Proceedings of LR4SSHOC: Workshop about Language Resources for the SSH Cloud, pages 10–18
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

From the attic to the cloud: mobilization of endangered language resources with linked data

Sebastian Nordhoff
Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS Berlin)
Schützenstr. 18, 10117 Berlin
nordhoff@leibniz-zas.de

Abstract

As an important example of the need to provide hosting and publication facilities for highly specific data types and the role thematic centres can play, this paper describes a collection of 20k ELAN annotation files harvested from five different endangered language archives. The ELAN files form a very heterogeneous set, but the hierarchical configuration of their tiers allow, in conjunction with the tier content, to identify transcriptions, translations, and glosses. These transcriptions, translations, and glosses are queryable across archives. Small analyses of graphemes (transcription tier), grammatical and lexical glosses (gloss tier), and semantic concepts (translation tier) show the viability of the approach. The use of identifiers from OLAC, Wikidata and Glottolog allows for a better integration of the data from these archives into the Linguistic Linked Open Data Cloud.
Keywords: endangered languages, corpus, ELAN, text mining, Linked Data

1. Introduction

One of the goals of linguistics is to gain insight into human cognition and culture. There are over 7 000 languages spoken in the world (Hammarström et al., 2019), varying wildly in structure, so we must have a large and diverse sample in order to gain any meaningful insight into what all human languages have in common. The amount of data to process is too large for one human brain, so that machine support is required. Unfortunately, NLP largely focuses on a very small number of languages spoken in the industrialized world. The wiki of the Association for Computational Linguistics lists NLP tools for 76 different languages,[1] i.e. about 1% of the worlds languages. It is true that there are text, audio, and video resources in other languages available, but these are often small, difficult to access, and even more difficult to reuse. Many of the resources for these lesser studied languages reside in endangered language archives such as TLA,[2] ELAR,[3] or PARADISEC.[4] While much of the content found in these archives is available for inspection in principle, there are significant issues of findability and interoperability, rendering its exploitation for NLP purposes difficult. This paper describes a workflow to identify, colllect and query the resources from five different endangered language archives from the DELAMAN network, giving access to 2 500 000 words in a structured format.

2. DELAMAN archives

DELAMAN (Digital Endangered Languages and Musics Archives Network) "is an international network of archives of data on linguistic and cultural diversity, in particular on small languages and cultures under pressure" (www.delaman.org). As such, DELAMAN is a very interesting starting point for the collection of processable resources for lesser studied languages. There are currently 12 member archives and 5 associated members, which hold content in 2420 different languages. For the purpose of this project, 5 archives were chosen for inclusion:

- AILLA (Archive of the Indigenous Languages of Latin America)
- ANLA (Alaska Native Language Archive)
- ELAR (Endangered Languages Archive at SOAS)
- PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures)
- TLA (The Language Archive at the Max Planck Institute for Psycholinguistics)

These archives vary in size, backend software, funding structure, and coverage of geographical areas. They have in common that their main focus has been on ingestion, and less so on mobilization. There are some query interfaces to identify resources of interest, but none of the archives offers an API or bulk downloads for instance.

3. Research with language archives: *The Language Archive* at the Max Planck Institute for Psycholinguistics

TLA used to be the "home archive" for the DoBeS programm (funded by the Volkswagen foundation), which funded 67 documentation projects for endangered languages. The last project funded started in 2011. In the course of these documentation projects, very interesting and important language data was collected and deposited in the archive. To this day, the researchers from these projects continue using the archive, still funded by the Max Planck Society. However, it is also true that there are only very few "third party" researchers, not involved in the original projects, which interact with the data. The Volkswagen foundation initiated so called phase-2 projects for theoretical research on language data stored in the archive, but only 5 such projects[5] have been awarded and as of today there seems to be no major research community interacting with archive data they have not deposited themselves. A continuation of these phase-2 efforts is the DoReCo project.[6] DoReCo "brings together spoken lan-

[1] https://aclweb.org/aclwiki/List_of_resources_by_language
[2] https://archive.mpi.nl/tla/
[3] https://elar.soas.ac.uk/
[4] http://www.paradisec.org.au/
[5] http://dobes.mpi.nl/research-projects
[6] http://doreco.info

guage corpora from about 50 languages, extracted from documentations of small and often endangered languages." But for this project, the original corpus creators are typically involved in the creation of an extra layer of annotation. It thus seems fair to say that the existing language archives are currently not available for inspection to researchers outside of the core community of language documenters.[7] Compare this with research on the SwitchBoard corpus (Godfrey and Holliman, 1993) or the Penn TreeBank (Marcus et al., 1999), where a lively community has grown around the initial resources and where most researchers are not in direct contact with the initial creators. Looking at possible reasons as to why the uptake of this vast resource of endangered language material is slow, we can come up with an unsurprising set of issues: findability, accessibility, interoperability, and reusability. For a given research question, researchers often need a resource which is a) in a particular format (text, audio, video) b) in a particular language (family) covering c) particular content and is d) accessible. The OLAC[8] (Simons and Bird, 2003) service provides querying capabilities for language and media type, but OLAC cannot guarantee that the resources it lists are indeed available. Since OLAC does not host the files, querying for content strings is not possible either.

A clear desideratum would be the possibility to query language resources based on metadata (region, language format, genre, as currently already possible via OLAC), but also on content. Content includes grammatical categories (give me all files with antipassive in them) but also semantic categories (give me all texts relating to agriculture). This paper will discuss a prototype which allows for such queries. OLAC is already part of the Linked Open Data Cloud (Chiarcos et al., 2012). The task is now to complement the metadata available from OLAC with information about grammatical categories and lexical and topical information which can be extracted from the transcriptions found in the archives. In order to do that, the relevant files have to be retrieved from the archives. An understanding of the structure of these archives is a prerequisite for that.

4. Structure of endangered language archives: *PARADISEC*

Endangered language archives share very similar underlying structures. An *archive* consists of several *collections*. Each collection is about one project, most often covering one particular language, but occasionally, more than one language can be part of a documentation project. A collection in turn consists of *session bundles*, which contain a coherent set of *files* (audio, video, transcription, photos). Files found in a session typically share the same time, location and participants. There can be multiple files of the same type, e.g. very long sessions might have several audio files, with associated transcriptions. The levels of collection, bundle, and file may or may not have their dedicated landing pages, where metadata is displayed. Metadata relevant for a given text is thus often distributed across the various levels. The separation between collections and bundles is not always very clear-cut and is sometimes only available via implicit file naming conventions. The content typically offered consists of audio files, video files, and transcription files. Less common file types include photographs, pdfs, FLEx,[9] Toolbox,[10] praat,[11] and MS Office files. There are typically several levels of access control, which we can enumerate from 1-4:

1. freely available
2. registration and acceptance of terms and conditions required
3. available upon request from depositor
4. unavailable (privacy or other legal issues) (Figure 1, https://catalog.paradisec.org.au/collections/AA1)

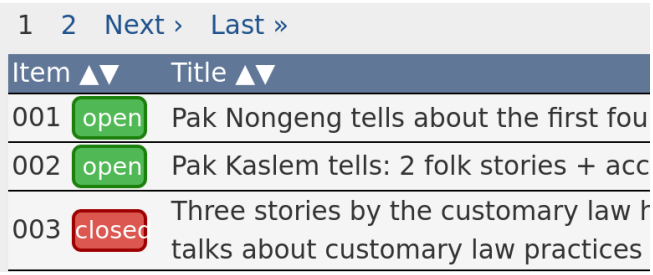


Figure 1: Access levels at PARADISEC.

Turning to findability and reusability, the following picture emerges: The querying possibilities for selected metadata are good. PARADISEC for instance offers nice faceting for country, language, and depositor (Figure 2). Other archives are similar. However, there are no ways to search for a particular language other than scrolling, and the value of metadata fields such as "depositor" or "source university" is not obvious. Other potentially relevant fields are absent from the querying interface, such as "access level" or "media type". I have not been able to formulate a query for "give me a collection which has at least one ELAN file and to which I have access". The only way to perform this query is to visit each and every collection, see whether there are ELAN files and try to download them.

Most archives provide an OAI-PMH[12] interface or have done so in the past.[13] This allows for a uniform query via OLAC.[14] Interestingly, while a query via media type is not possible on the PARADISEC site itself, it is possible on OLAC. The query https://bit.ly/39HueQE returns all sound files for the Namakura language which are available online. Unfortunately, the first bundle listed (*Two Namakura stories*) does indeed contain sound files, but they are not accessible.

Access to linguistic data is a sensitive topic. Next to the domains of privacy and copyright, there are also issues pertaining to language ownership and colonialism, which

[7] The MulitCAST project (https://multicast.aspra.uni-bamberg.de) is similar in setup to DoReCo.

[8] http://search.language-archives.org

[9] https://software.sil.org/fieldworks

[10] https://software.sil.org/toolbox/

[11] http://www.fon.hum.uva.nl/praat/

[12] https://www.openarchives.org/pmh

[13] ANLA and AILLA stopped in 2013 and 2017, respectively, see http://www.language-archives.org/archives

[14] http://search.language-archives.org/index.html

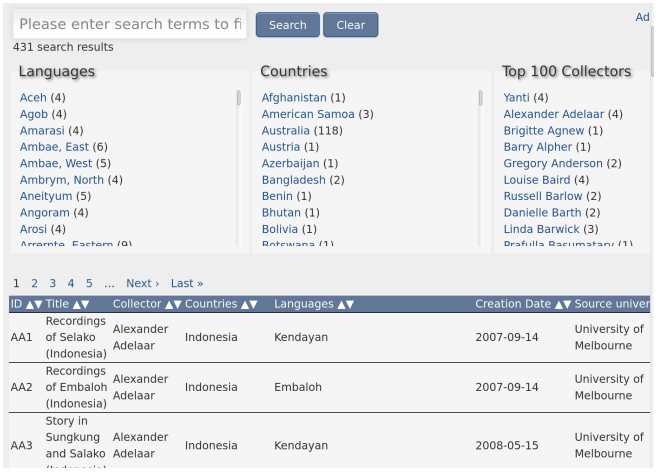


Figure 2: The PARADISEC querying interface.

have different levels of importance in different areas of the world (Holton, 2009). Therefore, archives often have custom terms and conditions, which diverge from better known licensing practices such as Creative Commons. The terms and conditions[15] for the PARADISEC archive for instance include:

> Not to copy the data in whole or in part except insofar as this may be necessary for security purposes or for my own personal use. Not to distribute the data to third parties, nor to publish or reproduce it in any way.
> ...
> To give access to the data only to persons directly associated with me or working under my control

The language here is very clear: do not copy, do not distribute.

5. The QUEST project

The QUEST (QUality-ESTablished)[16] project has as its stated goal to facilitate the interaction with and mobilization of (endangered) language data via the specification of standards and interfaces. One aspect is the standardization of future input during ingestion, which will facilitate subsequent retrieval. The other aspect is the development of querying tools with uniform interfaces working on extant data. This paper focuses on the latter of these two aspects. To this end, metadata were harvested from OLAC and the five archive websites. All referenced ELAN files were identified and downloaded as far as access restrictions permitted. The resulting set of 20k ELAN files was analysed for internal file structure and a converter into a common backend format was written. A couple of analyses were run on that backend format to prove the viability of the approach. Scripts for harvesting and analysis will be made available together with this paper, but access terms require each researcher collect the data individually from the archives (See §4.).

6. Description of the resources

In the context of this project, data satisfying the following criteria were considered:

1. The data must be programmatically **accessible** via command line tools. Many files in the archives are available "upon request", which means that a formal email has been written to the depositor. This setup does not scale and cannot be handled with the resources currently available. Authentication can be accomplished via the command line so that resources on the "registered user" level could be included.
2. The data must be **interoperable**. For all practical purposes, this means that data has to be in ELAN format.[17] Other file types are found in the archives, but they are either not suitable for data extraction (pdf), or their numbers are too low to justify the time to write an import script.

Current technology does not allow us to search directly in audio (e.g. by humming a melody), let alone in video. This means that querying audio or video boils down to querying transcriptions. The ELAN format is again very suitable, as the text content contained in ELAN is time-linked to multimedia files.

Of the 12 existing DELAMAN archives, 5 were chosen, as they show a variety of setups while at the same time providing a large enough sample of ELAN files to allow for an evaluation of the generic structure of the scripts developed. Table 1 gives a breakdown of the files which could be retrieved from the archives.[18]

ELAN as a file format links audio and video files to transcriptions. Transcription is organised in so-called tiers. Tiers are of a certain type ("translation", "gloss", "POS", etc.) and are hierarchically organised. The hierarchical relation between tiers is typically one of 1) time subdivision (a text is split into time-aligned sentences); 2) symbolic subdivision (a sentence is split into n words, but the words are not time-aligned themselves); and 3) association (a gloss is associated to a word). ELAN can accommodate multiple speakers. These then typically all have their own set of tiers. von Prince and Nordhoff (2020) contain more information about the ELAN file format as used in endangered language projects. Figure 3 shows the XML-representation of an ELAN file. The tier with the TIER_ID "ref@dam" is of the type "ref" and establishes time subdivisions. The tier with the TIER_ID "ut@DAM" references "ref@dam" and is of type "ut" (like 'utterance'). Annotations in tiers of the type "ut" are symbolically associated to the annotations in the parent tier. The tiers of type "ut" have further child tiers, which contain tokenized words ("tx"), morpheme segmentations ("mb") and glosses ("ge"). The tier "ft" contains a free translation for each utterance. Unfortunately for our purposes, the tier types and tier hierarchies are not defined in a specified standard, but are de-

[15] https://catalog.paradisec.org.au/collections/AA1/items/002/essences/967951/show_terms. Apparently, one has to sign in to access the terms and conditions.

[16] https://www.leibniz-zas.de/de/forschung/forschungsbereiche/syntax-lexikon/quest

[17] https://tla.mpi.nl/tools/tla-tools/elan/elan-description

[18] Scripts are available at https://github.com/ZAS-QUEST/eldpy

Table 1: The accessible holdings of the five DELAMAN archives surveyed.

	total		transcriptions			translations	
	files	file size	files	hours	words	files	words
AILLA	2 867	801M	2402	1054:59:29	1 120 059	85	14 284
ANLA	76	14M	48	12:49:40	6 906	45	6 463
ELAR	12 955	3.1G	7189	1470:28:23	1 074 463	706	298 457
PARADISEC	888	167M	706	132:56:03	94 962	153	15 335
TLA	3 473	1 002M	1062	217:20:54	155 476	1 497	72 014

```
<?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>abə</ANNOTATION_VALUE>
                </REF_ANNOTATION>
            </ANNOTATION>
            <ANNOTATION>
                <REF_ANNOTATION ANNOTATION_ID="ann9" ANNOTATION_REF="ann8">
                    <ANNOTATION_VALUE>kunəi pudza tukle hon lə mələk</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>
```

Figure 3: The XML structure of ELAN files with tiers referencing each other. Orange lines show references to timeslots, the green line shows the reference to a parent tier, purple shows reference to tier type definitions.

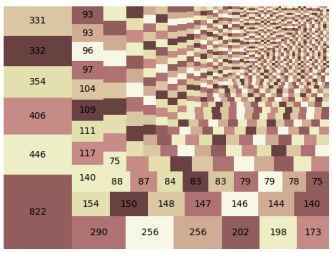
fined on a per-file basis at the very bottom of the XML-file. While ELAN makes sure that annotations are *syntactically* interoperable, *semantic* interoperability is not enforced by ELAN. The tier type containing the translation could be called any of "Translation", "English", "ft" (for free translation), "translation (eng.)" etc. The same goes for transcriptions and glosses. I have compiled a set of all names for tier types (several hundred) and have sorted them into the categories of translation, transcription, gloss, and unknown. This gives some hints about the content in a given tier, but this is not sufficient. The types as indications have to be complemented by information from the tier hierarchy.

The tier hierarchies used in ELAN files are also very heterogeneous. Some files have 3 tiers, some have 4, some have more than 20, and the parent-child relations can be of time subdivision, symbolic subdivision or association. We can establish a fingerprint of the hierarchy via a graph representing the parent-child relations with labelled edges. Table 2 gives an overview of the different configurations found. Among 7 189 ELAN files with transcriptions in ELAR, we find no less than 1 564 different ELAR tier hierarchies. Note that these hierarchies are agnostic of the names given to the tier types; if we included the names, the number would be much higher still.

Finally, some additional tests can be used to ascertain the status of a tier. A tier with English translation should pass

Table 2: Number of different tier hierarchies per archive and the distribution of 7,189 files from ELAR on 1 564 hierarchies.

AILLA	171
ANLA	21
ELAR	1 564
PARADISEC	162
TLA	537



a language detection test for English. A tier with vernacular transcription should fail a language detection test for English. A gloss tier should have close to no white space in its elements; presence of -, = or ALLCAPS words are, on the other hand, good evidence for a tier being a gloss tier and so on.

Based on the names of the tier types, their configurations, and the heuristics/sanity checks just described, we can give the numbers in Table 1 for the accessible holdings of the archives under discussion. The holdings are very different: ELAR has more than two orders of magnitude more available ELAN files than ANLA. and about five times the number of AILLA. But when it comes to transcibed time, AILLA with 1054 hours is not so far behind ELAR with 1470 hours. Apparently, ELAN files hosted at AILLA are more often transcribed than ELAN files at ELAR, which lack retrievable transcriptions in about 40% of the cases. This is reflected in the number of transcribed words, where AILLA with 1.12 million has slightly more than ELAR with 1.07 million, despite having fewer files to begin with. Looking at translations, the picture reverses again: ELAR has now 20 times more translated words than AILLA. The likely interpretation is that for AILLA, transcription is very important, but translation is less of a focus. Compare this to ANLA, where nearly every transcribed file contains a translation. To be fair, many of the projects in AILLA are from Latin America, so that English translations might be absent, but translations in to Spanish or Portuguese might be used instead.

In §3. I mentioned the 67 documentation projects funded by DoBeS. Looking at the 217 hours of transcribed material, this seems very little. Obviously, each of these 67 projects has done more than 3 hours of transcription, and the real amount of transcribed files stored in TLA is much higher. But many of these files are access level 2 (on request) or 3 (unavailable) and are therefore not available for general inspection and analysis. This presents a legal barrier to access. Another reason might be that tier hierarchies or tier type names are very idiosyncratic so that the tiers containing the transcription could not be identified. This would be a technical barrier to interoperability. It is hoped that TLA will address these two barriers to reuse to make sure that the valuable holdings in the "dark repos" can be incorporated into larger research enterprises in the future.

7. Analysis of the resources

Up to now, this article has described the structure of the archives, the structure of ELAN files and the strategies for identifying, retrieving and analyzing ELAN files. In the remainder, I will show some small analyses which can be performed via the uniform access. The analyses presented here have as their main goal the proof-of-concept of a programmatic and uniform access to a large number of ELAN files from diverse locations. They are very simple (even simplistic) on purpose, as the goal here is not to further our understanding of linguistics, but to further our understanding of research infrastructure

7.1. Proof-of-concept: accessing the transcription tier

As a proof of concept, I have computed the most frequent graphemes found in each archive. The plot of the findings is given in Figure 4. The total number of graphemes is 46.5 million. We find, unsurprisingly, that <a> is the most frequent grapheme, and <n> is the most frequent nasal. The order of <e> and <i>, however, is different between archives, suggesting that the languages contained in the respective archives use different orthographies. This is particularly obvious for ANLA, which includes <ł> in the top list.

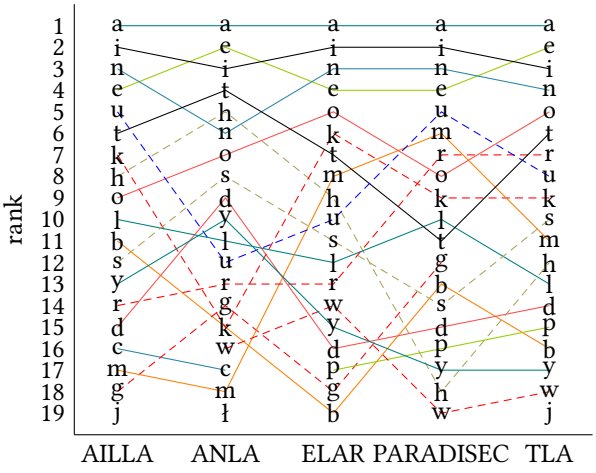


Figure 4: Most frequent graphemes in archives

7.2. Proof-of-concept: accessing the gloss tier

Within the gloss tiers, a total number of 3 274 394 morphemes could be retrieved. Figure 5 gives a breakdown of the most frequent grammatical glosses, while Figure 6 gives a breakdown of the most frequent lexical glosses. ANLA is excluded from both statistics because the amount of gloss material was not sufficient.

Some interesting observations can be made about the most frequent categories here: the number categories singular (SG) and plural (PL) are the most frequent grammatical glosses in ELAR, PARADISEC and TLA. This is not the case for AILLA, where apparently different conventions hold, and P is presumably used for plural instead. This highlights the need for a shared vocabulary, e.g. the Leipzig Glossing

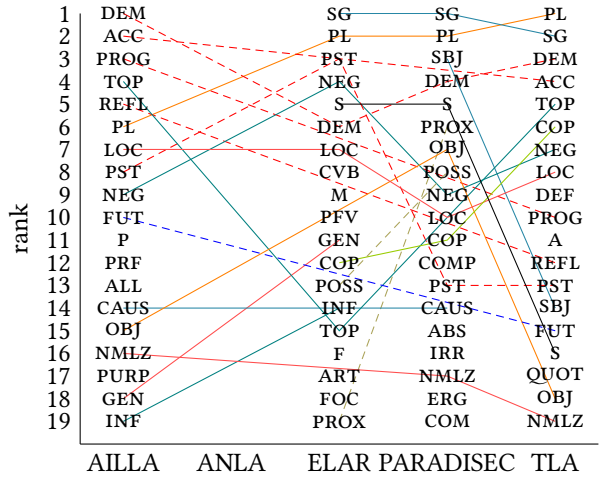


Figure 5: Most frequent grammatical glosses per archive

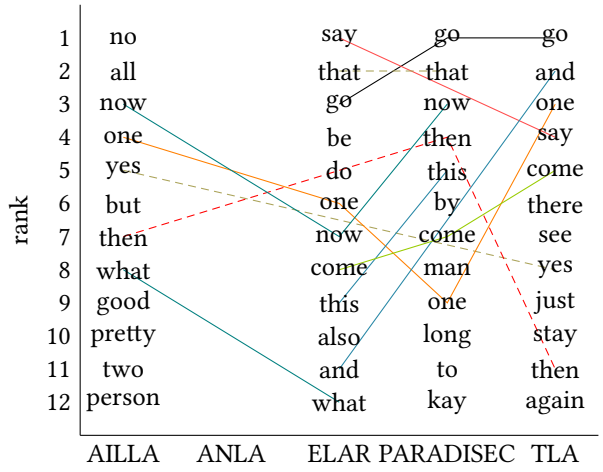


Figure 6: Most frequent lexical glosses per archive

Rules.[19] Another observation is that PST for 'past' is more frequent than FUT for future. This might be due to glossing conventions, the languages observed, or the types of text collected, but it is an interesting observation warranting further inspection.

Another obvious use of these resources would be the extraction of word-gloss-pairs for dictionary bootstrapping, which would be a particular type of text-mining. Table 3 gives words which have been glossed as 'sun' and 'moon' in languages of the corpus, respectively.

These translation data can further be included in a bridge towards Lemon.[20] The full interlinear representation of texts will also be made available in LIGT (Chiarcos and Ionov, 2019) in due course. Data refinements can be achieved with the pyigt library (List and Sims, 2019).

7.3. Proof-of-concept: accessing the translation tier

The proof-of-concept for the extraction of the translations from an ELAN tier involves Named Entity recognition via the NERD/GROBID online service.[21] Table 4 gives a breakdown of the entities retrieved. Tables 5 and 6 show the most frequent entities retrieved and the entities retrieved exactly 20 times. Taking a look at the concepts retrieved, we find a strong focus on agriculture, and on the Svan people from Georgia. The latter is a clear indication that the corpus is skewed and that there is an exceedingly large amount of well-transcribed files from a documentation project in the Caucasus, from where entities could easily be retrieved. But while this shows that one cannot simply run a quantitative analysis on the archives and be done, it also shows that the very high quality of the Caucasian data make the data much more findable and interoperable, giving them automatically a greater weight in scientific knowledge production. The "Caucasus bias" is obvious from the data, but at the same time, the "agriculture bias" is also something to take into account. Apparently, documentation projects more often focus on rural communities and crops/livestock than on urban settings and technology for instance. This must be borne in mind when drawing conclusions from the files stored in endangered language archives.

NERD/GROBID returns a Wikidata-ID (Vrandečić and Krötzsch, 2014), which allows to include the endangered language data in the wider Linked Open Data Cloud. This can be leveraged for semantic queries of the sort "give me all texts with a passive in them which deal with crops". For this query, we do not have to query for "maize", "rice", "wheat", "millet", etc. since Wikidata stores the in-

[19] https://www.eva.mpg.de/lingua/resources/glossing-rules.php. The GOLD ontology (Farrar and Langendoen, 2003) is often cited in this context, but has failed to develop any big impact due to a number of conceptual problems.

Table 3: Strings glossed as 'sun' and 'moon' in the corpus, which can be used for dictionary bootstrapping.

sun	moon
ane; eelo; hiisiis; hɛ; iisiis; indi;	biikousiis;
koyaš; künɣaraɣï; lainta; lezha;	bulan; cəlauni;
lénjí; lo'aa; mijiri; mə̃13;	din; goe-; hi; ilu;
mə̃13læ31; mə̃13lɛ33; mə̃31;	kàru; luna; lɔ13;
mə̃31la31; mə̃31læ31; mə̃33;	maham; moon;
mə̃33læ55; mə̃35; mə̃55; niel;	miŋgramin;
p'ûûs; siβu; siβun; siβuŋ; sool; sun;	owniv; oːlɛː;
sunə; tèle; tse/imá; t^ha55ia53; uni;	sahr; t'aar; turu;
vala'; was; yaal; yal; yaro; ^{3}tini;	tún; wula; ōl;
âftâw; čelɨ; ŋar; şımş; ŋɔ13; ʔawá	ótɔ'

[20] https://lemon-model.net/

[21] http://cloud.science-miner.com/nerd

Table 4: Entities retrieved from DELAMAN archives.

	total entities	different entities
AILLA	1 532	592
ANLA	301	142
ELAR	20 991	6 091
PARADISEC	1 163	568
TLA	10 346	3 281

Table 5: Most frequent retrieved entities across all archives.

#	Wikidata-ID	meaning
537	Q830	cattle
281	Q144	dog
271	Q11575	maize
270	Q7368	domestic sheep
250	Q5090	rice
239	Q383126	chronic condition
230	Q34067	Svan
212	Q5113	bird
209	Q2934	goat
204	Q19044	Svaneti
204	Q1364	fruit
187	Q626136	Arapaho people
184	Q8495	milk
184	Q532	village
177	Q190	God
166	Q7802	bread
163	Q13187	*Cocos nucifera* (coconut)
159	Q43238	*Poaceae* (grass)
158	Q503	banana
154	Q10798	pig
146	Q670887	*Bambusoideae* (bamboos)
145	Q11254	table salt
144	Q10998	potato
131	Q127980	fat
129	Q35808	firewood
120	Q846578	Svan people
117	Q1029907	stomach
115	Q10943	cheese
113	Q780	chicken
113	Q35409	family
101	Q41415	soup

formation all of these concepts are subclasses of https://www.wikidata.org/wiki/Q12117 "cereal", which in turn is subsubclass of https://www.wikidata.org/wiki/Q235352 "crop". Wikidata can furthermore be utilized for localization of queries: The data available about the concept https://www.wikidata.org/wiki/Q5090 "rice" contain translations into 171 languages, among which we find the translation into Swahili, *wali*. A constantly resurfacing requirement for archive mobilization is the accessibility to the speaker communities themselves. Being able to accept queries in a local language of wider communication, such as Swahili, is a crucial step for making the data *about* an ethnic group also being usable *by* that ethnic group.

8. Discussion

I have surveyed the existing language archives, and I have shown how a large corpus of ELAN files can be retrieved from these archives. These ELAN files are amenable to programmatic access, allowing to aggregate transcriptions, translations, and glosses, which can then be further analysed with regard to graphemes, grammatical categories or semantic fields. Two strands of research can be distinguished here. The first one is linguistics proper ("Which categories are used?"). The other one is closer to the sociology of science ("Which categories are used in which archives, and why? Which archives have more transcriptions, which ones have more translations, and why?"). Linguistics is often seen as a science bridging the gap between the natural sciences and the humanities. The first strand mentioned above is closer to the empirical approach, while the second strand is more a question typically asked within the humanities. The language resource assembled here can be used for both.

Table 6: Some medium frequency retrieved entities

#	Wikidata-ID	meaning
20	Q102192	freshwater
20	Q103459	livestock
20	Q107434	Sioux
20	Q11995	human pregnancy
20	Q125525	jackal
20	Q159334	secondary school
20	Q164088	*Metroxylon sagu* (sago palm)
20	Q184418	coffin
20	Q193110	floodplain
20	Q39861	*Hirundinidae* (swallows)
20	Q41692	mule
20	Q42302	clay
20	Q6450151	Kwande (district in Nigeria)
20	Q7632586	success

For purely quantitative research, the resource is obviously not suitable in its current state, as the "Caucasian bias" discussed in §7.3. shows. But a parametrization taking into account collections, languages, or even language families via genealogical data available from Glottolog is reasonably trivial.

But what can we do with the data? As mentioned above in §4., the ELAN files themselves cannot be shared due to the terms of access. In a linked data context (Chiarcos et al., 2012), however, this is not necessary. Once we have proper URIs which resolve to a given resource, we can use these as variables in our predicates. We can say that https://catalog.paradisec.org.au/collections/DLGP1/items/053 is a session which is about `glottolog:nama1268`, the URI for Namakura on Glottolog. We can say that a given session includes a file, which includes a tier, which includes a gloss which is the same as one of the Leipzig Glossing Rules glosses. The structures of the archives with collections, bundles, and files were discussed in §4.. In a linked data context, each collection, bundle, and file should have a different URI, but not all archives provide landing pages for all of those (Simons and Bird, 2020). Things get more difficult when using tiers or their parts (annotations) in Linked Data predicates, as the tiers and annotation will have to get URIs as well. A good solution for a resolver service will have to be developed, which will allow the use of these elements in assertions without requiring read or write access to the archives where the primary resources are hosted. This resolver will also help make the data findable by being citable, with exact location of the element in question in archive, collection,

file, and tier. Using such a resolver service will also allow the incorporation of sensitive data into the Linguistic Linked Open Data Cloud. We can say that the session with a given URI contains information about human sexual activity (https://www.wikidata.org/wiki/Q608), but we do not have to provide the session itself. This has obvious use cases in linguistics, but also in related fields of the humanities, such as anthropology or musicology. In the field of material culture, for instance, anthropologists look at items and appliances produced and used by given groups. Depending on the nature of the research question, broader or narrower concepts will be appropriate. In the domain of boat building, some researcher might be interested in all seafaring vessels, while for another one, only boats, only canoes, or only dugouts are relevant. A well-defined and ontologically grounded vocabulary for material culture can help the formulation of sensitive queries then (see e.g. eHRAF[22]).

What we can share, however, are download scripts for harvesting the archives. These scripts can be run by third party researchers and will provide the same files we have on our computers, but the third party researchers themselves have to agree to the terms and conditions before the download.

Interested researchers can request access from the relevant archive. Using Wikidata as a "semantic broker" also helps discoverability via the different language labels provided for the concepts, as described in §7.3..[23]

9. Outlook

Nordhoff et al. (2016) describe the Alaskan Athabascan Grammar Database (AAGD), which is also concerned with the findability of resources for endangered languages. For that project, texts from a number of native Alaskan languages were collected and made retrievable via a SOLR store. This SOLR store allowed faceted searches for metadata, but also for content categories such as semantic concepts and grammatical categories contained. While background and technology used are different, the requirements for the AAGD and this project are very similar. At the time of writing, the main focus is still the data model and the backend, but the repurposing of some of the frontend materials from the AAGD project should not be too difficult. The next step ahead will be the adaptation of the AAGD frontend to the QUEST datamodel. This adaptation will also allow for an easy integration of a "recommendation system". Such as system can use the texts a researcher has stated their interest in and propose new transcribed texts based on similarity in grammatical or semantic categories contained.

The main challenge ahead is the minting of URIs which adequately identify collections, sessions, files and tiers. This must be complemented by a useful ontology. Dublin Core isPartOf is used as an umbrella term for the time being, but more explicit relations would be useful.

Another technical challenge is the realization of federated queries. We need information from OLAC, Glottolog, Wikidata, and our own QUEST data. Ideally, OLAC as a central hub should provide the content searches described in this paper next to the metadata searches. If this is not to happen, a choice must be made whether one wants to go for some federated structure,[24] or whether a new service should be set up, which will periodically be updated with dumps from the other knowledge bases.

10. Acknowledgements

This research is funded by the German BMBF, as part of the QUEST project. I would like to thank Frank Seifart, Felix Kopecky, Hanna Hedeland, Mandana Seyfeddinipur, Kilu von Prince, and two anonymous reviewers for comments on previous versions of this paper.

11. Bibliographical References

Chiarcos, C. and Ionov, M. (2019). Ligt: An LLOD-Native Vocabulary for Representing Interlinear Glossed Text as RDF. In Maria Eskevich, et al., editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 3:1–3:15, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2012). Linking linguistic resources: Examples from the Open Linguistics Working Group. In Christian Chiarcos, et al., editors, *Linked Data in Linguistics. Representing Language Data and Metadata*, pages 201–216. Springer, Heidelberg.

Farrar, S. and Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT International*, 7(3):97–100.

Holton, G. (2009). Relatively ethical: A comparison of linguistic research paradigms in Alaska and Indonesia. *Language Documentation & Conservation*, pages 161–175.

List, J.-M. and Sims, N. A. (2019). Towards a sustainable handling of inter-linear-glossed text in language documentation. Preprint under review. Not peer-reviewed.

Nordhoff, S., Tuttle, S., and Lovick, O. (2016). The Alaskan Athabascan Grammar Database. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Simons, G. F. and Bird, S. (2020). Expressing language resource metadata as linked data: The case of the open language archives community. In Antonio Pareja-Lora, et al., editors, *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*. MIT Press, Cambridge.

von Prince, K. and Nordhoff, S. (2020). An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of LREC 2020*. Marseille.

[22] https://ehrafworldcultures.yale.edu

[23] Providing labels in different languages is a first step towards interculturally adequate discoverability. Wikidata itself probably has a significant Western bias in the selection and organisation of the concepts it contains. This bias cannot be resolved here.

[24] For instance in the context of CLARIN Federated Content Search architecture.

12. Language Resource References

Godfrey, J. J. and Holliman, E. (1993). *Switchboard-1 Release 2 LDC97S62*. Linguistic Data Consortium, Philadelphia.

Harald Hammarström, et al., editors. (2019). *Glottolog 4.0*. Max Planck Institute for the Science of Human History, Jena.

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., and Taylor, A. (1999). *Treebank-3 LDC99T42*. Linguistic Data Consortium, Philadelphia.

Simons, G. and Bird, S. (2003). The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *CoRR*, cs.CL/0306040.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.