Linguistic Typology for NLP researchers: Methods and Resources in the 21st century

Harald Hammarström
harald.hammarstrom@lingfil.uu.se

24 Jan 2026 Yerevan

Today: Truly diverse NLP resources

§1 Bible Corpora

§2 DoReCo

§3 Dream Corpus

§4 Experiment: Prefixing and Suffixing in the Languages of the World

§1 Bible Corpora

- The Bible is by far the most translated text (Wycliffe Bible Translators 2025)

Coverage	# languages
Full Bible	776
New Testament	1 798
Scripture Portions	1 433
	4007

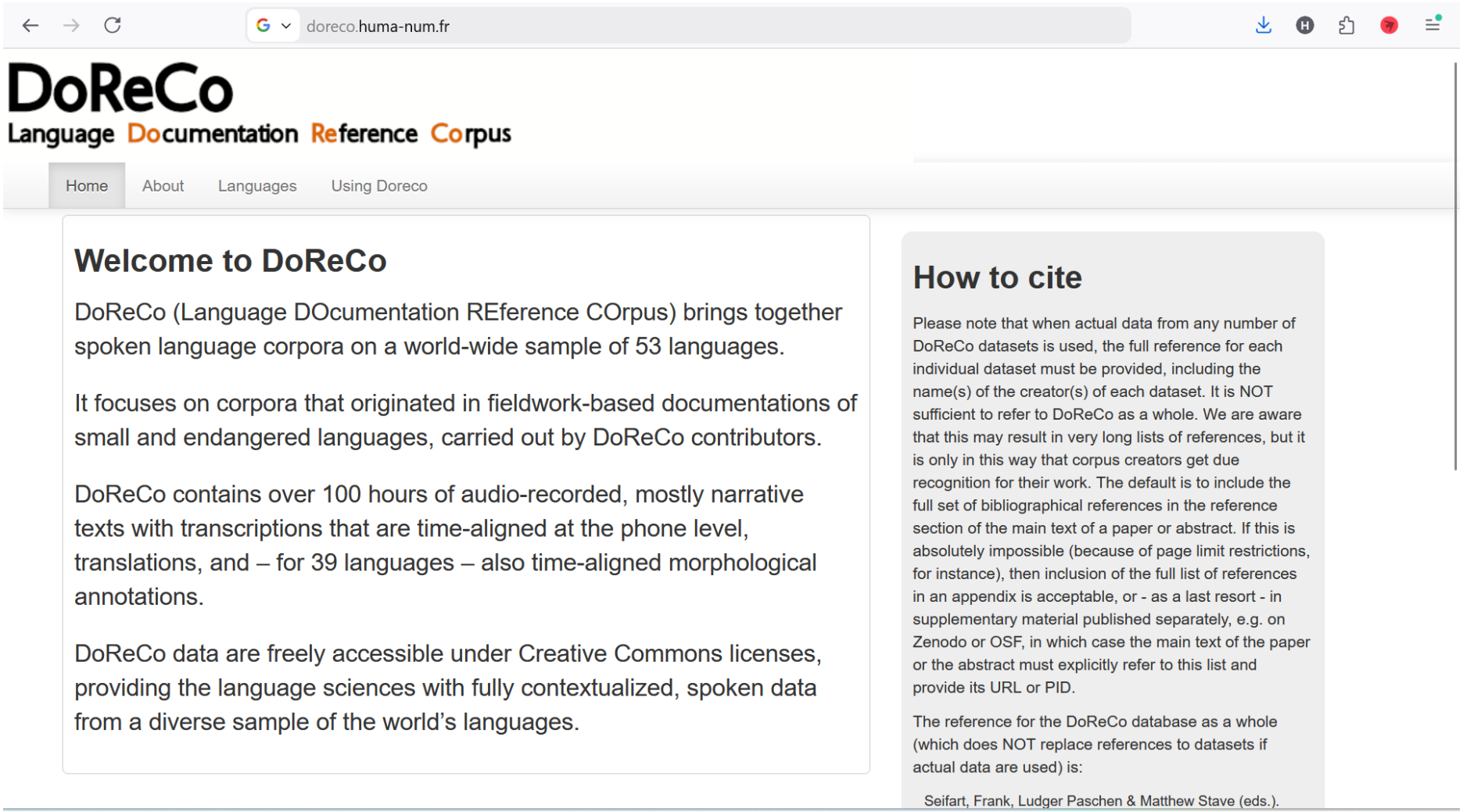
- New Testament translations in digital raw text form
  - 1 030 languages with a near-complete New Testament translation (McCarthy et al. 2020)
  - Also earlier/parallell efforts of similar size (Mayer and Cysouw 2014)

§1 Bible Corpora: Properties

- Collections indexed by ISO 639-3
- Aligned at the verse level
  - Surprisingly many translations do minor violations to the verse-to-verse principle
- Each translation is written in some orthography which is almost never IPA
- Example Abua [aau] of Papua New Guinea

```
40001012    enekwei israel hom babilon mon ma lwayr so meyki mon hokwe ,
40001017    ney - nona abraham hiy ma kekie nona nona hain mokwe ,  seme ,
            devit hiy lousne ,  ney - nona somokwe sankaw ,  14 .  sawk
            devit hiy ma sahre kekie nona nona ha mokwe ,  seme ,  israel
            hom babilon mon ma lwayr hom non ,  ney - nona somokwe seyr
            sankaw ,  14 .  enekwei israel hom babilon mon ma lwayr mokwe ,
            sehe ,  krais hiy ma lousne hiy non ,  ney - nona mokwe sankaw
            kekie nona ha ,  14 .  [ luk 2 : 17 ]
40001018    jisas krais se ma liwak enekwei hokwe ,  senkin lousne . hyopouh
            maria ke kokwe josep se me peykyay iwak .  sawk peyr hoh non
...         ...
```

§2 DoReCo: A well-curated diverse corpus



`https://doreco.huma-num.fr/` (Seifart et al. 2024, Seifart 2021)

§2 DoReCo: Data and Transcription

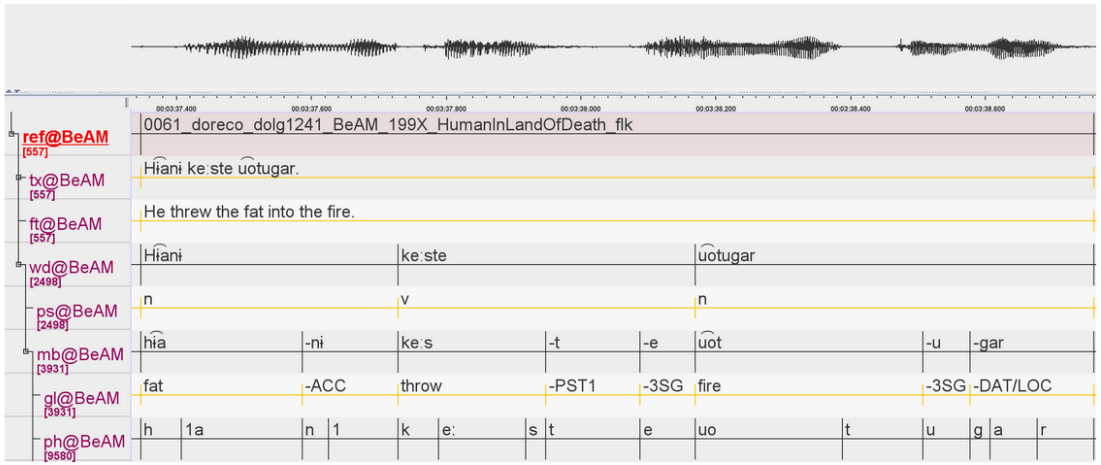
doreco.huma-num.fr/about

Data contained in DoReCo

The DoReCo database contains corpora on 53 languages from 33 top-level language families (as classified in Glottolog), covering languages from all inhabited continents and all linguistic macro-areas. Most of these data were originally collected in the context of language documentation projects focusing on preserving linguistic practices and traditions. They contain mostly monological, narrative texts, though some texts also represent conversations and stimulus retelling. Most datasets were extracted from larger collections archived in repositories such as TLA or ELAR.

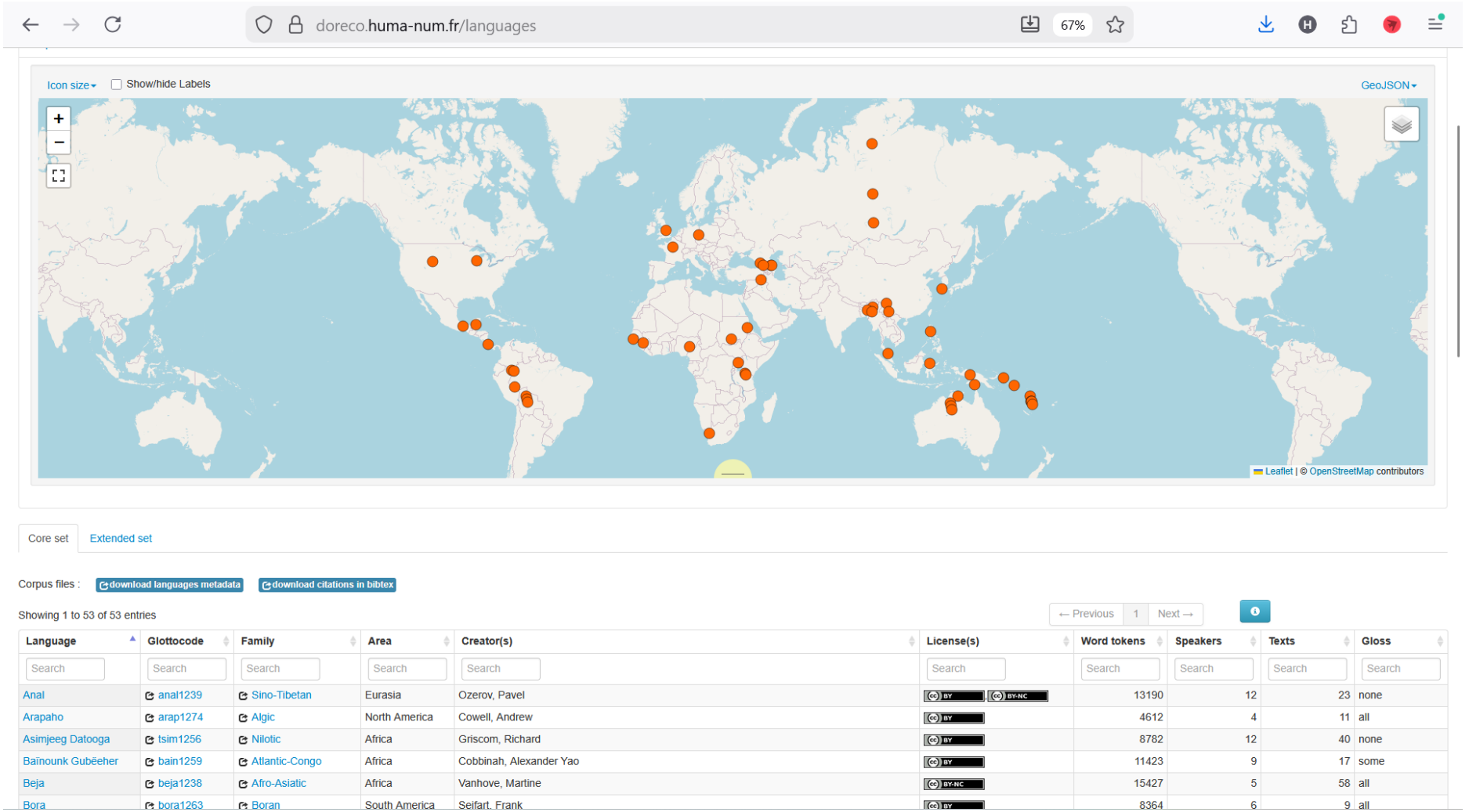
In total, DoReCo contains over 100 hours of recordings with almost half a million transcribed words that are time-aligned at the word and phone levels. The minimum amount of data per language is 35,000 phones (although some datasets are slightly below that mark), corresponding to more than 10,000 word tokens for isolating languages. The total number of core texts is 934, equivalent to 17 texts on average per language. Numbers of unique speakers per core dataset range from 1 (Kamas, Texistepec Popoluca, Yongning Na) to 30 (Urum). All texts are also translated, mostly into English, but in some cases also Portuguese, German, Russian, Swahili and other languages.

For 39 languages, DoReCo provides time-aligned interlinear morpheme glosses. For most of these 39 languages, additional texts with interlinear glosses that are not time-aligned are contained in the DoReCo extended set. In total, DoReCo provides over 300,000 word tokens of time-aligned interlinear glossed text and another 300,000 word tokens of glossed texts without time alignment. Each DoReDo dataset is accompanied by extensive corpus documentation on orthographic conventions, abbreviations used in glosses, and other useful information.



Example of time alignment at the word, morph and phone levels from the DoReCo Dolgan dataset (Däbritz, Kudryakova, Stapert & Arkhipov 2024).

§2 DoReCo: Diversity and Size



§3 DReaM Corpus: Digitized OCRed descriptive grammars

- Collection of digitized works in **descriptive linguistics** originally written for humans (Virk et al. 2020)
- Focus on minority languages
- Named after the **The Dictionary/Grammar Reading Machine** JPICH-ERC project 2017-2020
- Continually updated (numbers here are as of 2022/02/28)
- Amassed via going through handbooks/overviews systematically to find relevant literature (Hammarström and Nordhoff 2011)

§3 DReaM Corpus: Examples

Language	Description Type	Bibliographic Reference
Tauya [tya]	long grammar	MacDonald, Lorna. (1990) *A Grammar of Tauya* (Mouton Grammar Library 6). Berlin: Mouton de Gruyter. xiii + 385 pp.
Bolon [bof]	grammar	Zoungrana, Ambroise. (1987) *Esquisse phonologique et grammaticale du Bolon (Burkina-Faso)- contribution à la dialectologie mandé*. Université de la Sorbonne Nouvelle (Paris 3) doctoral dissertation. 336pp.
Usila Chinantec [cuc]	grammar sketch, dictionary	a Skinner, Leonard E. & Marlene B. Skinner. (2000) *Diccionario Chinanteco de San Felipe Usila, Oaxaca* (Serie de vocabularios y diccionarios indígenas Mariano Silva y Aceves 43). Coyoacán, México: Instituto Lingüístico de Verano. xxix + 602.
Norwegian Sign Language [nsl]	specific feature	Slowikowska Schröder, Bogumila. (2010) *Imperativ i norsk tegnspråk — en eksplorerende studie av et fenomen innen et visuelt-gestuelt språk [Imperative in Norwegian sign language an explorating study of a phenomenon in a visual-gestural language]*. Univ. of Oslo MA thesis.
Sobei [sob]	phonology	Sterner, Joyce K. (1975) Sobei phonology. Oceanic Linguistics 14. 146–167.
Northern Tujia [tji]	dictionary	Zhang, Weiquan. (2006) *Hàn yǔ tǔjiā yǔ cídiǎn [Chinese-Tujia dictionary]*. Guiyang Shi: Guizhou Minzu Chubanshe. 6 + 20 + 3 + 436pp.
Nisga'a [ncg]	text	Boas, Franz. (1902) *Tsimshian Texts* (Bulletin of American Ethnology 27). Washington: Government Printing Office. 254pp.
Asháninka [cni], Yine [pib], Shipibo-Conibo [shp]	wordlist	Carrasco, Francisco. (1901) *Principales palabras del idioma de las tribus de infieles antis, piros, conibos, sipibos*. Boletín de la Sociedad Geográfica de Lima 11. 204–211.
Dizin [mdx]	minimal	Conti Rossini, Carlo. (1937) *Il Popolo dei Magi nell'Etiopia Meridionale e il suo linguaggio*. In V Sezione: Etnografica-Filologica-Sociologica (Atti del Terzo congresso di Studi Coloniali VI), 108–118. Firenze: Instituto Coloniale Fascista.
Busuu [bju], Bishuo [bwh], Bikya [byb] Kutep [kub], Yukuben [ybl], Akum [aku], Beezen [bnz], Naki [mff]	overview	Breton, Roland. (1995) *Les Furu et leurs Voisins: Découverte et essai de classification d'un groupe de langues en voie d'extinction au Cameroun*. Cahiers des Sciences Humaines 31(1). 17–48.

§3: A Typology of Description Types

points	type	
5	long grammar	extensive description of most elements of the grammar $\sim$ 300 pages and beyond
4	grammar	a description of most elements of the grammar $\sim$ 150 pages and beyond
3	grammar sketch	a less extensive description of many elements of the grammar $\sim$ 50 pages
2	dictionary	$\sim$ 75 pages and beyond
2	text	text material
2	specific feature	description of some element of grammar (i.e., noun class system, verb morphology etc)
1	wordlist	$\sim$ a couple of hundred words
0	minimal	A small number of morphemes
0	overview	Document with meta-information about the language (i.e., where spoken, non-intelligibility to other languages etc.)

§3 DReaM Corpus: Number of Documents

Type	# Documents	# Target Languages
grammar	5 342	2 663
grammar sketch	8 374	3 891
dictionary	3 075	1 932
specific feature	3 789	2 581
phonology	2 159	1 752
text	1 154	1 004
wordlist	4 661	5 712
comparative	7 087	6 258
minimal	2 311	2 570
socling	1 861	1 753
dialectology	279	320
overview	9 107	7 015
ethnographic	5 479	3 232
bibliographical	680	1 075
Total	**41 483**	**7 533**

§3 DReaM Corpus: Document Metadata

- For each document, the following is known (manually curated):
  - Bibliographical details (year, author, place, etc.)
  - Description type (as per previous slides)
  - The target language, i.e., the minority language(s) described
  - The meta-language, i.e., the language used for description (English, Chinese, French, …)
  - The # pages
- This data is available in full via the open-access bibliography `glottolog.org` (Hammarström et al. 2021)

§3 DReaM Corpus: Provenance

1. Out-of-copyright texts digitized by national libraries, archives, scientific societies and other similar entities,
2. Texts posted online with a license to use for research, usually by university libraries and non-profit organizations (notably the Summer Institute of Linguistics), and
3. Texts under publisher copyright where data mining, but not dissemination, is legal.

I don't know the proportions but my estimate would be something like 15%, 15% and 70%

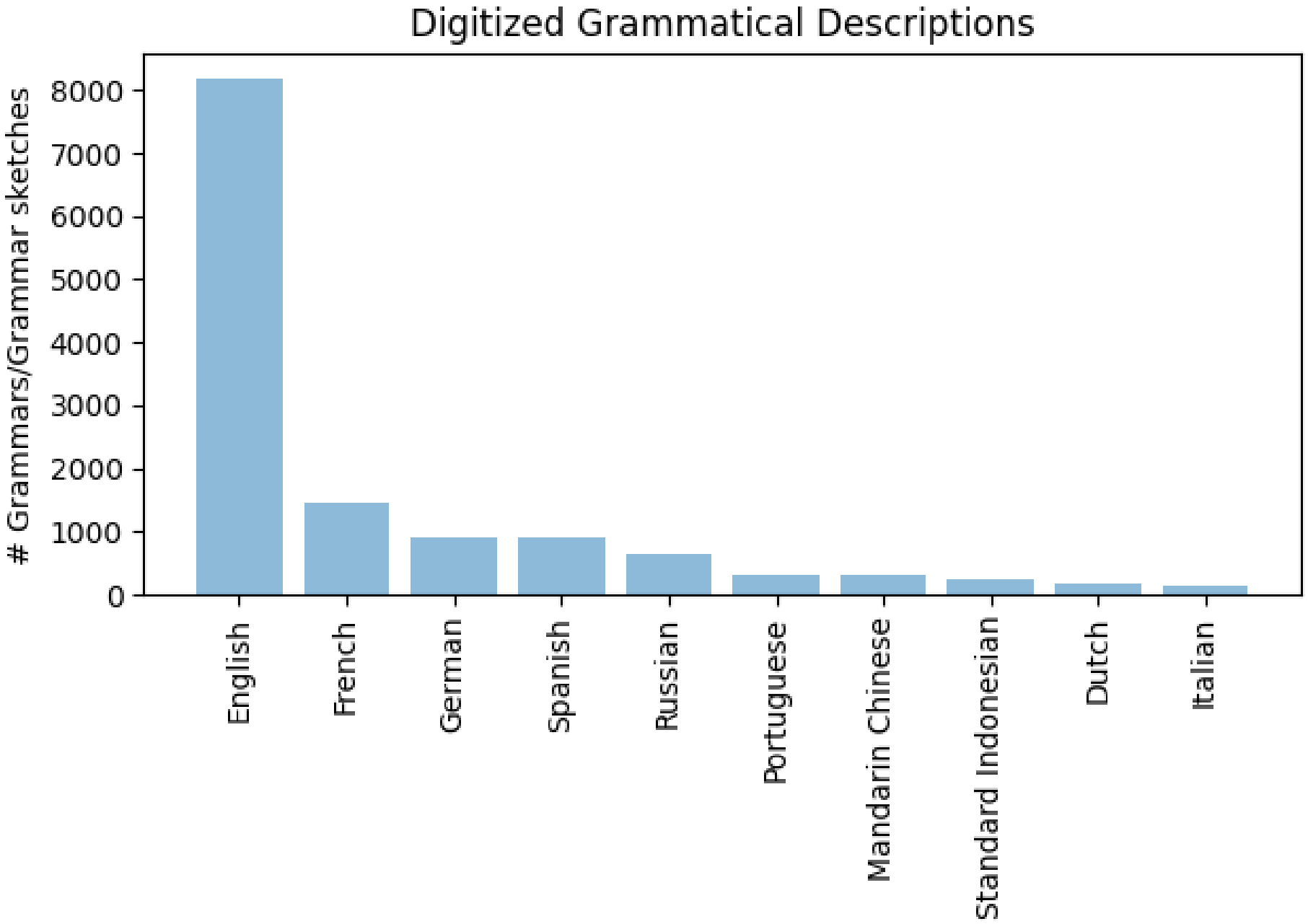
§3 DReaM Corpus: Grammar Subset

- The subset consisting of `grammar` and `grammar sketch` are of special value as they are typically focussing on one language each and seek to cover most/all aspects of that language

Type	# Documents	# Languages
`grammar`	5 342	2 663
`grammar sketch`	8 374	3 891
	13 716	4 613

- The total number of languages for which a grammar/sketch exists at all is 4694, so coverage is $\frac{4613}{4694} \approx 98.3\%$

§3: Meta-Languages of Grammar Subset



§3: Meta-Languages of Grammar Subset: Numbers

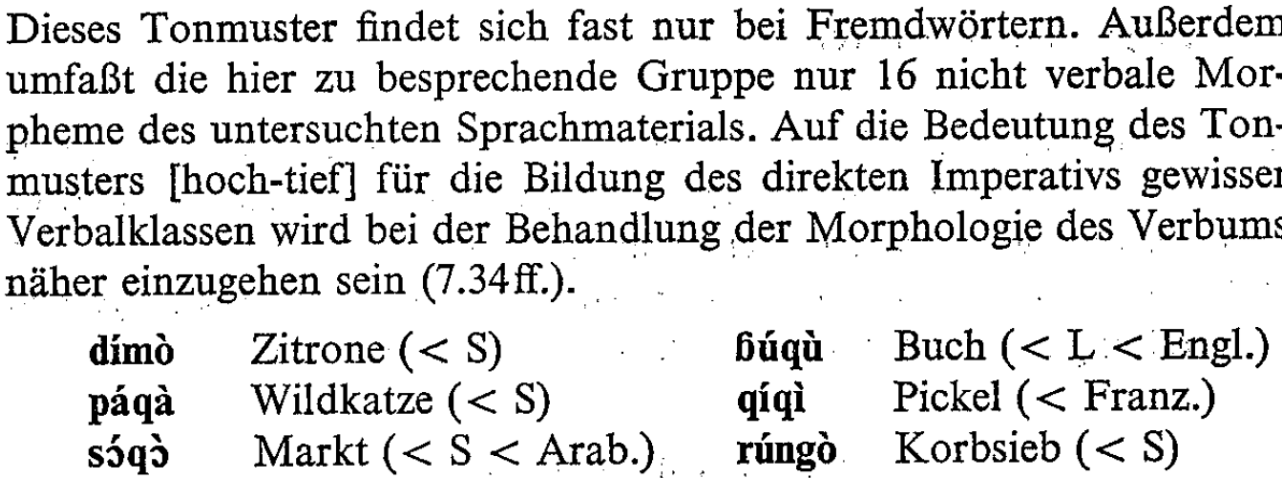
- 92 total meta-languages grammars (125 total all documents)

Meta-language		# languages	# documents
English	eng	3 662	8 162
French	fra	878	1446
German	deu	661	922
Spanish	spa	413	910
Russian	rus	336	639
Mandarin Chinese	cmn	202	302
Portuguese	por	152	303
Standard Indonesian	ind	140	228
Dutch	nld	114	185
Italian	ita	93	151
…	…	…	…

§3: Optical Character Recognition (OCR)

- Subset of documents born digital (proportion not known, guess 10%)
- These and other documents OCRed
  - ABBYY Finereader 14
  - Recognition language set to meta-language (59/92 meta-languages available)
- Result:
  - Most tokens of the meta-language correctly recognized
  - Low-quality recognition of tokens of the target language
  - For approximately 1% of the documents, OCR not possible (poor quality, handwriting, script not available for OCR, …)

§3: OCR Quality Example (Though Quality Varies)



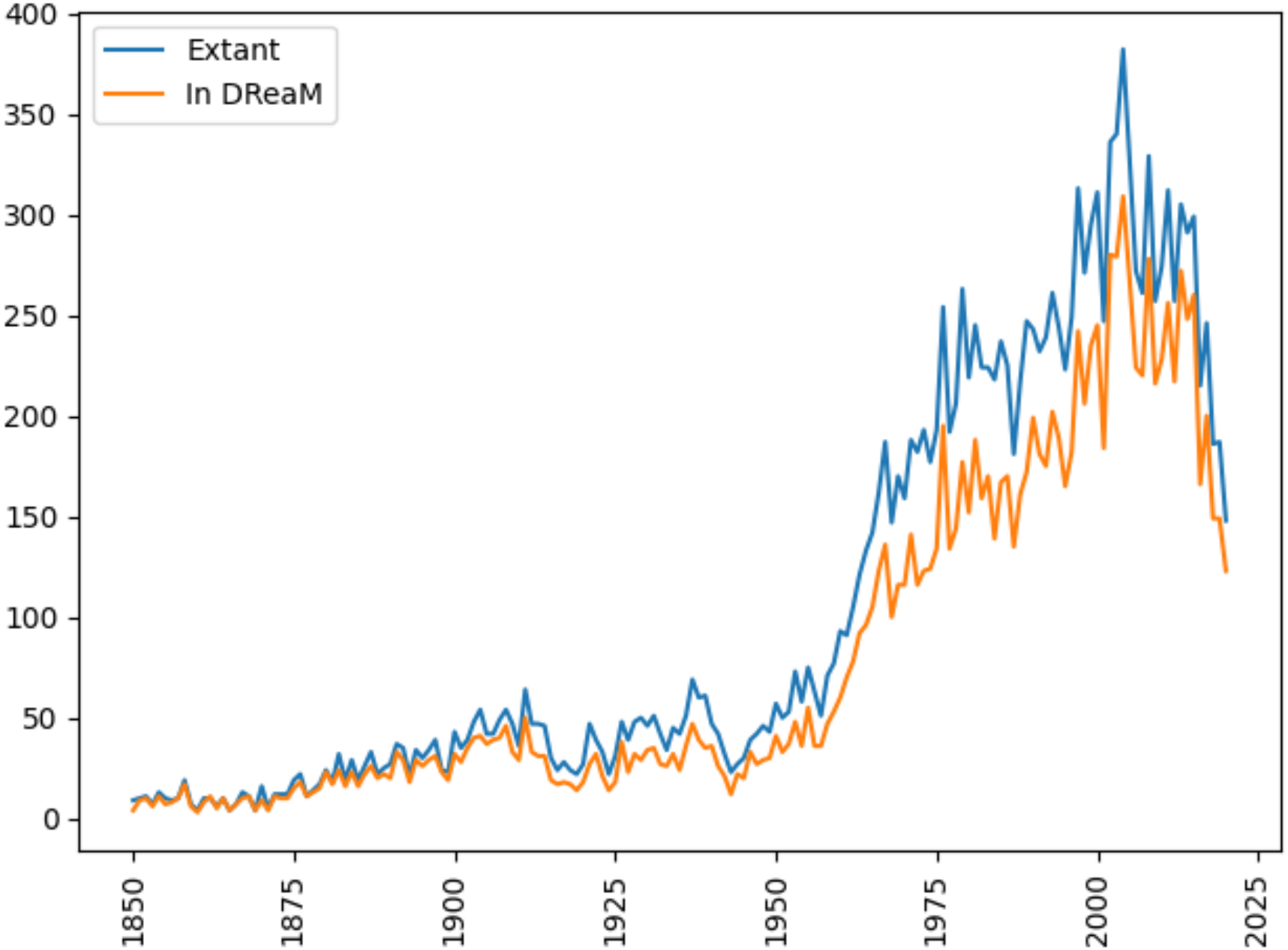
Dieses Tonmuster findet sich fast nur bei Fremdwörtern. Außerdem umfaßt die hier zu besprechende Gruppe nur 16 nicht verbale Morpheme des untersuchten Sprachmaterials. Auf die Bedeutung des Tonmusters [hoch-tief] für die Bildung des direkten Imperativs gewisser Verbalklassen wird bei der Behandlung der Morphologie des Verbums näher einzugehen sein (7.34ff.).

dímò	Zitrone (< S)	**ɓúqù**	Buch (< L < Engl.)
páqà	Wildkatze (< S)	**qíqì**	Pickel (< Franz.)
sɔ́qɔ̀	Markt (< S < Arab.)	**rúngò**	Korbsieb (< S)

⇓

```
Dieses Tonmuster findet sich fast nur bei Fremdwörtern. Außerdem
umfaßt die hier zu besprechende Gruppe m1r 16 nicht verbale
Morpheme des untersuchten Sprachmaterials. Auf die Bedeutun& des
Tonmus.ters [hoch-tief] für die Bildung des direkten Imperativs gewisser
Verbalklassen wird bei der Behandlung .der Morphologie des Verbums
nähereinzugehen sein (7.34ff.). Â·
Â·
Â·
dimo
paqa
s~q;,
```

§3: Growth in Grammar Production



§3: DReaM Challenge

- Can we instead machine-read the same grammars with accuracy comparable to human collection?
- Is there a combined human-machine approach that saves time/money? E.g., for typologists:
  - Does the language have tone?
  - Does the language have prepositions?
  - Does the language have SOV basic constituent order?
  - …
- It is of value that the machine-reading of grammar can explain its results, i.e., no black box neural network
  - (Naive) keyword-spotting techniques (Hammarström et al. 2021, Hammarström 2013, Macklin-Cordes et al. 2017, Virk et al. 2019, 2017, Wichmann and Rama 2019)
- Can an LLM read a grammar written for a human and improve?
  - Small improvments on translation abilities (Tanzer et al. 2023) — but maybe more work to come?

§3: Keyword Spotting

Simplest possible approach is to look for keywords that signal the presence of the features in question, e.g., **preposition(s)**, **dual**, **tone(s|me)**, …

- Does not work when the feature is expressed in a myriad of different ways across grammars, e.g., whether the verb agrees with the agent in person
- Simple but not completely trivial because of spurious occurrences:
  - Explicit absense: "there is no X"
  - Disparate target: "another relevant language/temporal stage has X"
  - Sample occurrence: X occurs in an example, a reference title etc.
  - …
- Genuine occurrences should be more frequent than spurious occurrences, but **how frequent is frequent enough**?

§3: Terms in a Grammatical Description

Genuine keywords: Terms that describe the language in question

Noise keywords: Descriptive terms that do not accurately describe the language in question

$$\Updownarrow \text{ rarely overlap}$$

Meta-language words: Words in the meta-language, e.g., *the*, *a*, *run*, that are not linguistic descriptive terms

Language-specific words: Words that are specific to the language being described but which do not describe its grammar, e.g., morphemes of the language, place names in the language area, etc.

§3: Terms in a Grammatical Description: Model

$$G(t) = \alpha \cdot L(t) + (1 - \alpha) \cdot N(t)$$

- $G(t)$: Frequency distribution of the keywords of a descriptive grammar composed of
  - the "**true**" underlying descriptive terms according to their functional load $L(t)$ and
  - a "**noise**" term $N(t)$
- with a **weight** α balancing the two

§3: Estimating Noise Via Multiple Grammars

$$
\begin{aligned}
G_1(t) &= \alpha_1 \cdot L(t) + (1 - \alpha_1) \cdot N_1(t) \\
G_2(t) &= \alpha_2 \cdot L(t) + (1 - \alpha_2) \cdot N_2(t) \\
&\ldots\ldots \\
G_n(t) &= \alpha_n \cdot L(t) + (1 - \alpha_n) \cdot N_n(t)
\end{aligned}
$$

- If we have many grammars for the **same** language we can estimate the noise levels α_i

$$\alpha_i = \frac{\sum_t g_L^i(t)}{\sum_t G_i(t)}$$

- where $g_L^i(t)$ is the **generality** of the term t

$$g_L^i(t) = \frac{\frac{1}{n-1} \sum_{j \neq i} G_j(t)}{G_i(t)}$$

§3: Estimating Noise: Example

t	`cojocaru`	`triphthongs`	`gender`	`stress`	`ghe`
Cojocaru 2004	0.00002	0.00004	0.00052	0.00025	0.00006
Agard 1958	0.00000	0.00002	0.00012	0.00078	0.00000
Gönczöl 2008	0.00002	0.00015	0.00046	0.00013	0.00002
Mallinson 1986	0.00000	0.00000	0.00103	0.00036	0.00000
Mallinson 1988	0.00000	0.00000	0.00055	0.00036	0.00000
Murrell 1970	0.00000	0.00004	0.00042	0.00027	0.00000
$g_{ron}^{\text{Cojocaru 2004}}(t)$	0.18	1.20	0.99	1.51	0.07

- So terms like `cojocaru`, `ghe`, … have poor generality vs `triphthongs`, `gender`, `stress`, … have better generality
- $\alpha_i = \frac{\sum_t g_L^i(t)}{\sum_t G_i(t)}$ is the average generality of all the terms of a grammar

§3: How frequent is frequent enough?

- Does the frequency of a term in a grammar exceed its noise level (1-α)?
- Removing the (1-α_i) of least frequent tokens effectively generates a threshold $\bar{t}$
- Example: Does Romanian [ron] have m/f/n gender?

Grammar	α	$\sum G_i(t)$	$\bar{t}$	masculine	feminine	neuter
Cojocaru 2004	0.81	83365	9	240 0.40 (74/184)	259 0.46 (84/184)	124 0.23 (43/184)
Murrell and Ştefănescu Drăgăneşti 1970	0.72	95226	13	3 0.01 (3/424)	5 0.01 (5/424)	4 0.01 (3/424)
Gönczöl-Davies 2008	0.68	45423	9	63 0.13 (30/233)	75 0.15 (34/233)	23 0.06 (13/233)
Agard 1958	0.68	51239	9	23 0.08 (10/123)	28 0.08 (10/123)	0 0.00 (0/123)
Mallinson 1988	0.66	11019	4	18 0.30 (9/30)	18 0.23 (7/30)	18 0.17 (5/30)
Mallinson 1986	0.82	105018	6	119 0.15 (57/375)	110 0.12 (46/375)	25 0.03 (11/375)
Majority consensus				**TRUE**	**TRUE**	**TRUE**

§3: Evaluation: Classifiers

- Grammars in English for 3 220 languages keyword-spotted for `classifier(s)`
- Evaluated against Gold Standard by Marc Tang and One-Soon Her (Her et al. 2022)

Gold Standard	**Keyword-Spotting**	**# lgs**	
False	False	2 357	73.2%
True	True	512	15.9%
True	False	317	9.8%
False	True	34	1.1%
		3 220	

- Overall accuracy is **89.1%**

§3: Manually Curated Databases: Accuracy

- On the WALS database
  - Wälchli 2005 checked every Latvian feature and found $102/112 \approx \mathbf{91.1}\%$ correct
  - Donohue 2006 checked every Tukang Besi feature and found $122/142 \approx \mathbf{85.9}\%$ correct
  - Plank 2009:67-68 checked every German feature and found

… for over a quarter, perhaps almost a third of the features mapped, the values assigned are erroneous, arbitrary, or uncertain in view of analytic alternatives, or would have been different if one or the other variety of the language summarily located at 52°N 10°E had been chosen for coding

  - Hammarström 2013 checked every language for one feature (basic word order in the transitive clause, WALS 81A) and found $1028/1228 \approx \mathbf{83.7}\%$ correct

- On the Grambank database (checking 3x20 languages in 2016)
  - Average % two coders use the same sign on the same source document: $\frac{2203}{3116} = \mathbf{70.7\%}$
  - Average % two agree when both are non-? on the same source document: $\frac{1514}{1682} = \mathbf{90.0\%}$

§4 Experiment: Prefixing and Suffixing in the Languages of the World

"the suffixing preference"

- Suffixing is more common than prefixing in the languages of the world
- Good empirical support and interesting possible explanations (see Himmelmann 2014:927 and references therein)
- But more detailed statistics on this tendency are needed to do a more fine-grained analysis of genealogical and areal factors (cf. Murawaki and Yamauchi 2018)
- With some 7 000 languages in the world, gathering these data can be a gargantuan task

§4 Three Approaches to Measure Affixation

Humans read grammars (HRG): Classic approach whereby a human reads grammatical descriptions

- Time-consuming and involves discretization judgments
- 948 languages

Machines read grammars (MRG): A machine reads digitized grammatical descriptions and simply counts occurrences of "prefix" vs. "suffix"

- Quick and dirty
- 4 287 languages

Machines Read Raw Text (MRT): A machine looks at raw text data and estimates prefixation vs suffixation using techniques from Unsupervised Learning of Morphology (Hammarström and Borin 2011)

- Correct segmentation likely not needed to gauge prefixation vs suffixation
- 1 030 languages

§4 Humans Read Grammars (HRG) #1

- Dryer (2005)'s database, reflected in
- WALS Feature `26A Prefixing vs. Suffixing in Inflectional Morphology https://wals.info/feature/26A.`

		Swedish [swe]		Swahili [swh]		Turkish [tur]		Nuaulu [nxl]	
		P	S	P	S	P	S	P	S
(i)	**case affixes on nouns**	-	-	-	-	-	2	-	-
(ii)	**pronominal subject affixes on verbs**	-	-	2	-	-	2	2	-
(iii)	**tense-aspect affixes on verbs**	-	2	2	-	-	2	-	-
(iv)	plural affixes on nouns	-	1	1	-	-	1	-	1
(v)	pronominal possessive affixes on nouns	-	-	-	-	-	1	0.5	0.5
(vi)	definite or indefinite affixes on nouns	-	1	-	-	-	-	-	-
(vii)	pronominal object affixes on verbs	-	-	1	-	-	-	-	-
(viii)	negative affixes on verbs	-	-	1	-	-	1	-	-
(ix)	interrogative affixes on verbs	-	-	-	-	-	1	-	-
(x)	adverbial subordinator affixes on verbs	-	-	-	-	-	1	-	-
Affixing index (AI)		0 / 3	$\frac{3}{3+0} = 1.0$	7 / 0	$\frac{0}{0+7} = 0.0$	0 / 11	$\frac{11}{11+0} = 1.0$	2.5 / 1.5	$\frac{1.5}{1.5+2.5} = 0.375$

§4 Humans Read Grammars (HRG) #2

- Affixing Index $AI = \frac{S}{S+P}$
- Discretized into five categories
- We only have access to the languages labeled with the discretized labels, not the underlying counts or AI
- We project SR_{HRG} back using the midpoint of the range associated with each label, i.e., 0.1, 0.3, 0.5, 0.7, 0.9

	Label	# lgs	Examples
$P + S \leq 2$	**Little or no inflectional morphology**	141	Thai [tha] (0+0), Vai [vai] (0+2), …
$0.8 \leq AI$	**Strongly suffixing**	406	Swedish [swe] (3/3)
$0.6 \leq AI < 0.8$	**Weakly suffixing**	123	Beja [bej] (10/13)
$0.4 \leq AI < 0.6$	**Equal prefixing and suffixing**	147	Ubykh [uby] (5/10)
$0.2 \leq AI < 0.4$	**Weakly prefixing**	94	Mohawk [moh] (3/9)
$AI < 0.2$	**Strongly Prefixing**	58	Hunde [hke] (0.5/10)
		948	

§4 Machines Read Grammars (MRG): #1

- Over 12 032 raw text grammatical descriptions digitally available for computational processing (Virk et al. 2020)
- Describes 4 287 languages

Meta-language		**# lgs**	**# docs**
English	eng	3 345	7 451
French	fra	792	1 323
German	deu	561	815
Spanish	spa	388	849
Russian	rus	288	537
Mandarin Chinese	cmn	166	249
Portuguese	por	136	285
Indonesian	ind	131	217
Dutch	nld	88	165
Italian	ita	81	139
			12 032

§4 Machines Read Grammars (MRG): #2

Heading	Chinese [cmn]	German [deu]	English [eng]	French [fra]
Prefix	字首|词头	\W[Pp]r[eä]fix	\W[Pp]refix	\W[Pp]r..?fix
Suffix	后缀|字尾|词尾	\W[Ss]uffix	\W[Ss]uffix	\W[Ss]uffix
Heading	**Italian [ita]**	**Portuguese [por]**	**Russian [rus]**	**Spanish [spa]**
Prefix	\W[Pp]refiss	\W[Pp]refix	\Wпрефикс	\W[Pp]refij
Suffix	\W[SS]ufiss	\W[Ss]ufix	\Wсуффикс	\W[Ss]ufij
Heading	**Indonesian [ind]**	**Dutch [nld]**		
Prefix	\W[Pp]refiks|	\W[Pp]refix|		
	\W[Aa]walan	\W[Vv]oorvoegsel|		
Suffix	\W[Ss]ufiks|	\W[Ss]uffix|		
	\W[Aa]khiran	\W[Aa]chtervoegsel|		

- The suffix ratio for Machines Read Grammars is $SR_{MRG} = \frac{S}{S+P}$ if $S + P > 0$ and conventionally set to 0.5 otherwise.

§4 Machines Read Grammars (MRG): #3

Mbo (Cameroon) [mbo]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Hedinger, Ekandjoum and Hedinger 1981	S	0.56	9	11515	9	0
Éwané 2016	G	0.70	11	73042	138	48
Majority					True	True

Hedinger, Robert, Joseph Ekandjoum & Sylvia Hedinger. (1981) *Petite grammaire de la langue mboó*. Yaoundé: Association des Etudiants Mboó, Université de Yaoundé. [hedinger_mboo1981_o.pdf hedinger_mboo1981.pdf]
Show hits
Éwané, Christiane Félicité. (2016) *Description systématique du Mbo (langue bantoue A.15)*. Bordeaux: Presses Universitaires de Bordeaux. [ewane_mbo2016_o.pdf ewane_mbo2016.pdf]
Show hits

Mbere-Mbamba [mdt]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Engouale 1980	S	0.71	1	20942	0	1
Okoudowa 2005	S	0.64	4	18514	34	0
Okoudowa 2010	S	0.64	13	50014	92	87
Majority					True	True

Engouale, Jean Pierre. (1980) Towards a contrastive study of English and Mbere. Université de la Sorbonne Nouvelle (Paris IV) MA thesis. [engouale_mbere1980_o.pdf engouale_mbere1980.pdf]
Show hits
Okoudowa, Bruno. (2005) Descrição preliminar de aspectos da fonologia e da morfologia do lembaama. Universidade de São Paulo MA thesis. [okoudowa_lembaama2005v2_o.pdf okoudowa_lembaama2005v2.pdf okoudowa_lembaama2005.pdf]
Show hits
Okoudowa, Bruno. (2010) Morfologia verbal do lembaama. Universidade de São Paulo MA thesis. [okoudowa_lembaama2010_o.pdf okoudowa_lembaama2010.pdf]
Show hits

Mbe [mfo]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Pohlig 1981	S	0.71	12	31764	13	324
Majority					True	True

Pohlig, James. (1981) The Mbe Verb: A description of the verb system of Mbe, a language of Northern Cross River State, Nigeria. Ms. [pohlig_mbe1981_o.pdf pohlig_mbe1981.pdf]

§4 Machines Read Raw Text #1

- New Testaments for 1 030 languages (McCarthy et al. 2020)
- Given a set W of word types of a corpus of a language
- Let for a string x, let $RA(-x) = \frac{\text{Probability word final occurrence in } W}{\text{Probability non-word final occurrence in } W}$ and $RA(x-)$ analogously
- For example, $RA(-ing) \approx 35.1$ and $RA(ing-) \approx 0.01$ in the English New Testament
- Keep only $-x/x-$ that are the best parse for some word in W, to get two sets S and P
- It is known that thiis procedure select "true" affixes with possible extra characters on them ("prolongation", cf. Hammarström and Borin 2011:322-326)
- Hopefully the "prolongation" affects P and S uniformly and the ratio between the two is nevertheless preserved

$$SR_{MRT} = \frac{|S|}{|S| + |P|}$$

§4 Machines Read Raw Text #2

	Swedish [swe]		English [eng]		Swahili [swh]	
	x	$RA(x)$	x	$RA(x)$	x	$RA(x)$
1	*-igt*	814.7	*-ned*	556.3	*nili-*	1655.0
2	*-ades*	362.8	*-teth*	475.9	*hawa-*	1365.8
3	*förb-*	343.7	*-ions*	407.9	*wame-*	1341.7
4	*upp-*	316.6	*-nts*	339.9	*-okea*	1261.3
5	*fram-*	248.2	*-ity*	321.4	*walio-*	1140.8
6	*-ligen*	222.7	*-ered*	290.5	*-ieni*	1124.8
7	*förh-*	216.4	*-ied*	284.3	*-zwa*	1108.7
8	*tills-*	203.6	*-neth*	259.6	*nina-*	1100.7
9	*förk-*	203.6	*-tly*	253.4	*wanao-*	1012.3
10	*förm-*	197.3	*-ias*	253.4	*nim-*	988.2
…	…		…		…	

§4 Example Results

Language		SR_{MRT}	SR_{HRG}	SR_{MRG}
Adamawa Fulfulde	fub	0.60	0.70	0.91
Alekano	gah	0.39	0.90	0.85
Amharic	amh	0.53	0.70	0.67
Burarra	bvr	0.55	0.30	0.25
Nogai	nog	0.70	0.90	0.75
Nyankole	nyn	0.31	0.10	0.14
Páez	pbb	0.61	0.90	0.67
Uighur	uig	0.76	0.90	0.79
Woun Meu	noa	0.68	0.90	0.91
Wubuy	nuy	0.51	0.50	0.38
…	…	…	…	…

§4 Comparison: Overall Suffix Ratio

- Average $SR_{HRG} \approx 0.65$
- Average $SR_{MRT} \approx 0.51$ — only a minimal suffix preference!
- Average $SR_{MRG} \approx 0.59$ — average of sources per language
  - Source grammars for the same language differ quite a lot in their suffix ratio, on average $|SR_{MRG}(s_1) - SR_{MRG}(s_2)| \approx 0.24$
  - Discrepancies due to differences in scope and attention to functional load across descriptions of the same language, but also relate to differences in author style
- Average $SR_{MRG} \approx 0.61$ — source with the most hits per language

§4 Comparison: Correlation

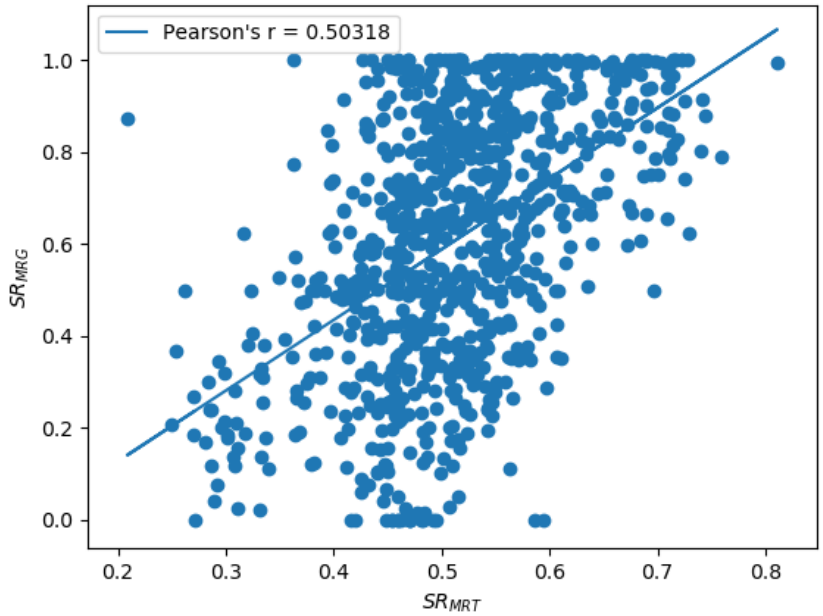


Figure: The correlation between *MRG* and *MRT*.

Dataset	# lgs	Pearson's *r*	*SR* polarity agreement
MRT ∩ HRG	306	0.54	0.73
MRT ∩ MRG	880	0.50	0.67
HRG ∩ MRG	917	0.67	0.75

Summary: Linguistically diverse NLP corpora

- # languages is approximately known
- State of grammatical description is approximately known
- Large potential for (semi-)automatic use of digitized resources now in the digital age
- Notes on the availability some resources mentioned today:
  - DoReCo: Free and open
  - Bible Corpus: Email curators (McCarthy et al. 2020)
  - DReaM corpus (text verions): Will be available in late 2026 in a web-search interface
  - PDF collection underlying DReaM corpus: Email me for login
  - Glottolog: Free and open

Agard, Frederick B. 1958. A Structural Sketch of Rumanian. Language 34(3). 7–127. Language Dissertation No. 26.

Cojocaru, Dana. 2004. Romanian Grammar. Durham: SEELRC.

Donohue, Mark. 2006. Review of the The World Atlas of Language Structures. LINGUIST LIST 17(1055). 1–20.

Dryer, Matthew S. 2005. Prefixing Versus Suffixing in Inflectional Morphology. In Bernard Comrie, Matthew S. Dryer, David Gil & Martin Haspelmath (eds.), World Atlas of Language Structures, 110-113. Oxford: Oxford University Press.

Gönczöl-Davies, Ramona. 2008. Romanian: an essential grammar. New York: Routledge.

Hammarström, Harald & Lars Borin. 2011. Unsupervised Learning of Morphology. Computational Linguistics 37(2). 309–350.

Hammarström, Harald & Sebastian Nordhoff. 2011. LangDoc: Bibliographic Infrastructure for Linguistic Typology. Oslo Studies in Language 3(2). 31–43.

Hammarström, Harald, One-Soon Her & Marc Tang. 2021. Term-Spotting: A quick-and-dirty method for extracting typological

features of language from grammatical descriptions. In Simon Dobnik, Richard Johansson & Peter Ljunglöf (eds.), Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020), 25-27 November 2020, 27-34. Linköping: Linköping Electronic Press.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2021. Glottolog 4.5. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at `http://glottolog.org`. Accessed on 2021-12-10.

Hammarström, Harald. 2013. Three Approaches to Prefix and Suffix Statistics in the Languages of the World. Paper presented at the Workshop on Corpus-based Quantitative Typology (CoQuaT 2013).

Her, One-Soon, Harald Hammarström & Marc Allassonnière-Tang. 2022. Defining numeral classifiers and identifying classifier languages of the world. Linguistics Vanguard 8(6). 1–14.

Himmelmann, Nikolaus. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. Language 90(4). 927–960.

Macklin-Cordes, Jayden L., Nathaniel L. Blackbourne, Thomas J. Bott, Jacqueline Cook, T. Mark Ellison, Jordan Hollis, Edith E. Kirlew, Genevieve C. Richards, Sanle Zhao & Erich R. Round. 2017. Robots who read grammars. Poster presented at CoEDL Fest 2017, Alexandra Park Conference Centre, Alexandra Headlands, QLD.

Mallinson, Graham. 1986. Rumanian (Croom Helm Descriptive Grammars). London: Croom Helm.

Mallinson, Graham. 1988. Rumanian. In Martin Harris & Nigel Vincent (eds.), The Romance Languages, 391-419. London: Croom Helm.

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel bible corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 3158-3163. Reykjavik, Iceland: European Language Resources Association (ELRA).

McCarthy, Arya D. , Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post & David

Yarowsky. 2020. The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In Proceedings of The 12th Language Resources and Evaluation Conference, 2877–2885. Marseille, France: European Language Resources Association.

Murawaki, Yugo & Kenji Yamauchi. 2018. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. Journal of Language Evolution 3(1). 13–25.

Murrell, Martin & Virgiliu Ştefănescu-Drăgăneşti. 1970. Romanian (Teach Yourself Books). London: English Universities Press.

Plank, Frank. 2009. WALS values evaluated. Linguistic Typology 13(1). 41–75.

Seifart, Frank, Ludger Paschen & Matthew Stave. 2024. Language Documentation Reference Corpus (DoReCo) 2.0. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Seifart, Frank. 2021. Combining documentary linguistics and corpus phonetics to advance corpus-based typology. In Geoffrey Haig, Stefan Schnell & Frank Seifart (eds.), Doing corpus-based typology with spoken language corpora. State of the art (Language

Documentation & Conservation Special Publication 25), 115-139. Honolulu: University of Hawai'i Press.

Tanzer, Garrett, Mirac Suzgun, Eline Visser, Dan Jurafsky & Luke Melas-Kyriazi. 2023. A Benchmark for Learning to Translate a New Language from One Grammar Book. Arxiv.

Virk, Shafqat Mumtaz , Harald Hammarström, Markus Forsberg & Søren Wichmann. 2020. The DReaM Corpus: A Multilingual Annotated Corpus of Grammars for the World's Languages. In Proceedings of The 12th Language Resources and Evaluation Conference, 871–877. Marseille, France: European Language Resources Association.

Virk, Shafqat Mumtaz, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal & Nazia Khurram. 2019. Exploiting Frame-Semantics and Frame-Semantic Parsing for Automatic Extraction of Typological Information from Descriptive Grammars of Natural Languages. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 1247–1256. Varna, Bulgaria: NCOMA Ltd.

Virk, Shafqat Mumtaz, Lars Borin, Anju Saxena & Harald Hammarström. 2017. Automatic Extraction of Typological Linguistic Features from Descriptive Grammars. In Kamil Ekštein & Václav Matoušek (eds.), Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings (Lecture Notes in Computer Science 10415), 111-119. Berlin: Springer.

Wichmann, Søren & Taraka Rama. 2019. Towards unsupervised extraction of linguistic typological features from language descriptions. First Workshop on Typology for Polyglot NLP, Florence, Aug. 1, 2019 (Co-located with ACL, July 28-Aug. 2, 2019).