Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2778–2787
Marseille, 11–16 May 2020

An empirical evaluation of annotation practices in corpora from language documentation

Kilu von Prince, Sebastian Nordhoff
Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS Berlin)
Schützenstr. 18, 10117 Berlin
{vonprince, nordhoff}@leibniz-zas.de

Abstract

For most of the world's languages, no primary data are available, even as many languages are disappearing. Throughout the last two decades, however, language documentation projects have produced substantial amounts of primary data from a wide variety of endangered languages. These resources are still in the early days of their exploration. One of the factors that makes them hard to use is a relative lack of standardized annotation conventions. In this paper, we will describe common practices in existing corpora in order to facilitate their future processing. After a brief introduction of the main formats used for annotation files, we will focus on commonly used tiers in the widespread ELAN and Toolbox formats. Minimally, corpora from language documentation contain a transcription tier and an aligned translation tier, which means they constitute parallel corpora. Additional common annotations include named references, morpheme separation, morpheme-by-morpheme glosses, part-of-speech tags and notes.

Keywords: endangered languages, annotations, parallel corpora

1. Background

1.1. Introduction

Our knowledge about the world's languages is heavily biased towards large, official, literate languages (Dahl, 2015) from western, educated, industrialized, rich, democratic societies (Henrich et al., 2010). For most of the languages spoken today, no substantial records of primary data are available (Simons and Fennig, 2017). Throughout the last two decades, language documentation has emerged as a new subfield of linguistics that is concerned with providing repositories of primary data from less resourced and endangered languages. The corpora produced within this field represent some of the richest records of spoken language available today. As audio and video recordings with time-aligned transcriptions, translations into one or more widespread languages, detailed metadata, and often morpheme-by-morpheme glosses and part-of-speech tags, these resources hold an immense potential both for linguistic research and for the development of advanced natural language processing (Good, 2011; Cox, 2011; Seifart, 2012).

However, the exploration of these resources is still far from trivial. Even when deposited in one of the major archives, they can be hard to locate and hard to access due to technical and legal barriers. Once located, users will have to make sense of the annotations, which do not adhere to a fixed set of conventions. There are, of course, good reasons for this lack of conventions. The corpus creators have been faced for the first time with an extreme variety of sociolinguistic settings, with structurally very different languages, and with speaker communities that each have their own needs and restrictions. At the same time, most language documenters have been working with a limited set of tools and workflows, and had similar goals in mind, so that certain annotation practices are sufficiently widespread to serve as *de facto* conventions and standards. The recommendations and guidelines developed within major funding programs such as the DoBeS initiative[1] and the ELDP[2] have further fostered the emergence of certain standards.

We have carried out a detailed investigation of corpora from 32 different languages. Von Prince created two of those corpora in the context of the DoBeS project on West Ambrym languages[3] and worked extensively with seven corpora, as principal investigator of the MelaTAMP project.[4] We obtained the others from the DoReCo project.[5] This was complemented by an automated analysis of 20k ELAN files available from five different DELAMAN[6] archives conducted by Nordhoff.

In this paper, we outline the main conventions we could identify in these corpora regarding, in particular, the logical dependencies between different aspects of annotation, embodied in so-called annotation tiers. The purpose of this study is to help computational linguists to explore corpora from language documentation; and to help corpus creators to make their corpora maximally accessible by providing relevant documentation, and by making informed choices about their workflows and annotation levels with respect to existing conventions. A shared understanding of standards in the domain of linguistic annotation for less-resourced languages between data-producing field linguists and data consuming computational linguists will help formulate and address research questions in a more quantitative way, complementing the qualitative approaches which still predominate as of today.

1.2. Basis for this study

The 32 corpora investigated in detail are a highly curated set of resources which were selected in the context of different projects for their size and the quality and detail of their

[1] http://dobes.mpi.nl

[2] https://www.eldp.net

[3] https://www.leibniz-zas.de/en/research/research-areas/project-archive/2010-2020/ambrym

[4] https://www.projekte.hu-berlin.de/en/melatamp

[5] http://doreco.info

[6] http://www.delaman.org

annotations. All of them are based on audio and/or video recordings. The seven corpora in the MelaTAMP project comprise between 30k and 150k tokens of fully transcribed, translated, glossed and pos-tagged text. The additional corpora from the DoReCo project comprise a minimum of 10k tokens with full translations and optionally additional annotations. Table 2 gives an overview of the languages documented by these collections. Most corpora from language documentation are archived in one of the major repositories dedicated to this purpose. Table 1 indicates the main archives in which the collections of this study are deposited.

Archive	URL	#
TLA	https://archive.mpi.nl/	12
ELAR	https://elar.soas.ac.uk	10
Pangloss	http://lacito.vjf.cnrs.fr/pangloss/	2
Paradisec	http://paradisec.org.au	2
Other		6
Total		32

Table 1: Numbers of collections hosted by different archives.

The automated analysis was performed on ELAN files archived in DELAMAN archives accessible to all registered users. The URLs for these files were extracted from records harvested via OAI-PMH.[7] A custom Python script then took care of authentication and download. All code is available from https://github.com/ZAS-QUEST.

1.3. Formats and workflows

Most written resources in language-documentation data are based on oral recordings and their transcriptions. The two annotation levels that are usually thought to be indispensable in documentary records of endangered languages are transcriptions of audio recordings and their translations into a metalanguage (cf. Wittenburg et al. (2002)).

The most common tools for creating transcriptions in language documentation are Praat, Transcriber and ELAN. The translation can be added in the same program and format. Morpheme separation, glosses and part-of-speech tags are often added in a specialized interlinearization program, typically SIL Shoebox/Toolbox or, more recently SIL FLEx. These programs create a lexical database that is linked to the text database and compute possible parses of the text from the lexicon. SIL Shoebox/Toolbox create text files with minimal mark-up. Boundaries between annotation units are indicated by white space, which makes this format prone to errors. SIL FLEx and ELAN are XML-based.

As far as we could assess, the most common format in the archives is ELAN, but there are also significant amounts of collections with files created by SIL Toolbox and SIL FLEx. Many language-documentation projects use ELAN for transcription, then export the data for interlinearization in one of the SIL programs, and then reimport the annotations back into ELAN for archiving, and sometimes for additional annotation. Many projects that have Toolbox/FLEx files also have ELAN files as a result – this is the case for all but one collection in TLA. In the following, we will therefore focus on ELAN, with a few examples from Toolbox.

2. Annotation tiers

2.1. Introduction to tiers

Annotation software makes use of so-called tiers for annotation. A tier is a linear ordered set of annotations, which can be linked to time codes or to other tiers. A document consists of several tiers, which can have parent tiers, thus establishing a hierarchy (a forest, in graph theory). Thus, morpheme glossing is dependent on the transcription tier, and parts-of-speech are dependent on the interlinear morpheme glossing tier. A sample tier hierarchy in ELAN is given in Figure 1.

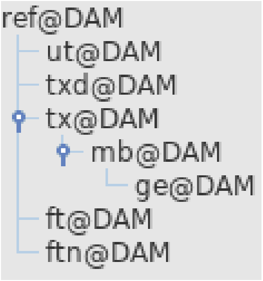


Figure 1: Tier hierarchy of a sample ELAN document

2.2. XML representation of Tiers

ELAN 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.

2.3. Overview of annotations

In order to qualify for this study, corpora are based on oral recordings and need to have a transcription. Generally speaking, all other annotations are based on the transcription tier, so if an oral corpus has annotations, this usually implies that it has a transcription. Translations into more widely accessible languages are usually deemed indispensable for further processing by non-native speakers, and are accordingly very widespread. In addition to this, highly curated corpora often also include a tier that indicates morpheme breaks, a layer in which morphemes or words are glossed, and a layer that assigns part-of-speech (POS) tags to morphemes and/or words. These annotation tiers, which are at the same time widespread and highly useful for third users, are analyzed in more detail in the following sections.

2.4. Transcription

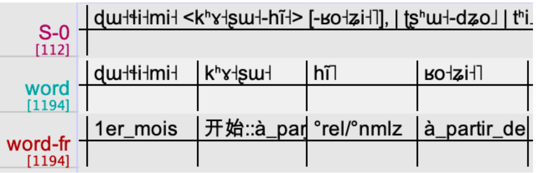
The transcription tier contains an orthographic transcription of the recording. The character set and orthographic rules depend on language-specific conventions, which may have existed prior to the documentation, or may have been developed in this context. In our sample, the character sets used for transcriptions are mostly based on ASCII characters and IPA characters, but also contain additional symbols. The main character encoding scheme used is UTF-8.

[7] https://lat1.lis.soas.ac.uk/ds/oaiprovider/oai2

	Language	Family	Macro-area	Glottocode	Donator/Reference
1	Mavea	Austronesian	Papunesia	mafe1237	(Guérin, 2006)
2	Nafsan	Austronesian	Papunesia	sout2856	(Thieberger and Brickell, 2019)
3	Daakaka	Austronesian	Papunesia	daka1243	(von Prince, 2013)
4	Daakie	Austronesian	Papunesia	port1286	(Krifka, 2016)
5	Dalkalaen	Austronesian	Papunesia	none	(von Prince, 2013)
6	Saliba-Logea	Austronesian	Papunesia	sali1295	(Margetts et al., 2017)
7	North Ambrym	Austronesian	Papunesia	nort2839	(Franjieh, 2013)
8	Anal	Sino-Tibetan	Eurasia	anal1239	(Ozerov, 2018)
9	Arapaho	Algic	N America	arap1274	(Cowell, 2019)
10	Asimjeeg Datooga	Nilotic	Africa	isim1234	(Griscom, 2018)
11	Bora	Boran	S America	bora1263	(Seifart, 2009)
12	Goemai	Afro-Asiatic	Africa	goem1240	(Hellwig, 2003)
13	Gorwaa	Afro-Asiatic	Africa	goro1270	(Harvey, 2017)
14	Jakarta Indonesian	Austronesian	Papunesia	cjin1234	(Gil et al., 2015)
15	Kakabe	Atlantic-Congo	Africa	kaka1265	(Vydrina, 2013)
16	Kamas	Uralic	Eurasia	kama1378	(Gusev and Klooster, 2018)
17	Katla	Atlantic-Congo	Africa	katl1237	(Hellwig, 2007)
18	Komnzo	Yam	Papunesia	wara1294	(Döhler, 2019)
19	Mojeño Trinitario	Arawakan	S America	trin1274	(Rose, 2018)
20	Movima	(isolate)	S America	movi1243	(Haude and Beuse, 2016)
21	Resígaro	Arawakan	S America	resi1247	(Seifart, 2019)
22	Ruuli	Atlantic-Congo	Africa	ruul1235	(Witzlack-Makarevich et al., 2019)
23	Sadu	Sino-Tibetan	Eurasia	sadu1234	(Xu et al., 2012a)
24	Savosavo	Austronesian	Papunesia	savo1255	(Wegener, 2016)
25	Tabaq (Karko)	Nubian	Africa	kark1256	(Hellwig, 2007)
26	Totoli	Austronesian	Papunesia	toto1304	(Leto et al., 2010)
27	Tunisian Arabic	Afro-Asiatic	Africa	tuni1259	Fatma Messaoudi, unpublished
28	Urum	Turkic	Eurasia	urum1249	(Skopeteas, 2018)
29	Vera'a	Austronesian	Papunesia	vera1241	(Schnell, 2015)
30	Yali	Trans-New-Guinea	Papunesia	angg1239	(Riesberg et al., 2016)
31	Yongning Na	Sino-Tibetan	Eurasia	yong1270	(Michaud, 2017)
32	Yurakaré	(isolate)	S America	yura1255	(Van Gijn et al., 2012)

Table 2: The corpora used in this study. Classification based on Hammarström et al. (2019).

Common transcription tiers follow orthographic conventions and do not constitute detailed phonetic transcriptions. For most purposes, this is an important level of abstraction, which allows, among other things, for readable texts, a compilation of word lists, measures of type-token ratios and similar. Words are generally separated by white space. In ELAN, words may additionally have annotation boundaries. In many ELAN files, there are two transcription tiers, one where words are separated by white space, and one in which each word has is its own annotation unit. This is mostly motivated by technical necessity and slightly redundant, except in languages with significant sandhi phenomena.



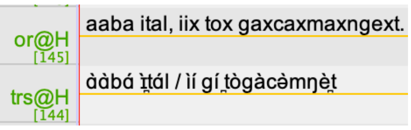
S-0 [112]	ɖɯ˧ɬi˧mi˧ <kʰɤ˧ʂɯ˧-hĩ˧> [-ʁo˧ʑi˧˥], | ʈʂʰɯ˧-dʑo˩ | tʰi.			
word [1194]	ɖɯ˧ɬi˧mi˧	kʰɤ˧ʂɯ˧	hĩ˥	ʁo˧ʑi˧˥
word-fr [1194]	1er_mois	开始::à_par	°rel/°nmlz	à_partir_de

Figure 3: Tone sandhi transcription in the Yongning Na corpus. `S-0`: sentence-level transcription; `word`: word-level transcription.

Thus, in the Yongning Na corpus, the sentence-based transcription tier traces tone sandhi phenomena, while the word-level transcription tier shows the underlying lexical tones (Figure 3).

Following the conventions introduced by the desktop program *SIL Toolbox* from the 1990s, the transcription tier is often labeled as `tx`, sometimes also as `transcription`.

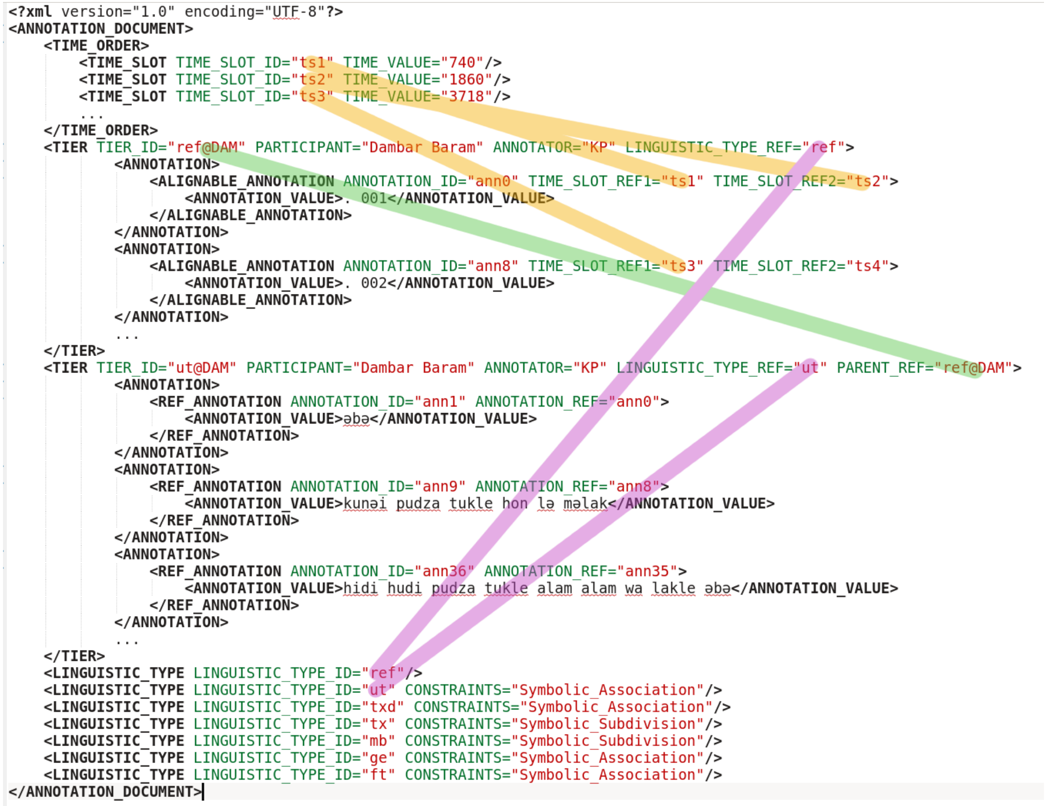
In addition to the orthographic transcription, which is a prerequisite for the corpora in our sample, three corpora also contain a phonetic transcription tier. The following example is from the Katla corpus (Hellwig, 2007), where `trs` is the name of the phonetic annotation tier, including transcriptions of tones, and `or` is the name of the orthographic transcription tier, as shown in Figure 4. This figure also illustrates the conventions for indicating different speakers in elan files, as in `TierName@Speaker`.



or@H [145]	aaba ital, iix tox gaxcaxmaxngext.
trs@H [144]	ààbá ɪ̰̀tál / ìí gí ̰tògàcə̀mŋḛ̀t

Figure 4: Orthographic and phonetic transcription in the Katla corpus.

Tones can be encoded by diacritics as in Figure 4, by super-



```
<?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 mark-up 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).

script numbers (in the Sadu corpus) or by tonal IPA characters (in the Yongning Na corpus).

Collections vary in terms of how they handle speech disfluencies, errors, intonation boundaries within transcription units and similar.

Summing up, transcription tiers commonly have the following properties:

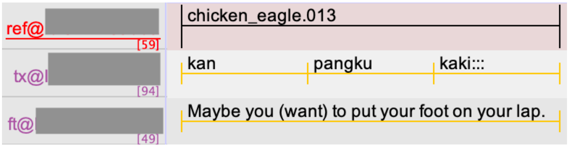
- they are orthographic rather than phonetic;
- words are separated by white space and/or annotation boundaries;
- sentences and clauses are separated by annotation boundaries (possibly of superordinate tiers) and/or punctuation.
- disfluencies may be indicated by ellipses '…' or project-specific conventions.

2.5. Translation

The translation tier is crucial for the usability of the corpus data (Wittenburg et al., 2002). Accordingly, corpora from language documentation generally include translations into at least one language of wider communication. The only exception in our sample is the corpus on Tunisian Arabic, which may be argued to be accessible to a reasonably large number of users by virtue of being an Arabic language. While corpora without translations are not usable for most research purposes, corpora *with* translations constitute parallel corpora with fine-grained alignment and therefore a very rich resource, which can be used for translation mining (Wälchli, 2007) and other methods.

The main language of translation is English, but other metalanguages are also widespread. Some corpora even contain several translations tiers in different languages. Thus, the Kakabe corpus is translated into Russian, French and English (Vydrina, 2013); the Kamas corpus contains translations into Russian, English and German (Gusev and Klooster, 2018); the Sadu corpus is systematically translated into Mandarin Chinese and English (Xu et al., 2012b), etc. Multiple translation tiers not only open the resource to a wider range of users, they also create more parallel data, with all the applications they offer. Following the convention established by SIL Toolbox, translation tiers are often labeled as `ft` ('free translation'), as `fn` (translation into the national language), or follow the pattern `fe`, `fg` for translations into English and German respectively.

The translation tier is aligned with the transcription tier. In Toolbox format, this would mean that both belong to the same reference unit. In ELAN, this could mean merely that they have the same temporal extension, although in most cases the dependency between them will be implemented more explicitly. Thus, the translation tier may be a child of the transcription tier, or both might be child tiers of a 'reference' tier. Reference tiers often result from import from interlinearization software. They may define the time span of an annotation unit and give it a unique label. For example, in the Totoli corpus (Leto et al., 2010), the transcription tier is a child tier of the reference tier and segmented into individual words. The translation tier is also a child tier of the reference tier. Figure 5 shows all three tiers displayed in ELAN.



ref@[redacted] [59]	chicken_eagle.013		
tx@l[redacted] [94]	kan	pangku	kaki:::
ft@[redacted] [49]	Maybe you (want) to put your foot on your lap.		

Figure 5: Alignment of translation and transcription in the Totoli corpus. Additional tiers are hidden from view.

Importantly, free translations generally constitute translations of entire sentences. They interpret the utterance, not its individual words. Of course, translations need to be used with caution, since they may always contain considerable amounts of translationese (Gellerstam, 1986). This also applies to corpora from language documentation, possibly more so than with some other resources, since the researchers or language experts in charge are typically not trained as translators and may not be native speakers of the language they translate into.

In sum:

- A translation tier contains free translations of entire sentences.
- Its units are aligned with the transcription units.
- The metalanguage is an internationally widely used language.
- There may be more than one translation tier with a different metalanguage each.

2.6. Morpheme separation and glosses

While the only obligatory tiers are the transcription and translation tiers, highly curated corpora often contain additional information about morpheme boundaries, glosses, and POS tags. Out of the sample of 32 highly curated corpora, 27 contain a tier in which words from the transcription tier are split into separate morphemes. With one exception, corpora that contain a morpheme tier also provide a tier with morpheme-by-morpheme glosses. Some of the corpora are only partially glossed.

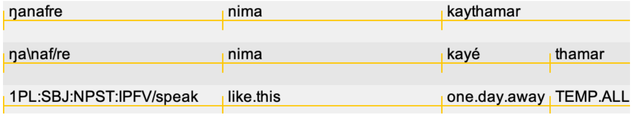
The gloss tier typically requires the existence of a morpheme tier, since glosses generally provide information about morphemes. There are several ways in which morpheme boundaries can be encoded, and in which they can be aligned with the transcription tier and with their glosses. In SIL Toolbox, the interlinearization tool creates a morpheme tier where morphemes are separated by white space, and bound morphemes are affixed with separating symbols such as '-' for affixes and '=' for clitics. The same boundary markers are recreated on the gloss tier. Boundaries that contain only white space and no other separating symbols correspond to word boundaries in the transcription tier.

This is illustrated by the following example from the Nafsan (South Efate) corpus (Thieberger, 2006), where `tx` is the transcription tier, `mr` is the morpheme tier and `mg` is the gloss tier:

```
\tx Rapan rasoki asler.
\mr ra= pan ra= sok -ki asel -e -r
\mg 3D.RS= go 3D.RS= jump -TR friend -V -3P.DP
```

In ELAN, there are several different ways in which morpheme boundaries can be marked and these three tiers can be aligned. The choice between them depends not only on workflows and software, but also on language specific properties.

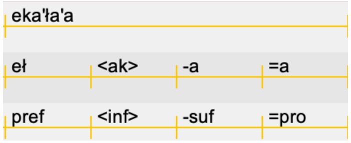
Komnzo, a Yam language of PNG, is characterized by distributed exponence, which means that morphological information is not encoded in a simple linear fashion. Instead, specific combinations of prefixes and suffixes, or circumfixes, encode features such as 'plural' (Döhler, 2018). Nonlinear morphology constitutes a special challenge to large-scale morpheme-by-morpheme glossing. In the Komnzo corpus (Döhler, 2019), a substantial set of texts has been manually annotated in ELAN. Here, in the morpheme tier, linear combinations of morphemes are separated by an annotation boundary – all morphemes of one word are aligned with this word in a word-by-word transcription tier. Nonlinear inflected forms are contained within a single annotation unit. In the morpheme tier, boundaries between the root and the affixes are indicated by slashes – AFFIX\ROOT/AFFIX. In the gloss tier, the information encoded by the combination of affixes is given first, while the information encoded by the root comes second and is indicated by a slash – AFFIX_MEANING/ROOT_MEANING. Figure 6 illustrates this solution. This type of morphology is very hard to annotate semi-automatically with existing interlinearization software and is generally added manually instead.[8] It also presents challenges to automated ingestion for further processing in a toolchain (Figure 6).



ŋanafre	nima	kaythamar	
ŋa\naf/re	nima	kayé	thamar
1PL:SBJ:NPST:IPFV/speak	like.this	one.day.away	TEMP.ALL

Figure 6: Morpheme separation and glossing in Komnzo.

Another language in the sample that has highly complex, and sometimes non-linear, morphology is Movima. In the corpus (Haude and Beuse, 2016), the individual morphemes of a word constitute separate annotation units, which are subdivisions of the word unit. The nature of the morphological boundary is indicated by different separator characters: '<INFIX>', '-AFFIX', '=CLITIC'. These symbols correspond to the conventions in the Leipzig Glossing Rules (Comrie et al., 2008). They are replicated in the gloss tier, labeled as g and the POS tier, labeled as p. This is illustrated in Figure 7.



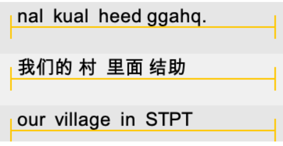
eka'ła'a			
eł	<ak>	-a	=a
pref	<inf>	-suf	=pro

Figure 7: Morpheme separation and glossing in Movima.

In highly isolating languages, the difference between words

[8] The Giellatekno project creates parsing software for morphologically complex languages (http://giellatekno.uit.no/index.eng.html).

and morphemes largely collapses, so that word-by-word glosses are a feasible option. In our sample, this is the case for Sadu, a Sino-Tibetan language of China (Xu et al., 2012b). Figure 8 shows how the transcription is glossed in Mandarin Chinese and English. In this particular instance, units in both gloss tiers are coextensive with the transcription units, words are separated by white space only, so that cross-tier correspondences between units can only be inferred indirectly. The Sino-Tibetan language Yongning Na also has word-level glosses.



nal kual heed ggahq.
我们的 村 里面 结助
our village in STPT

Figure 8: Word-by-word glosses in Sadu.

As we have seen from the above examples, the metalanguage for glosses is usually English, but other languages, too, can be used for glosses. Thus, the Kakabe corpus has glosses in English, Russian and French. The only corpus in our sample which has no English glosses is the Yongning Na corpus (Michaud, 2017) – the only metalanguage for glosses here is French.

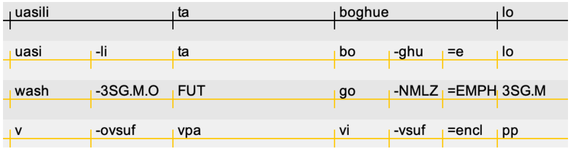
Glosses are often systematically divided into ‘content morphemes’ such as *table* and ‘grammatical morphemes’ such as PLURAL by typographic means. Thus, grammatical morphemes may be in all-caps or be prefixed by a special symbol such as ‘°’ in the Yongning Na corpus. Common labels for the morpheme tier are `mb` and `morph`. For the gloss tier, labels usually start with a *g*, frequent labels are `gl`, `gloss` and `ge` for English glosses.

In some, the following conventions apply to morpheme and gloss tiers:

- Morphemes are aligned with word units.
- Relations between morphemes are encoded by specific separator symbols.
- Glosses can be assigned to morphemes (default) or to words.
- Glosses are usually in English, but other metalanguages are not uncommon. As with translations, there may be several gloss tiers in different metalanguages.
- Content morphemes may be differentiated from others by project-specific typographic conventions.

2.7. POS tags

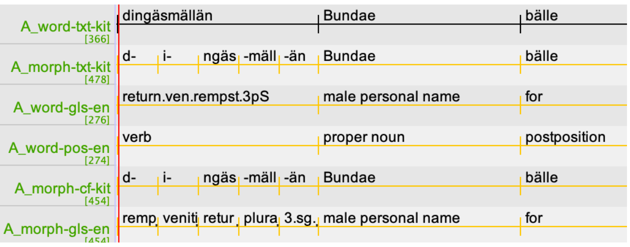
Many of the corpora that are glossed also have a tier for part-of-speech tags (POS). When using SIL Toolbox for interlinear glossing, a POS tier is automatically created along with the morpheme separation and the glosses. The POS tier is then a child of the morpheme tier, and POS tags label individual morphemes rather than whole words. This is an important difference to most other corpora with POS tags, where POS tags indicate properties of words. The example from Savosavo (Wegener, 2016) in Figure 9 indicates that the category of the word cannot be trivially determined from the morpheme-wise POS tags.



uasili		ta	boghue			lo
uasi	-li	ta	bo	-ghu	=e	lo
wash	-3SG.M.O	FUT	go	-NMLZ	=EMPH	3SG.M
v	-ovsuf	vpa	vi	-vsuf	=encl	pp

Figure 9: Morpheme-wise POS tags in the Savosavo corpus.

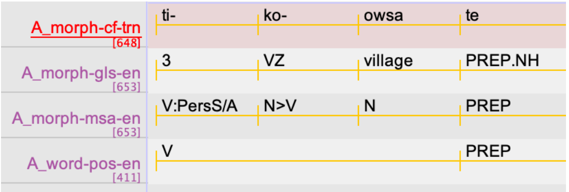
In SIL FLEx,[9] users can choose to add word-level POS tags. In our sample, a subset of the Ende corpus (Lindsey, 2007) has such word-level POS tags, created by the workflow with SIL FLEx as shown in Figure 10.



A_word-txt-kit [366]	dingäsmällän					Bundae	bälle
A_morph-txt-kit [478]	d-	i-	ngäs	-mäll	-än	Bundae	bälle
A_word-gls-en [276]	return.ven.rempst.3pS					male personal name	for
A_word-pos-en [274]	verb					proper noun	postposition
A_morph-cf-kit [454]	d-	i-	ngäs	-mäll	-än	Bundae	bälle
A_morph-gls-en [454]	remp	veniti	retur	plura	3.sg.	male personal name	for

Figure 10: Word-level POS tags in the Ende corpus (tier label `A_word-pos-en`).

Some projects even combine morpheme-level POS tags and word level POS tags. The only such case in our sample is the Mojeño corpus (Rose, 2018), which also relies on SIL FLEx for interlinearlization – see Figure 11.



A_morph-cf-trn [648]	ti-	ko-	owsa	te
A_morph-gls-en [653]	3	VZ	village	PREP.NH
A_morph-msa-en [653]	V:PersS/A	N>V	N	PREP
A_word-pos-en [411]	V			PREP

Figure 11: Morpheme-level POS tags (`A_morph-msa-en`) and word-level POS tags (`A_word-pos-en`) in the Mojeño corpus.

The inventories of POS tags are highly language specific and differ in terms of their granularity, for example with respect to whether transitive verbs are differentiated from intransitive ones. Syntactic dependencies are not typically annotated in corpora from language documentation.

In short, POS tiers are aligned either with word units or with morpheme units and label the properties of the corresponding units.

2.8. Other tiers

Apart from the tiers discussed above, the only other two frequent annotation tiers are reference tiers and note tiers. Reference tiers were briefly addressed in section 2.5.. The content of note tiers is entirely arbitrary. Some collections have more than one note tier, e. g. notes on the transcription vs. notes on the translation. Note tiers are often labeled

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

as `nt`, according to the SIL Toolbox convention, or more explicitly as `Note` or similar.

Many projects additionally have tiers for their own purposes. Some collections contain several session-wide tiers, each of which contains metadata on the circumstances of the recording, such as time and date. Apart from the tiers discussed so far, there are no significant overlaps between different collections.

2.9. Sets of tiers for different speakers

The entire hierarchy of annotation tiers may exist not just once but several times per recording, if there is more than one speaker. Tier names are then identical across speakers except for a speaker-specific, often anonymized affix. A common convention is for speaker references to be suffixed to the tier name, with the `@` as a separator, as in `tx@a`, `ft@a@`, `tx@b`, `ft@b` etc. Two of the collections in this survey use alphabetical prefixes with an underscore instead as in `A_phrase-gls-en`.

2.10. Summary

The most frequently used tiers that are also likely to be useful for uses by third parties are listed in Table 3. This table also lists the main parameters along which each tier type differs between corpora. Third-party users of the corpora will for most purposes need to verify how to fill in this table for each corpus. Creators of such corpora can facilitate the use of their data by providing this information.

Tier	Name	Language	Other parameters
Transcription	tx, …		**Clause-wise** vs. word-wise segmentation, word-wise segmentation; **orthographic**, phonetic
Translation	ft, …	English, …	
Morphemes	mb, …		separated by **annotation boundaries** vs. special characters; allomorphs, **underlying morphemes**
Glosses	ge, …	English, …	**morpheme-wise**, word-wise;
POS tags	ps, …	English, …	**morpheme-wise**, word-wise;

Table 3: The main annotation levels and their parameters, with defaults in bold.

An overview of how the corpora in this study relate to these parameters is given in table 4.

3. Automated analysis

Next to the manual inspection of the 32 corpora discussed above, we also performed an automated analysis of more than 20 000 *eaf files harvested from five different endangered language archives of the DELAMAN network:

- 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).

At this point, no effort was made to calibrate the collection, but future research could look into how consistent or divergent particular collections are. The two quantitative analyses we ran covered tier names and tier structures.

3.1. Tier names

This analysis was run on the ELAR archive only. We find a total of 652 different tier names, with 554 occurring more than once, so that accidental typos are unlikely. Table 5 gives a breakdown of the most frequent tiers.

We see that the identifiers `tx`, `ft` or `ge` discussed above are well represented, but there are also other tier names like `SA` which do not have an intuitive interpretation. For those cases, researchers would have to look at the individual files themselves to arrive at an interpretation. A recommendation would be for depositors to stick to the tier names which have emerged as the most frequent ones, i.e. the ones given in Table 5.

3.2. Tier structure

Disregarding the names, we can also analyze the hierarchical structure of tiers in ELAN files. In ELAN, tiers can have dependent tiers (child tiers). These can have the relations “default”, “time subdivision”, “symbolic subdivision”, “symbolic association”, and “included in”. Table 6 gives an overview of the relations which hold between tiers. We see that “symbolic association” is used most frequently with 111 385 instances (mainly for glossing and parts-of-speech, see above), but that “included in” is hardly ever used as a relation between a child tier and its parent.

Based on the part-whole-relations, we can give each file a “fingerprint”, e.g. [x[saa]x] for a file with five tiers of which the second, third, and fourth tier are children of the first one, and are of type “subdivision” and “association” (2x), respectively. This allows us to disregard the actual naming conventions for tiers and still arrive at meaningful comparisons. Based on these fingerprints, we can show the heterogeneity of ELAN files (Figure 12). There are 2 187 different fingerprints. The most frequent one with meaningful hierarchies[10] is “[x[saa]x]”, which only used in 910 out of 20,089 files, i.e. in 4.5%. This means that applications wanting to tap into ELAN corpora as a resource will only achieve a coverage of less than 5% if they focus on the most common configuration. In order to arrive at a coverage of 50%, 38 configurations have to be supported (Figure 13); for 90% coverage, 584 different tree structures will have to be supported.

There might be some prospects for reuse of code for automated analysis since there are files of the type “[x[saa]xx[saa]x]”, which are found often in files covering dialogues. These could be separated into twice “[x[saa]x]”, adding a couple of percentage points to the coverage.

[10] The configurations [xx] (1 081) and [xxx] (974) are more frequent, but their structure does not give any hints as to the semantics of tiers.

	Transcription	Translation	Morphemes	Glosses	POS tags
1	tx	ft	mb	ge	ps
2	tx	fg	mb	mg	gd
3	tx	ft	mb	ge	ps
4	tx	ft	mb	ge	ps
5	tx	ft	mb	ge	ps
6	tx	f	mb	g	p
7	lexical-data	trans	morphological-data	gls	type
8	[various]	[various]	morph_an		
9	tx	ft	mb	ge	ps
10	[various]	FT, [various]	MORPH	GLOSS	
11					
12	or	ft	mb	gl	ps
13	Text	Free Translation	Morpheme Break	Gloss	
14	\tx, \ph*	\ft	\mb	\ge	\ps
15	tx	ft, ftf†, ftr†			
16	ts	fe, fr†, fg†	mb, mp‡	ge	ps
17	or, trs*	ft	mb	gl	ps
18	tx	ft	mb	gl	pos
19	Transcription-trn	A_phrase-gls-{en|es†|fr†}	morph-cf-trn	morph-gls-en	morph-msa-en, word-pos-en**
20	tx1	fe, fn†, ft†	m	g	p
21	t	f	mb	gl	ps
22	tx	fte	mb	gl	ps
23	ft, wp*	te, tn†	tx, fn†		
24	st	ft	mb	gl	wc
25	trs	ft, fn†			
26	tx††	ft, ftn†	mr	ge, gn†	
27	[various]				
28	tx-a	ft	mr	ge, gn†	
29	utterance	utterance_translation	grammatical_words	gloss	graid**
30	tx††	ft, ftn†	mr	ge, gn†	
31	S-0	S-0-fr†, S-0-zh	word-fr		
32	ts	tl			

Table 4: Defaults are: orthographic transcription, English as metalanguage, morpheme-level glosses, morpheme-level POS tags, clause-wise transcription. Other options: *phonetic transcription; †other metalanguage than English; ‡reconstruction of underlying morphemes; **word-level POS tags; ††word-wise transcription; ‡‡word-level glosses.

Table 5: The most frequent tier names in 12k eaf files from ELAR.

8488	`default-lt`	1917	`ge`
7275	`Note`	1900	`UtteranceType`
5179	`ref`	1756	`Transcription`
4429	`morph-item`	1698	`mb`
4185	`tx`	1561	`text`
4108	`Words`	1396	`Free Translation (English)`
3454	`Phrases`	1289	`word`
3161	`ft`	1232	`ps`
2815	`translation`	1173	`phrase`
2696	`SA`	1142	`word-item`
2617	`Text`	1119	`notes`
2131	`nt`	1070	`Symbolic Association`
2000	`Translation`	1024	`morph`
1951	`phrase-item`	998	`one-to-one`

default	59 286
symbolic subdivision	26 167
symbolic association	111 385
included in	1 218
time subdivision	3 248

Table 6: Frequency of tier relations in 20k ELAN files.

4. Conclusions

We have described common annotation practices in corpora from language documentation. We have shown that highly curated collections represent very rich resources, that include finely aligned parallel corpus data with translations in one or more languages; and often with manually added information about morphological structure, glosses and POS information. As such, they are a highly valuable resource both for theoretical research, a range of practical applications especially for speaker communities, and as a potential test bed for NLP technologies.

Even though there is a lot of variation between different projects in terms of their annotations, we could identify a certain core of common conventions. We hope that this description will help with the exploration of existing corpora and facilitate the documentation and creation of future data collections.

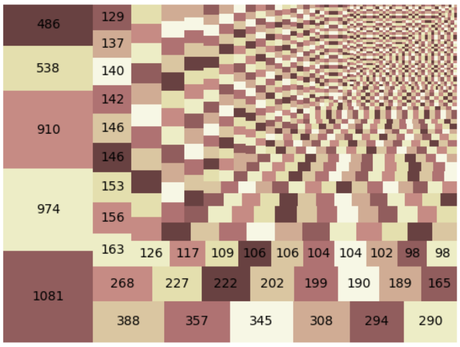


Figure 12: Users can configure the relation of annotation tiers in ELAN files. This graphics shows 2 187 different tier configurations found in the 20 089 *eaf files openly available from the five DELAMAN archives. Each box represents a different logical configuration, with the number of files making use of that configuration. The names of tiers are disregarded

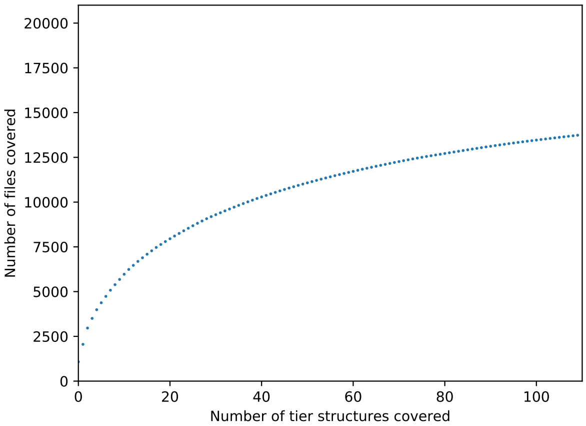


Figure 13: Coverage of the different ELAN tier structures increases as more different tier structures are taken into account.

5. Acknowledgements

This research is funded by the German BMBF, as part of the QUEST project. We thank Frank Seifart and Ludger Paschen for discussion and feedback.

6. Bibliographical References

Comrie, B., Haspelmath, M., and Bickel, B. (2008). The leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.

Cox, C. (2011). Corpus linguistics and language documentation: challenges for collaboration. In *Corpus-based studies in language use, language learning, and language documentation*, pages 239–264. Brill Rodopi.

Dahl, Ö. (2015). How WEIRD are WALS languages? paper at the MPI EVA Linguistics closing conference, May.

Döhler, C. (2018). *A grammar of Komnzo*. Number 22 in Studies in Diversity Linguistics. Language Science Press, Berlin.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, 1:88–95.

Good, J. (2011). Data and language documentation. In Peter Austin et al., editors, *Handbook of Endangered Languages*, page 212–234. Cambridge University Press, Cambridge.

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

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302):29.

Seifart, F. (2012). The threefold potential of language documentation. In *Potentials of language documentation: Methods, analyses, and utilization*, pages 1–6. University. of Hawai'i Press.

Gary F. Simons et al., editors. (2017). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 20 edition.

Wälchli, B. (2007). Advantages and disadvantages of using parallel texts in typological investigations. *Sprachtypologie und Universalienforschung*, 60:118–134.

Wittenburg, P., Mosel, U., and Dwyer, A. (2002). Methods of language documentation in the DOBES program. In *The 3rd International Conference on Language Resources and Evaluation (LREC 2002). Workshop on Tools and Resources in Field Lingusitics*, pages 36–42. European Language Resources Association.

7. Language Resource References

Cowell, James Andrew. (2019). *Arapaho*. University of Colorado.

Döhler, Christian. (2019). *Komnzo text corpus*. Zenodo. http://doi.org/10.5281/zenodo.2634978.

Franjieh, Michael. (2013). *A documentation of North Ambrym, a language of Vanuatu*. SOAS, Endangered Languages Archive. https://elar.soas.ac.uk/Collection/MPI67426. [Accessed on 2017/10/04].

Gil, David and Tadmor, Uri and Bowden, John and Taylor, Bradley. (2015). *Data from the Jakarta Field Station*. Department of Linguistics, Max Planck Institute for Evolutionary Anthropology.

Griscom, Richard. (2018). *Documenting Isimjeeg Datooga*. London, SOAS: Endangered Languages Archive.

Guérin, Valérie. (2006). *Documentation of Mavea*. London, SOAS: Endangered Languages Archive.

Gusev, Valentin and Klooster, Tiina. (2018). *INEL Kamas Corpus*. Hamburger Zentrum für Sprachkorpora.

Harvey, Andrew. (2017). *Gorwaa: an archive of language and cultural material from the Gorwaa people of Babati (Manyara region, Tanzania)*. London, SOAS: Endangered Languages Archive.

Haude, Katharina and Beuse, Silke Angelika. (2016). *Movima*. Nijmegen: TLA.

Hellwig, Birgit. (2003). *Goemai texts*. London, SOAS: Endangered Languages Archive.

Hellwig, Birgit. (2007). *A documentation of Tabaq, a Hill Nubina language of the Sudan, in its sociolinguistic context*. London, SOAS: Endangered Languages Archive.

Krifka, Manfred. (2016). *Daakie*. Nijmegen: TLA.

Leto, Claudia and Alamudi, Winarno S. and Himmelmann, Nikolaus P. and Kunht-Saptodewo and Riesberg, Sonja and Basri, Hasan. (2010). *DoBeS Totoli Documentation*. Nijmegen: TLA.

Lindsey, Kate L. (2007). *Language Corpus of Ende and other Pahoturi River Languages (LSNG08). Digital collection managed by PARADISEC*.

Margetts, Anna and Margetts, Andrew and Dawuda, Carmen. (2017). *Saliba/Logea*. The Language Archive. http://dobes.mpi.nl/projects/saliba.

Michaud, Alexis. (2017). *Na Corpus*. Pangloss collection, LACITO-CNRS.

Ozerov, Pavel. (2018). *A community-driven documentation of natural discourse in Anal, an endangered Tibeto-Burman language*. London, SOAS: Endangered Languages Archive.

von Prince, Kilu. (2013). *Dalkalaen, The Language Archive*. MPI for Psycholinguistics.

Riesberg, Sonja and Walianggen, Kristian and Zöllner, Siegfried. (2016). *DoBeS Documentation Summits in the Central Mountains of Papua*. Nijmegen: TLA.

Rose, Françoise. (2018). *Corpus mojeño trinitario*. Online database.

Schnell, Stefan. (2015). *Multi-CAST Vera'a*. Haig, Geoffrey and Schnell, Stefan (eds.), Multi-CAST: Multilingual corpus of annotated spoken texts.

Seifart, Frank. (2009). *Bora documentation*. Nijmegen: TLA.

Seifart, Frank. (2019). *Resígaro*. Nijmegen: TLA.

Skopeteas, Stavros. (2018). *Urum*. Universität Göttingen.

Thieberger, Nick and Brickell, Timothy. (2019). *Multi-CAST Vera'a*. Haig, Geoffrey and Schnell, Stefan (eds.), Multi-CAST: Multilingual corpus of annotated spoken texts.

Thieberger, Nick. (2006). *Dictionary and texts in South Efate*. Digital collection managed by PARADISEC. DOI: 10.4225/72/56FA0C5A7C98F.

Van Gijn, Rik and Hirtzel, Vincent and Gipper, Sonja. (2012). *The Yurakaré archive*. Nijmegen: TLA.

von Prince, Kilu. (2013). *Daakaka*. Nijmegen: TLA.

Vydrina, Alexandra. (2013). *Description and documentation of the Kakabe language*. London, SOAS: Endangered Languages Archive.

Wegener, Claudia. (2016). *Savosavo*. Nijmegen: TLA.

Witzlack-Makarevich, Alena and Namyalo, Saudah and Kiriggwajjo, Anatol and Molochieva, Zarina and Atuhairwe, Amos. (2019). *A corpus of spoken Ruuli*. Makerere University & Hebrew University of Jerusalem.

Xu, Xianming and Bai, Bibo and Yang, Yan. (2012a). *Linguistic and cultural documentation of Sadu*. London, SOAS: Endangered Languages Archive.

Xu, Xianming and Bai, Bibo and Yang, Yan. (2012b). *Linguistic and cultural documentation of Sadu*. SOAS, Endangered Languages Archive.