20 000 files under the sea

What can distant reading of language archives tell us?

by Sebastian Nordhoff
(Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS))
on September 24, 2020

» Depth of data analysis

1. Formalized data
   $[\ \text{+NASAL}\ ] \rightarrow [\ \text{+LABIAL}\ ] / \_\_ [\ \text{+LABIAL}\ ]$
2. Descriptive data "There is regressive assimilation"
3. Raw data "kithampe"

» Depth of data analysis

1. Formalized data
   $[\ \text{+NASAL}\ ] \rightarrow [\ \text{+LABIAL}\ ] / \_\_ [\ \text{+LABIAL}\ ]$
2. Descriptive data "There is regressive assimilation"
3. Raw data "kithampe"

4. Annotated data: kitham=pe
   1PL=POSS
   'our'

» The QUEST project

- Quality - Established: Testing and application of curation criteria and quality standards for audiovisual annotated language data (Krifka, Seifart, Seyfeddinipur)
- 2019–2022
- reuse of digital annotated language data
  - analysis: what is being held in an archive?
  - mobilization: how can we make third parties interact with the archive?
- perspectives
  - prospective: development of standards, curation criteria, and workflows
  - retrospective: enrichment of existing legacy data
- today: retrospective analysis of data from 4 different DELAMAN archives
- what types of answers can we get from their holdings?

» The DELAMAN archives

- Digital Endangered Languages and Musics Archives Network (DELAMAN)
- 12 full members; 7 associate members
- President: Mandana Seyfeddinipur
- Archives considered in this talk:
  - The Archive of the Indigenous Languages of Latin America (AILLA)
  - Endangered Languages Archive at SOAS University of London (ELAR)
  - Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)
  - The Language Archive/Dobes (TLA)
  - ~~Alaska Native Language Archive (ANLA)~~

» AILLA

* The Archive of the Indigenous Languages of Latin America
* 226 collections

» ELAR

- Endangered Languages Archive at SOAS University of London
- 544 collections

Endangered Languages Archive at SOAS University of London

All Fields Search

Search:

Level: Deposit ✖

Level

Deposit ✖

Funding body

ELDP	(443)
National Science Foundation	(21)
Leverhulme Trust	(6)
AHRC	(3)
British Academy	(2)
DEL	(1)

more ...

Status

Collection online	(455)
Forthcoming	(89)

Country

Australia	(36)
Papua New Guinea	(29)
China	(28)
Mexico	(28)
India	(23)
Indonesia	(22)

more ...

Language

Showing **1** - **544** of **544**

Title	Depositor	Language	Country
A Conversational Database of the Arapaho Language in Video Format	Andrew Cowell	Arapaho (ISO639-3:arp)	United States
A Digital Documentation of Panará	Bernat Bardagil Mas	Panará	Brazil
A Documentation and Description of the Arta language	Yukinori Kimoto	Arta (ISO639-3:atz)	Philippines
A Documentation of Cashibo-Cacataibo of San Alejandro (Pano) with a Focus on Information Structure	Daniel Valle	Cashibo-Cacataibo (ISO639-3:cbr)	Peru
A Documentation of Gurene Folk Tales, Riddles, Songs, Palace Genres and other Oral Genres in Bolga	Samuel Atintono	Gurene	Ghana
A Documentation of Saaroa, a Moribund Austronesian Language of Taiwan	Chiajung Pan	Saaroa (ISO639-3:sxr)	Taiwan
A Documentation of the vDab-pa Tibetan	Mian Zhang	vDab-pa (ISO639-3:)	China
A Documention of Tabaq, a Hill Nubian language of the Sudan, in its sociolinguistic context	Gerrit J. Dimmendaal Birgit Hellwig	Tabaq	Sudan
A Video and Text Documentation of Mangetti Dune !Xung	Amanda Miller	Mangetti Dune !Xung , Vasakela , !Xun	Namibia
A community-driven documentation of natural discourse in Anal, an endangered Tibeto-Burman language	Pavel Ozerov	Anal	India
A comprehensive documentation of Bine - a language of Southern New Guinea	Christian Döhler	Bine (ISO639-3:bon)	Papua New Guinea

» PARADISEC

* Pacific And Regional Archive for Digital Sources in Endangered Cultures
* 445 collections

PARADISEC

PARADISEC Catalog

Sebastian Nordhoff | Sign out

Home | Dashboard | Collections | Items | Contact

Please note that, due to attempted cyber-attack on our catalog, we have temporarily blocked new user registration. Please contact us if you need to register and we can arrange it manually. Our data and systems remain secure and intact.

Collections

Please enter search terms to fi | Search | Clear | Advanced Search

445 search results

Languages

Aceh (6)
Agob (4)
Amarasi (5)
Ambae, East (6)
Ambae, West (5)
Ambrym, North (4)
Aneityum (5)
Angoram (4)
Apma (4)
Arosi (4)

Countries

Afghanistan (1)
American Samoa (3)
Australia (121)
Austria (1)
Azerbaijan (1)
Bangladesh (2)
Benin (1)
Bhutan (3)
Bolivia (1)
Botswana (1)

Top 100 Collectors

Yanti (8)
Rob AMERY (2)
Alexander Adelaar (4)
Brigitte Agnew (1)
Barry Alpher (1)
Gregory Anderson (2)
Louise Baird (4)
Russell Barlow (2)
Danielle Barth (2)
Linda Barwick (3)

1 2 3 4 5 ... Next › Last »

ID ▲▼	Title ▲▼	Collector ▲▼	Countries ▲▼	Languages ▲▼	Creation Date ▲▼	Source university ▲▼	Actions
AA1	Recordings of Selako (Indonesia)	Alexander Adelaar	Indonesia	Kendayan	2007-09-14	University of Melbourne	View
AA2	Recordings of Embaloh (Indonesia)	Alexander Adelaar	Indonesia	Embaloh	2007-09-14	University of Melbourne	View
AA3	Story in Sungkung and Salako (Indonesia)	Alexander Adelaar	Indonesia	Kendayan	2008-05-15	University of Melbourne	View
AA4	Ma'anyan narratives	Alexander Adelaar	Indonesia	Benyadu' Ma'anyan Siang	2014-07-28	University of Melbourne	View
	Recordings of various texts in Amarasi, a language						

» TLA/Dobes archive

* 68 collections

The Language Archive ELAN Forums Help Login MPI Archive »

Browse Archive Browse by

Search

Filters

Access Level
(number of bundles containing)
info

Restricted (14514) + -
Registered (5417) + -
Open (4139) + -
Academic (26) + -

Contributor

- Frank Christian Seifart (921) + -
- Gabriele Mueller (784) + -
- Doris Fagua (676) + -
- Volker Heeschen (613) + -
- Christfried Naumann (500) + -
- Claudia Wegener (451) + -
- Raquel Guirardello-Damian (436) + -
- Silke Angelika Beuse (420) + -
- Gertrud Boden (414) + -
- Bruce Birch (343) + -

Show more

Language

- Spanish (2732) + -
- English (2661) + -
- Yurakaré (1645) + -
- Beaver (1042) + -
- Portuguese (885) + -
- Bora (701) + -
- Ocaina (652) + -

ARCHIVE / DOBES ARCHIVE

DOBES Archive

1 2 3 4 next last

Aché
The Aché Documentation Project (ADOP) documents the language, practices, and cultural knowledge of the indigenous Aché groups of eastern Paraguay. This project focuses on the traditional language of the communities in all its varieties. Data include audio and video recordings of mythology, life history narratives, cultural practices, and traditional songs. Recordings are annotated in ELAN and Toolbox-files. A second project—the Aché Language Studies Project (ALSP) —investigates language contact in history and as the result of present-day cultural contact and change. The coordinators of both projects are Jost Gippert and Sebastian Drude. Research is being conducted by Eva-Maria Roessler, Jan...

Akie
The Akie of Tanzania are a traditional hunter-gatherer society whose language is seriously endangered. The language, presumably a member of the Kalenjin branch of the Southern Nilotic languages, is still actively spoken in three villages of northeastern Tanzania, but the majority language and culture in the Akie-speaking area is Maa (speaking the Maasai dialect), which belongs to the Eastern Nilotic branch of the Nilotic family. The total number of Akie people is estimated at roughly 2500 people, but the number of people still speaking the language is presumably below 200. The massive impact of Maa language, culture, and life style plus the increasing influence of Bantu languages, including...

Aru languages

Awetí
The main objective of the Awetí Language Documentation Project is the comprehensive documentation of the language spoken by the Awetí, an indigenous tribe of just over a hundred people living in the Xingú headwater area of central Brazil. This comprehensive multimedia corpus is suitable as a basis for further research on the language and culture of the Awetí even if and when the language and culture become extinct. The Awetí Project closely co-operated with two other DOBES projects on languages in the area (Trumai and Kuikuro / Upper Xinguan Karib): all these languages are genetically unrelated. Co-operation was to ensure, among other things, the creation of analogous corpora for the

» OLAC

* Open Language Archives Community

Participatin

OLAC Language Resource Catalog

Search for language resources go

Navigating the Catalog
- Catalog Home
- Search Strategies
- Advanced Search
- New: Records recently added or modified

Quick Links
- Browse by Language
- Browse by Country
- Browse by Linguistic Field
- Browse by Linguistic Type
- Browse by Language Family

Contacts
- Email Us

More information
- OLAC Homepage
- OLAC FAQ
- Participating Archives

This catalog, developed by the **Open Language Archives Community (OLAC)**, provides access to a wealth of information about thousands of languages, including details of text collections, audio recordings, dictionaries, and software, sourced from dozens of digital and traditional archives.

Browse the OLAC records by Geographic region or by Language:

- Undetermined (40207)
- English (23053)
- Dutch (19460)
- German (8814)
- Spanish (8458)
- Russian (5423)
- Japanese (4345)
- Turkish (3785)
- Bathari (3429)
- French (3229)
- Indonesian (3120)
- Yuracare (1895)
- Mandarin Chinese (1870)
- Kachin (1827)
- Yele (1754)
- Khmu (1698)
- Tzeltal (1640)
- Portuguese (1589)
- No linguistic content (1459)
- Tok Pisin (1436)

View more..

» Acquisition

* enumerate: what do we have?
* authenticate: provide identity and credentials
* harvest: download resources we have access to

» Enumerate

* partially possible via OLAC and the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)
* but information is not complete, so we have to extract (scrape) the websites of the archives, eg ELAR below.

Access	Name	Type	Resource
O U S	D04NKapa.eaf	elan	Download
O U S	D04NKapa.txt	document	Download
O U S	D04NKapa.wav	audio	00:00 06:37 Download

Showing 1 to 3 of 3 entries

Previous 1 Next

Inspector Console Debugger Style Editor Performance Memory Network Storage Accessibility Application What's New

Search HTML

```
<div id="record-tabs" class="tab-content">
  <div id="holdings-tab" class="tab-pane active">
    <!--CUSTOM CODE FOR SOAS LIBRARY ELAR @author Simon Barron sb174@soas.ac.uk-->
    whitespace
    <div id="example_wrapper" class="dataTables_wrapper form-inline dt-bootstrap no-footer">
      <div class="row">…</div>
      <div class="row">
        ::before
        <div class="col-sm-12">
          <table id="example" class="table table-striped table-bordered dataTable no-footer" role="grid" aria-describedby="example_info" style="width: 95%;" width="95%"
            cellspacing="0"> event
            <thead>…</thead>
            <tbody>
              <tr class="odd" role="row">
                <td class="table-align sorting_1">…</td>
                <td class="table-align">
                  <input type="hidden" name="nodeid" value="">
                  <input type="hidden" name="userid" value="">
                  <input type="hidden" name="filename" value="">
                  <a onclick="insertStat('Sebastian','Nordhoff','snordhoff@soas.ac.uk','00… 08:01:14','','57484','V','U','elan','{ELDP}','1350340','0')"
                  href="https://elar.soas.ac.uk/MyResearch/Resource/0003-Goemai_Texts-b-Annotations-b-D04NKapa.eaf" download="">D04NKapa.eaf</a> event
                </td>
                <td class="table-align">elan</td>
                <td class="table-align">…</td>
```

```
Filter Styles :hov .cls
Pseudo-elements
This Element
element { inline
}
bootstrap-custom.css:802
a:hover, a:focus {
  //background-color: #FFF !important;
  //color: black;
}
a:hover, a:focus { compiled.css:1
  color: #092b47;
  text-decoration: underline;
}
a:active, a:hover { compiled.css:1
  outline: 0;
}
a { compiled.css:1
  color: #12538b;
  text-decoration: none;
```

» Authenticate

* Resources are either available to
  0) everybody,
  1) to registered users,
  2) upon request, or
  3) never.
* Register with all archives in order to get access to levels 0 and 1.
* Find a way to log in via the command line, bypassing the browser.
  * OK for AILLA, ELAR, TLA, more cumbersome for PARADISEC.

» Log in and download

```
with requests.Session() as s:
    s = requests.Session()
    un_name = "name"
    pw_name = "pass"
    values = {
        un_name: username.strip(),
        pw_name: password,
        "op": "log+in",
        "form_id": "user_login_block",
    }
    s.post(base_url, data=values)
    session_id = s.cookies.get_dict().get("SSESS64f35ecaf4903fe271ed0b0c15ee2bce")
    b_root = url2root(s, base_url)
    collection_links = b_root.findall(".//div/dl/dd/a")
    collection_urls = [
        "https://ailla.utexas.org/%s" % a.attrib["href"] for a in collection_links
    ]
```

* try to access each enumerated resource
* if access is denied, skip to the next item in list
* otherwise, download the resource and store it locally
* script runs for a couple of hours per archive, altogether several days
* Result: ~20 000 ELAN files could be retrieved

» ELAN

- de facto standard for annotation of linguistic multimedia recordings
- supports a variety of audio and video formats
- multi-speaker support
- time-aligned annotations
- annotations are organized in tiers
- tiers are organized hierarchically:
  - `translation` is a child tier of `transcription`

» Multi-speaker

File Edit Annotation Tier Type Search View Options Window Help

Grid | Text | Subtitles | Lexicon | Comments | Recognizers | Metadata | Controls

Volume: 100 0

Rate: 100 0

00:57:26.000 Selection: 00:00:00.000 - 00:00:00.000 0

Selection Mode Loop Mode

00:57:18.000 00:57:19.000 00:57:20.000 00:57:21.

WAD phoneti [15] — sivererei pasi pau

translation s [0]

EAG [0]

translationEA [0]

sp1bi [12] — bintang-bintang banyak

trans sp1bi [0]

notes [1]

WAD BI wd [0]

sp1 BI free [0]

» Translation

File Edit Annotation Tier Type Search View Options Window Help

Grid | Text | Subtitles | Lexicon | Comments | Recognizers | Metadata | Controls

Volume: 100 0

Rate: 100 0

00:00:08.266 Selection: 00:00:00.000 - 00:00:03.002 3002

Selection Mode Loop Mode

00.000 00:00:01.000 00:00:02.000 00:00:03.000 00:00:04.000 00:00:05.000

A_Transcription-txt-mos [18] — saʔ ni vicʰi beʔ | ibə bəruʔ dɤ ta ni ŋa imə saʔ ni vicʰi abə tani

A_Words-txt-mos [0]

A_Translation-gls-en [18] — The story of tiger and armadillo | If we say how it happened, the tiger and the

A_Participant-note-en [0]

Interlinear-title-mos [0]

» Annotation

File Edit Annotation Tier Type Search View Options Window Help

Grid | Text | Subtitles | Lexicon | Comments | Recognizers | Metadata | Controls

Volume: 100 0

Rate: 100 0

00:02:08.746 Selection: 00:02:06.120 - 00:02:08.420 2300

Selection Mode Loop Mode

00:02:06.000 00:02:06.500 00:02:07.000 00:02:07.500 00:02:08.000 00:02:08.500

A_paragraph [47]											
A_phrase-segnum-en [47]	43										
A_phrase-gls-en [47]	this Kagate language will not be lost, speaking										
A_phrase-gls-ne [47]											
A_word-txt-syw-Latn-NP [187]	dì	ɕùba	tám	tórtɕudʑe			mèòŋge			làpti	
A_morph-txt-syw-Latn-NP [261]	dì	ɕúuba	tám	tór	-tɕu	-dʑe	mè	òŋ	-ge	làp	-ti
A_morph-cf-syw-Latn-NP [253]	dì	ɕúuba	tám	tór	-tɕu	-tɕe	mè	òŋ	-ke	làp	-di
A_morph-gls-en [250]	this	Kagate	talk	lose	CAUS	INF	COP.NE	come	PRES	speak	IPFV
A_morph-hn-en [94]		1		2		1	2				1
A_morph-type [253]	root	stem	stem	stem	suffix	suffix	stem	stem	suffix	stem	suffix
A_morph-variantTypes-en [40]											
A_word-gls-en [118]	this	Kagate	talk								
A_word-txt-qaa-x-SYW [10]											
interlinear-text-title-syw-Latn-NP [1]											
interlinear-text-comment-en [1]											
interlinear-text-title-en [1]											

» ELAN in the archives

	collections	eaf files	transcribed	duration
AILLA	10	1 674	1 447	532.7h
ELAR	201	13 758	10 139	2588.3h
PARADISEC	78	2 619	1 776	302.0h
TLA	68	1 695	1 441	506.5h
total	289	19 746	14 803	3929.5h

* note the difference between the total number of collections given for the archives in the introduction and the collections with at least one accessible ELAN file, given here

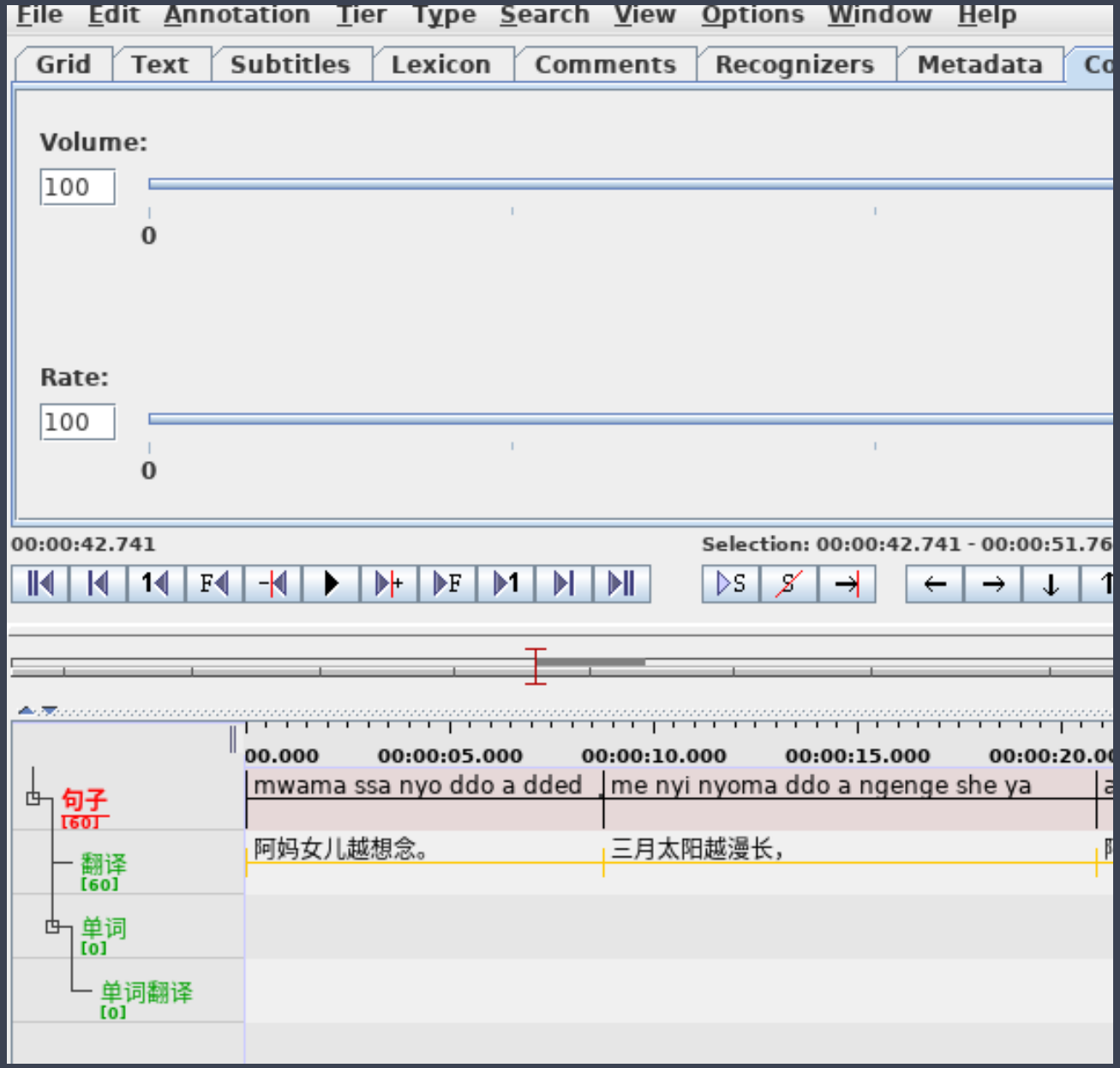
» XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<ANNOTATION_DOCUMENT>
    <TIME_ORDER>
        <TIME_SLOT TIME_SLOT_ID="ts1" TIME_VALUE="740"/>
        <TIME_SLOT TIME_SLOT_ID="ts2" TIME_VALUE="1860"/>
        <TIME_SLOT TIME_SLOT_ID="ts3" TIME_VALUE="3718"/>
        ...
    </TIME_ORDER>
    <TIER TIER_ID="ref@DAM" PARTICIPANT="Dambar Baram" ANNOTATOR="KP" LINGUISTIC_TYPE_REF="ref">
            <ANNOTATION>
                <ALIGNABLE_ANNOTATION ANNOTATION_ID="ann0" TIME_SLOT_REF1="ts1" TIME_SLOT_REF2="ts2">
                    <ANNOTATION_VALUE>. 001</ANNOTATION_VALUE>
                </ALIGNABLE_ANNOTATION>
            </ANNOTATION>
            <ANNOTATION>
                <ALIGNABLE_ANNOTATION ANNOTATION_ID="ann8" TIME_SLOT_REF1="ts3" TIME_SLOT_REF2="ts4">
                    <ANNOTATION_VALUE>. 002</ANNOTATION_VALUE>
                </ALIGNABLE_ANNOTATION>
            </ANNOTATION>
            ...
    </TIER>
    <TIER TIER_ID="ut@DAM" PARTICIPANT="Dambar Baram" ANNOTATOR="KP" LINGUISTIC_TYPE_REF="ut" PARENT_REF="ref@DAM">
            <ANNOTATION>
                <REF_ANNOTATION ANNOTATION_ID="ann1" ANNOTATION_REF="ann0">
                    <ANNOTATION_VALUE>əbə</ANNOTATION_VALUE>
                </REF_ANNOTATION>
            </ANNOTATION>
            <ANNOTATION>
                <REF_ANNOTATION ANNOTATION_ID="ann9" ANNOTATION_REF="ann8">
                    <ANNOTATION_VALUE>kunəi pudza tukle hon lə məlak</ANNOTATION_VALUE>
                </REF_ANNOTATION>
            </ANNOTATION>
            <ANNOTATION>
                <REF_ANNOTATION ANNOTATION_ID="ann36" ANNOTATION_REF="ann35">
                    <ANNOTATION_VALUE>hidi hudi pudza tukle alam alam wa lakle əbə</ANNOTATION_VALUE>
                </REF_ANNOTATION>
            </ANNOTATION>
            ...
    </TIER>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="ref"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="ut" CONSTRAINTS="Symbolic_Association"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="txd" CONSTRAINTS="Symbolic_Association"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="tx" CONSTRAINTS="Symbolic_Subdivision"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="mb" CONSTRAINTS="Symbolic_Subdivision"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="ge" CONSTRAINTS="Symbolic_Association"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="ft" CONSTRAINTS="Symbolic_Association"/>
</ANNOTATION_DOCUMENT>
```

» Analysis of the XML

* simply look up the tiers called ut (utterance), mb (morpheme breaks), ge (gloss English), ft (free translation) and be done!

» Free choice of tier labels



» Tier label inventories

glosses ge, morph-item, gl, Gloss, gloss, glosses, word-gls, gl (interlinear gloss).
translations eng, english translation, English translation, fe, fg, fn, fr, free translation, Free Translation, Free-translation, Free Translation (English), ft, fte, tf (free translation), Translation, tl, tn, tn (translation in lingua franca), tf_eng (free english translation), trad1, Traducción Español, Tradución, Traduccion, Translate, trad, traduccion, traducción, traducción , Traducción, Traducción español, Traduction, translation, translations, Translation, xe, 翻译.

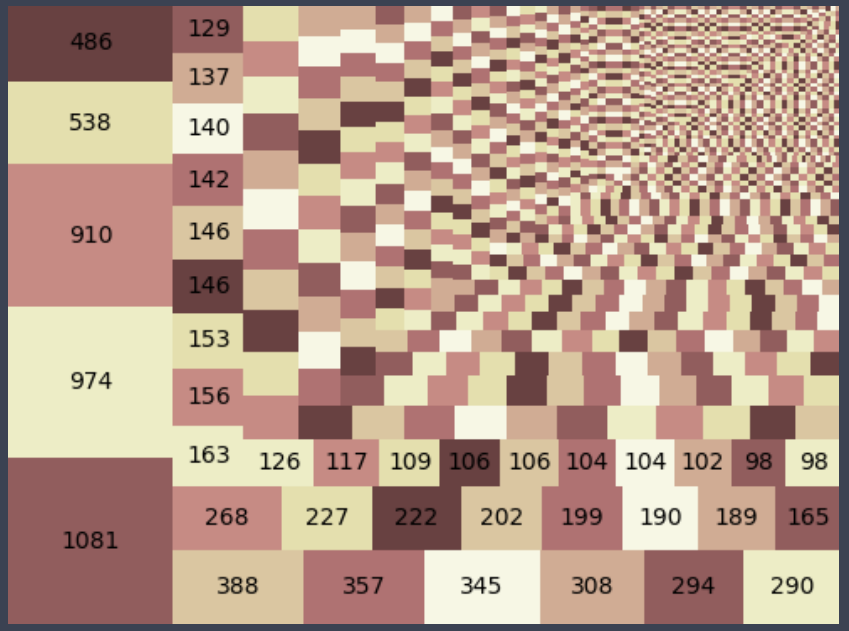
» Tier label inventories

transcriptions: arta, Arta, conversación, default-lt, default-lt, Dusun, Fonética, Frases, Hablado, Hakhun orthography, Hija, hija, ilokano, interlinear-text-item, Ikaan sentences, Khanty Speech, main-tier, Madre, madre, Matanvat text, Matanvat Text, Nese Utterances, o, or, orth, orthT, orthografia, orthografía, orthography, othography, po, po (practical orthography), phrase, phrase-item, Phrases, Practical Orthography, sentence, sentences, speech, Standardised-phonology, Sumi, t, Tamang, texo , text, Text, Text , texto, Texto, texto , Texto principal, Texto Principal, tl, time aligned, timed chunk, tl, Transcribe, Transcrição, TRANSCRIÇÃO, Transcript, Transcripción chol, transcripción chol, Transcripción, Transcripcion, transcripción, Transcripcion chol, transcript, Transcription, transcription, transcription_orthography, trs, trs@, trs1, tx, tx2, txt, type_utterance, unit, ut, utt, Utterance, utterance, uterrances, utterances, utterrances, Utterances, utterance transcription, UtteranceType, vernacular, Vernacular, vilela, Vilela, word-txt, word_orthography, xv, default transcript, 句子, 句子 , 句子 .

» Configurations

- Instead of tier names, we can look at the tier hierarchies
- Relation between tiers can be "time subdivision", "symbolic subdivision", "symbolic association"
- We can establish file fingerprints based on the tier hierarchies
- The root tier of a speaker is labelled "x"
- A file with the fingerprint [xx] would have two speakers and no further tiers
- A file with a [x[aa]] would have one speaker with two dependent tiers of type "symbolic association"
- A file with a [x[aa]x[aa]] would have two speakers with two dependent tiers of type "symbolic association" each
- Maybe some tier hierarchies are very frequent, and we can take advantage of this for our analyses? This would allow us to disregard the tier labels.

» Found tier hierarchies



- there are 2 187 different fingerprints
- each box represents a different tier hierarchy/fingerprint
- numbers show how often the given hierarchy occurs in the dataset
- conclusion: the tier hierarchies are very heterogeneous
- not a very promising route

» Semantic content analysis

- What are the holdings of the archives about?
- Most files have transcriptions, but not all files have translations
- Translation is a faithful rendering of the content of an utterance in another language
- Topics in the source language should be found in the translated language as well
- Named Entity Recognition with Grobid/NERD based on translations
  - "Gestern haben wir Karneval gefeiert und ich war als Bär verkleidet"
  - "Yesterday we celebrated carnival and I was disguised as a bear"

» Close reading/distant reading

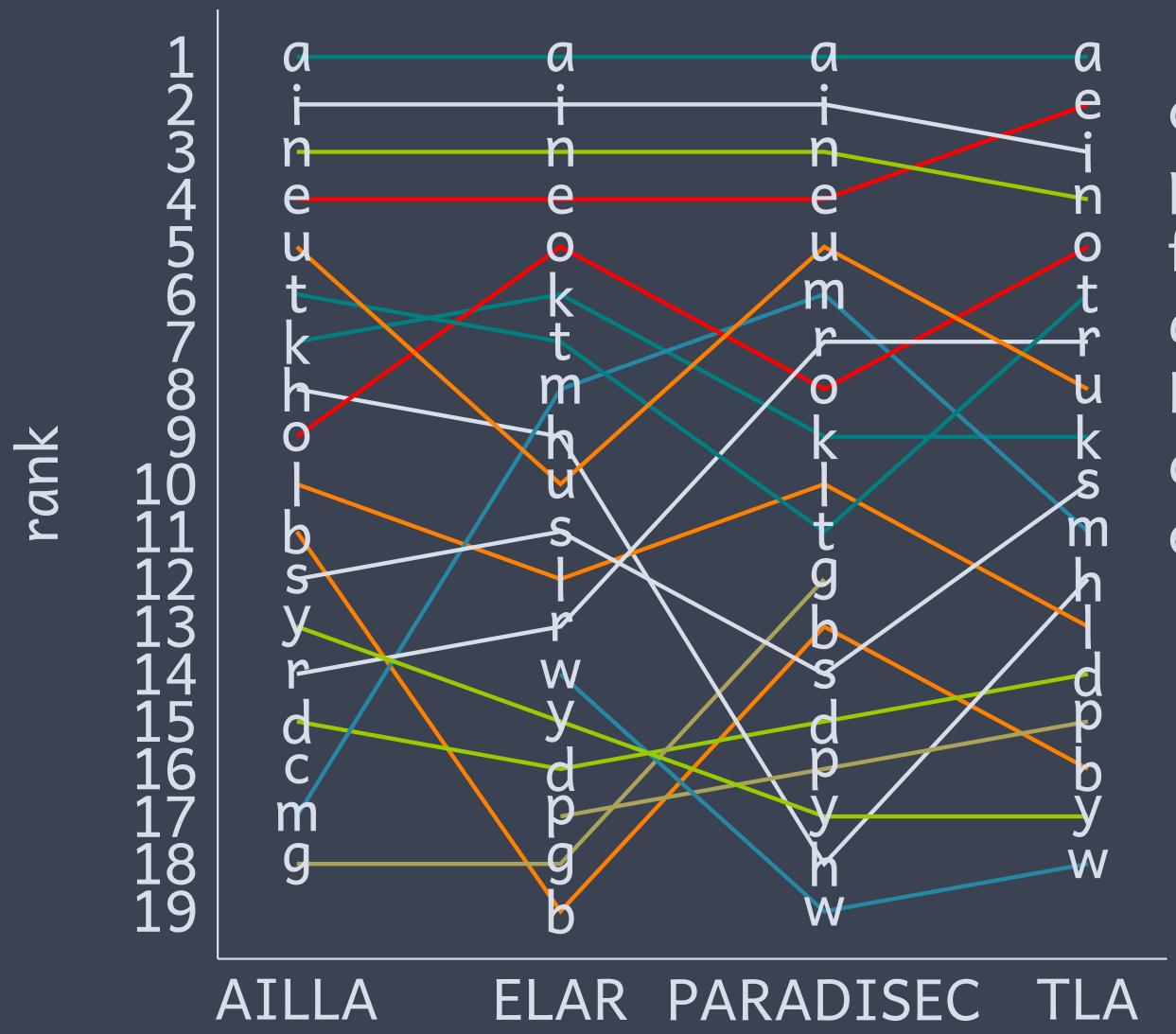
- Named Entity Recognition allows for Distant Reading of a collection
- Close reading: examine one text very thoroughly
- Distant reading: examine a wealth of texts, less thoroughly
  - originated in literature studies to cover "the great unread", not only the 1% of canonised texts
  - used in library science for instance to describe holdings
    - eg "78% of our holdings deal with Europe; 50% of our holdings deal with other continents; 15% of our holdings have no identified geographical coverage"

» Elements

	utterances	words	translations	glossed words	entity types	tokens
AILLA	633 520	1 957 913	*14 248	*57 371	1 532	703
ELAR	2 221 543	8 119 023	306 836	2 628 943	20 991	8 192
TLA	675 934	2 102 332	247 946	474 705	10 346	4 249
PARADISEC	224 923	942 615	49 641	105 243	1 163	683
total	3 755 920	13 121 883	*618 671	*3 266 262	34 032	(11 715)

* What kind of analyses can we run on these data?

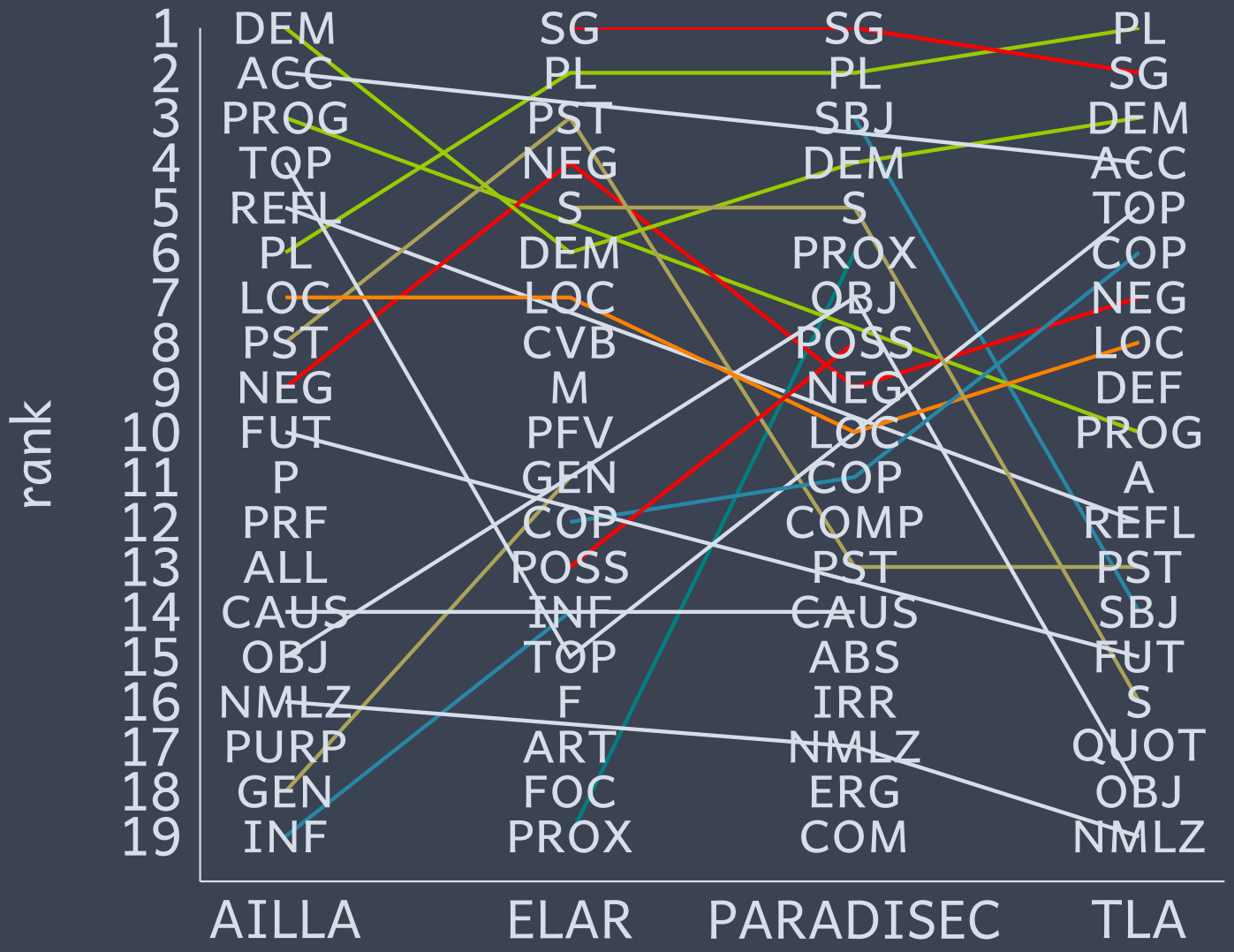
» Graphemes



compare phoneme type frequency according to Donohue database (6950 entries):

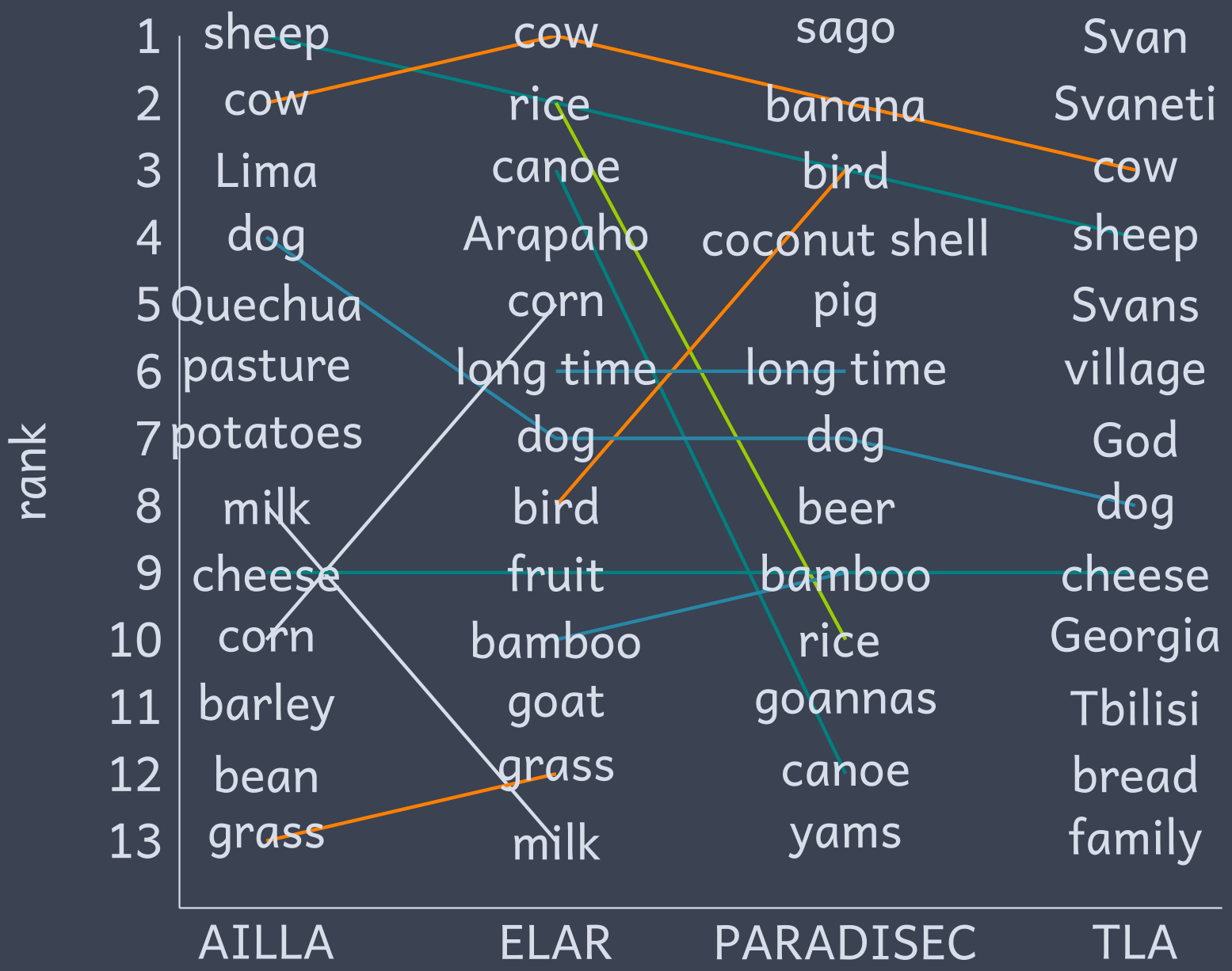
i	97%
a	96%
u	93%
e,ɛ	86%
o,ɔ	83%

» Grammatical categories



» Lexical glosses

» Named entities



» Biases in semantic domains

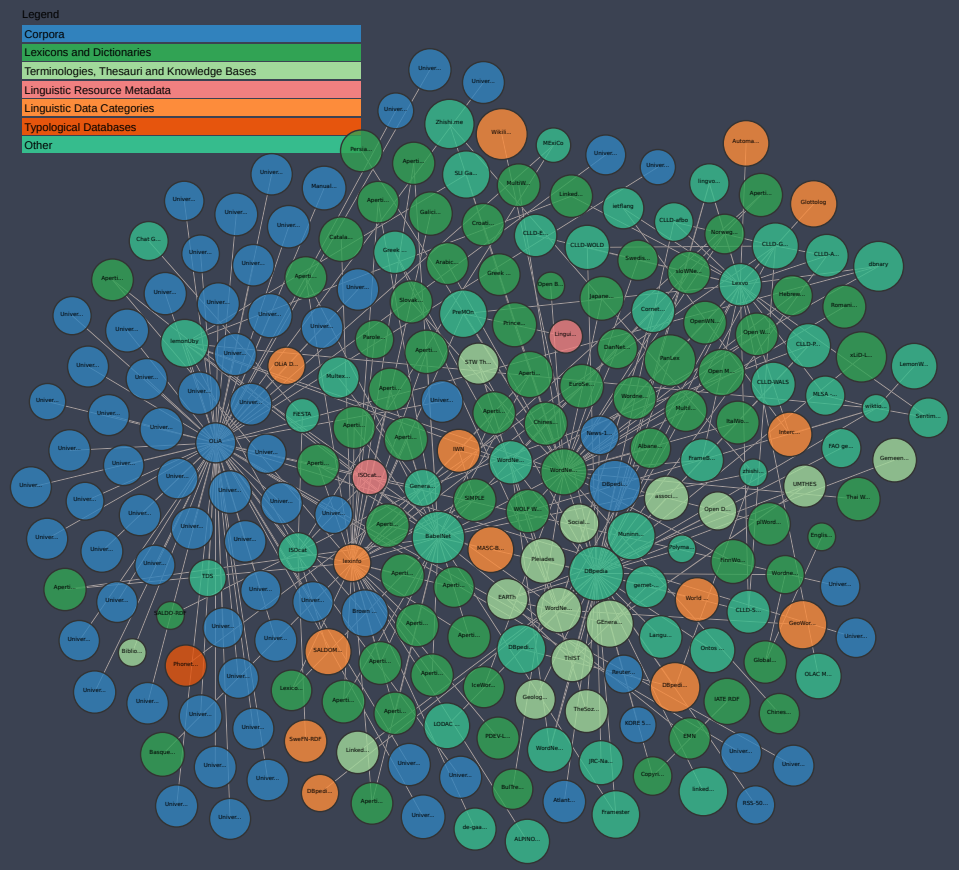
- Definitely too many Caucasus-related concepts in TLA (Svan, Svaneti, Svans, Georgia, Tbilisi)
  - One very large and thorough collection dominates the rest
  - TLA is biased towards the Caucasus
- But why are almost all other concepts about crops and livestock?
- Do the populations really talk about agriculture the whole day?
- When did you last talk about agriculture?
- Or is it maybe the linguists who happen to ask only about those domains?
- What does this tell us about the way Western academia envisions language documentation? Peasants all over the place?

» Shortcomings of the approach

- Dirty data
  - An unknown amount of the 20k files are faulty/erroneous
    - empty files
    - syntactically invalid files
    - empty tiers
    - semantic nonsense
    - semantic underspecification
  - But estimations are possible based on sampling
  - calibration
- Type/token problem
  - do we have more ⟨a⟩ because more lexemes have ⟨a⟩ or because the lexemes with ⟨a⟩ are more frequent? Or because there is one large collection of an ⟨a⟩-heavy language which dominates the rest?
  - areal/genealogical distribution
    - This can be controlled for by integrating data from other knowledge bases, but has not been done yet.

» Linguistic Linked Open Data Cloud

- Connect different knowledge bases via shared vocabularies.
- Every blob is a knowledge base, every edge is a connection
- Connections are typed, eg Work123 $\xrightarrow{\text{dc:language}}$ Lg456



» Three types of metadata/annotations

1. context (who, when, where, in what medium)
   * Dublin Core (Creator, Topic, Title, Language)
   * WGS84 (Geo coordinates)
   * Glottolog (Genealogical relationships)
2. structure (part-whole relations of the resource)
   * Linked Interlinear Glossed Text (LIGT; Utterances, Words, Glosses)
3. content (what)
   * Wikidata (Concepts)
4. We can integrate existing knowledge from these other repositories to enrich our own knowledge base.

» Usage

- discovery
  - integration of the Wikidata concept hierarchy (`instance_of`, `sublass_of`)
  - search for "bird" yields "cardinal woodpecker" as well
- merger
  - Dataset can be merged with the APICS dataset (not done yet)
- enrichment
  - Additional data from Glottolog, OLAC
    - control for genealogical, areal bias
  - Gloss translations from DBnary (Chiarcos et al. 2017)
    - can mediate between English/Spanish metalanguages, eg AILLA

» So, how far can you get with "unanalyzed" data?

- Different types of data:
  - raw
  - annotated
  - analyzed
- the data at hand show no traces of syntactic analyses
- possible research questions about:
  - languages: which languages have value X for feature F?
  - texts: which texts from which languages/cultures contain X?
  - scientific communities: who works on what?

» Semantic questions: foodstuff

- why is the cow more frequent than the pig in AILLA, ELAR & PARADISEC but it's the other way round in PARADISEC?
  - easy answer: PARADISEC contains mainly content from cultures from the Pacific, where the pig is a much more important animal than the cow.
  - the same is true for the dominance of potatoes (AILLA, Andes), sago (PARADISEC, Pacific), rice (ELAR, PARADISEC, Asia), and cheese (TLA, Caucasus)

» Number questions: why the frequency difference between SG/PL?

- in some archives, SG is more frequent as a gloss, in others PL
- explanations favouring SG:
  - there are more singular referents than plural referents in the world
  - text genres archived are more often monological
- explanations favouring PL:
  - optional number marking in many languages
  - even when number is obligatory, SG is often zero-marked
- which one is the right explanation?

» Graphemes: why the frequency differences between <a,i,e,u,o>?

- favouring ⟨a⟩
  - many orthographies use ⟨a⟩ for schwa as well, and schwa is a frequent sound.
- favouring ⟨e⟩
  - historical reasons: English and French colonizers use ⟨e⟩ for schwa, not ⟨a⟩
- favouring ⟨i,u⟩ over ⟨e,o⟩
  - maximize vocalic space
- favouring ⟨i,e⟩ over ⟨u,o⟩
  - frontness preferred (but why?)
- which one is the right explanation?
  - analogous approach for consonants

» Disentangling hypotheses/biases

- universal preferences
  - anatomy (phonemes), cognition (number?)
- language specific preferences
  - areal and genetic bias; “historical accidents”; extralinguistic factors
- collectors’ preferences
  - “agricultural bias”
- should apply to “analyzed” data as well

» Prospects for distant reading

- data in endangered language archives
  - nominal (X is present)
  - ordinal (There is more X than Y)
  - ~~scalar (X is factor 1.5 more than Y)~~
- how far can you get in phonology:
  - somewhere
- how far can you get in grammatical categories:
  - somewhere
- how far can you get in syntax:
  - nowhere
- how far can you get in (formal) semantics:
  - nowhere
- how far can you get in sociology of science:
  - good prospects

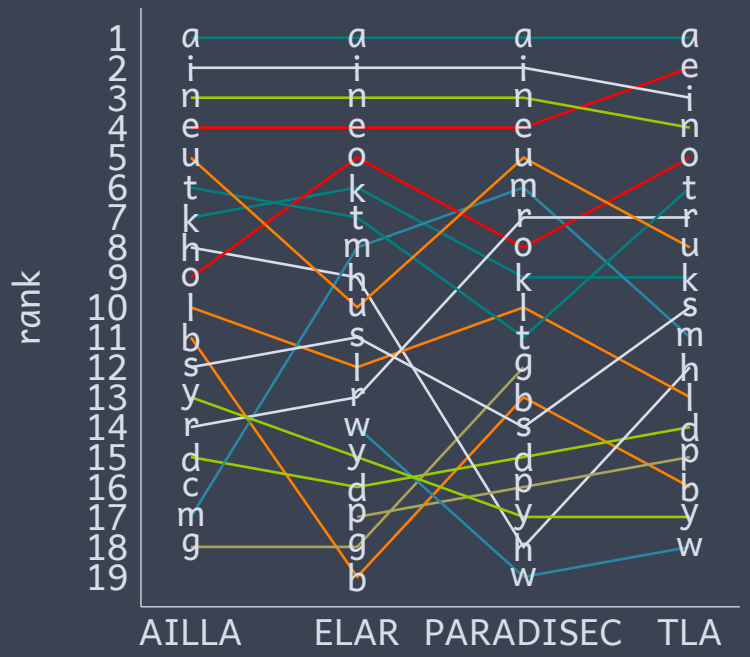
» Why do distant reading?

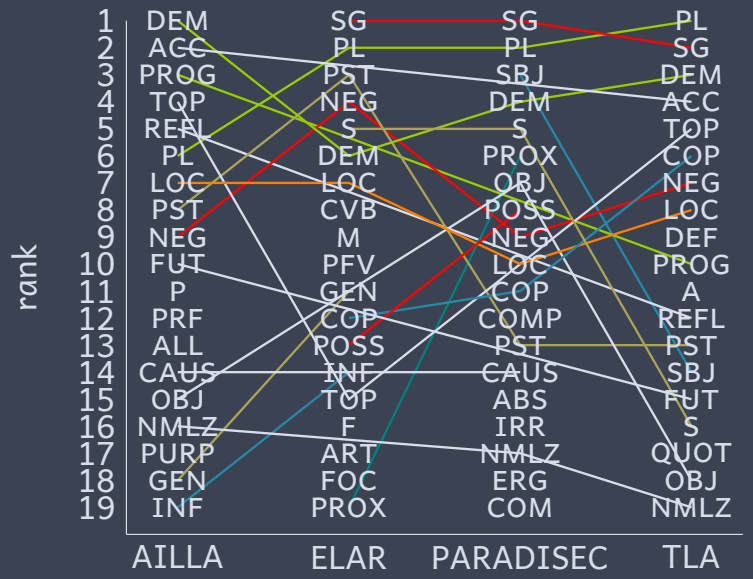
- costs for unanalyzed data
  - data acquisition: several days computing time
  - analysis: couple of hours computing time
  - colexification analysis of whole archive: 30 seconds
- costs for "analyzed data":
  - data acquisition: weeks to months to years
  - data analysis: months/years
- tradeoff between depth of analysis and time/cost to produce a resource

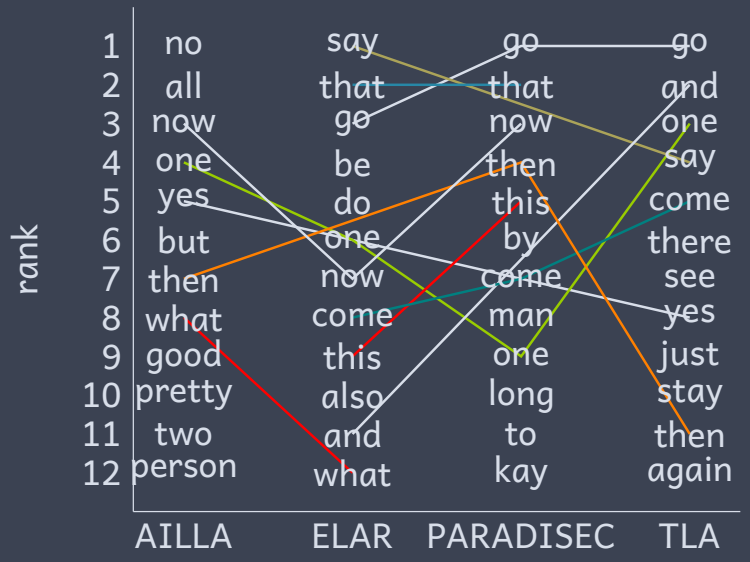
» Conclusion

* distant reading is a coarse approach
* can answer some questions
* can generate some hypotheses
* opens up surprising new fields for research

Thank you


rank
AILLA ELAR PARADISEC TLA


rank
AILLA ELAR PARADISEC TLA


rank
AILLA ELAR PARADISEC TLA


rank
AILLA ELAR PARADISEC TLA