Modelling and annotating interlinear glossed text from 280 different endangered languages as Linked Data with LIGT

Sebastian Nordhoff
Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)
nordhoff@leibniz-zas.de

Abstract

This paper reports on the harvesting, analysis, and annotation of 20k documents from 4 different endangered language archives in 280 different low-resource languages. The documents are heterogeneous as to their provenance (holding archive, language, geographical area, creator) and internal structure (annotation types, metalanguages), but they have the ELAN-XML format in common. Typical annotations include sentence-level translations, morpheme-segmentation, morpheme-level translations, and parts-of-speech. The ELAN format gives a lot of freedom to document creators, and hence the data set is very heterogeneous. We use regularities in the ELAN format to arrive at a common internal representation of sentences, words, and morphemes, with translations into one or more additional languages. Building upon the paradigm of Linguistic Linked Open Data (LLOD, Chiarcos et al. (2012b)), the document elements receive unique identifiers and are linked to other resources such as Glottolog for languages, Wikidata for semantic concepts, and the Leipzig Glossing Rules list for category abbreviations. We provide an RDF export in the LIGT format (Chiarcos and Ionov (2019)), enabling uniform and interoperable access with some semantic enrichments to a formerly disparate resource type difficult to access. Two use cases (semantic search and colexification) are presented to show the viability of the approach.

1 Introduction

1.1 Understudied languages

A couple of major languages, most predominantly English, have been the mainstay of research and development in computational linguistics. Recently Joshi et al. (2020) have analysed the representation of different languages in the research world. They established 6 classes for 2 485 languages of the world. Table 1 gives their classification

We see that, next to English, there are only 6 further languages for which the resources can be considered satisfying (Class 5). As we relax requirements for labeled and unlabeled data, we arrive at

						criteria	
	Class	example	# lgs	#spks	%	unlabeled data	labeled data
5	winners	Spanish	7	2.5B	0.28	good	good
4	underdogs	Russian	18	2.2B	1.07	good	insufficient
3	rising stars	Indonesian	28	1.8B	4.42	good	none
2	hopefuls	Zulu	19	5.7M	0.36	?	smallish sets
1	scraping-bys	Fijian	222	30M	5.49	smallish	none
0	left-behinds	Warlpiri	2 191	1.2B	88.38	none	none

Table 1: Joshi et al's classes.

The 14th Linguistic Annotation Workshop, pages 93–104
Barcelona, Spain, December 12, 2020.

classes 4, 3, and 2, with each one numbering languages in the low two-digit range. In class 1, only some unlabeled data are available, which Joshi et al find to be true for 222 languages. For 2 191, they find no noteworthy data whatsoever. Class 0 alone holds 7 times more languages than all of the other classes combined. Of course, there is more data available than Joshi et al include in their study. The Glottlog glottoscope[1] lists 7 794 languages, of which 2,088 have only a short word list or less, and 5 706 languages have better data available than "word list". This being said, virtually all those additional languages mentioned in Glottolog will be in Joshi et al's class 0: the resources listed are in the range of 10^2 to 10^4 tokens, whereas NLP applications typically require at least 10^6 datapoints to work properly, possibly much more.

But the human language faculty is not restricted to the 7 languages where we have a good data situation. It is due to historical accidents that these languages are predominant today, and they are actually not very protoyypical representatives of the class "natural language".[2] By focusing our attention, and our research, on these 7 larger languages, we are making a commercially sensible choice, but we might be missing important insights into the nature of the human mind.

But not all hope is lost, since for the last 25 years, data from many Class-0 languages have been collected, largely unnoticed by the NLP communities. These data reside tucked away in endangered language archives, waiting to be discovered.

1.1.1 Endangered language archives

Following upon a 1992 article by Hale et al. (1992) alerting to the danger of languages disappearing at an alarming rate, a number of endangered language documentation programs were set up (see Seifart et al. (2018) for an overview). Field linguists collected textual, audio, and video data from the speaker communities, annotated them and stored them in a number of archives. These archives set up the umbrella organisation Digital Endangered Languages and Musics Archives Network (DELAMAN) in 2003. There are currently 12 full members, of which the archives ELAR (UK), TLA (NL), AILLA (US), and PARADISEC (AU) host a very large number of documents, organised in over 1250 "collections" and are thus the most interesting to try computational approaches on large data sets of class 0 languages.[3] Schalley (2019) already points to the value of the data stored there: "If these data could be linked and integrated, current computational tools would allow for targeted searches across the body of knowledge", hinting at the same time at the major issue: integration and discoverability. This paper describes how data from the various archives can be programmatically accessed, integrated, and made searchable, thus establishing "endangered language archive scraping" as a novel collection method, an offshoot of conventional web scraping.

1.2 Novel types

While audio and video data also offer interesting use cases for automated approaches (Kisler et al. (2012), Paschen et al. (2020)), currently available technologies clearly focus on text. Endangered language archives typically contain interlinear glossed text (IGT, von Prince and Nordhoff (2020)), see Figure 1. Various formalisations for IGT have been proposed in the past (Nickles (2001), Drude (2002), Lewis (2006), Goodman et al. (2015), Chiarcos and Ionov (2019)). The different documentation projects contributing to the DELAMAN archives had different foci (phonetics vs. discourse; monologic data vs. dialogic data; etc.) but the great majority of projects build their textual data around the mappings of a vernacular utterance to a free translation, and a mapping of the component morphemes of that utterance to morphological glosses (Figure 1). For a given computational task, one has to model and

[1] https://glottolog.org/langdoc/status

[2] Cysouw (2011) concludes "The most fascinating [quantitative] result was that the northwestern European area, centred on Continental West Germanic, turned out to be one of the most linguistically unusual geographical areas word-wide. Many of the rare characteristics as attested in this area might have been considered the norm from a European perspective, though the typological data show that these characteristics are to be considered special structures of European languages, and not of human language in general."

[3] AILLA focuses on the Americas, PARADISEC has a focus on the Pacific. The other two archives have no particular regional focus. Together, they provide a very good coverage of work in the domain of language documentation of the last 25 years.

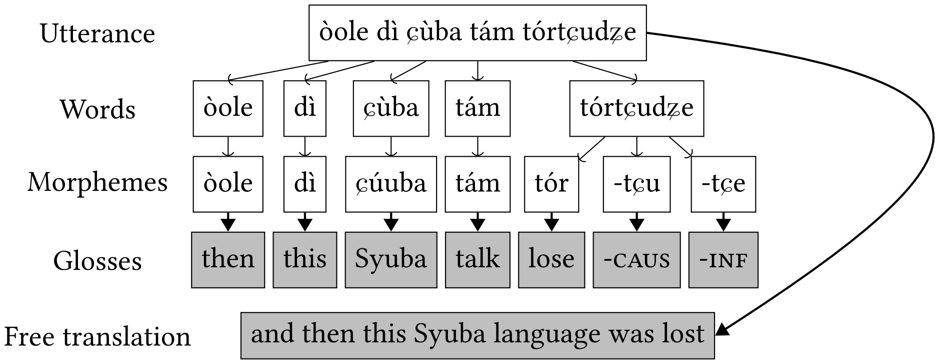


Figure 1: An example of interlinear text (Annotations-b-SUY1-140126-07, Gawne (2015)). Light arrows denote part-whole relations; thick arrows denote translational equivalents. Note that there is no translation for the word level, and that the citation form of the morphemes differs from their realization in a given word (*ɕùba*/*ɕúuba*; *dʑe*/*tɕe*).

access the part-whole relations between utterances and morphemes in the documents; and the correspondence relations between vernacular language and translation. This then allows for querying and analysis.

2 Endangered language archives

2.1 State of the art in discovery, querying, and analysis of endangered language archive holdings.

In a recent overview, Cimiano et al. (2020) state:

> Language resources (dictionaries, terminologies, corpora, etc.) developed in the fields of corpus linguistics, computational linguistics and natural language processing (NLP) are often encoded in heterogeneous formats and developed in isolation from one another. This makes their discovery, reuse and integration for both the development of NLP tools and daily linguistic research a difficult and cumbersome task.

This could not be more true for endangered language archives. The value of cross-queryable IGT from distributed holdings was pointed out as early as 2006 by Lewis (2006), who setup the ODIN platform[4], which lists 2017 documents containing IGT from 1274 languages. The only querying possibility, however, is by language name. There is no possibility to find all IGT documents which mention 'boat', '2PL', or 'birth'. There used to be a download facility, but this is broken as of today.

All endangered language archives have some metadata search functionalities (title, language, keyword), but no content search. They all require users to register and sign a code-of-conduct in order to download documents. That code of conduct typically restricts the distribution of downloaded resources. See Seyfeddinipur et al. (2019) for context.

The EOPAS project (Schroeter and Thieberger. (2006))[5] provided a very nice web interface for searching across various annotated video files and display the content in the browser. This project was down for a while, but the technology is being integrated into the PARADISEC catalog (Nick Thieberger, personal communication 2020-08-20). It is possible to search for a string like "house", which yields 4 different documents mentioning the word "house". The corresponding media files can be played in the browser, and the XML can be displayed, but a download facility for the XML files seems to be missing for now.

The ELAN desktop software allows a search across multiple XML files in a local directory (thus not in a remote archive). It does not allow the specification of particular tiers or relations, however.[6]

[4] http://odin.linguistlist.org

[5] https://github.com/CoEDL/modpdsc

[6] https://www.mpi.nl/corpus/html/elan/ch07s02.html

Cimiano et al. (2020) state:

> [W]ith respect to the provision of access to data, most of the available repositories lack at least one of the following features: Provision of domain specific linguistic/language data, which is open for re-use and free of charge[;] Search functionality that facilitates finding specific resources[;] Possibility to narrow search to open data and to resources in linked data formats as well as to directly download the data.

The three aspects of free re-use, search functionalities, and bulk downloads mentioned by Cimiano et al are indeed also what we find in the domain of endangered language archives. We cannot address the legal issues here, but we show ways how to improve search and retrieval.

2.2 Provisos

The underlying data for this project come from four different endangered language archives (AILLA, ELAR, PARADISEC, TLA) and are collected opportunistically.[7] Everything which is listed, available, accessible, and parseable has been taken into account. No effort has been made to ensure a balanced representation of genealogical or geographical factors, but such aspects can easily be factored in based on available metadata (OLAC[8], Glottolog).

Different field linguists have different conventions (von Prince and Nordhoff (2020)), which do not always map neatly on each other. Furthermore, the archives also often contain files in a relatively incomplete state, with empty tiers, empty slots or placeholders like '???' or '***'. The data are thus relatively dirty, but they happen to be the best of what we have for the group-0 languages. It will probably not be possible to run very advanced statistical or stochastic methods on the data, but the following section will give a glimpse of what kinds of analyses the data do permit to be run.

2.3 Former results

Nordhoff (2020) performed analyses of the structure and the content of annotations. As for structure, he found that there is a wide variety of tier configurations in the surveyed files (see below for the concept of "tier"). While all files share the ELAN-XML Format, the configuration options are nearly never identical between two files.

As for content, Nordhoff was able to run meaningful analyses on the frequency of graphemes, grammatical categories, and semantic concepts found in the texts. He found for instance that the most frequent graphemes employed are ⟨a⟩, ⟨i⟩, ⟨n⟩, and ⟨e⟩, in that order. This makes some intuitive sense, but is different from the phonemic (type) frequencies reported in PHOIBLE[9] (Moran and McCloy (2019)), where we find /m/, /i/, /k/, /j/, /u/, in that order.[10]

The most frequent grammatical categories are "singular" and "plural", next to common verbal tense categories. This is also in line what would be expected, but the interesting question emerges which one of the two number categories should be more frequent than the other. Finally, Nordhoff found that the texts have a semantic bias towards agriculture.

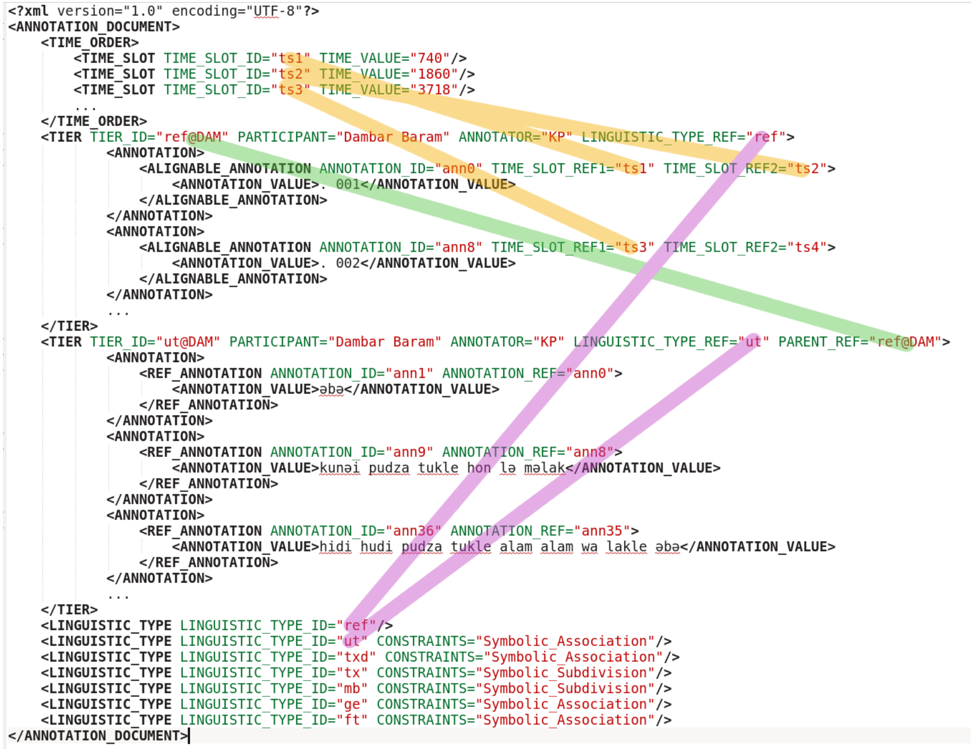
Psychological (*What do people use?*), typological (*What do people use where and why?*), sociological (*What are Western research projects interested in and why?*), and text-genre based explanations (*What kinds of text do field workers collect and why?*) can be explored for these grammatical and semantic findings, but this paper will focus on technical refinements and representations rather than theoretical explanations.

[7] The Alaska Native Language Archive (ANLA) was included in a pilot but was later dropped due to the very low number of usable files.

[8] language-archives.org

[9] https://phoible.org/parameters

[10] Note that PHOIBLE is based on the phonology sections of grammatical descriptions, while Nordhoff is agnostic to phonological values and compares graphemes. It is reasonable to assume that ⟨m⟩ will represent /m/, but this approach has obvious limitations for other graphemes, like ⟨j⟩ or ⟨c⟩.



```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>
```

Figure 2: The ELAN-XML format with time slots, tiers, constraints, and the links between these elements. Didactically unnecessary markup removed. Parent relations are given in green. Tiers can be of a certain linguistic type (linked in purple), and they can be linked to specified time slots (orange).

3 Data

In the context of this study, we restrict ourselves to documents in the ELAN-XML format (*.eaf). The reason for this is that this is a well-structured format used for many documents. The competing Shoebox/Toolbox[11] format is idiosyncratic and undocumented, while the FLEx format[12] is not very well represented in the archives under discussion.

3.1 The ELAN-XML format

ELAN is an annotation software developed at the MPI for Psycholinguistics in Nijmegen[13] (Wittenburg et al. (2006)) and uses XML as a storage format. The central element type are tiers, time slots, and linguistic types defining constraints. Figure 2 gives a simplified sample document highlighting the relations between different elements.

There is no registry for tier names or tier types. Users are free to define the semantics and labels of linguistic types as they see fit. A linguistic type named "ut" as in Figure 2 could theoretically contain anything. In practice, however, a number of common naming patterns emerge, so "ut" is for instance commonly used for ‘utterance’, and never for ‘free translation’. We have identified 99 common names for the transcription tier, 26 common names for the translation tier, and 8 common names for the gloss tier (cf. Figure 1, full lists are given in the appendix).

3.2 Data collection and preparation

We wrote a harvester to download all available ELAN files from the following four archives: ELAR, TLA, AILLA, PARADISEC. This yielded a total of roughly 20k ELAN files. We also wrote a parser which, for each file, identified the tiers containing transcription (=vernacular text), free translation, morpheme

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

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

[13] https://archive.mpi.nl/tla/elan

segmentation, and morpheme translation, using information from the tier names, tier content, and the relation between tiers. The morpheme translation tier for instance has to be a daughter of the morpheme segmentation tier, the translation tier should pass a language identification test as “English” and so on. All code is freely available at https://github.com/ZAS-QUEST/eldpy.

3.3 Annotations

Representations of sentences, words,[14] and morphemes on the one hand, and their respective translations on the other was stored and cached as JSON. The result is an ordered list of sentences linked to their translations, and of the component morphemes of these sentences linked to their respective translations/glosses as well.

The list of morpheme translations is thus a list of 2-tuples (vernacular morpheme, English translation), which allows us to create a look-up function like “give me all words which are translated as ‘moon’.” Each tuple is in turn associated with a given language, so that words translated as ‘moon’ can be compared across languages (see §6).

4 Data enrichment

For the vernacular languages, there are no NLP tools available, as explained above, but an explanation on how to carve a canoe with an axe should yield the relevant concepts ‘canoe’ and ‘axe’ via the English translation just as well. In order to tackle the task of discoverability, we ran the grobid-ner named entity recognition service[15] on the English translation. For this, we collated all translations of all the sentences in a given document.

4.1 Wikidata concept lookup

Grobid-ner service outputs Wikidata IDs of the type ‘Q12345’. A sample document in the Ju|ʼhoansi language for instance yielded the following concepts via the English translation route: Q1029907: "stomach", Q25312: "flies", Q25439: "woodpecker", Q506131: "carrying pole", Q606886: "cardinal woodpecker", Q6842999: "midriff". Concepts like "cardinal woodpecker" and "carrying pole" show that a pretty granular analysis of semantic content can be achieved in this way. 8 457 different concepts could be identified and linked to Wikidata. It is of course true that only few people would actively search "cardinal woodpecker"; a search term like "bird" would be much more likely. But the document is not marked up for "bird". So we add this information via Wikidata.

4.2 Transitive closure on Wikidata concepts

Wikidata hosts over 95 million concepts. Of these, we only need the 8 457 found in the documents (like "woodpecker"), and their (transitive) parent concepts, like "bird" and "animal". Parent concepts can be either the instanceOf relation ("Woody Woodpecker" is an instance of "woodpecker") or the subclassOf relation ("woodpecker" is a subclass of "bird") and should form a directed acyclic graph.[16] We recursively retrieved all parent concepts of the 8457 initial concepts and stored the transitive closure in a lookup table. This lookup table allows us to retrieve all documents with information about "bird" or any of its transitive child concepts. The parent concept for "carrying pole" is "utensil". The following list gives a list of the most frequently found child concepts thereunder, with document frequencies in parentheses: Q18341850: "pestle" (20), Q381155: "sieve" (18), Q193358: "ladle" (18), Q207763: "rolling pin" (14), Q32489: "knives" (9), Q154038: "cooking pot" (8), Q127666: "frying pan" (8), Q208364: "wok" (6).

Anthropologists interested in material culture now have a very simple entry point into the data; the same is true for ethnozoologist/ornithologists and their kin. In the RDF output (see §5), we use the Dublin Core “subject” property to link the Wikidata concepts to the documents for the time being.

[14] The word level is needed as an intermediate helper category, but does not contribute to the analyses conducted here. While sentences and morphemes are typically accompanied by translations/glosses, word-level elements typically lack translations of their own. For more on the cognitive reality of “word” as a concept, see Schiering et al. (2010).

[15] https://github.com/kermitt2/grobid-ner

[16] Some loops in Wikidata were found during the course of the research.

4.3 Leipzig Glossing Rules

We have analysed morpheme translations as to their components, and linked glosses like "2SG.ACC" to the list provided by the Leipzig Glossing Rules (Comrie et al. (2008)), in this particular case, "2", "SG", and "ACC". While there are a variety of concept registries for grammatical categories, such as GOLD (Farrar and Langendoen (2003),Farrar and Langendoen (2010), ISOCAT (Kemps-Snijders et al. (2008)), and Clarin Concept Registry,[17] the original documents were not created with these registries in mind.[18] The Leipzig Glossing Rule provide a fairly short an simple list of common abbreviations, and linguist tend to adhere to them. Note that the LGR are a list of *abbreviations*, not a list of categories. For the present purposes, linking to LGR allows us to state that IMP is to be expanded as 'imperative', and not as 'imperfective'. The latter would be IPFV under the Leipzig Glossing Rules. This statement is a lot less powerful than linking IMP to something like gold:imperative. But cross-linguistic categories are fraught with conceptual difficulties, and there is an ongoing debate as to whether they can be defined in a meaningful way at all (see Haspelmath (2007) for discussion). For this reason, we prefer to err on the side of caution and only integrate with the list of abbreviations.

5 Models for IGT

Next to external links to the existing knowledge bases like Wikidata, and, in a less evolved fashion, LGR, we also provide a semantic model of the internal structure our IGT data based on the LIGT model (Chiarcos and Ionov (2019)). The LIGT model expands on existing ontologies and vocabularies, such as Dublin Core and NIF (Hellmann et al. (2013)). Figure 3 provides a visualization of this model. The model is based on a ligt:Document, a subclass of dc:Dataset. This document has a text, whose subcomponents are modelled recursively as substrings down to the layer of utterance. Within an utterance, LIGT distinguishes the elements of Word and Morph, which are ligt:Item's arranged in the respective tiers. The vernacular words and glosses are realized as labels, e.g. '"tám"@syw' and '"talk"@eng'. This maps neatly on the structure given in Figure 1.

For LGR glosses, we use a BCP47[19] private use subtag 'lgr' to specify the restricted variety of English we are using here. This yields '"ACC"@en-x-lgr' for the gloss ACC for 'accusative' for instance.

6 Integration

Our export as LIGT-RDF has, among other things, the advantage of an easy integration into the Linguistic Linked Open Data Cloud (LLODC, Chiarcos et al. (2012a; Cimiano et al. (2020)). Next to metadata about the languages from Glottolog, IGT data from endangered language archives can be integrated with various other resources with the Linked Data approach used here.

Chiarcos et al. (2017) used DBnary (Sérasset (2014)), based on Wiktionary, to provide glosses in 15 additional languages, beyond the glosses supplied in the original document.

Chiarcos and Ionov (2019) expanded on this and model IGT based on DublinCore, NIF (Hellmann et al. (2013)) and WebAnnotation (Sanderson et al. (2017)) vocabularies. This allows them to integrate IGT from Toolbox, FLEx, and the Xigt format into LIGT. To this set, we can now add the ELAN data from the endangered language archives.

Nordhoff (2020) enriched the documents with semantic concept annotations linked to Wikidata. He used this for an analysis of the most frequent semantic domains found in documents in endangered language archives (to wit, the Caucasus and agriculture). We have now added the transitive closure of the parent relations of these concepts.

Another integration is the generation of candidates for colexification (François (2008)). Concepts do not map alike on lexemes in the languages of the world. Some languages use different words for the concepts HAND and ARM for instance, while others use the same word. Those languages which use the same word are said to "colexify" the two concepts. The study of colexification is complicated by homonymous words which do not share any semantics, like *arm*$_{upper\ limb}$ and *arm*$_{weapon}$.

[17] https://www.clarin.eu/ccr

[18] For the checkered history of GOLD, ISOCAT, and CCR, see Cimiano et al. (2020).

[19] https://tools.ietf.org/html/bcp47

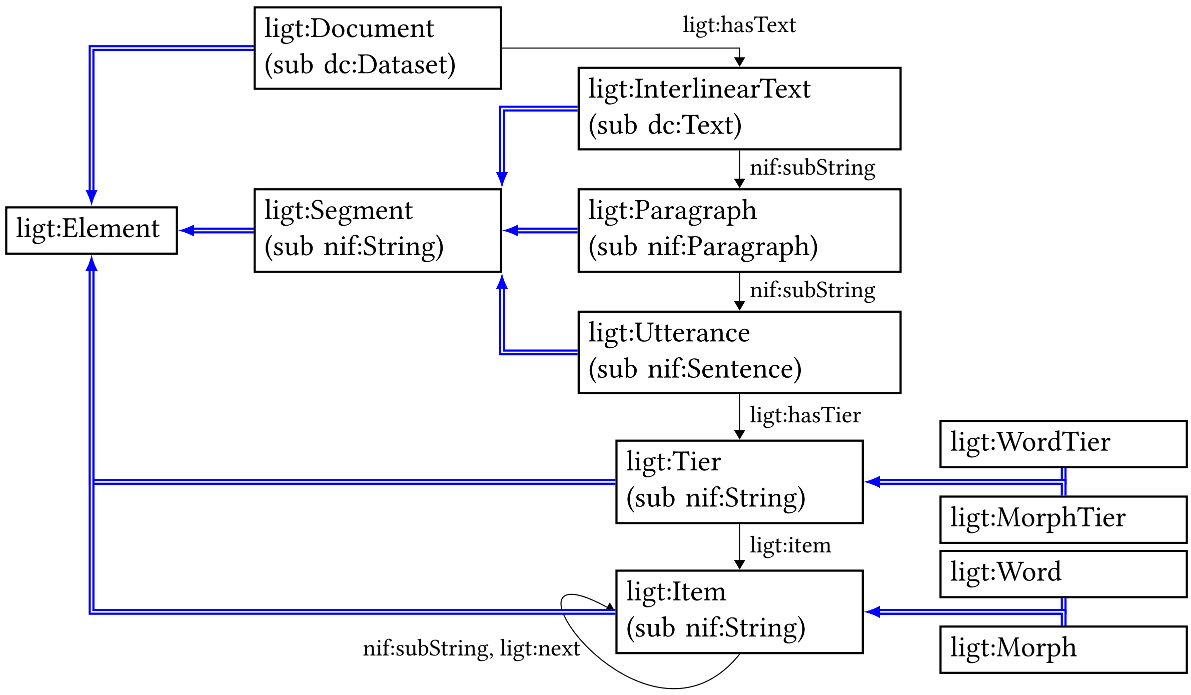


Figure 3: LIGT data model from Chiarcos and Ionov (2019) (adapted). Double arrows indicate subclass relations. Single arrows are labelled for the relations they express.

Figure 4: Retrieved colexifications (subset).

again:come	anvil:fireplace	ball:fist	bear.fruit:child	become.wet:cold	beg:pray	broom:saw	build:cook
cultivate:field	deceiver:liar	dive:follow	earth:soil	elder:old.man	enclose:enter	father:seedbed	field:yard
first:start	follow:plunge	friendship:please	give:pleasant	govern:power	leave_for:meet	light:outside	
literate:student	manure:rot	owner:self	plant:send	pleasant:weep	prepare:stop	reap:snake	run:speed
seize:take	small:younger	speak:speech					

Studies of colexification have been undertaken with (modified) Swadesh lists and questionnaires as an input (List et al. (2018)), as these are available for many languages. The corpus of endangered language data allows us to go beyond these short lists and generate hypotheses about colexification. For this, we look for all words in a given collection which have two different English translations and list them as potential colexification candidates for the two concepts. If a given 2-tuple is found in more than one language, further investigation of the semantic of this is warranted to check whether we are dealing indeed with colexification. Figure 4 gives a list of pairs thus identified.

7 Sharing of resources

The legal status of the underlying documents is shaky. The legal domains concerned are privacy law on the one hand and intellectual property law on the other. Depending on where the data were collected, different legal frameworks might apply. Some data might not enjoy any protection, like someone counting from 1 to 10, while other documents have clearly copyrightable content, e.g. a narrative. The same holds true for translations of said content.

Things are different for linguistic analysis. The fact that in a given document, we find a word which has been glossed as "give-1PL.PRES" can probably be shared without too much of a legal risk. The same is true for lists of such annotations, i.e. morpheme-by-morpheme glosses of whole sentences. These

	collections	eaf files	transcribed	duration	words transcribed	export triples
AILLA	10	1 674	1 447	532.7h	648 908	2.6M
ELAR	201	13 758	10 139	2588.3h	2 498 122	24.2M
PARADISEC	78	2 619	1 776	302.0h	225 974	10.7M
TLA	(68)[20]	1 695	1 441	506.5h	677 959	1.6M
total	289+	19 746	14 803	3929.5h	4 050 963	39.1M

Table 2: Holdings of the four investigated DELAMAN archives. Only collections with at least one accessible ELAN file are counted.

can be made available for further inspection.

Even less problematic are the semantic enrichments. The fact that a given document is a about a certain kind of woodpecker is unproblematic and can be shared. The Linked Data approach used here allows us to elegantly circumvent the problem: We use the landing page of a document in a given archive as our URI variable for linked data purposes. We can now predicate over this variable, and say things like that it contains 34 sentences, that it contains information about woodpeckers, and that the 2nd morpheme of the 4th word in the 3rd sentence is glossed as ACC. Interested users can look up the page on the archive and request access to the original document if they so wish. One remaining issue is that the semantic structuring of the archive web sites is slightly different. We have the logical levels of "collection", consisting of "bundles/sessions", in turn containing a number of "files", but not all of these have landing pages for each archive or are cross-linked in an obvious manner with a transparent URL. We are currently working on a resolver service which will allow the URIs used in RDF predicates to be resolved to the URLs in the endangered language archives.

8 Evaluation

Table 2 gives a breakdown of the holdings of the different archives as far as known.

Collections typically contain one language, but some projects working in multilingual settings have more than one language. On the other hand, some languages are found in more than one archive. Still, we can say that the number of collections is a good approximation of the number of different languages we know have structured and annotated data for.

For 4 100 files, at least one concept could be retrieved. In total, 34 336 concepts were retrieved from the texts, of which 8 457 are distinct.

As for LGR abbreviations, there are 438 874 instances of 80 distinct types (There are 84 abbreviations listed in the LGR).

9 Outlook

Rzymski et al. (2019) describe progress in the Database of Cross-Linguistic Colexifications (CLICS^3). They set up a data format following Cross-Linguistic Data Formats (CLDF, Forkel et al. (2018)) and detail how other projects can add to their colexifcation data in a well-defined workflow. Colexification candidates generated from endangered language documentation archives would be a very good candidate for such a workflow.

10 Conclusion

Around 90% of all human languages fall into Joshi et al. (2020)'s 'group-0' languages with no noteworthy data capable of providing a starting point for NLP application development. In this paper, we have shown that structured data are indeed available from endangered language archives, a resource by and large ignored by the NLP community up to now. To the extent that legal and technical barriers can be overcome, analyzable and annotated data from almost 300 different languages could be downloaded,

[20] The current way of retrieving TLA metadata does not allow a grouping by collection, but grouping by sessions is possible.

parsed, processed, enriched, interlinked, and exported as RDF, making use of the LIGT model, so that we now have structurally interoperable data. The enriched data allow for semantic querying and provide a good starting point to connect pipelines such as the CLICS³ framework. The main issues to address in future work will be the precise legal conditions for sharing and reuse of the data as well as a good resolver service allowing a stable and precise dereferencing of the Linked Data URIs used.

Appendix

Tier names indicating translation tiers "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".

Tier names indicating transcription tiers "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".

Tier names indicating gloss tiers "ge", "morph-item", "gl", "Gloss", "gloss", "glosses", "word-gls", "gl (interlinear gloss)".

References

Christian Chiarcos and 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, and Milan Dojchinovski, editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, number 70 in OpenAccess Series in Informatics (OASIcs), pages 3:1–3:15, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2012a. Linking linguistic resources: Examples from the Open Linguistics Working Group. In Chiarcos et al. (Chiarcos et al., 2012b), pages 201–216.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. 2012b. *Linked Data in Linguistics. Representing Language Data and Metadata.* Springer, Heidelberg.

Christian Chiarcos, Maxim Ionov, Monika Rind-Pawlowski, Christian Fäth, Jesse Wichers Schreur, and Irina Nevskaya. 2017. LLODifying linguistic glosses.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications.* Springer, Cham.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.* Max Planck Institute for Evolutionary Anthropology, Leipzig.

Michael Cysouw. 2011. Quantitative explorations of the worldwide distribution of rare characteristics, or: the exceptionality of northwestern European languages. In Horst J. Simon and Heike Wiese, editors, *Expecting the Unexpected: Exceptions in Grammar.* De Gruyter Mouton, Berlin, Boston.

Sebastian Drude. 2002. Advanced glossing: A language documentation format and its implementation with Shoebox. In Peter Austin, Helen Dry, and Peter Wittenburg, editors, *Proceedings of the International LREC workshop on Resources and Tools in Field Linguistics*.

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

Scott Farrar and D. Terence Langendoen. 2010. An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In A. Witt and D. Metzing, editors, *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*. Springer, Dordrecht.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205.

Alex François. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Martine Vanhove, editor, *From Polysemy to Semantic Change*, number 106 in Studies in Language Companion Series, pages 163–216. John Benjamins, Amsterdam.

Lauren Gawne. 2015+. *Kagate (Syuba), an endangered Tibeto-Burman language of Nepal.* ELAR, London.

Michael Wayne Goodman, Joshua Crowgey, Fei Xia, and Emily M. Bender. 2015. Xigt: extensible interlinear glossed text for natural language processing. *LREC*, 49(2):455–485.

Kenneth Hale, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, and LaVerne M. Jeanne. 1992. Endangered languages. *Language*, 68(1):1–42.

Martin Haspelmath. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, 11.1:119–132.

S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. 2013. Integrating NLP using linked data. In *Proceedings of the 12th International Semantic Web Conference, 21–25 October 2013, Sydney.*

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and 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*, page 6282–6293.

M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S. E. Wright. 2008. ISOcat: Corralling data categories in the wild. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, page 887–891.

T. Kisler, F. Schiel, and H. Sloetjes. 2012. Signal processing via web services: The use case WebMAUS. In *Proceedings Digital Humanities*, page 30–34. Hamburg.

William D. Lewis. 2006. *ODIN: A Model for Adapting and Enriching Legacy Infrastructure.* 2nd IEEE International Conference on E-Science and Grid Computing, Amsterdam.

Johann-Mattis List, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. 2018. CLICS2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0.* Max Planck Institute for the Science of Human History, Jena.

M. Nickles. 2001. *Systematics - Ein XML-basiertes Internet-Datenbanksystem für klassifikationsgestützte Sprachbeschreibungen.* Centrum für Informations- und Sprachverarbeitung, München.

Sebastian Nordhoff. 2020. 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*, pages 10–18, Marseille, France, May. European Language Resources Association.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, page 2657–2666.

Christoph Rzymski, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List. 2019. *The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies*, volume 7.

Robert Sanderson, Paolo Ciccarese, and Benjamin Young. 2017. Web Annotation data model. Technical report, W3C Recommendation.

Andrea C. Schalley. 2019. Ontologies and ontological methods in linguistics. *Lang Linguist Compass*, 13(e12356).

R. Schiering, B. Bickel, and K. Hildebrandt. 2010. The prosodic word is not universal, but emergent. *Journal of Linguistics*, 46(3):657–709.

Ronald Schroeter and Nick Thieberger. 2006. Eopas, the ethnoer online representation of interlinear text. in sustainable data from digital fieldwork. In Linda Barwick and Nick Thieberger, editors, *Sustainable data from digital fieldwork*, pages 99–124. University of Sydney, Sydney.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation 25 years on. *Language*, 94(4):e324–e345.

G. Sérasset. 2014. DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web Journal*, 648.

Mandana Seyfeddinipur, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L. Epps, Vera Ferreira, Ana Vilacy Galucio, Brigit Hellwig, Oliver Hinte, Gary Holton, Dagmar Jung, Irmgarda Kasinskaite Buddeberg, Manfred Krifka, Susan Kung, Miyuki Monroig, Ayu'nwi Ngwabe Neba, Sebastian Nordhoff, Brigitte Pakendorf, Kilu von Prince, Felix Rau, Keren Rice, Michael Rießler, Vera Szoelloesi Brenig, Nick Thieberger, Paul Trilsbeek, Hein van der Voort, and Tony Woodbury. 2019. Public access to research data in language documentation. *Language Documentation & Conservation*, 13:545–563, 10.

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

P. Wittenburg, Brugman H., A. Russel, A. Klassmann, and H. Sloetjes. 2006. *ELAN: A Professional Framework for Multimodality Research.*