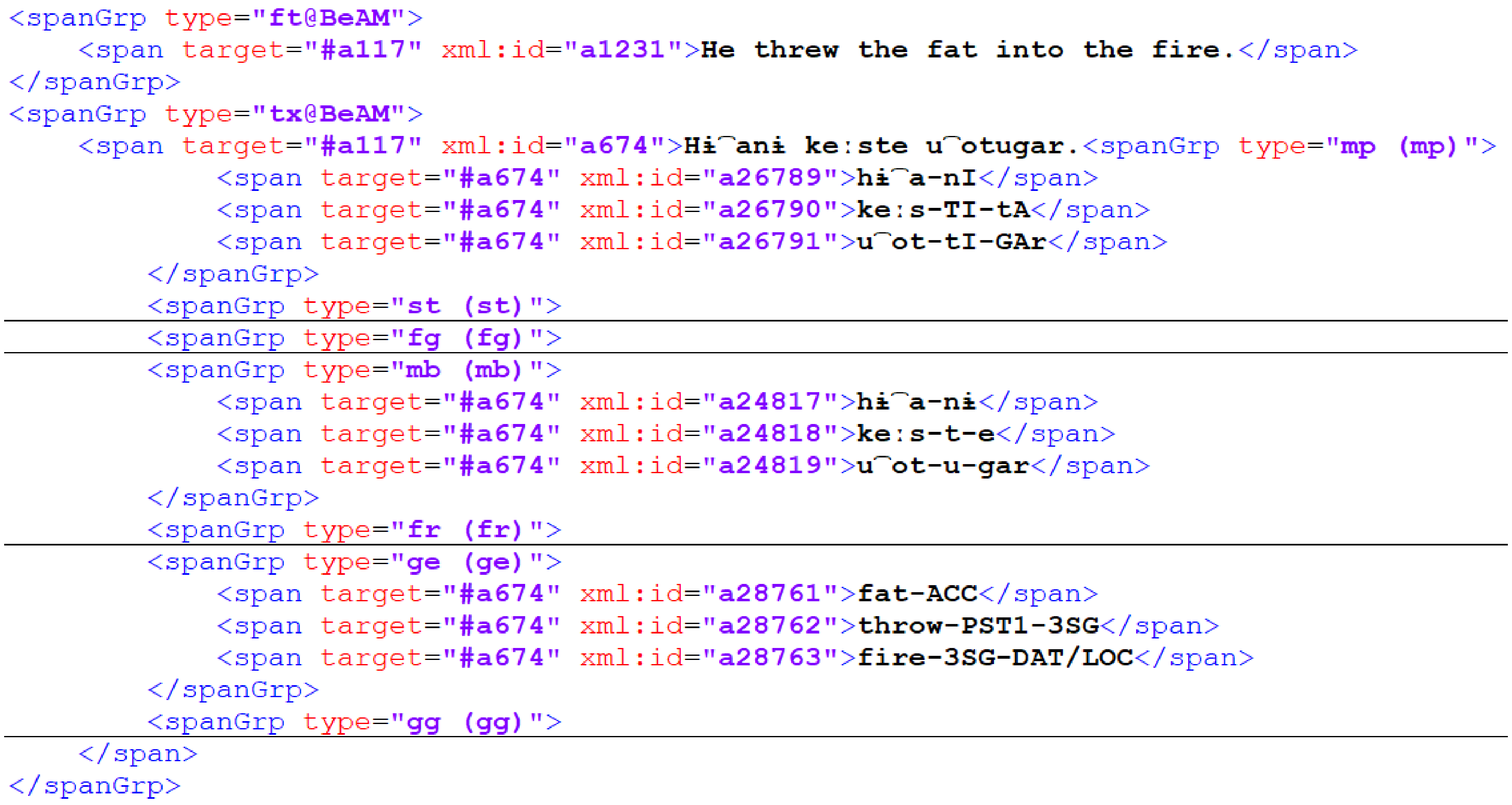
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	lang	file	speaker	ph	start	end	ref	tx	ft	wd_ID	wd	ml	mb	d	ps	gl
8034	Dolgan	doreco_dolg1241_BeAM_199	BeAM	h	217.347	217.407	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2202	Hɨ͡anɨ	a7	hɨ͡a		n	fat
8035	Dolgan	doreco_dolg1241_BeAM_199	BeAM	1a	217.407	217.587	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2202	Hɨ͡anɨ	a7	hɨ͡a		n	fat
8036	Dolgan	doreco_dolg1241_BeAM_199	BeAM	n	217.587	217.627	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2202	Hɨ͡anɨ	a7	nɨ		n	ACC
8037	Dolgan	doreco_dolg1241_BeAM_199	BeAM	1	217.627	217.73	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2202	Hɨ͡anɨ	a7	nɨ		n	ACC
8038	Dolgan	doreco_dolg1241_BeAM_199	BeAM	k	217.73	217.79	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2203	ke:ste	a7	ke:s		v	throw
8039	Dolgan	doreco_dolg1241_BeAM_199	BeAM	e:	217.79	217.92	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2203	ke:ste	a7	ke:s		v	throw
8040	Dolgan	doreco_dolg1241_BeAM_199	BeAM	s	217.92	217.95	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2203	ke:ste	a7	ke:s		v	throw
8041	Dolgan	doreco_dolg1241_BeAM_199	BeAM	t	217.95	218.09	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2203	ke:ste	a7	t		v	PST1
8042	Dolgan	doreco_dolg1241_BeAM_199	BeAM	e	218.09	218.172	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2203	ke:ste	a7	e		v	3SG
8043	Dolgan	doreco_dolg1241_BeAM_199	BeAM	uo	218.172	218.392	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2204	u͡otugar	a7	u͡ot		n	fire
8044	Dolgan	doreco_dolg1241_BeAM_199	BeAM	t	218.392	218.512	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2204	u͡otugar	a7	u͡ot		n	fire
8045	Dolgan	doreco_dolg1241_BeAM_199	BeAM	u	218.512	218.582	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2204	u͡otugar	a7	u		n	3SG
8046	Dolgan	doreco_dolg1241_BeAM_199	BeAM	g	218.582	218.612	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2204	u͡otugar	a7	gar		n	DAT/LOC
8047	Dolgan	doreco_dolg1241_BeAM_199	BeAM	a	218.612	218.682	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2204	u͡otugar	a7	gar		n	DAT/LOC
8048	Dolgan	doreco_dolg1241_BeAM_199	BeAM	r	218.682	218.766	0061_	Hɨ͡anɨ ke:ste u͡otug	He threw the fat into the fire.	a2204	u͡otugar	a7	gar		n	DAT/LOC
8049	Dolgan	doreco_dolg1241_BeAM_199	BeAM	<p:>	218.766	219.19	<p:>	<p:>	<p:>	a2205	<p:>	a7	<p:>		<p:>	<p:>

ANNOTATION FILES (2/4 CONT'D): CSV (WD LEVEL)



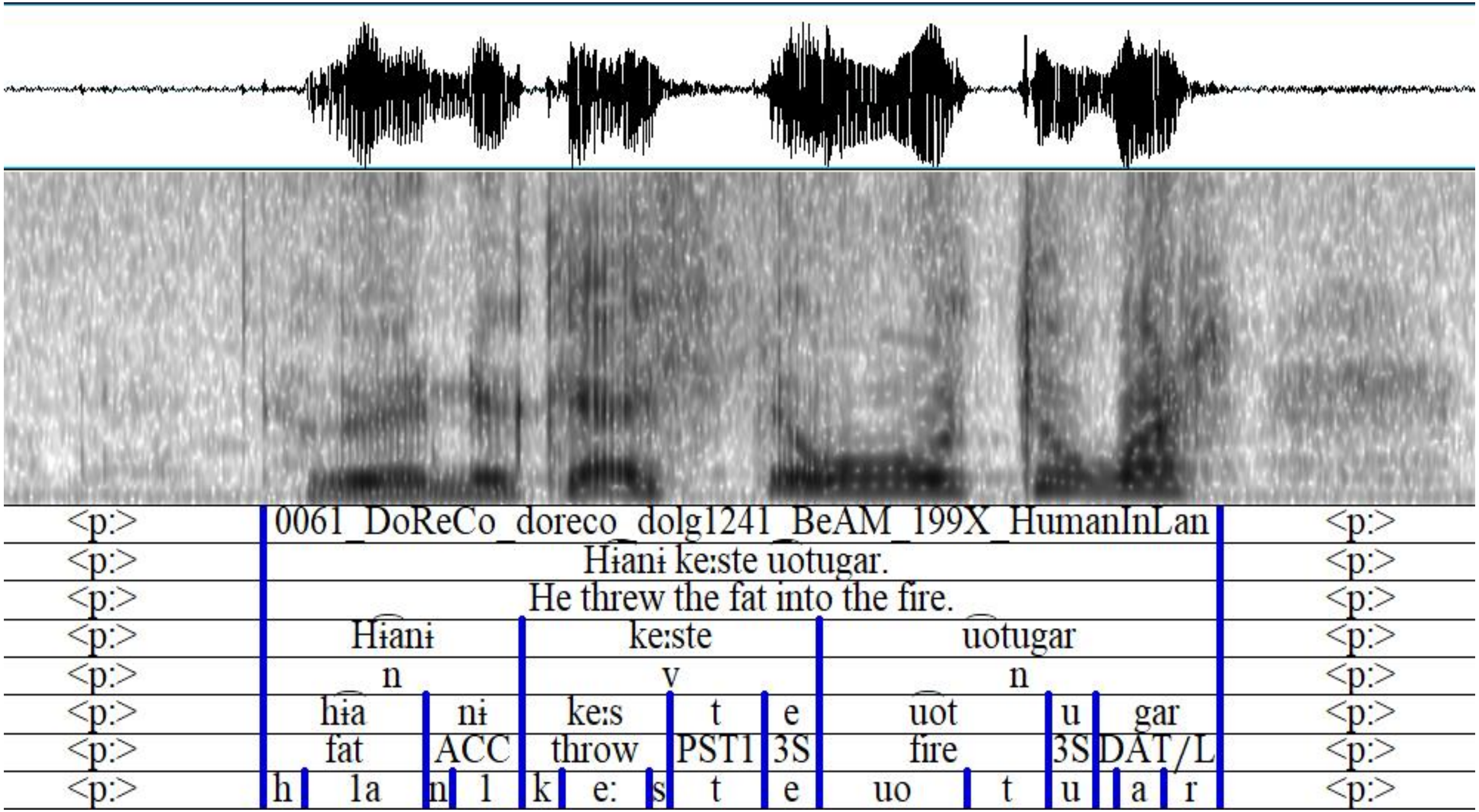
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	**lang**	**file**	**speaker**	**wd**	**start**	**end**	**ref**	**tx**	**ft**	**mb_**	**mb**	**do**	**ps**	**gl**	**ph**
2125	Dolgan	doreco_d	BeAM	Hɨ͡anɨ	217.347	217.73	0061_Do	Hɨ͡anɨ ke:ste u͡otugar.	He threw the fat into the fire.	a74	hɨ͡a nɨ		n	fat ACC	h 1a n 1
2126	Dolgan	doreco_d	BeAM	ke:ste	217.73	218.172	0061_Do	Hɨ͡anɨ ke:ste u͡otugar.	He threw the fat into the fire.	a74	ke:s t e		v	throw PST1 3SG	k e: s t e
2127	Dolgan	doreco_d	BeAM	u͡otugar	218.172	218.766	0061_Do	Hɨ͡anɨ ke:ste u͡otugar.	He threw the fat into the fire.	a74	u͡ot u gar		n	fire 3SG DAT/LOC	uo t u g a r
2128	Dolgan	doreco_d	BeAM	<p:>	218.766	219.19	<p:>	<p:>	<p:>	a74	<p:>		<p:>	<p:>	<p:>
2129	Dolgan	doreco_d	BeAM	Inn'en	219.19	219.57	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a74	inn'en		adv	so	i n: e n
2130	Dolgan	doreco_d	BeAM	bara:n	219.57	219.934	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a74	bara:n		post	after	b a r a: n
2131	Dolgan	doreco_d	BeAM	<p:>	219.934	220.16	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a74	<p:>		<p:>	<p:>	<p:>
2132	Dolgan	doreco_d	BeAM	mahɨ	220.16	220.47	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a74	mah ɨ		n	wood ACC	m a h\ 1
2133	Dolgan	doreco_d	BeAM	kiri͡esti:	220.47	221.293	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a74	kiri͡es ti:		adv	cross SIM	k i r ie s t i:
2134	Dolgan	doreco_d	BeAM	<p:>	221.293	221.597	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a75	<p:>		<p:>	<p:>	<p:>
2135	Dolgan	doreco_d	BeAM	u:ran	221.597	221.97	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a75	u:r an		v	lay CVB.SEQ	u: r a n
2136	Dolgan	doreco_d	BeAM	bara:n	221.97	222.273	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a75	bara:n		post	after	b a r a: n
2137	Dolgan	doreco_d	BeAM	<p:>	222.273	222.503	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a75	<p:>		<p:>	<p:>	<p:>
2138	Dolgan	doreco_d	BeAM	mahɨ	222.503	222.782	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a75	mah ɨ		n	wood ACC	m a h\ 1
2139	Dolgan	doreco_d	BeAM	otunna	222.782	223.516	0062_Do	Inn'en bara:n mahɨ kiri͡esti:	After that he placed some wo	a75	otun a		v	heat 3SG	o t u n: a
2140	Dolgan	doreco_d	BeAM	<p:>	223.516	223.893	<p:>	<p:>	<p:>	a75	<p:>		<p:>	<p:>	<p:>
2141	Dolgan	doreco_d	BeAM	U͡ota	223.893	224.233	0063_Do	U͡ota basku͡oj bagajdɨk ubaj	His fire is burning very nicel	a75	u͡ot a		n	fire 3SG.[NOM]	uo t a
2142	Dolgan	doreco_d	BeAM	basku͡oj	224.233	224.653	0063_Do	U͡ota basku͡oj bagajdɨk ubaj	His fire is burning very nicel	a75	basku͡oj		adj	beautiful	b a s k uo j
2143	Dolgan	doreco_d	BeAM	bagajdɨk	224.653	225.136	0063_Do	U͡ota basku͡oj bagajdɨk ubaj	His fire is burning very nicel	a75	bagaj dɨk		adv	very ADVZ	b a G a j d 1 k
2144	Dolgan	doreco_d	BeAM	ubajar	225.136	225.65	0063_Do	U͡ota basku͡oj bagajdɨk ubaj	His fire is burning very nicel	a75	ubaj ar		v	flame.up PRS.[3SG]	u b a j a r

ANNOTATION FILES (3/4): TEI XML



```
<spanGrp type="ft@BeAM">
    <span target="#a117" xml:id="a1231">He threw the fat into the fire.</span>
</spanGrp>
<spanGrp type="tx@BeAM">
    <span target="#a117" xml:id="a674">Hɨ⁀anɨ keːste u⁀otugar.<spanGrp type="mp (mp)">
            <span target="#a674" xml:id="a26789">hɨ⁀a-nI</span>
            <span target="#a674" xml:id="a26790">keːs-TI-tA</span>
            <span target="#a674" xml:id="a26791">u⁀ot-tI-GAr</span>
        </spanGrp>
        <spanGrp type="st (st)">
        <spanGrp type="fg (fg)">
        <spanGrp type="mb (mb)">
            <span target="#a674" xml:id="a24817">hɨ⁀a-nɨ</span>
            <span target="#a674" xml:id="a24818">keːs-t-e</span>
            <span target="#a674" xml:id="a24819">u⁀ot-u-gar</span>
        </spanGrp>
        <spanGrp type="fr (fr)">
        <spanGrp type="ge (ge)">
            <span target="#a674" xml:id="a28761">fat-ACC</span>
            <span target="#a674" xml:id="a28762">throw-PST1-3SG</span>
            <span target="#a674" xml:id="a28763">fire-3SG-DAT/LOC</span>
        </spanGrp>
        <spanGrp type="gg (gg)">
    </span>
</spanGrp>
```

ANNOTATION FILES (4/4): TEXTGRID



<p:> 0061_DoReCo_doreco_dolg1241_BeAM_199X_HumanInLan <p:>

<p:> Hɨanɨ ke:ste uotugar. <p:>

<p:> He threw the fat into the fire. <p:>

<p:> Hɨ͡anɨ ke:ste u͡otugar <p:>

<p:> n v n <p:>

<p:> hɨ͡a nɨ ke:s t e u͡ot u gar <p:>

<p:> fat ACC throw PST1 3S fire 3S DAT/L <p:>

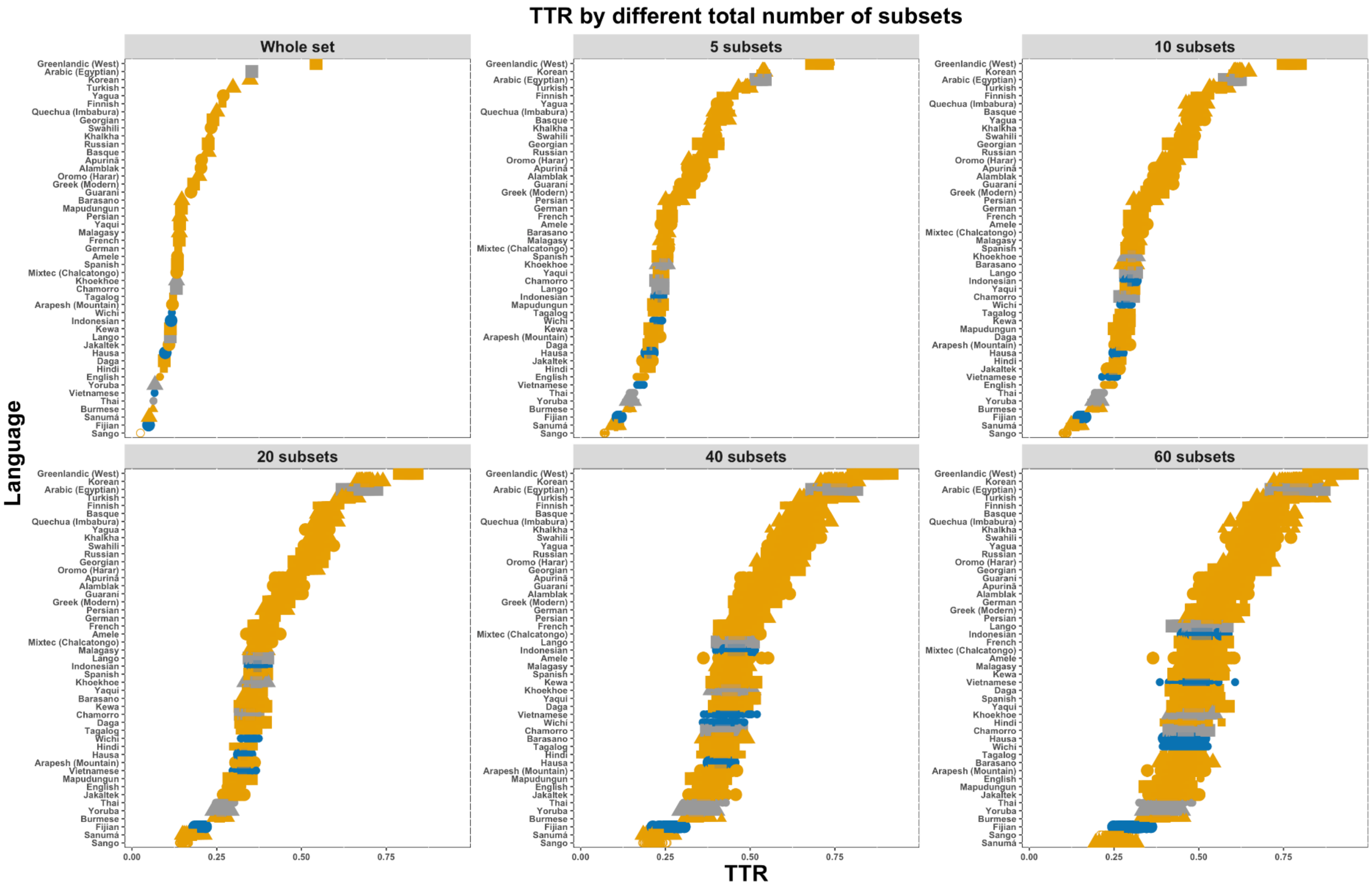
<p:> h ɨa n ɨ k e: s t e uo t u a r <p:>

OA Towards robust complexity indices in linguistic typology

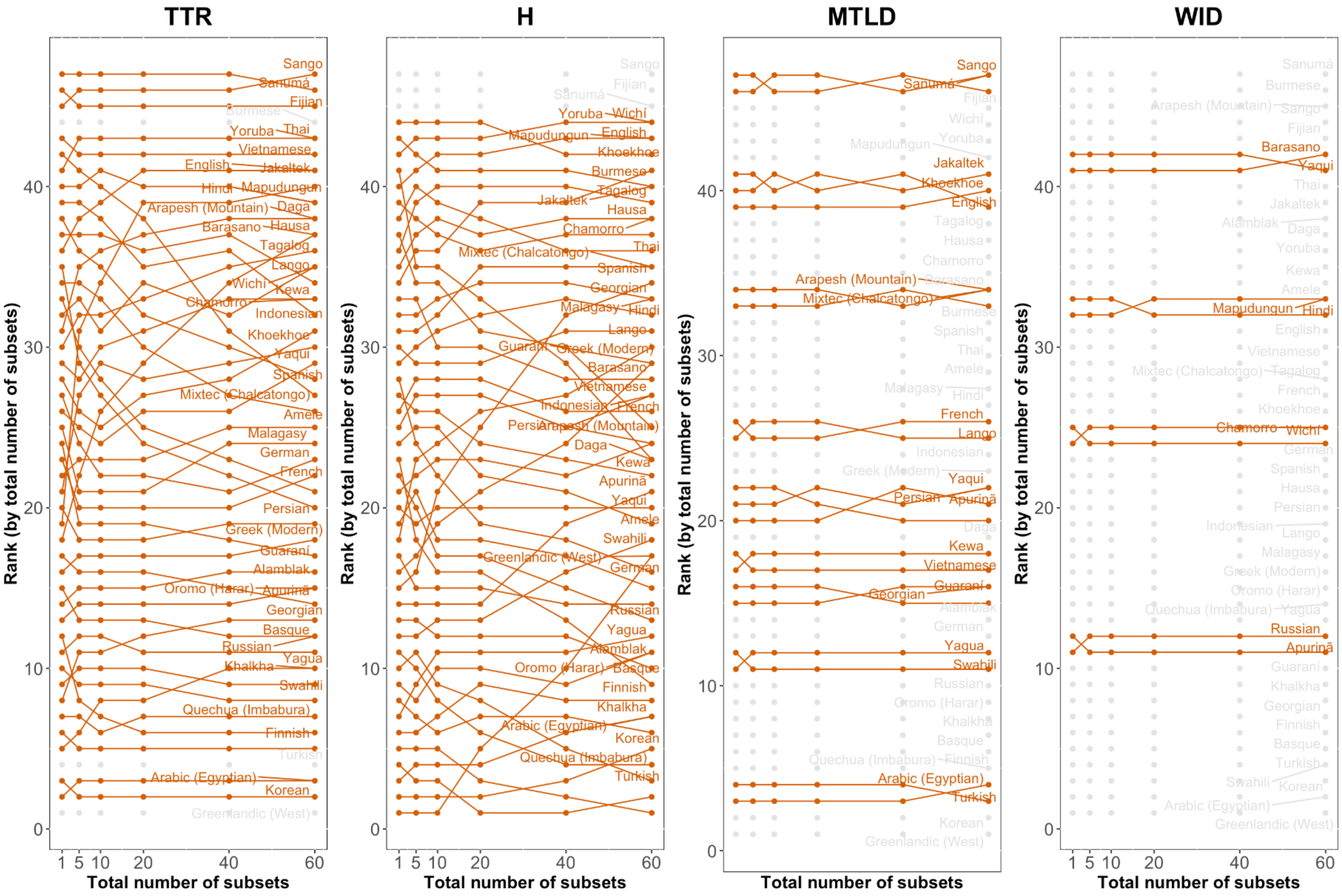
A corpus-based assessment

Author(s): Yoon Mi Oh[1], François Pellegrino[2]

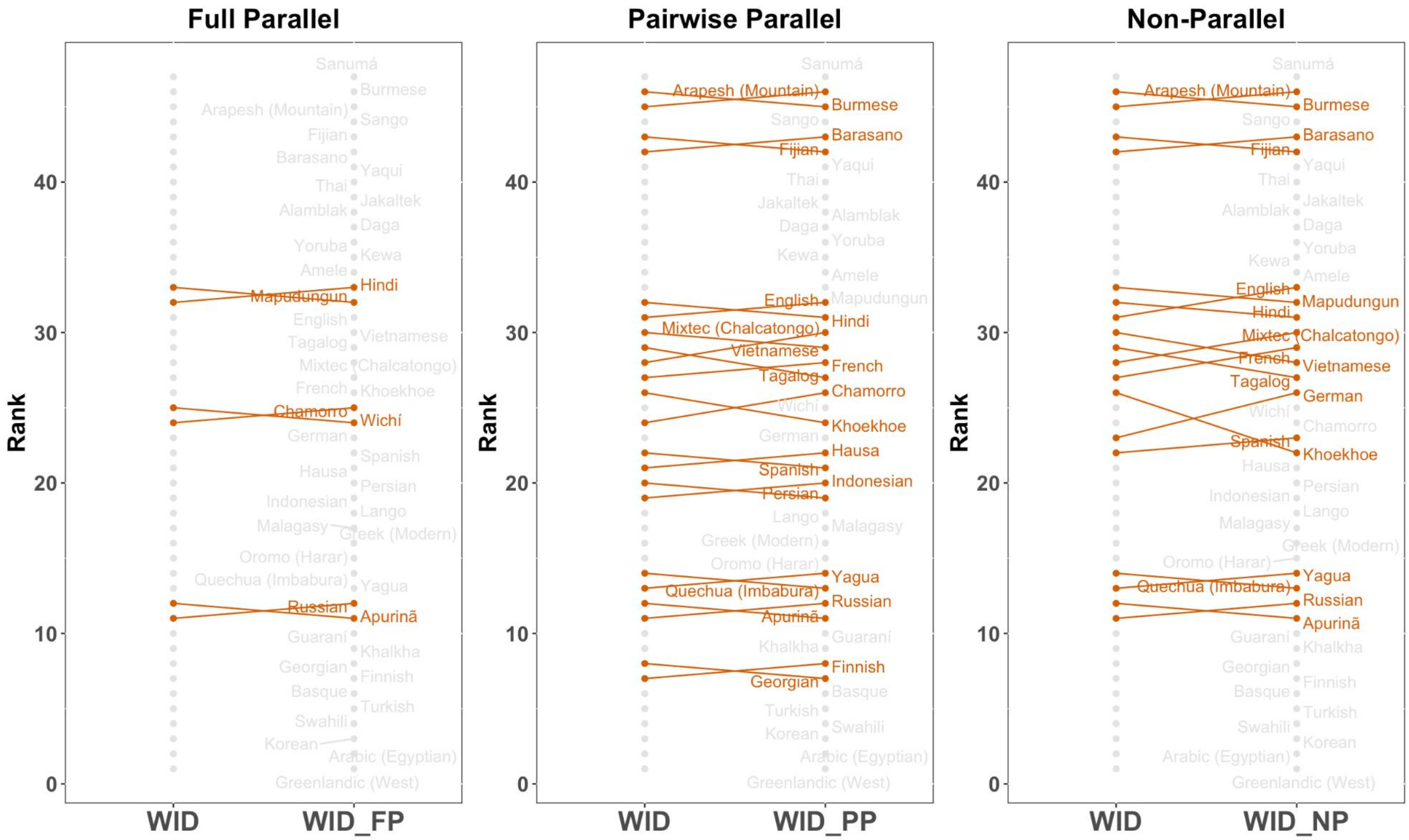
Figure 5. Languages ranked by Type-Token Ratio (TTR, x-axis). Each panel corresponds to a different corpus sampling configuration, from one unique sample (Whole set, top left panel) to 60 samples (bottom right panel). In each panel, languages are ranked by average TTR over the subsets, potentially leading to differences in ranking across the panels



Oh, Y. M., & Pellegrino, F. (2022). Towards robust complexity indices in linguistic typology: A corpus-based assessment. *Studies in Language*.



TTR
H
MTLD
WID
Rank (by total number of subsets)
Total number of subsets
0
10
20
30
40
1
5
10
20
40
60
Sango
Sanumá
Fijian
Burmese
Yoruba
Thai
Vietnamese
English
Jakaltek
Hindi
Mapudungun
Arapesh (Mountain)
Daga
Barasano
Hausa
Tagalog
Lango
Wichí
Kewa
Chamorro
Indonesian
Khoekhoe
Yaqui
Spanish
Mixtec (Chalcatongo)
Amele
Malagasy
German
French
Persian
Greek (Modern)
Guaraní
Alamblak
Oromo (Harar)
Apurinã
Georgian
Basque
Russian
Yagua
Khalkha
Swahili
Quechua (Imbabura)
Finnish
Turkish
Arabic (Egyptian)
Korean
Greenlandic (West)
Sango
Fijian
Sanumá
Yoruba
Wichí
Mapudungun
English
Khoekhoe
Burmese
Tagalog
Jakaltek
Hausa
Chamorro
Thai
Mixtec (Chalcatongo)
Spanish
Georgian
Malagasy
Hindi
Lango
Guaraní
Greek (Modern)
Barasano
Vietnamese
Indonesian
French
Persian
Arapesh (Mountain)
Daga
Kewa
Apurinã
Yaqui
Amele
Swahili
Greenlandic (West)
German
Russian
Yagua
Alamblak
Oromo (Harar)
Basque
Finnish
Khalkha
Arabic (Egyptian)
Korean
Quechua (Imbabura)
Turkish
Sango
Sanumá
Fijian
Wichí
Yoruba
Mapudungun
Jakaltek
Khoekhoe
English
Tagalog
Hausa
Chamorro
Arapesh (Mountain)
Barasano
Mixtec (Chalcatongo)
Burmese
Spanish
Thai
Amele
Malagasy
Hindi
French
Lango
Indonesian
Greek (Modern)
Yaqui
Persian
Apurinã
Daga
Kewa
Vietnamese
Georgian
Guaraní
Alamblak
German
Yagua
Swahili
Russian
Oromo (Harar)
Khalkha
Basque
Quechua (Imbabura)
Finnish
Arabic (Egyptian)
Turkish
Korean
Greenlandic (West)
Sanumá
Burmese
Arapesh (Mountain)
Sango
Fijian
Barasano
Yaqui
Thai
Jakaltek
Alamblak
Daga
Yoruba
Kewa
Amele
Mapudungun
Hindi
English
Vietnamese
Mixtec (Chalcatongo)
Tagalog
French
Khoekhoe
Chamorro
Wichí
German
Spanish
Hausa
Persian
Indonesian
Lango
Malagasy
Greek (Modern)
Oromo (Harar)
Quechua (Imbabura)
Yagua
Russian
Apurinã
Guaraní
Khalkha
Georgian
Finnish
Basque
Turkish
Swahili
Korean
Arabic (Egyptian)
Greenlandic (West)


Full Parallel
Pairwise Parallel
Non-Parallel
Rank
0
10
20
30
40
WID
WID_FP
WID_PP
WID_NP
Sanumá
Burmese
Arapesh (Mountain)
Sango
Fijian
Barasano
Yaqui
Thai
Jakaltek
Alamblak
Daga
Yoruba
Kewa
Amele
Hindi
Mapudungun
English
Vietnamese
Tagalog
Mixtec (Chalcatongo)
French
Khoekhoe
Chamorro
Wichí
German
Spanish
Hausa
Persian
Indonesian
Lango
Malagasy
Greek (Modern)
Oromo (Harar)
Quechua (Imbabura)
Yagua
Russian
Apurinã
Guaraní
Khalkha
Georgian
Finnish
Basque
Turkish
Swahili
Korean
Arabic (Egyptian)
Greenlandic (West)