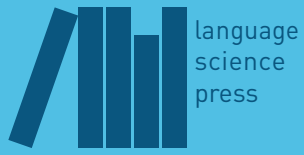
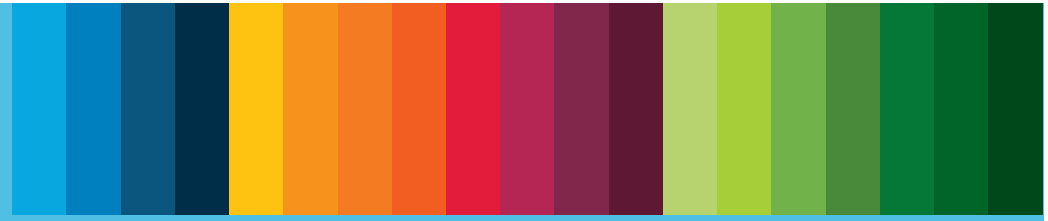


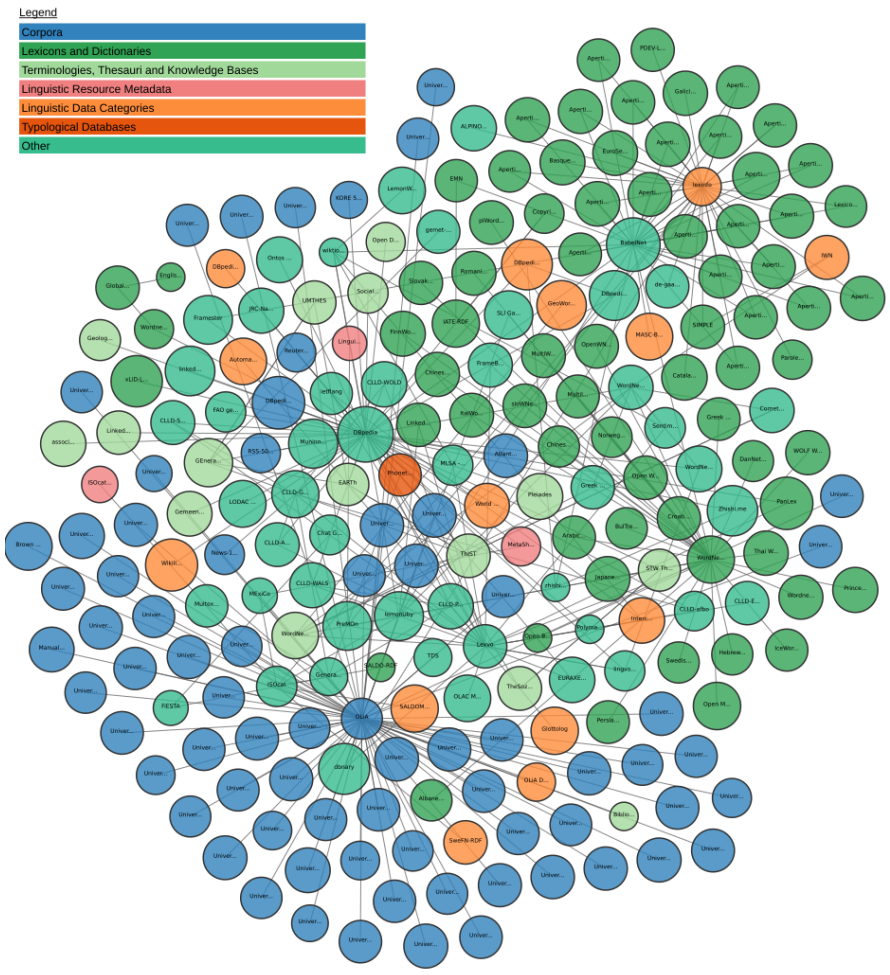


IMTVault: Linked Data from Open Access LaTeX books

Sebastian Nordhoff & Thomas Krämer

June 23, 2022

The LLOD cloud



- structured language data for NLP,
- WordNet,
- Thesauri,
- Dictionaries,
- Treebanks,
- Annotated corpora

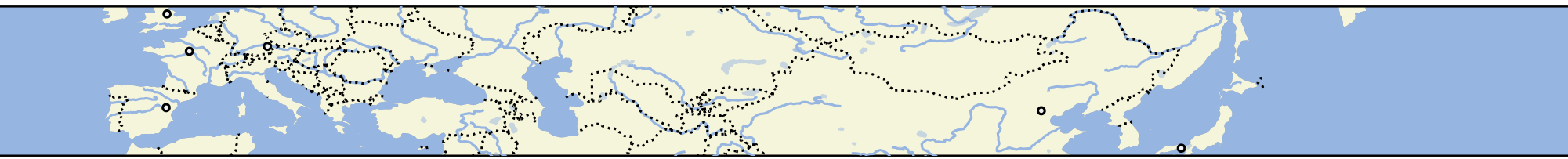
For how many languages?

Coverage of NLP

Joshi et al. (2020) analyzed the languages covered in NLP research and resources.

		criteria				
	Class	unlabeled data	labeled data	example	# lgs	%
5	winners	good	good	Spanish	7	0.28
4	underdogs	good	insufficient	Russian	18	1.07
3	rising stars	good	none	Indonesian	28	4.42
2	hopefuls	?	smallish sets	Zulu	19	0.36
1	scraping-bys	smallish	none	Fijian	222	5.49
0	left-behinds	none	none	Warlpiri	2 191	88.38
-1	not included	?	?	Komnzo	6 404	–

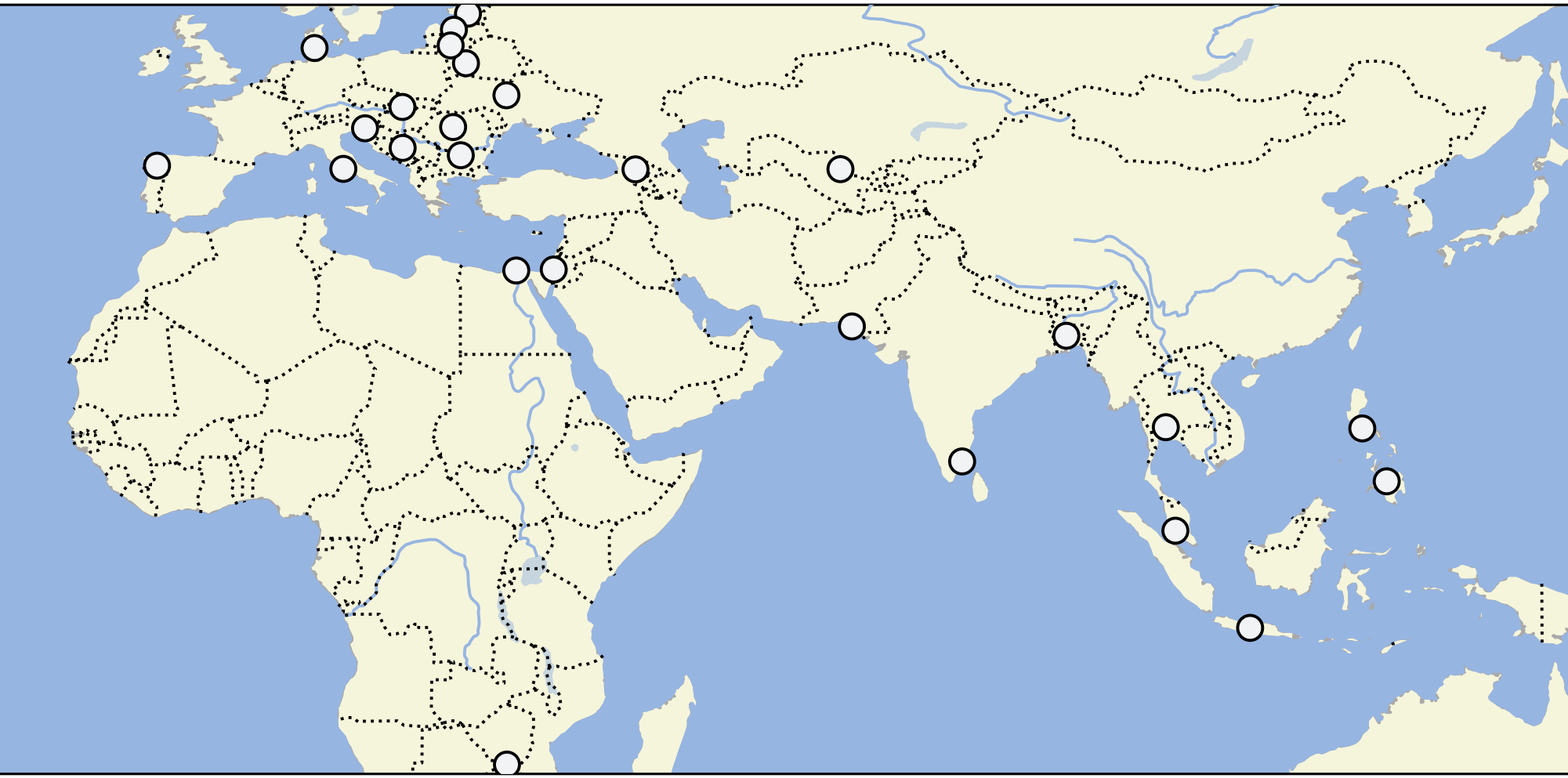
Group 5: winners



Group 4: underdogs

Group 3: rising stars

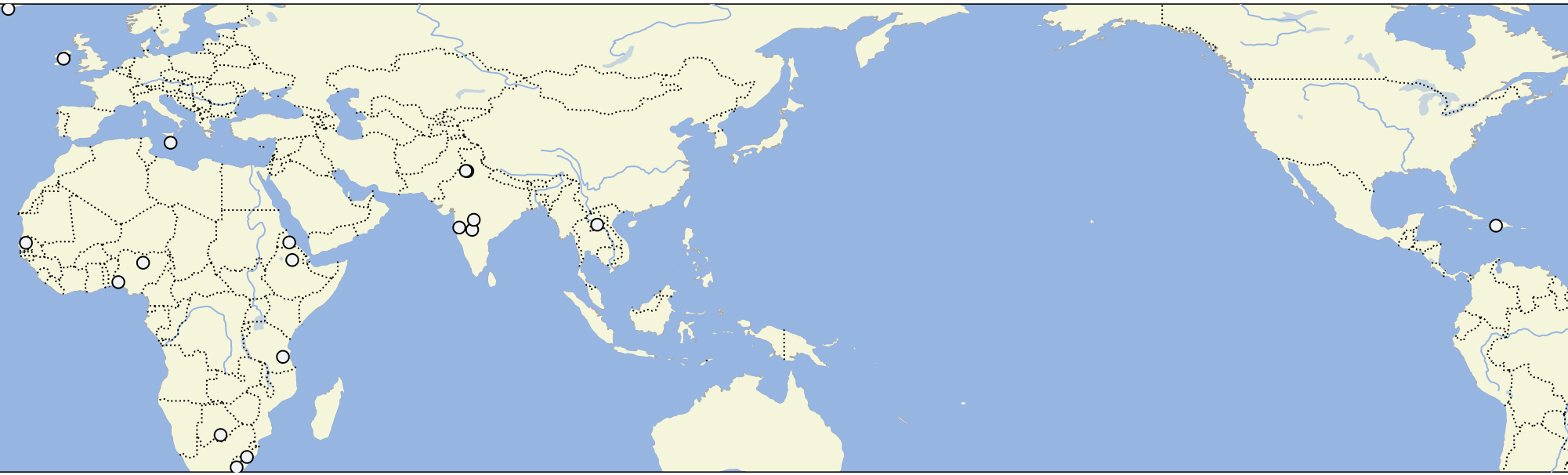
Joshi et al. (2020)



language science press

Group 2: hopefuls

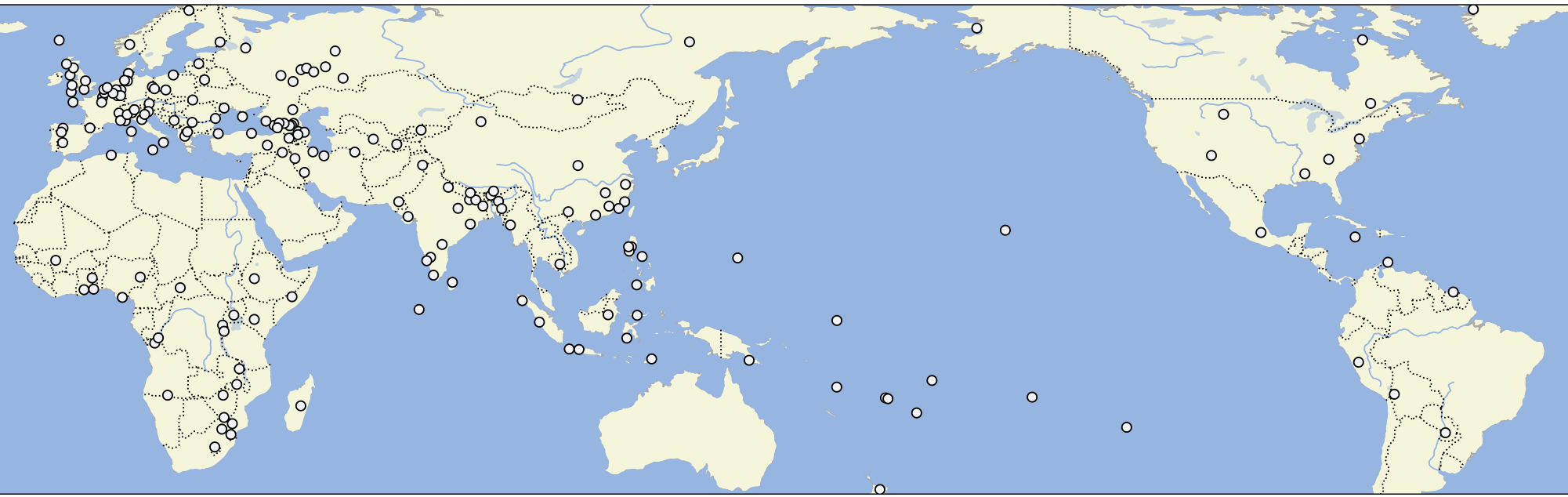
Joshi et al. (2020)



language science press

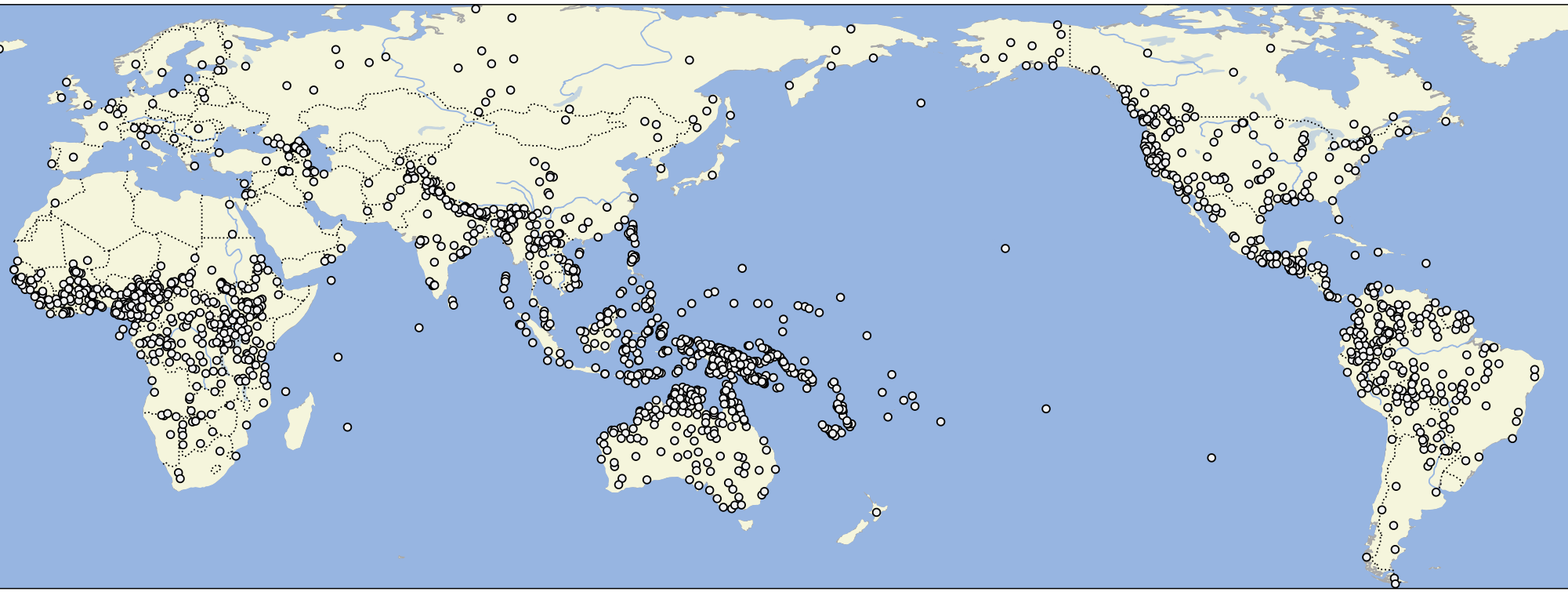
Group 1: scraping-bys

Joshi et al. (2020)



Group 0: left-behinds

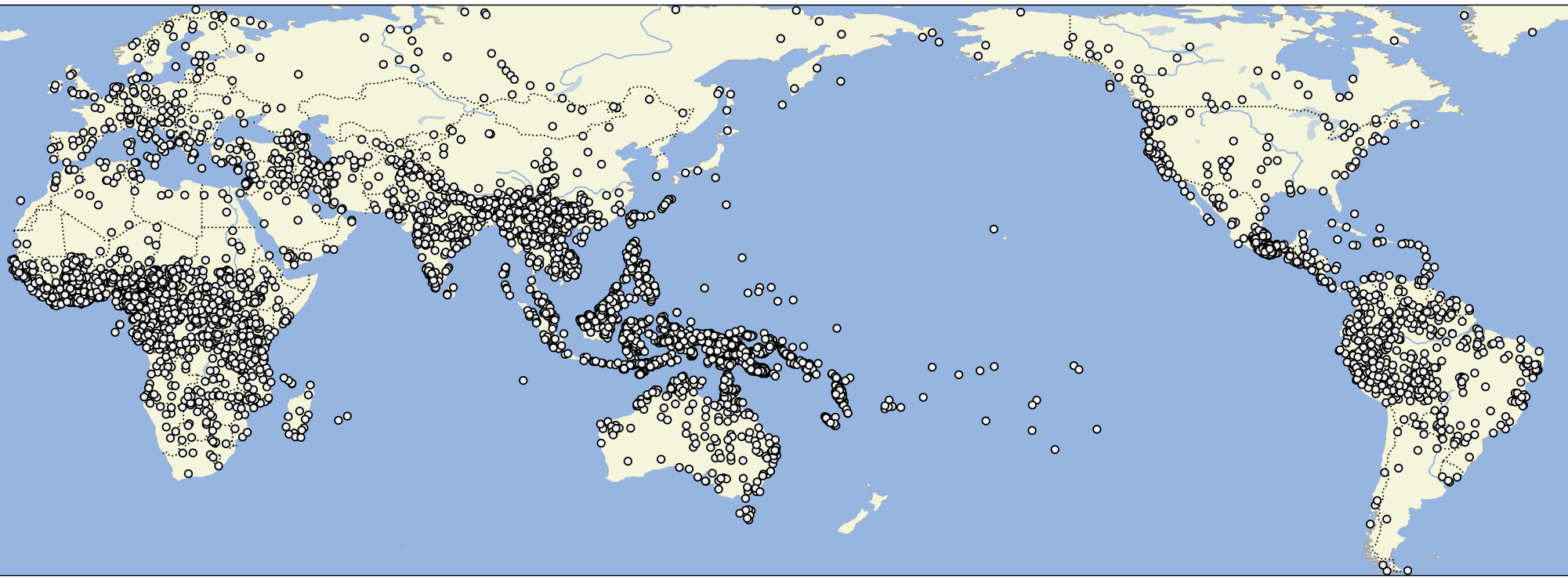
Joshi et al. (2020)



language science press

Group -1: there is such a language??

Joshi et al. (2020)



What do we have instead of "classic" NLP stuff?

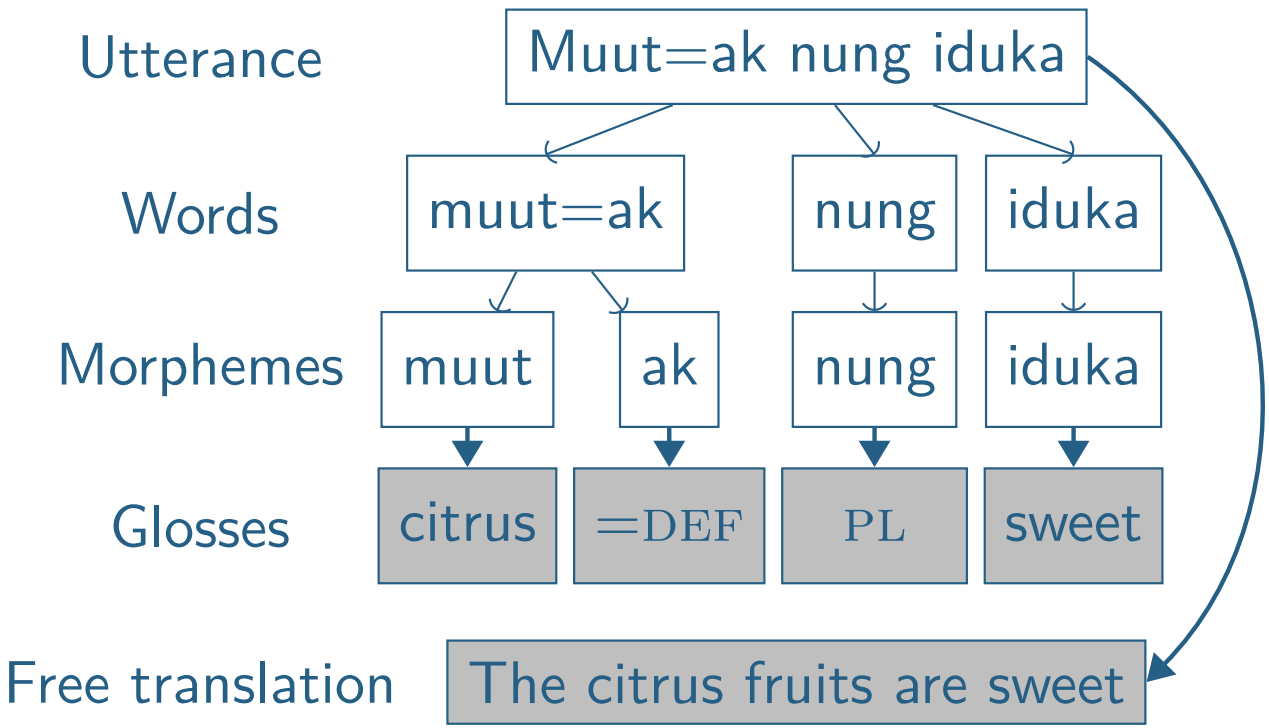
- More than 6000 languages are all but ignored by NLP
- For many of the languages of groups 0 and –1, we do indeed have some resources
  - typological databases (WALS, APICS)
  - Endangered Language Archives (ELAR, PARADISEC, TLA, AILLA)
  - scientific publications (books, articles)

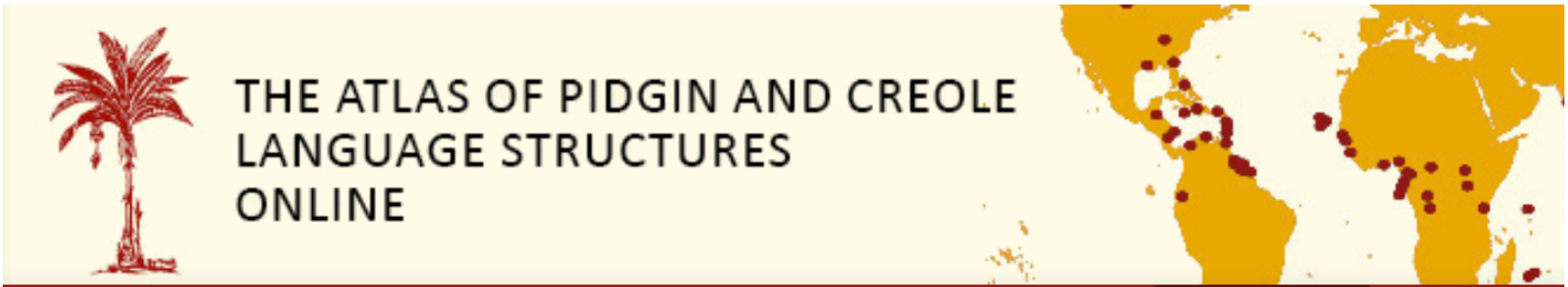
Common datastructure: IGT

⟩ IGT: interlinear glossed text

(1) Muut=ak nung iduka.
citrus=DEF PL sweet
'The citrus fruits are sweet.'

Common datastructure: IGT





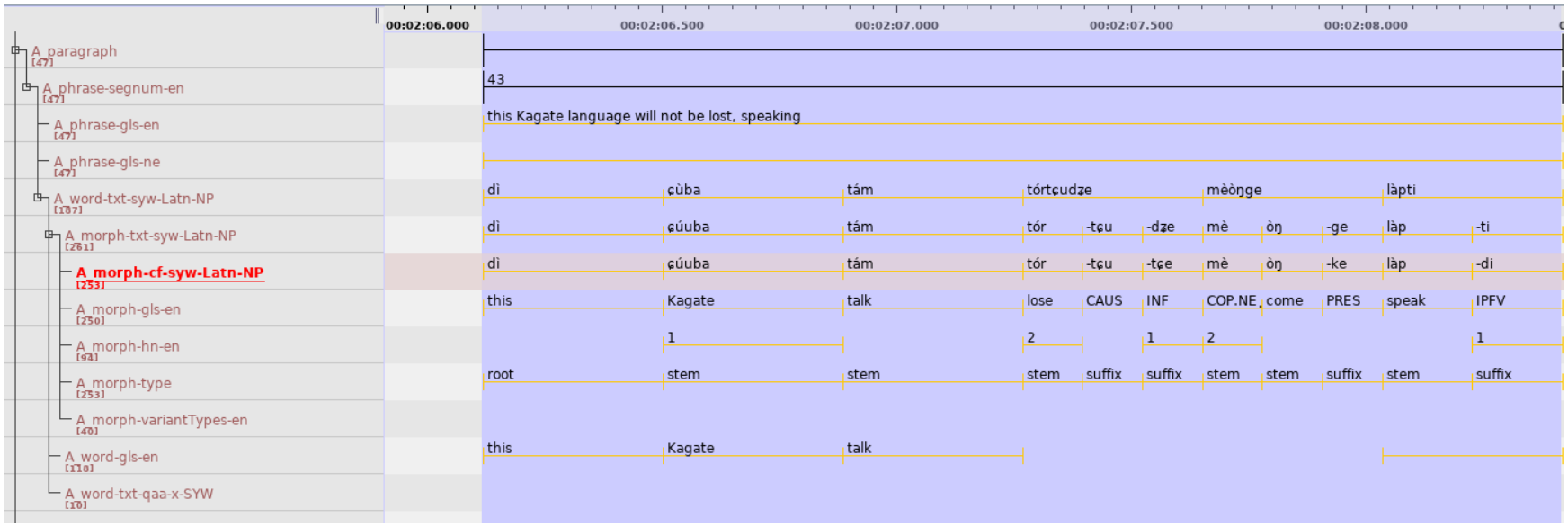
Example 22-78

Bai mitupela i ringim taksi.
Bai mitupela i ring-im taksi.
FUT 1DU.EXCL PM ring-TR taxi
'We'll ring a taxi.'

Type:
naturalistic spoken
Source:
Dutton and Thomas 1985: 88

(Chiarcos & Ionov 2019)

ELAN files from endangered language archives



von Prince & Nordhoff (2020); Nordhoff (2020a,b)

This talk: scientific publications


```
\langinfo{Kamang}{}{Schapper, fieldnotes} \\
\gll  Muut=ak nung iduka. \\
  citrus=\textsc{def} \textsc{pl} sweet   \\
\glt `The citrus fruits are sweet.'
```

Language Science Press

- open access publisher in linguistics founded 2014
- 180 published books
- all books licensed as CC-BY
- tex codes for all books available no GitHub
- LaTeX package gb4e used for interlinear glossed text

A grammar of Gyeli

A grammar of Gyeli

Nadine Grimm

Synopsis

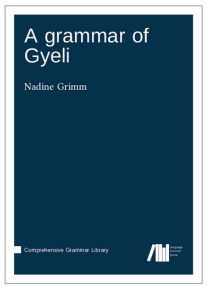
This grammar offers a grammatical description of the *Ngòlò* variety of Gyeli, an endangered Bantu (A80) language spoken by 4,000-5,000 "Pygmy" hunter-gatherers in southern Cameroon. It represents one of the most comprehensive descriptions of a northwestern Bantu language.

The grammatical description, which is couched in a form-to-function approach, covers all levels of language, ranging from Gyeli phonology to its information structure and complex clauses.

It draws on nineteen months of fieldwork carried out as part of the "Bagyeli/Bakola" DoBeS (Documentation of Endangered Languages) project between 2010 and 2014. The resulting multimodal corpus from that project, which includes texts of diverse genres such as traditional stories, narratives, multi-party conversations and dialogues, procedural texts, and songs, provides the empirical basis for the grammatical description. The documentary text collection, supplemented by data from elicitation work, questionnaires, and experiments, are accessible in the Bagyeli/Bakola collection of *The Language Archive*. With additional ethnographic, sociolinguistic, diachronic, and comparative remarks, the grammar may appeal to a wider audience in general linguistics, typology, Bantu studies, and anthropology.

In 2019, the grammar received the Pāṇini Award by the *Association for Linguistic Typology*.

Statistics



PDF

Bibliography

Buy from amazon.de

Buy from amazon.co.uk

Buy from amazon.com

Collaborative reading on PaperHive

LaTeX source on GitHub

langsci / 298 Public

Code | Issues | Pull requests | Actions | Projects | Wiki

main 298 / chapters / CH7.tex

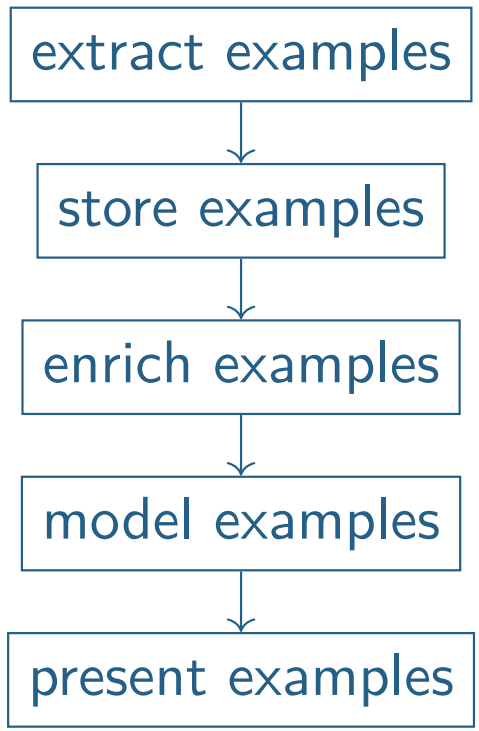
Glottotopia copy from overleaf

1 contributor

2871 lines (2042 sloc) | 186 KB

```
\chapter{Simple clauses}
\label{sec:SC}



In this chapter, I describe the different types of simple clauses in Gyeli.




\section[Copula constructions]{Non-verbal and verbal copula constructions}
\label{sec:nonverbalC}

Gyeli has copula clauses with both non-verbal and verbal copula constructio

\ea\label{John1}
\gll John {\bfseries ni} m-kubwa \\
 $\emptyset$1.{\PN} {\COP} 1-big \\
\trans `John is big'
\z
```

Pipeline


extract examples
store examples
enrich examples
model examples
present examples

Extract examples

Kamang (Schapper, fieldnotes)
Muut=ak nung iduka.
citrus=DEF PL sweet
'The citrus fruits are sweet.'

```
\langinfo{Kamang}{}{Schapper, fieldnotes}
\gll Muut=ak nung iduka.
citrus=\textsc{def} \textsc{pl} sweet
\glt `The citrus fruits are sweet.'
```

Store examples

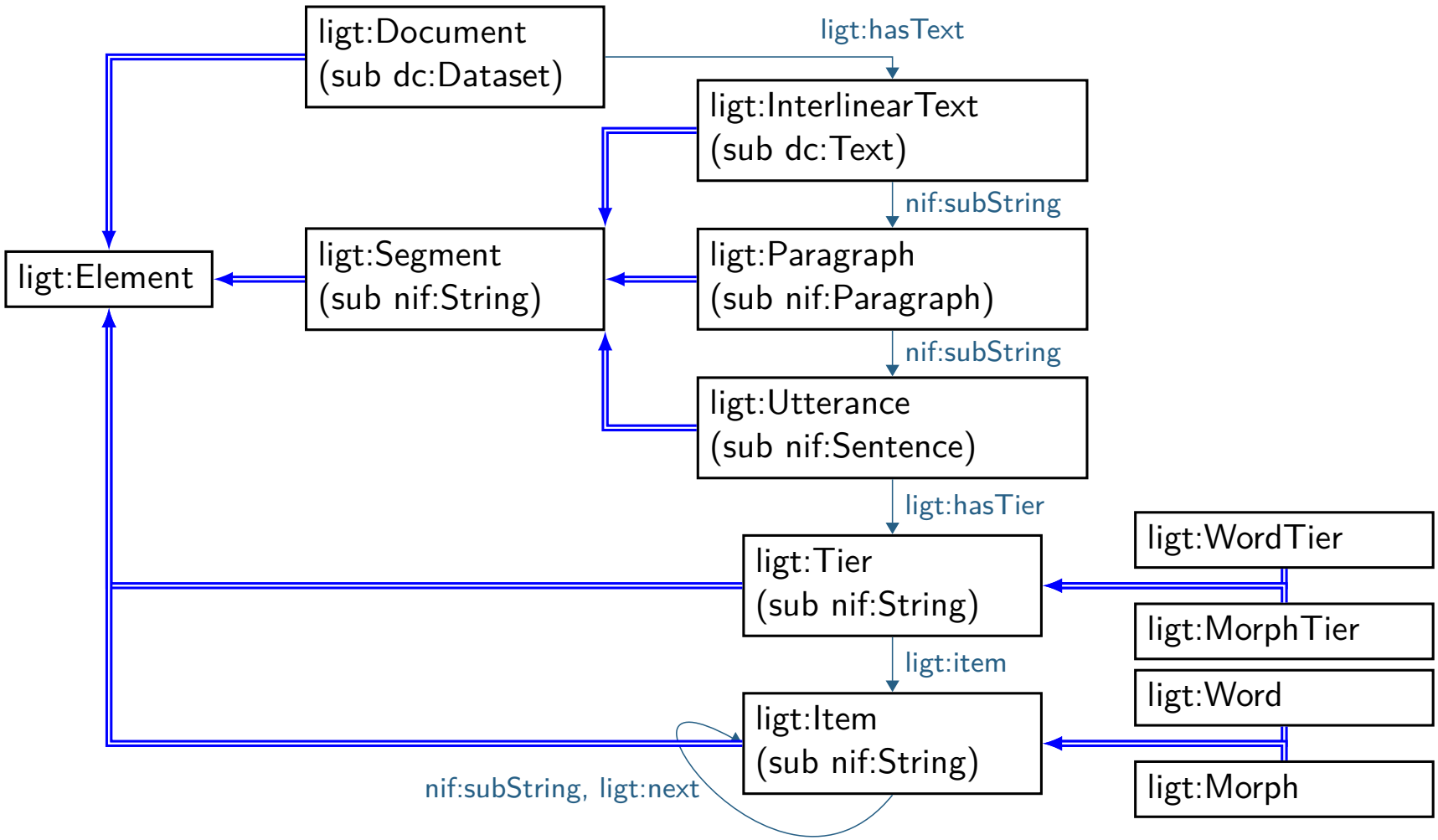

```
{"ID": "wl09-cb9806ea53",
 "book_ID": 157,
 "book_URL": "https://langsci-press.org/catalog/book/157",
 "book_metalanguage": "eng",
 "book_title": "The Alor-Pantar languages²",
 "categories": ["def","pl"],
 "imtwordsbare": [
     "citrus=DEF",
     "PL",
     "sweet"
 ],
 "label": "Muut=ak nung iduka.",
 "srcwordsbare": [
     "Muut=ak",
     "nung",
     "iduka."
 ],
 "trs": "The citrus fruits are sweet."
}
```

Enrich examples

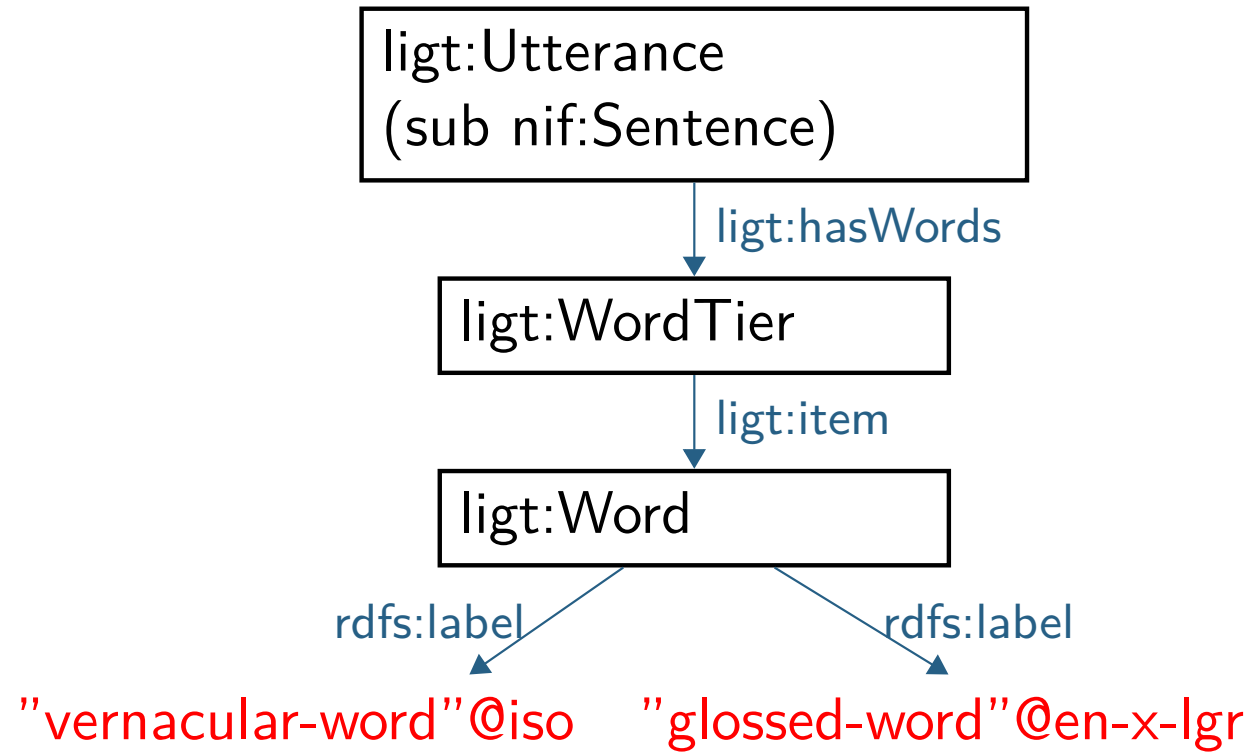
- Use Glottolog for additional language information and Science-Miner for NER based on Wikidata


```
{"ID": "wl09-cb9806ea53",
 ...
 "entities": [{
   "label": "citrus fruits",
   "wdid": "Q81513"
 }],
 ...
 "language": "https://glottolog.org/resource/languoid/id/kama1365",
 "language_iso6393": "woi",
 "language_name": "Kamang",
 ...
 "parententities": [{
     "label": "Fruit",
     "wdid": "Q1364"
   },{
     "label": "Food",
     "wdid": "Q2095"
     }
 ]
}
```

Model with LIGT



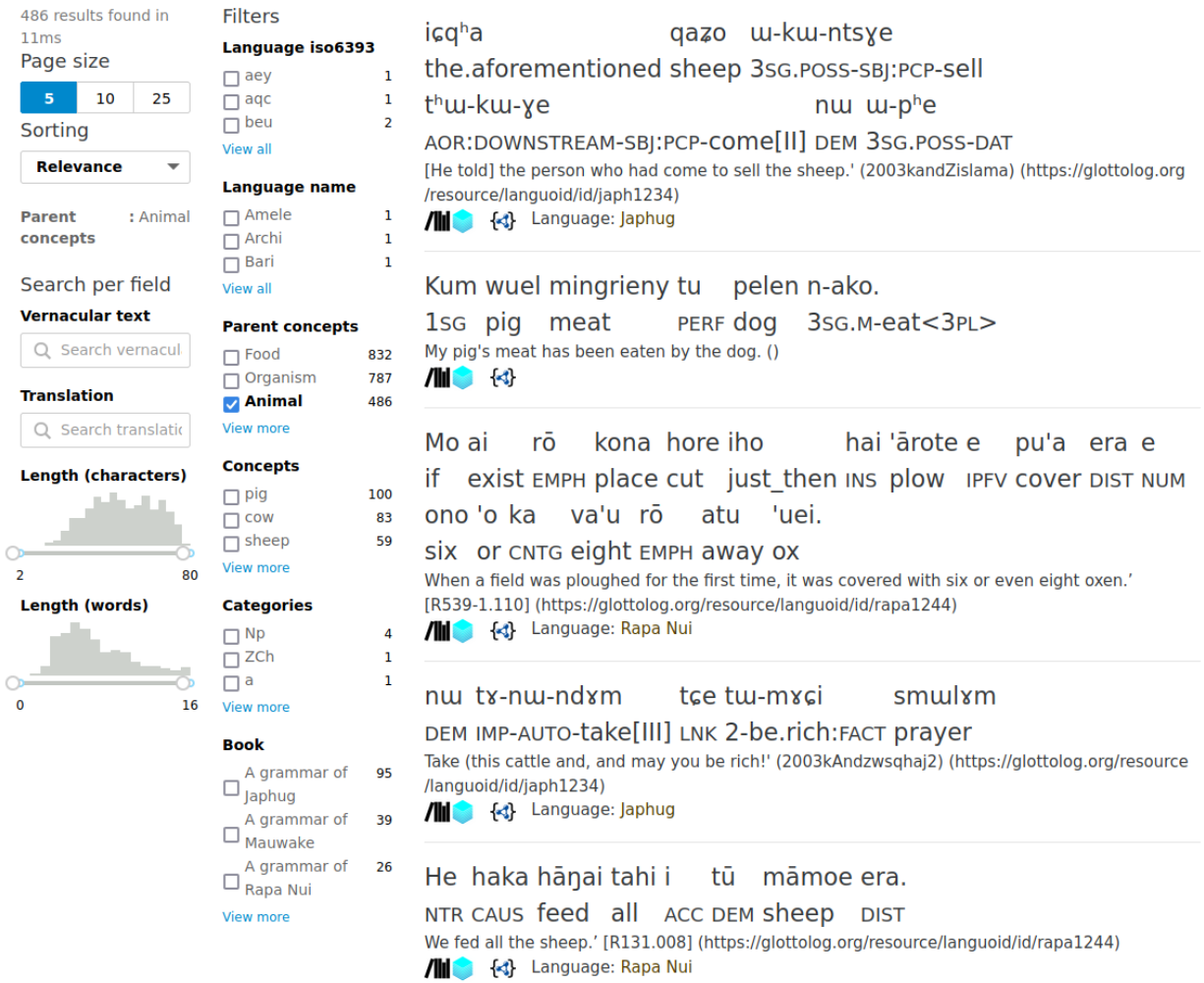
LIGT



- Word nodes have two labels: RFC 5646 label “en” with a private subtag “-x-lgr” for Leipzig Glossing Rules, following specifications in Section 2.2.7

Present examples: humans

imtvault.org



Present examples: machines

⟩ JSON-LD

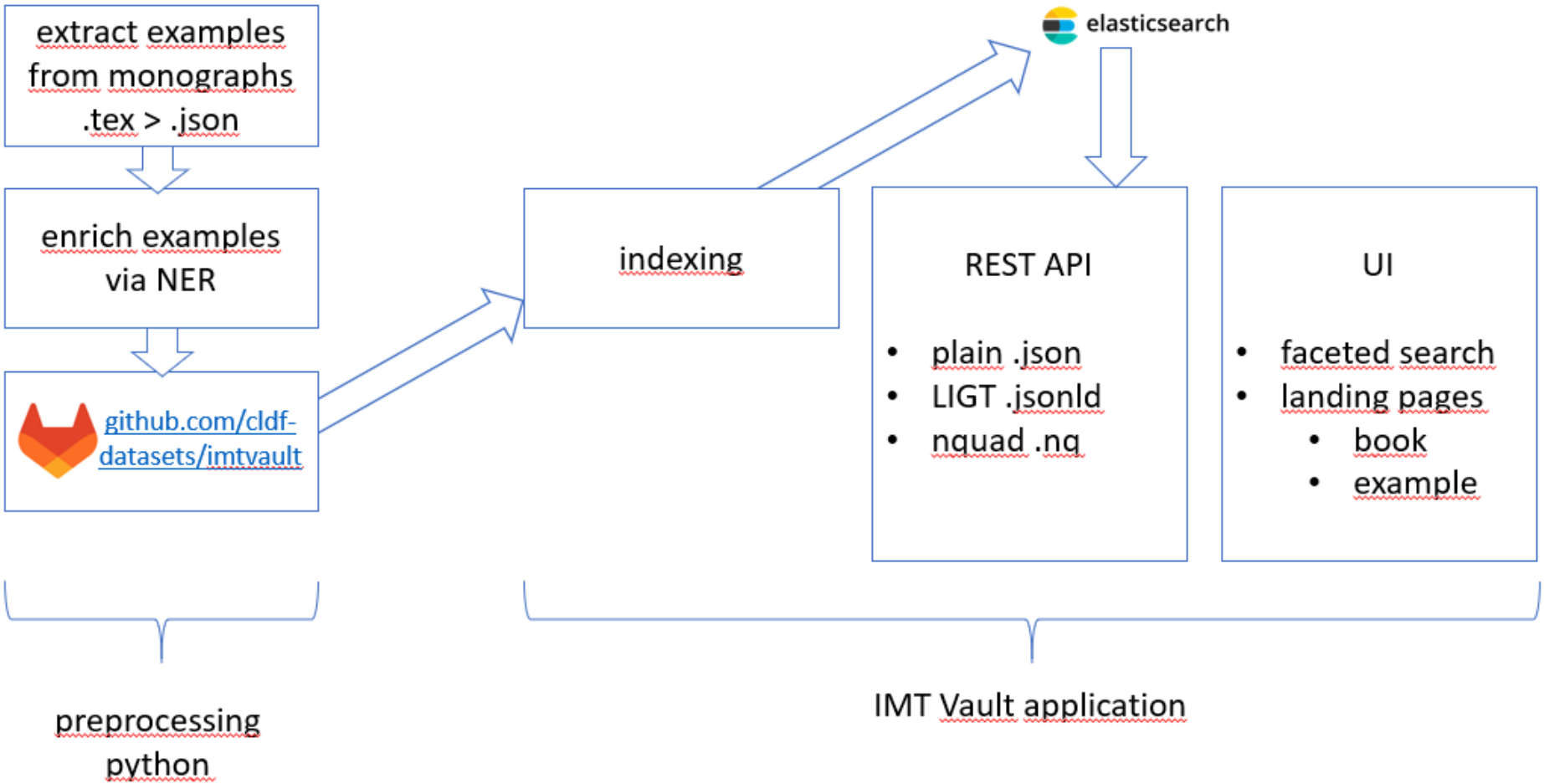

```
▼ http://purl.org/dc/terms/isPartOf:
  ▼ 0:
      @id:                                              "https://langsci-press.org/catalog/book/157"
▼ http://purl.org/dc/terms/hasPart:
  ▼ 0:
    ▶ @id:                                              "https://www.eva.mpg.de/l…s/glossing-rules.php#def"
  ▶ 1:                                                  {…}
▶ http://purl.org/dc/terms/license:                     […]
▶ http://purl.org/dc/elements/1.1/coverage:             […]
▼ https://imtvault.org/content/static/ligt-0.2.ttl#hasWords:
  ▼ 0:
      @id:                                              "https://imtvault.org/wl09-cb9806ea53_wt"
    ▼ @type:
        0:                                              "https://purl.org/liodi/ligt#WordTier"
    ▼ https://imtvault.org/content/static/ligt-0.2.ttl#item:
      ▼ 0:
          @id:                                          "_:wl09-cb9806ea53_0"
        ▼ https://imtvault.org/content/static/ligt-0.2.ttl#Word:
          ▼ 0:
              @language:                                "woi"
              @value:                                   "Muut=ak"
          ▼ 1:
              @language:                                "en-x-lgr"
              @value:                                   "citrus=DEF"
        ▶ https://imtvault.org/content/static/ligt-0.2.ttl#nextWord:  […]
      ▶ 1:                                              {…}
      ▶ 2:                                              {…}
```

Present examples: machines

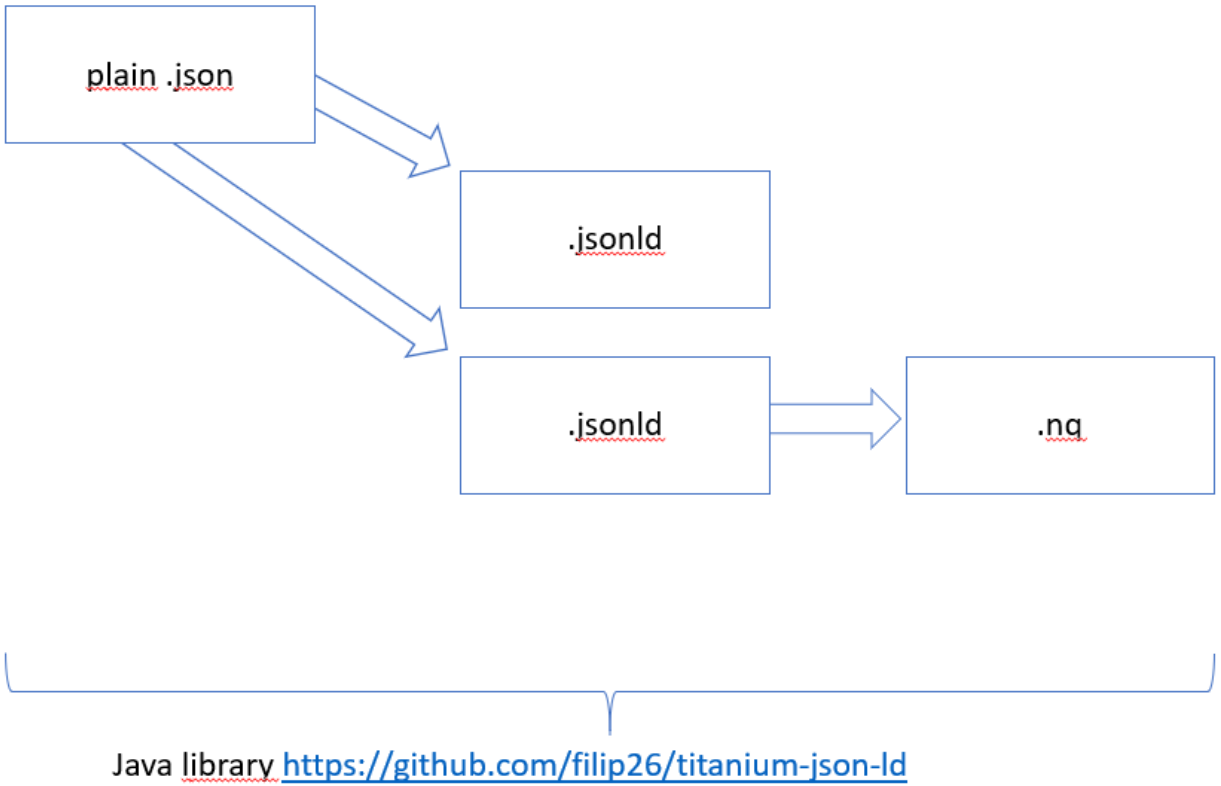
⟩ nquads

```
_:b0 <https://imtvault.org/content/static/ligt-0.2.ttl#Word> "Muut=ak"@woi .
_:b0 <https://imtvault.org/content/static/ligt-0.2.ttl#nextWord> _:b1 .
_:b1 <https://imtvault.org/content/static/ligt-0.2.ttl#Word> "nung"@woi .
_:b1 <https://imtvault.org/content/static/ligt-0.2.ttl#nextWord> _:b2 .
_:b2 <https://imtvault.org/content/static/ligt-0.2.ttl#Word> "iduka."@woi .
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <http://www.w3.org/1999/02/22-rdf
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <http://purl.org/dc/elements/1.1/
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <http://purl.org/dc/elements/1.1/
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <http://purl.org/dc/elements/1.1/
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <http://purl.org/dc/terms/SizeOrD
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <http://purl.org/dc/terms/hasPart
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <http://purl.org/dc/terms/hasPart
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <http://purl.org/dc/terms/isPartO
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <http://purl.org/dc/terms/license
<https://imtvault.org/b/157/ex/wl09-cb9806ea53> <https://imtvault.org/content/sta
<https://imtvault.org/wl09-cb9806ea53_wt> <http://www.w3.org/1999/02/22-rdf-synta
<https://imtvault.org/wl09-cb9806ea53_wt> <https://imtvault.org/content/static/li
<https://imtvault.org/wl09-cb9806ea53_wt> <https://imtvault.org/content/static/li
<https://imtvault.org/wl09-cb9806ea53_wt> <https://imtvault.org/content/static/li
<https://www.wikidata.org/wiki/Q1364> <http://purl.org/dc/elements/1.1/coverage>
<https://www.wikidata.org/wiki/Q2095> <http://purl.org/dc/elements/1.1/coverage>
```

Processing


extract examples from monographs .tex > .json
enrich examples via NER
github.com/cldf-datasets/imtvault
preprocessing python
indexing
elasticsearch
REST API
plain .json
LIGT .jsonld
nquad .nq
UI
faceted search
landing pages
book
example
IMT Vault application

Sequential transformation



REST API

⟩ SWAGGER UI

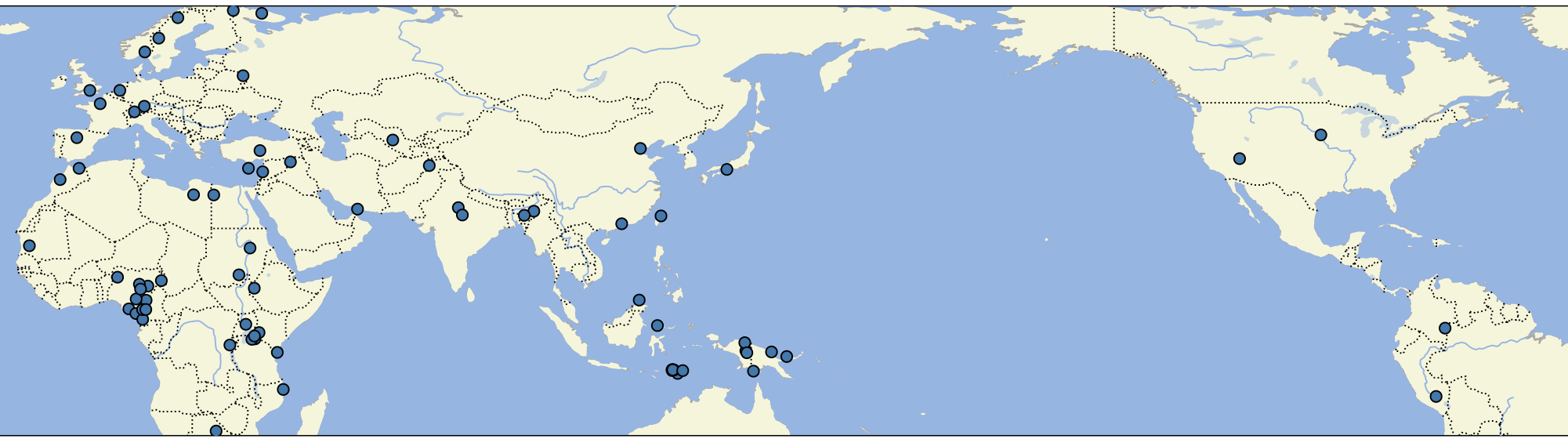
url-resolver URL Resolver

GET	**/b/{book_ID}**	redirectWithUsingRedirectView
GET	**/b/{book_ID}.htm**	getBookHtml
GET	**/b/{book_ID}.json**	getBook
GET	**/b/{book_ID}/ex/{exampleid}**	getExample
GET	**/b/{book_ID}/ex/{exampleid}.htm**	getExampleHtml
GET	**/b/{book_ID}/ex/{exampleid}.json**	getExampleSimpleJson
GET	**/b/{book_ID}/ex/{exampleid}.jsonld**	getExampleJSONLD
GET	**/b/{book_ID}/ex/{exampleid}.nq**	getExampleJSONLDRDFNQ
GET	**/book/{book_ID}**	redirectWithUsingRedirectView

Bugs, Feature requests, PRs?

⟩ Issue tracker https://github.com/cldf-datasets/imtvault/issues

Conclusion: added languages



Conclusions

- There are (semi-)structured sources for linguistic data around
  - typological databases, endangered language archives, publications
- these can be tapped into
- expanding the linguistic coverage of the LLOD-cloud
- still in the range of 10^2 datapoints, far away from 10^7 or so required for serious NLP.
- Glottolog will be a useful to aggregate the data
  - data is available in CLDF format as well

Thank you

IMT VAULT

Example in book 85

Tɛnk gɔ́d wé yu dɔ́n kán!
thank God SUB 2SG PRF come
Thank God that you have come!’

imtwordsbare	[thank, God, SUB, 2SG, PRF, come]
language_glottocode	fern1234
language	https://glottolog.org/resource/languoid /id/fern1234
book_ID	85
label	Tɛnk gɔ́d wé yu dɔ́n kán!
wlength	6
license	https://creativecommons.org/licenses/by/4.0
book_metalanguage	eng
language_iso6393	fpe
book_title	A grammar of Pichi
book_URL	https://langsci-press.org/catalog/book/85
ID	10-7c7dfc059a
categories	[prf, sub]
language_name	Pichi
srcwordsbare	[Tɛnk, gɔ́d, wé, yu, dɔ́n, kán!]
clength	57

Thank you


Chiarcos, Christian & Maxim Ionov. 2019. Ligt: an LLOD-native vocabulary for representing interlinear glossed text as RDF. In Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek & Milan Dojchinovski (eds.), *2nd conference on language, data and knowledge (ldk 2019)* (OpenAccess Series in Informatics (OASIcs) 70), 3:1–3:15. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. DOI: 10.4230/OASIcs.LDK.2019.3.


Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali & Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, 6282–6293.


Nordhoff, Sebastian. 2020a. From the attic to the cloud: mobilization of endangered language resources with linked data. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, 10–18. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lr4sshoc-1.3.


Nordhoff, Sebastian. 2020b. Modelling and annotating interlinear glossed text from 280 different endangered languages as linked data with LIGT. In *Proceedings of the 14th Linguistic Annotation Workshop*, 93–104. Barcelona: Association for Computational Linguistics. https://aclanthology.org/2020.law-1.9.

Thank you

von Prince, Kilu & Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of LREC 2020*. Marseille: LREC.