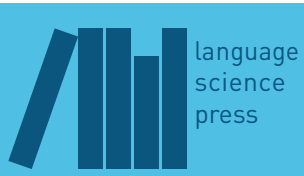
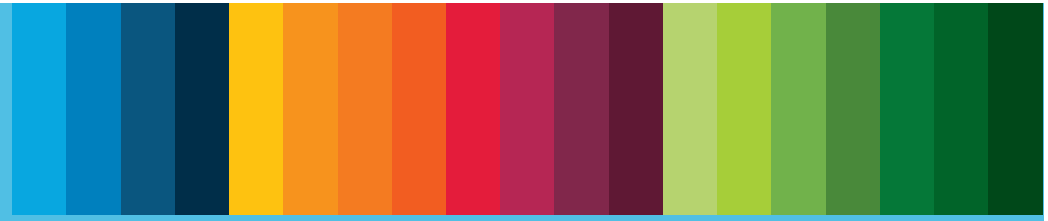




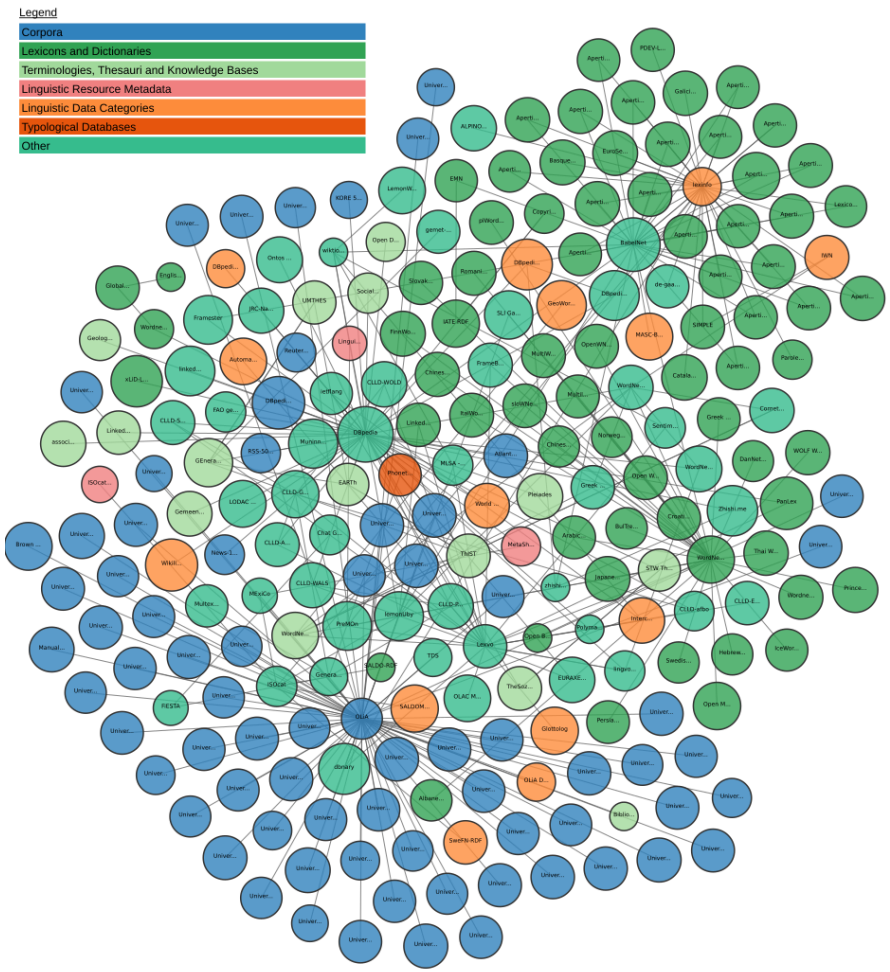
IMTVault: Linked Data from Open Access LaTeX books

SN & TK

June 2, 2022

language science press

The LLOD cloud



The Linguistic Linked Open Data Cloud from lod-cloud.net

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

For how many languages?

language science press

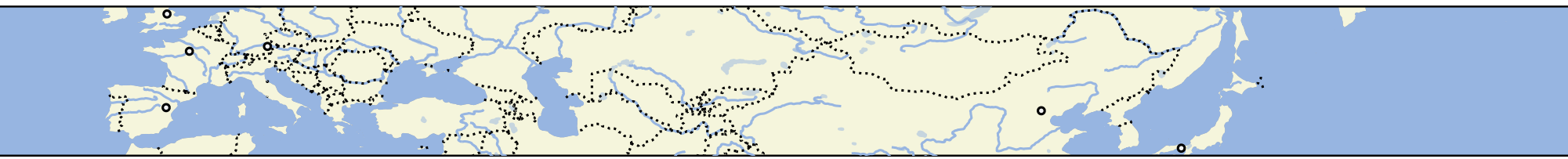
Coverage of NLP

Joshi 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	–

language
science
press

Joshi's group 5: winners

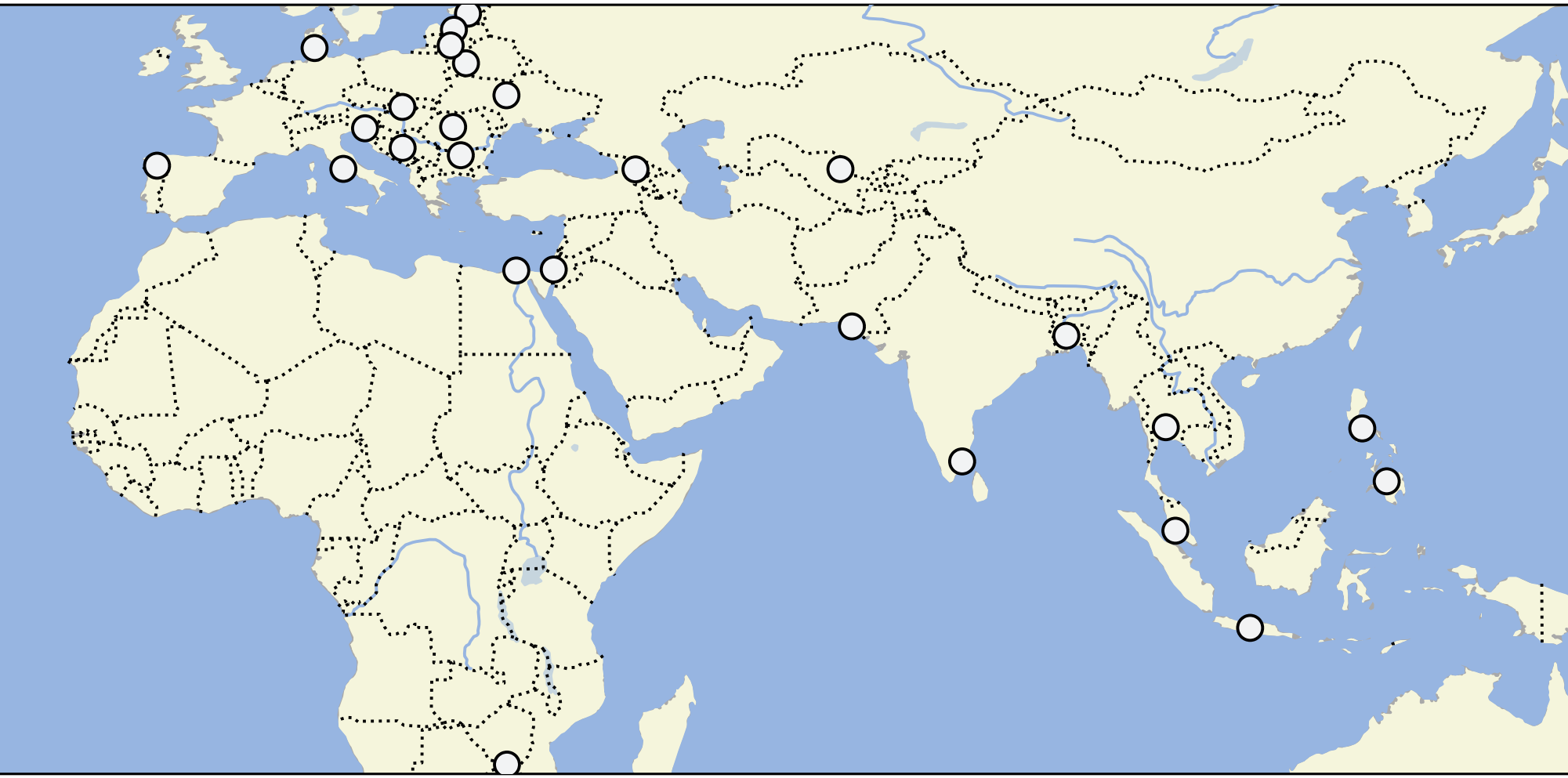


language
science
press

Joshi's group 4: underdogs

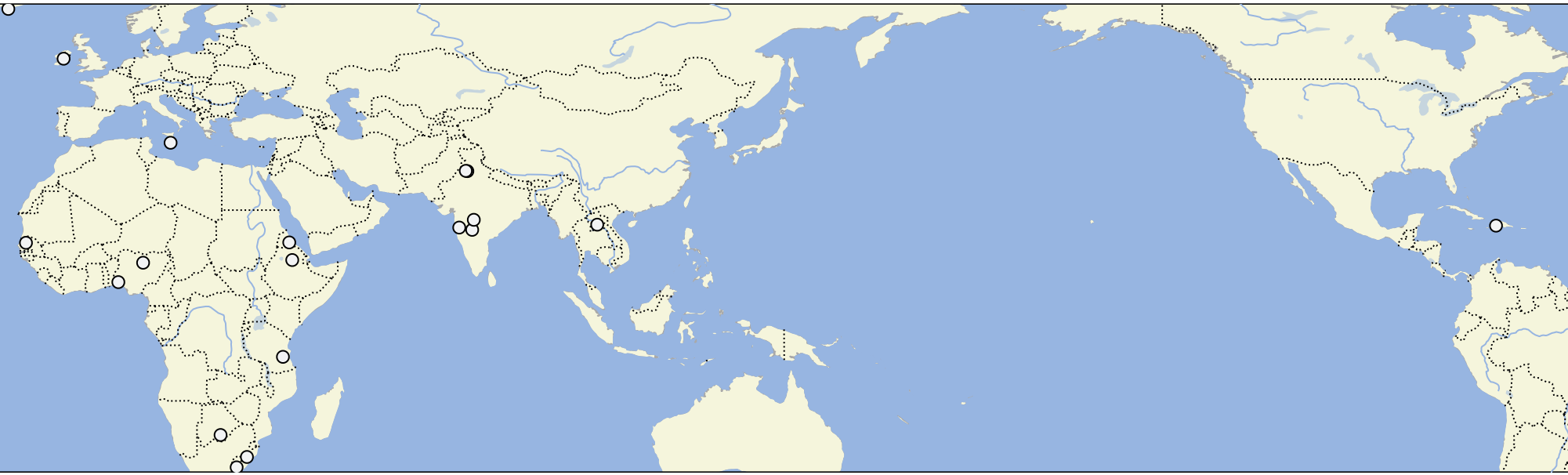
language
science
press

Joshi's group 3: rising stars



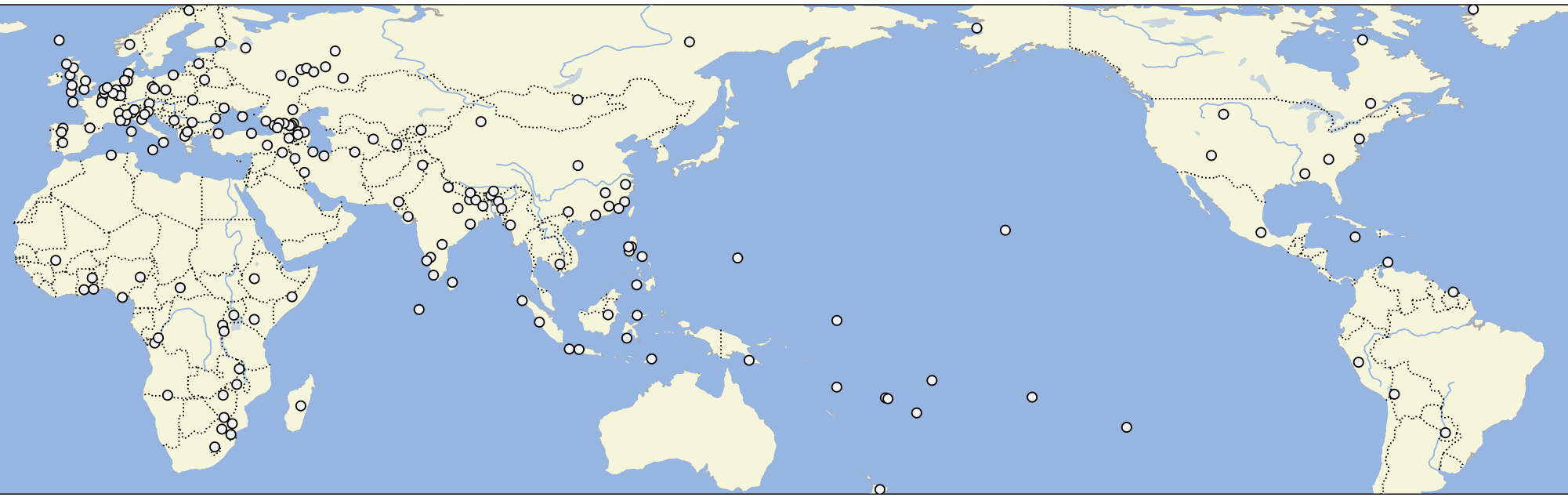
language
science
press

Joshi's group 2: hopefuls



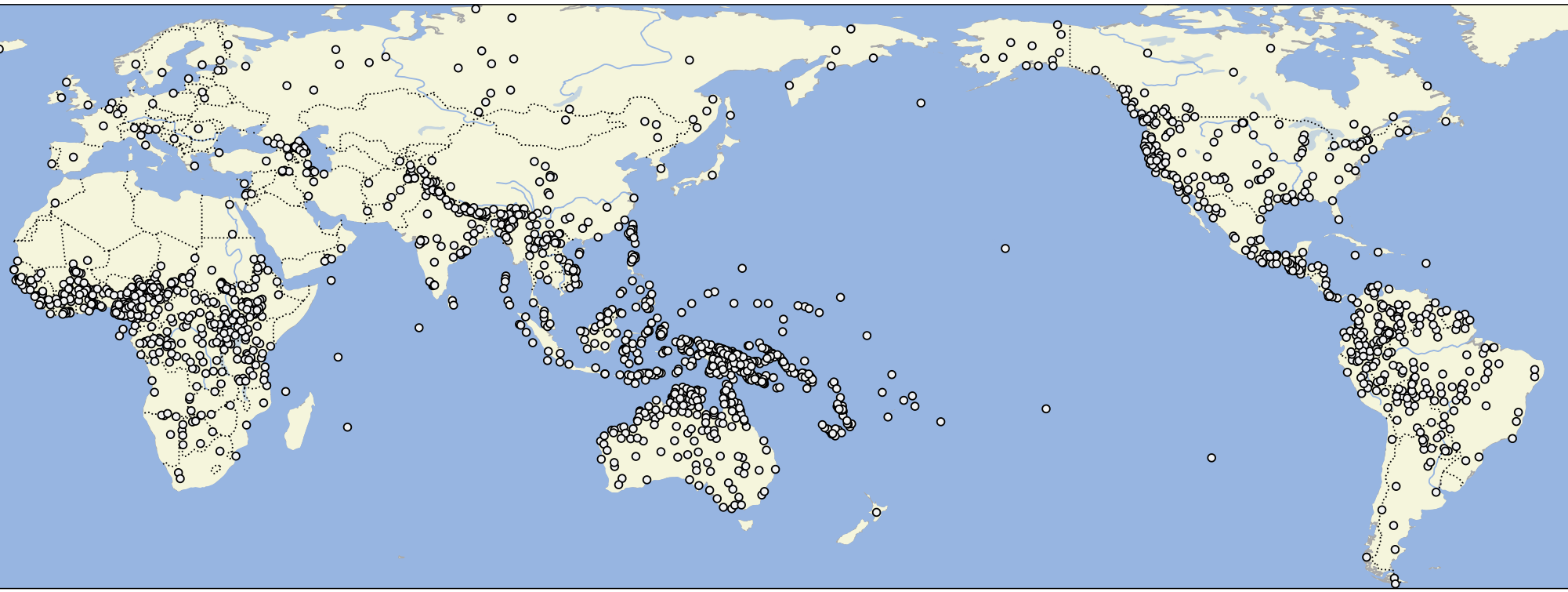
language science press

Joshi's group 1: scraping-bys



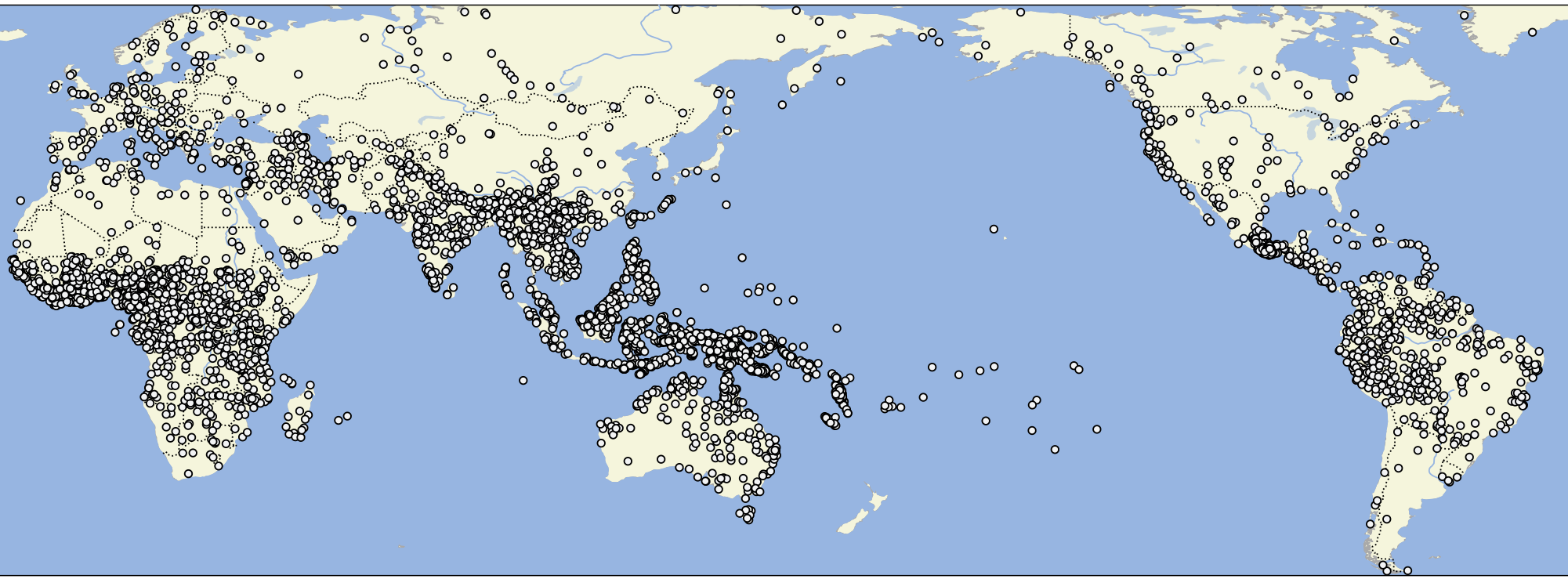
language science press

Joshi's group 0: left-behinds



language science press

Group -1: there is such a language??



language science press

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

- 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)

language science press

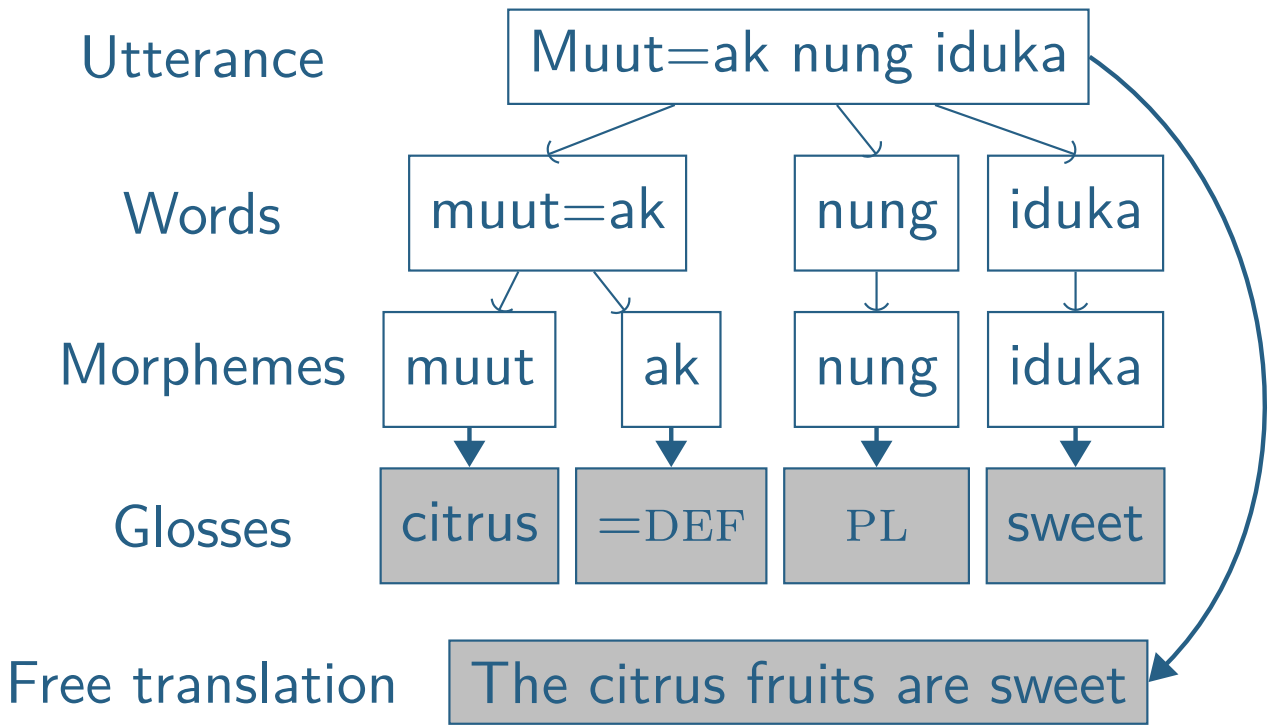
Common datastructure: IGT

⟩ IGT: interlinear glossed text

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

language science press

Common datastructure: IGT



THE ATLAS OF PIDGIN AND CREOLE LANGUAGE STRUCTURES ONLINE

Example 22-78

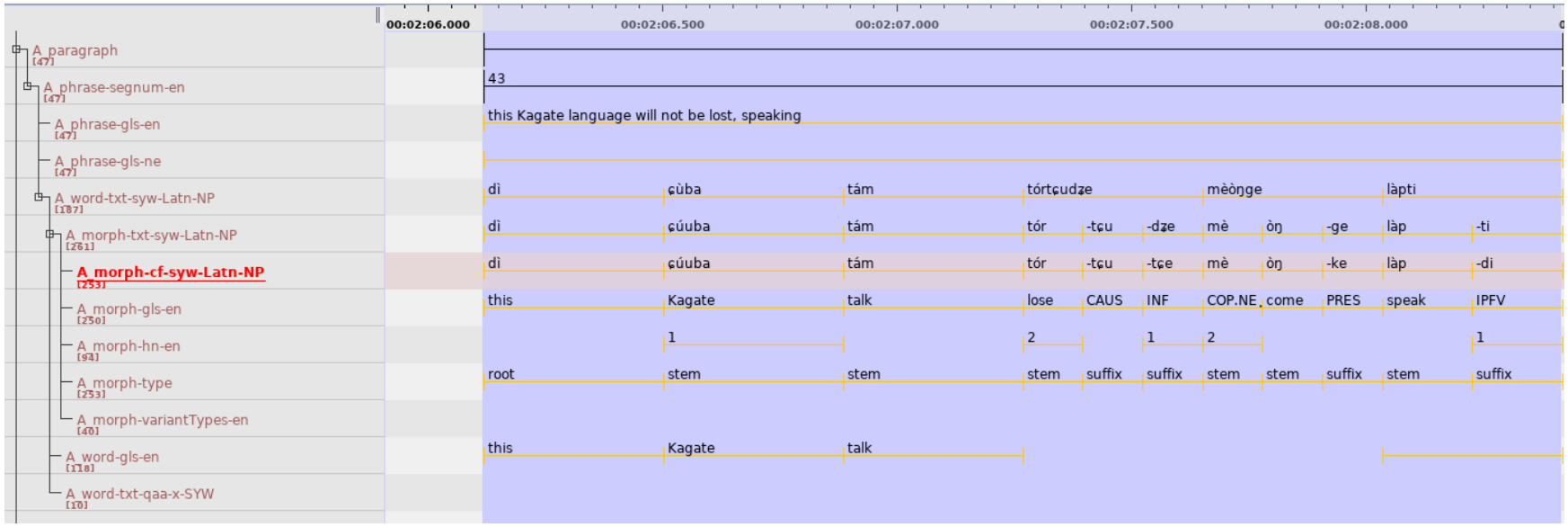
Bai mitupela i ringim taksi.
Bai mitupela i ring-im taksi.
FUT 1DU.EXCL PM ring-TR taxi
'We'll ring a taxi.'

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

(Chiarcoslonov)

language science press

ELAN files from endangered language archive



language science press

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

language science press

A grammar of Gyeli

A grammar of Gyeli

Nadine Grimm

Synopsis

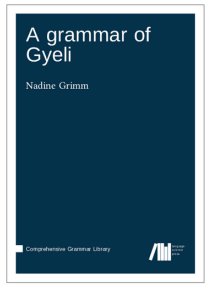
This grammar offers a grammatical description of the *Ngòló* variety of Gyeli, an endangered Bantu (A80) language spoken by 4,000-5,000 "Pygmy" hunter-gatherers in southern Cameroon. It represents one of the most comprehensive descriptions of a northwestern Bantu language.

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

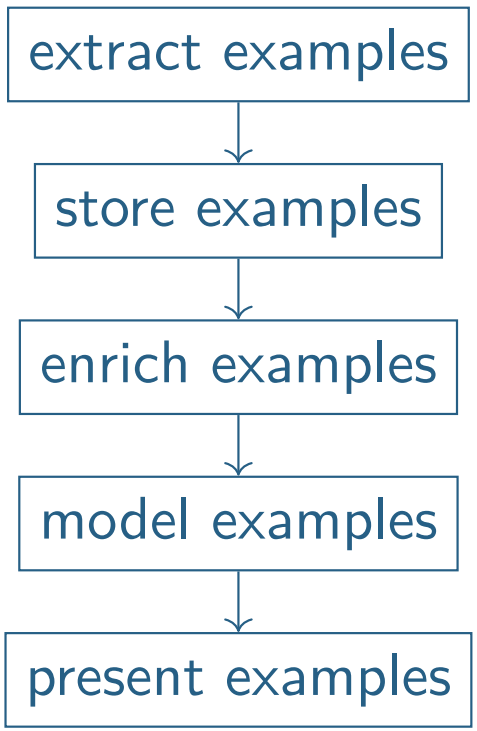
1 contributor

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
```

language science press

pipeline



extract examples

KamangSchapper, 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.'
```

language science press

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."
}
```

language science press

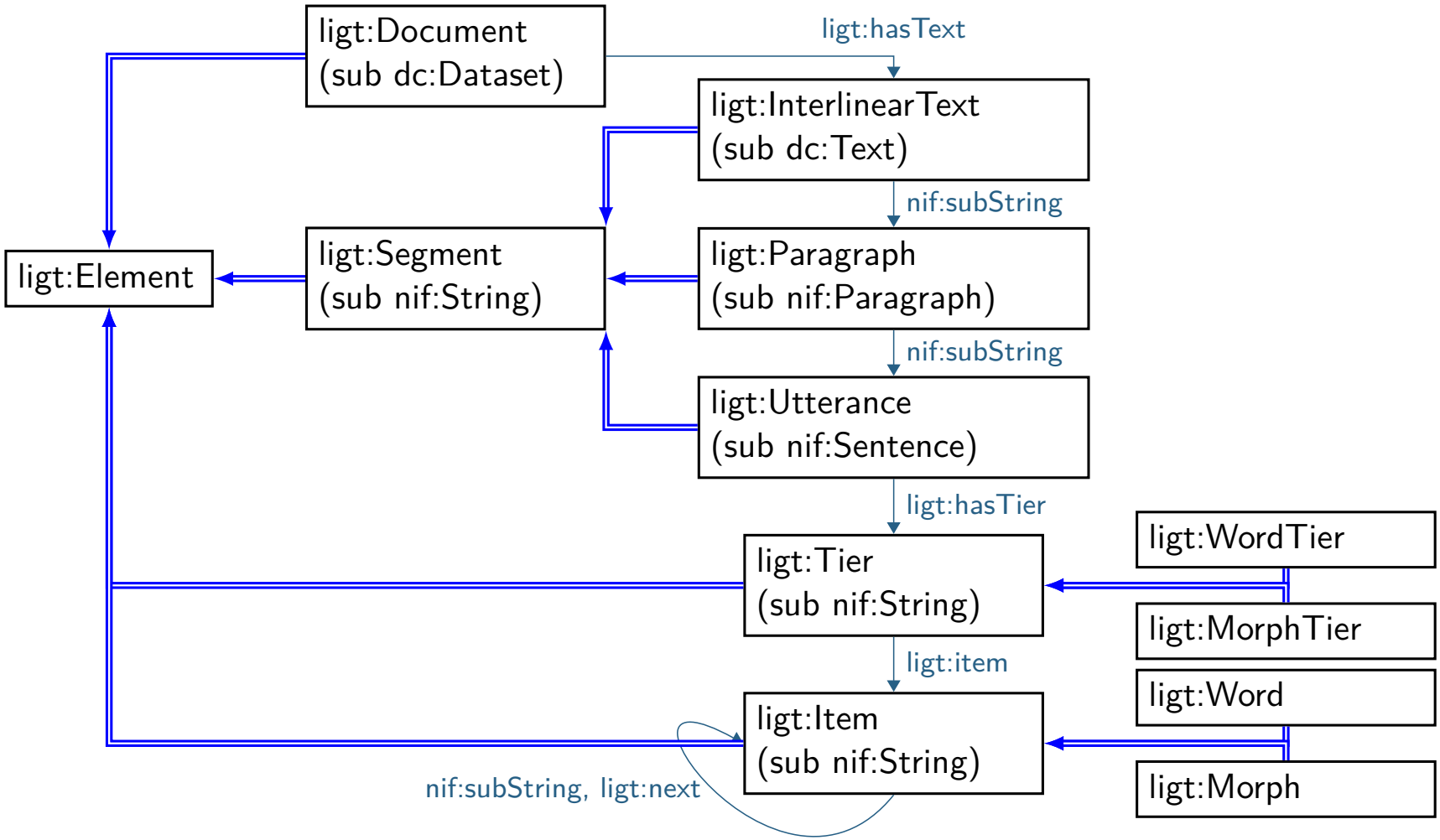
Enrich examples

- 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"
     }
 ]
}
```

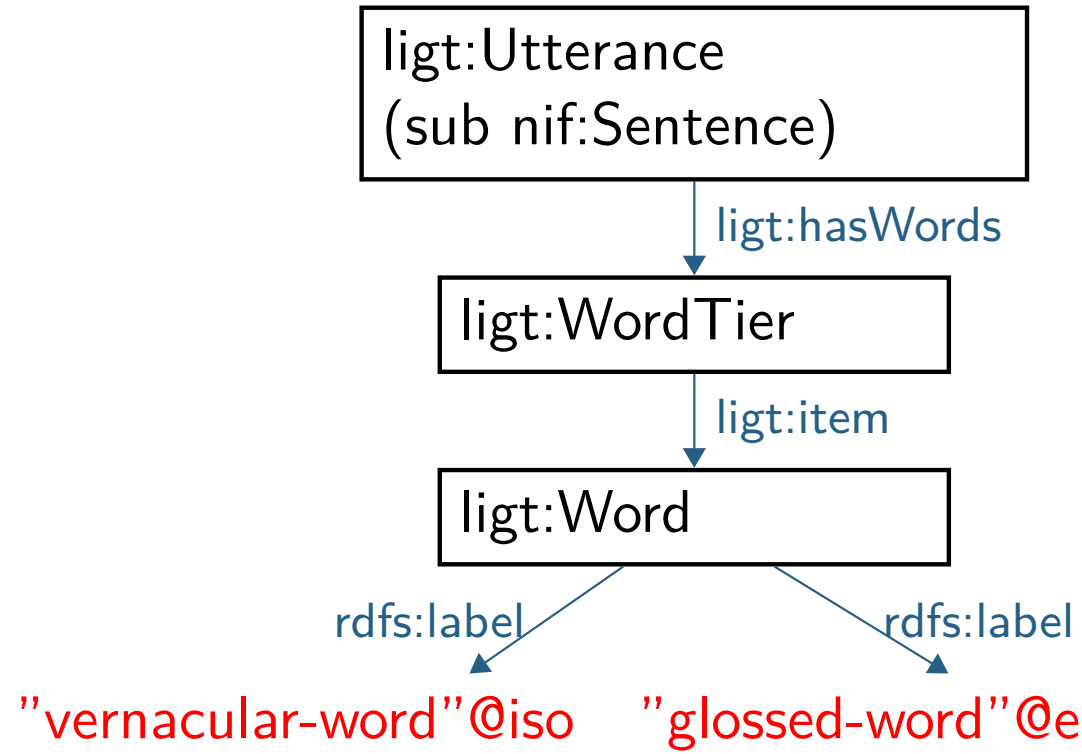
language science press

Model with LIGT



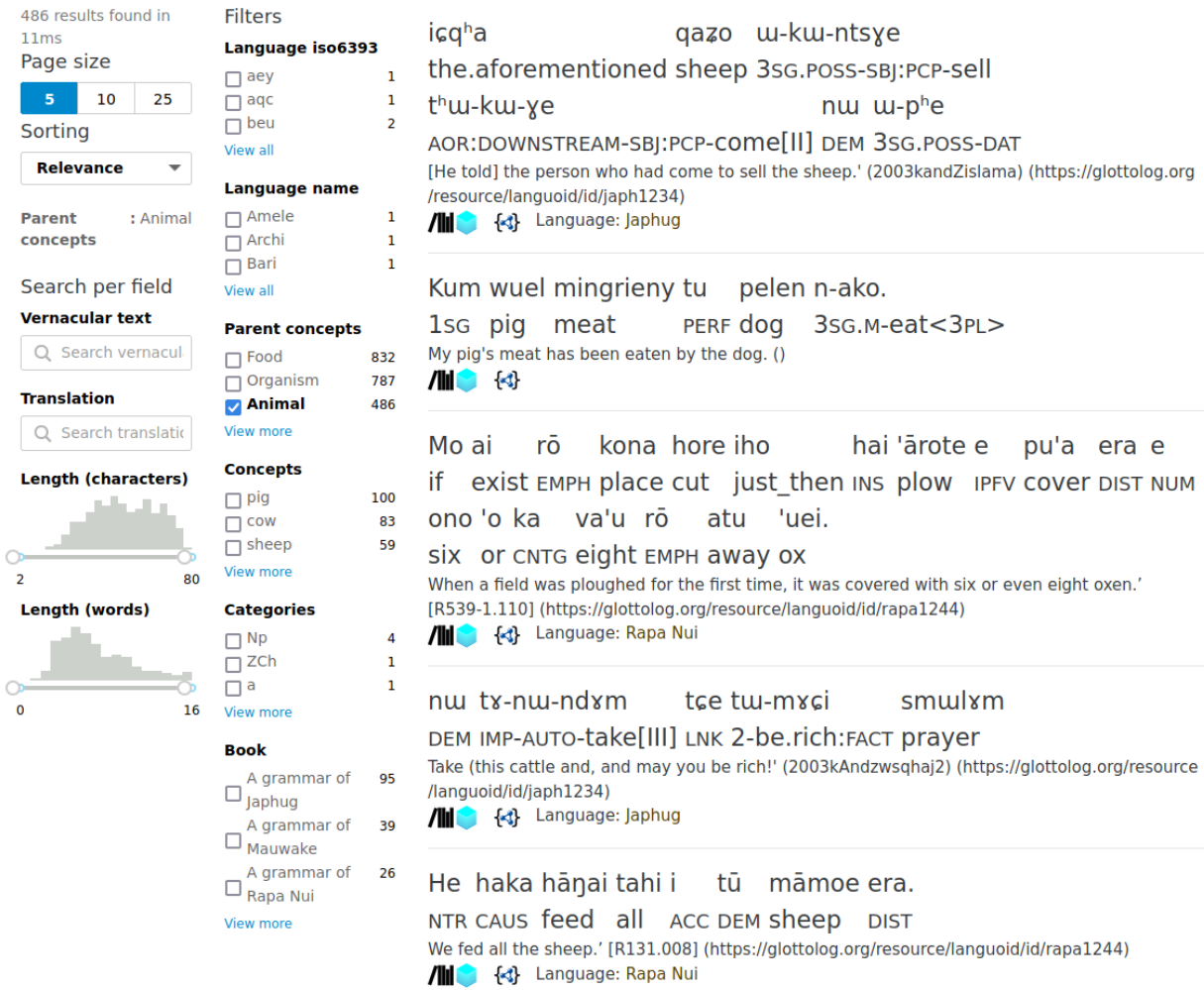
language science press

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



486 results found in 11ms

Page size: 5 | 10 | 25

Sorting: Relevance

Parent concepts : Animal

Search per field

Vernacular text: Search vernacul

Translation: Search translatic

Length (characters): 2 – 80

Length (words): 0 – 16

Filters

Language iso6393
- aey 1
- aqc 1
- beu 2

View all

Language name
- Amele 1
- Archi 1
- Bari 1

View all

Parent concepts
- Food 832
- Organism 787
- Animal 486

View more

Concepts
- pig 100
- cow 83
- sheep 59

View more

Categories
- Np 4
- ZCh 1
- a 1

View more

Book
- A grammar of Japhug 95
- A grammar of Mauwake 39
- A grammar of Rapa Nui 26

View more

iɕqʰa qaʑo ɯ-kɯ-ntsɣe
the.aforementioned sheep 3SG.POSS-SBJ:PCP-sell
tʰɯ-kɯ-ɣe nɯ ɯ-pʰe
AOR:DOWNSTREAM-SBJ:PCP-come[II] DEM 3SG.POSS-DAT
[He told] the person who had come to sell the sheep.' (2003kandZislama) (https://glottolog.org/resource/languoid/id/japh1234)
Language: Japhug

Kum wuel mingrieny tu pelen n-ako.
1SG pig meat PERF dog 3SG.M-eat<3PL>
My pig's meat has been eaten by the dog. ()

Mo ai rō kona hore iho hai 'ārote e pu'a era e
if exist EMPH place cut just_then INS plow IPFV cover DIST NUM
ono 'o ka va'u rō atu 'uei.
six or CNTG eight EMPH away ox
When a field was ploughed for the first time, it was covered with six or even eight oxen.' [R539-1.110] (https://glottolog.org/resource/languoid/id/rapa1244)
Language: Rapa Nui

nɯ tɤ-nɯ-ndɤm tɕe tɯ-mɤɕi smɯlɤm
DEM IMP-AUTO-take[III] LNK 2-be.rich:FACT prayer
Take (this cattle and, and may you be rich!' (2003kAndzwsqhaj2) (https://glottolog.org/resource/languoid/id/japh1234)
Language: Japhug

He haka hāŋai tahi i tū māmoe era.
NTR CAUS feed all ACC DEM sheep DIST
We fed all the sheep.' [R131.008] (https://glottolog.org/resource/languoid/id/rapa1244)
Language: Rapa Nui

present examples: machines

⟩ JSON-LD


```
▼ http://purl.org/dc/terms/isPartOf:
  ▼ 0:
      @id:                                                  "https://langsci-press.org/catalog/book/157"
  ▶ http://purl.org/dc/termWors/hasPart:                    […]
  ▶ http://purl.org/dc/terms/license:                       […]
  ▶ http://purl.org/dc/elements/1.1/subject:                […]
▼ https://imtvault.org/content/static/ligt-0.2.ttl#hasWords:
  ▼ 0:
      @id:                                                  "https://imtvault.org/wl09-cb9806ea53_wt"
    ▶ http://www.w3.org/2000/01/rdf-schema#label:           […]
    ▼ @type:
        0:                                                  "https://purl.org/liodi/ligt#WordTier"
    ▼ https://imtvault.org/content/static/ligt-0.2.ttl#item:
      ▼ 0:
          @id:                                              "_:wl09-cb9806ea53_0"
        ▼ http://www.w3.org/2000/01/rdf-schema#label:
          ▶ 0:                                              {…}
          ▶ 1:                                              {…}
        ▼ @type:
          ▶ 0:                                              "https://imtvault.org/con…static/ligt-0.2.ttl#Word"
        ▶ https://imtvault.org/content/static/ligt-0.2.ttl#nextWord:  […]
      ▼ 1:
          @id:                                              "_:wl09-cb9806ea53_1"
        ▶ http://www.w3.org/2000/01/rdf-schema#label:       […]
```

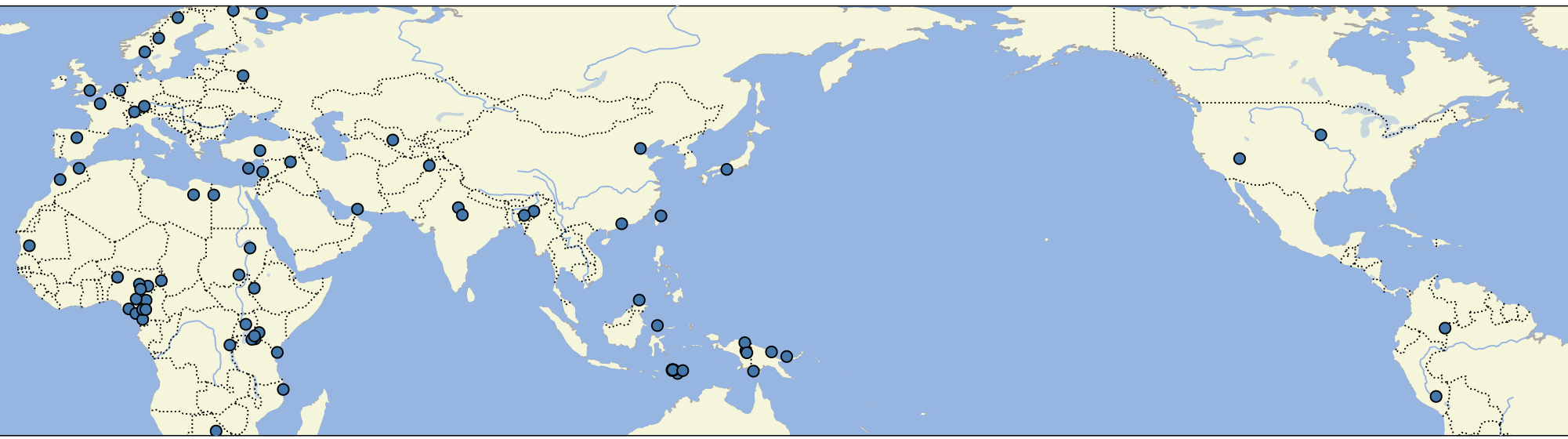
present examples: machines

⟩ nquads

```
_:b0 <http://purl.org/dc/terms/identifier> <https://www.wikidata.org/wiki/Q1364>
_:b0 <http://www.w3.org/2000/01/rdf-schema#label> <https://imtvault.org/Fruit> .
_:b1 <http://purl.org/dc/terms/identifier> <https://www.wikidata.org/wiki/Q2095>
_:b1 <http://www.w3.org/2000/01/rdf-schema#label> <https://imtvault.org/Food> .
_:b10 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/woi> .
_:b11 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/en-x-lgr>
_:b12 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://imtvault.org/con
_:b12 <http://www.w3.org/2000/01/rdf-schema#label> _:b13 .
_:b12 <http://www.w3.org/2000/01/rdf-schema#label> _:b14 .
_:b13 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/woi> .
_:b14 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/en-x-lgr>
_:b2 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/woi> .
_:b3 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/en-x-lgr> .
_:b4 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/woi> .
_:b5 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/en-x-lgr> .
_:b6 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://imtvault.org/cont
_:b6 <http://www.w3.org/2000/01/rdf-schema#label> _:b7 .
_:b6 <http://www.w3.org/2000/01/rdf-schema#label> _:b8 .
_:b6 <https://imtvault.org/content/static/ligt-0.2.ttl#nextWord> _:b9 .
_:b7 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/woi> .
_:b8 <http://purl.org/dc/elements/1.1/language> <https://imtvault.org/en-x-lgr> .
```

language
science
press

Conclusion: added languages



language science press

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 or so required for serious NLP.
- Glottolog will be a useful to aggregate the data
  - data is available in CLDF format as well

language science press

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