Creating and enriching a repository of 177k interlinearized examples in 1 611 mostly lesser-resourced languages

Sebastian Nordhoff
BBAW
nordhoff@bbaw.de

Thomas Krämer
GESIS - Leibniz-Institut für Sozialwissenschaften
thomas.kraemer@gesis.org

Abstract

Much of NLP is concerned with languages for which dictionaries, thesauri, word nets or treebanks are available. This contribution focuses on languages for which all we have might be some isolated examples with word-to-word translation. We detail the collection, aggregation, storage and querying of this database of 177k examples from 1 611 languages with a special eye on enrichment via Named Entity Recognition and links to the Wikidata ontology. We also discuss pitfalls of the approach and discuss the legal status of interlinear examples.

1 Introduction

1.1 Overview

While for major languages, linguistic resources are plentiful and available in breadth and depth, this is not the case for the majority of the languages of the world. Joshi et al. (2020) classified n languages of the world according to the materials they have available. This yielded 6 groups, given in Table 1, with Group 5 for the languages with the most resources and Group 0 for the languages with the least resources. Joshi et al. (2020) only used 2k languages. Nordhoff (2020b) expanded on Joshi et al.'s classfication, adding a group –1, where there is some data available, but it is even less than for group 0. Nordhoff (2020b) showed how data for languages from group 0 can be harvested from heterogeneous data found in endangered language archives (von Prince and Nordhoff, 2020).

Nordhoff and Krämer (2022) extended this approach to include data from open access books published by Language Science Press.[1] and also provided a modelling as Linked Data. This yielded 40 000 examples in 280 languages.

In this paper, we will discuss further improvements on the ingestion side, with the inclusion of Open Text Collections and the corpus of Indigenous Northern Eurasian Languages, yielding a total of 177k examples in 1 611 languages. The examples are enriched with metadata for geography, linguistic affiliation, and semantic content. We discuss challenges in ingestion, enrichment, and federated querying and compare the platform imtvault.org to extant platforms like ODIN, OLAC, or the Delaman archives.

2 Interlinear glossed text

In the context of linguistic typology and language documentation, the typical format is so called interlinear glossed text (IGT). An example is given in (1)

(1) Tayaĝu-x̂ qa-x̂ qa-ku-x̂.
man-SG fish-SG eat-PRES-3SG.
'The man is eating the fish.'

The first line contains the vernacular text, in this case in Aleut. The second line contains a word-to-word (or morpheme-to-morpheme) translation. The third line contains a free translation of the whole sentence.

While this is only very little information, some insights can readily be obtained: *x̂* for instance marks singular both on nouns and verbs, and the lexical items *tayaĝu* 'man' *qa* 'fish' and *qa* 'eat' can also be extracted.

Various formats have been proposed for the modeling of interlinear glossed text (Drude, 2003; Goodman et al., 2015; Chiarcos et al., 2017; Chiarcos and Ionov, 2019). For our purposes, we use the CLDF format (Forkel et al., 2018), which is csv-based and which has a whole ecology supporting a variety of websites and services, such as WALS, APiCS or Glottolog. A CLDF rendering of example 1 is given in Figure 1.

[1] https://langsci-press.org

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

Table 1: Joshi et al's classes

Analyzed_Word	Gloss	Translated_Text
Tayaĝu-x̂⇒qa-x̂⇒qa-ku-x̂.	man-SG⇒fish-SG⇒eat-PRES-3SG.	The man is eating the fish.

Figure 1: The CLDF representiation of example (1). ⇒ stands for a tab. Note that this tabular data is complemented by a json file describing the different column types.

3 Ingestion

3.1 Sources

Nordhoff and Krämer (2022) detail the ingestion of LangSci books via the CLDF format (Forkel et al., 2018), for 40k examples. This proved already useful for the training of automated glossing procedures for unknown text (Okabe and Yvon 2023, also see Ginn et al. 2023, 2024 for similar approaches). Since then, the number of examples extracted from LangSci books has been augmented to 66k, but the basic approach has remained the same. In addition to the provider LangSci, the CLDF `examples.csv` also contains 26 537 examples retrieved from the open access journal Glossa and examples used in various CLLD websites hosted by the Max Planck Institute for Evolutionary Anthropolopgy (apics: 15 805; wals: 3 907; malchukovditransitives: 2 071; uratyp 1 985; dictionaria: 3 957; igasttdir: 676; jacquesestimative: 32).

The project Open Text Collections (OTC, Nordhoff et al. 2024) collects narratives in lesser described languages and makes them available as pdf, printed books, but also as structured data, in CLDF format. One Open Text Collection has been published so far, of the language Komnzo spoken in Papua New Guinea, adding another 1 970 examples, with 15 more collections of comparable size in the pipeline.

Finally, the INEL project (Grammars, Corpora and Language Technology for Indigenous Northern Eurasian Languages) is an 18-year long-term project hosted by the Academy of Sciences and Humanities in Hamburg, which released extensive corpora of the Siberian languages Dolgan, Kamas, Selkup, Evenki, Enets, and Nenets, with altogether another 100k+ examples. Figure 2 shows the various levels of aggregation.

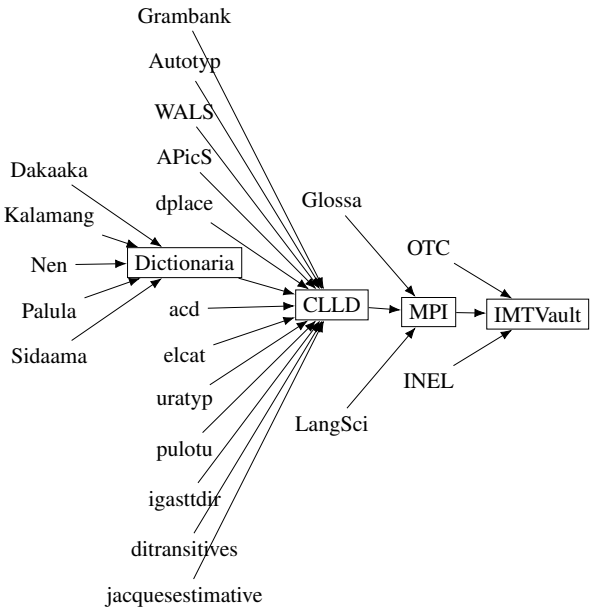

Figure 2: The aggregation of linguistic examples in several steps

3.2 File formats

The INEL project has a very deep and granular XML-based annotation, which differs consid-

(12) Acehnese
Hana lön-bloe saka sabab ***mantöng le*** ***di rumoh***.
NEG 1SG-buy sugar because still many in house
'I am not buying any sugar because there is **still much at home**.'
(Asyik 1987: 175)

(a) PDF

```
\begin{exe}
      \ex \ili{Acehnese}\label{exScalarAcehnese}\\
      \gll Hana lön-bloe saka sabab \textbf{mantöng} \textbf{le} \textbf{di} \textbf{rumoh}.\\
      \textsc{neg} 1\textsc{sg}-buy sugar because still many in house\\
      \glt \lq I am not buying any sugar because there is \textbf{still much at home}.'
      \\\parencite[175]{Asyik1987}
\end{exe}
```

(b) Corresponding LaTeX source code

Figure 3: An example in Acehnese (Language Science Press)

mìiir ă-mòr
giraffe.SG DECL.SG-fast
'The giraffe is fast.'

(a) HTML

```
<list list-type="sentence-gloss">
  <list-item>
    <list list-type="word">
      <list-item>
        <p>miiir</p>
      </list-item>
      <list-item>
        <p>giraffe.
        <sc>SG</sc></p>
      </list-item>
    </list>
    <list list-type="word">
      <list-item>
        <p>ă-mòr</p>
      </list-item>
      <list-item>
        <p>
        <sc>DECL.SG</sc>-fast</p>
      </list-item>
    </list>
  </list-item>
  <list-item>
    <list list-type="final-sentence">
      <list-item>
        <p>'The giraffe is fast.'</p>
      </list-item>
    </list>
  </list-item>
</list>
```

(b) Corresponding XML source code

Figure 4: An example in Dinka (Glossa)

erably from the rather shallow data structures we find in LangSci, Glossa or CLLD sites.

Figure 5 shows some of the tiers of the file `AnKA_2009_Story_nar.exb`, in the Dolgan language. Note that the tier with the ID `"ts"` establishes stretches T1–T6, T6–T13, T13–T17 etc and is exhaustive, but the tier `"ge"` has lapses: T1–T2 and T3–T4 are there, but T2–T3 is missing, corresponding to *(ha-)* in the tier `"ts"`. This makes the reconstitution of the correspondences more complicated than for the other cases.

4 Querying

The site imtvault.org offers various querying facilities. The site runs ElasticSearch, which can be accessed by humans through a responsive faceted search interface, or queried by machines using a well documented query language[2].

One main use case is the retrieval of examples based on strings found in the vernacular, the gloss, or the translation. This can be accomplished via a free text search for all three fields together, or via dedicated entry fields for the different lines. Next to this string-based search, examples can also be filtered by length. For syntactic research, for instance, more than 3 words could be required in order to arrive at any meaningful conclusions regarding syntax.

All glosses in ALLCAPS are seen as grammatical categories and matched against the Leipzig Glossing Rules (Comrie et al., 2008). The Leipzig Glossing Rules area a standardized set of common abbreviations, such as ACC(usative) or FUT(ure). Additionally, any lists of abbreviations contained in a LangSci book or a Glossa article are also made available. The categories are taken as strings, at face value. No efforts are made to match them to an ontology or to merge/reconcile/disambiguate them. It is up to the reader to interpret whether the IRR in an example would indeed match the reader's preferred definition of 'irrealis' for instance.

These querying facilities are basic and work on the content already available in the original dataset. Further querying possibilities are available via various enrichment procedures, which draw information from other datasets, link it, and make it available (Section 5). Figure 7 shows a complex query for the concept "vehicle", the cat-

[2] https://www.elastic.co/guide/en/elasticsearch/reference/6.8/full-text-queries.html

```
<tier id="ts" speaker="AnKA" category="ts" type="a" display-name="ts" >
  <event start="T1" end="T6">Bi:r (ha-) e bi:r hajin. </event>
  <event start="T6" end="T13">((LAUGH)) Bejebit balokka hild'a:ččibit každij den' otto ke. </event>
  <event start="T13" end="T17">Elbek bagaji ogo bǔ̈ola:ččibit. </event>
  <event start="T18" end="T27">Onton klassnij bagaji bǔ̈ola:ččï d'îe stroittammit etibit bǔ̈ mahinan ((LAUGH)) onton ke.</event>
  <event start="T28" end="T37">O ol d'îe maspitin ubatan ke:spippit, ubatan ke:spittere, onton. </event>
  <event start="T37" end="T42">Onton ke, ribaktartan balik kördönö:ččübüt. </event>
  <event start="T42" end="T49">Ribaktartan balik kördönö:ččübüt šašlik onosto:ččubut kastjordanan baran. </event>
  <event start="T50" end="T61">Hild'a:ččibit, palatka egelsteččibit iti Diana palatka egelste:ččï, onno onn'o:ččubut palatka ihiger. </event>
  <event start="T61" end="T62">Elete. </event>
</tier>
<tier id="tx" speaker="AnKA" category="tx" type="t" display-name="tx" >
</tier>
<tier id="mb" speaker="AnKA" category="mb" type="a" display-name="mb" >
</tier>
<tier id="mp" speaker="AnKA" category="mp" type="a" display-name="mp" >
</tier>
<tier id="ge" speaker="AnKA" category="ge" type="a" display-name="ge" >
  <event start="T1" end="T2">one</event>
  <event start="T3" end="T4">eh</event>
  <event start="T4" end="T5">one</event>
  <event start="T5" end="T6">summer.[NOM]</event>
```

Figure 5: Excerpt of the file `AnKA_2009_Story_nar.exb`

egory "past" and the language family "Atlantic-Congo". This query returns 4 hits in three languages (Fwe, Limbum, Mossi) from three different publications. While the information about the category "past" is present in the source files, information about concepts and languoids has to be added via enrichment procedures, discussed in the Section 5. Next to the HTML view given in Figure 7, the knowledge base can also be queried via the normal ElasticSearch API.

5 Enrichment

5.1 Languoids

A prime information for examples is the object language. What language is this? For examples from LangSci articles, this information is often available in the line immediately preceding the example (Figure 3).

The information that "Acehnese" is a string of relevance in (3) can be gleaned from its positional information (above the `\gll`) and from its being enclosed in `\ili`, which signal terms to be added to the language index.

"Acehnese" is then sent to a lookup service, which returns the glottocode, in this case `achi1257`.

For Glossa, the language can be retrieved from either the article title or the keywords given in the metadata. (4) shows an example from Dinka (title: "On the nature of adjectives: evidence from Dinka", keywords "Dinka, adjectives, property concepts, lexical categories, non-concatenative morphology").

While this approach yields a sizable number of linkings, false positives are also reported, for instance if, within the Dinka article, another related language is discussed, or if the source line just before an example contains a language name for unrelated reasons. Figure 6 shows the 1611 languoids for which at least one example is available.

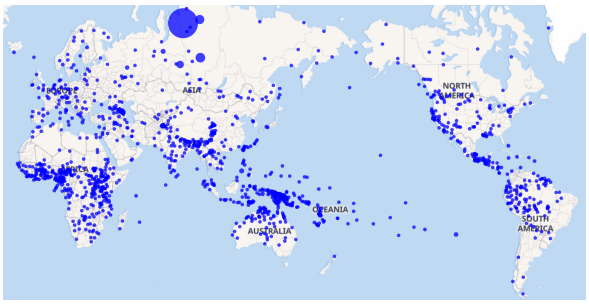

Figure 6: Provenance of examples in IMTVault. Every dot represents a languoid. Size corresponds to amount of examples, currently dominated by the Siberian INEL corpora.

5.2 Countries and macroareas

While it is nice to know that there are 5 examples tagged for "Acehnese" on IMTVault, often typologists are interested in languages from a particular region or of a particular language family. In order to accommodate these queries, the relevant country information (name and ISO 3166-2 code) is pulled from Glottolog (Hammarström et al., 2024).[3] In the case of Acehnese, this is "Indonesia"/"ID". For languages spoken in more than one country, several values are possible. These countries are then mapped to so-called linguistic macroareas, of which there are 6 (https://glottolog.org/parameters/macroarea).

[3] Next to the country information, the geographical point coordinates are also retrieved from Glottolog.

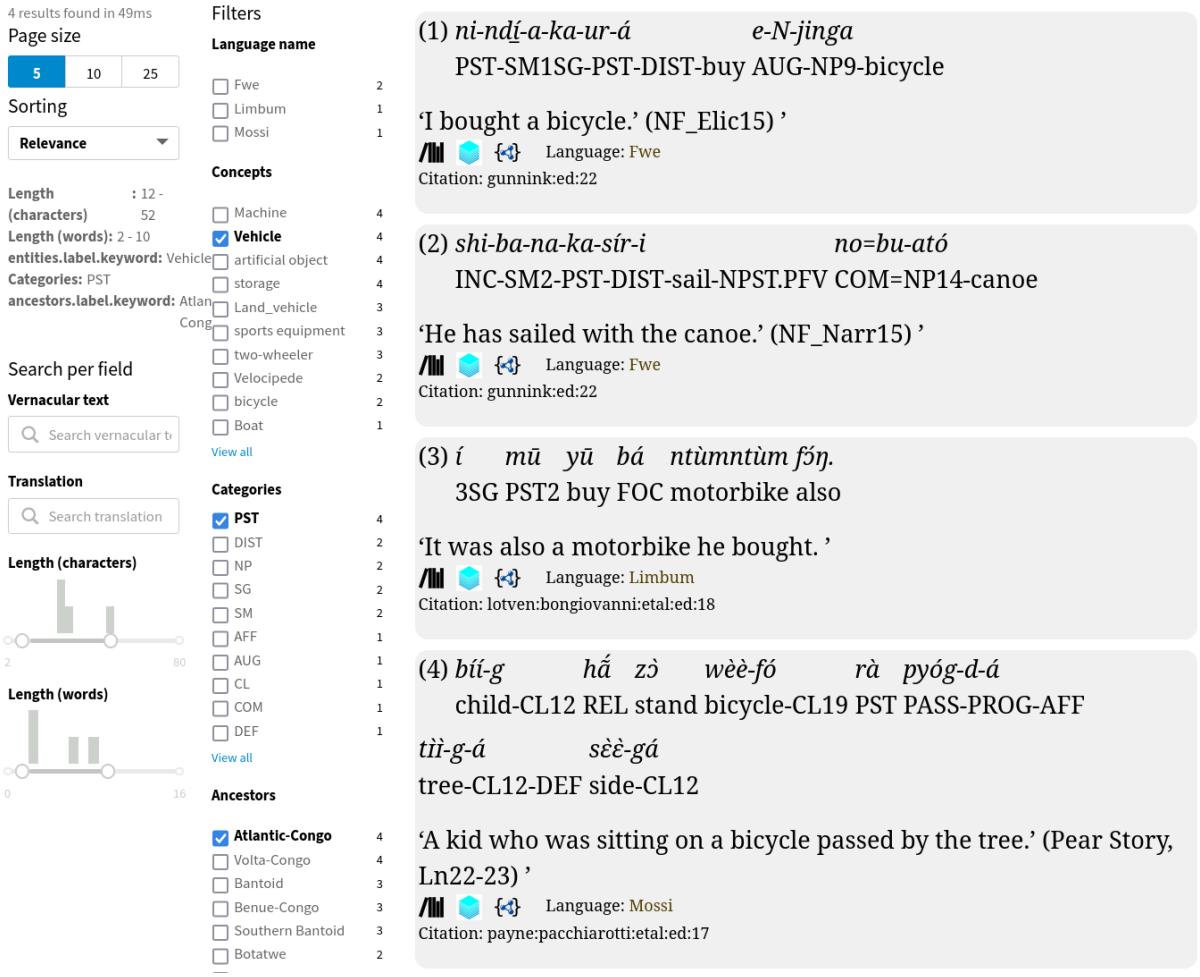

Figure 7: Complex query involving concept, category, and language family.

5.3 (Sub)families

This setup allows queries per country or per macroarea. In order to allow queries for language families and subfamilies, all nodes above a given languoid are retrieved and stored per languoid. This is the transitive closure of the mother-child relation in the genealogical language tree, which optimizes for speed of lookup with the trade-off of higher requirements for space. For "Acehnese/achi1257", the following additional ancestor languoids are stored:

(2) Aceh-Chamic (cham1327),
Malayo-Chamic (mala1554),
Malayo-Polynesian (mala1545),
Austronesian (aust1307)

This allows for selecting an arbitrary node in the genealogical tree and get all examples from languages which are part of that particular (sub)family.

5.4 Concepts

Anthropologists or oral historians are often interested in texts dealing with particular topics, such as birth and death, coming of age, or various aspects of material culture (Nordhoff, 2020a). It is of course true that for the languages at hand, we do not have the NLP tools available to do the relevant content analysis. But since we have translations, and the translations should faithfully render the object language in English (or another language of wider communication), we can use the tools developed for the larger languages to get insights into the concepts covered. We will first describe the general approach before we discuss some shortcomings we incur.

For every example, we used the GROBID-NER Named Entity Extraction.[4] The main advantage of GROBID is that it uses Wikidata-IDs for the concepts retrieved, which allow for the integration into a larger ontology (see Section 5.2).

[4] https://github.com/kermitt2/grobid

For instance, for the Kalamang example (3)

(3) warkin se laur et se pouk
tide iam rising_tide canoe iam float
'It's high tide; the canoes float.'
(Visser, 2022)

the following concepts were retrieved (with Wikidata ID):

(4) a. tide (Q23384)
b. canoe (Q171529)
c. buoyancy (Q6497624).

This would not have been possible based on the Kalamang words *warkin*, *laur* or *pouk*, but the English translation affords this inference.

Given that we now have the Wikidata IDs, we can use the relations `instance_of` (p31) and `subclass_of` (p279) to find more general concepts, traversing the tree to the root or a cut-off point (see Section 5.2). This is similar to the approach we took for languoids to arrive at (sub)families. For the case at hand, this yields the additional concepts:

(5) a. rowing equipment, Q43399738,
b. rowing boat, Q1195684,
c. Boat, Q35872,
d. watercraft, Q1229765,
e. sports equipment, Q768186,
f. Machine, Q11019,
g. Vehicle, Q42889,
h. artificial object, Q16686448,
i. floating_object, Q50380212,
j. storage, Q9158768,
k. fluid flow, Q28195494,
l. active motion, Q17988854,
m. phenomenon, Q16722960,

Linking the examples to these additional concepts allows for queries like "give me all examples relating to watercraft", which would return examples about canoes, but also yachts, dingis, dugouts, sailing ships etc. This is obviously useful.

This approach is, however, not without its problems. It is dubious, for instance, whether the Kalamang people see a canoe as a "sports equipment" (Q768186). Rather, the fact that a canoe is seen as related to leisure is due to the Western world view imbued into Wikidata. Other instances of this involve "witchcraft" (Q259745) to be "magic" (Q81741), "occultism" (Q178934) and then "pseudoscience" (Q483677); or "demons" (Q177413) being "fallen angels" (Q581450) and therefore a subclass of "angel in Judaism" (Q690175) "angel in Christianity" (Q10822464) and "angel in Islam" (Q1266031). These classifications assume a Western worldview, which does not necessarily reflect the content and semantic entailments of the context in which the utterance was produced.

A similar problem can already be seen one step earlier, in the named entity recognition. An utterance involving "Nevermind" is happily recognized as referring to the Nirvana album of the same name (Q17444), the affirmative or dubitative particle *mhm* is taken to refer to the Mill Hill Missionaries (Q119018), and *wasn't* is linked to the WASN radio station (Q7946755). This is often due to the fact that the utterances are very short, and hence only little context is available. The recognition algorithm does its best guess.

More problematic are instances of clearly biased views. For instance, the string "hoe" should be linked to the agricultural tool of the same name (Q131154), but is rather consistently misrecognized as "female sex worker" (Q107722369), based on the homographic slur.

While Western bias skews the representiations, non-Western annotations can also affect the usefulness. For instance, the concept Q7802 "bread" is a subclass of Q5004791 "bánh (Vietnamese term for a wide variety of prepared foods)". While the inclusion of non-Western food ontologies and conceptualization is in principle welcome, providing a whole array of different ontologies wil quickly overwhelm the interface.

The same is true for very detailed ontologies for goods and services, like

(6) a. "field crop and vegetable growers", Q108290536
b. "market gardeners and crop growers", Q108289653
c. "market-oriented skilled agricultural workers" Q108289043
d. "skilled_agricultural,_forestry_and _fishery_workers" Q108288352

This level of granularity is unlikely to be useful for the intended audience. At the same time, both "farmer" and "fisher" are useful concepts, as is the aggregation into "worker". Somehow related,

the integration of various specialist ontologies into Wikidata yields concepts like Q26902962 "products of manufacturing industries by OKPD and CPA 2002 (D), OKPD2 AND CPA 2008 (C)" which are unlikely to ever be queried.

Finally, Wikidata has some upper ontology, where the upper concepts can clutter the search space.

(7) gemstone (Q83437) < mineral (Q7946) < solid matter (Q11438) < matter (Q35758) < physical substance (Q28732711) < concrete object (Q4406616) < object (Q488383)

It is very unlikely that users will formulate queries about "physical substance", but excluding "matter".

So, the raw list of concepts recognized, augmented by the concepts added via Wikidata has to undergo some pruning for a) (Western) misextractions, b) too granular ontologies, and c) upper ontologies. We have manually compiled a list of 1 200 concepts which we remove, but this list is far from final.

One could of course think about other ontologies, which are more constrained than Wikidata and have stronger curation. The problems of too high granularity and upper ontology concepts will, however, still have to be addressed even when using a different ontology.

6 Comparison

IMTVault is not the only aggregator for information about lesser-resourced languages. The aggregators can be divided into aggregators for resource bundles above the sentence level (documents, corpora) on the one hand and aggregators on the sentence level on the other. For the sentence level, we can mention ODIN, for the document/corpus level, we can mention OLAC, Pangloss, and VLO.

6.1 ODIN

ODIN was started in the early 2000s (Lewis, 2006) with the aim to provide links to PDFs available online containing interlinear glossed text. There used to be a site online, but this seems to be down at the time of writing. It is possible to get access to the ODIN corpus in the XIGT format (Goodman et al., 2015) on request.

The corpus has CC-licence. It is unclear how examples with unclear license situation culled from the internet in the early 2000s can end up with a CC licence, though.

The files contained in the corpus are available as XML. There is no provenance or license data in the files, and the data quality is not convincing. A randomly drawn set of examples showed encoding errors, mix-up of data and metadata, and examples which are not interlinear text at all (Figure 8).

6.2 OLAC

OLAC is the metadata service run by the Open Language Archives Community [5]. At the time of writing the platform is undergoing a major overhaul, where Author 2 is a leading developer. The search is currently in beta status, and it aggregates more 467,000 records (text, audio, video) from 64 data providers over the OAI-PMH protocol. Records exist for more than 4,300 languages and can be filtered attributes such as language, media type, linguistic type, linguistice field, and provenance[6]. Records are linked back to the original source. The lesser-resourced language with the most records (2,905) is Southern Jinghpaw. It is not possible to search for language families or strings/concepts within a document.

6.3 Pangloss

In its own words, the "The Pangloss collection offers, in free access, linguistic audio documents, with a specialization in rare or less-studied languages." (https://pangloss.cnrs.fr) Languoid information is available as strings in French, e.g. "Inuktitut_(dialecte_du_Nunavik)". It is thus difficult to query/match this information.

The focus is thus on audio, but some of the audio documents have an XML representation for the interlinear text (Figure 9). It is possible to filter the resources on whether they have any annotation (e.g. translation), but it is not possible to specify that one is interested only in resources which do have interlinear data. Filtering on the sub-text level (i.e. sentences) is not possible either. To be fair, the main aim of Pangloss is to provide audio, with interlinear data as a kind of by-product, so the lack of querying facilities for this cannot really be held against them.

[5] http://language-archives.org
[6] https://search.language-archives.org

```
21    <tier id="n" type="odin" alignment="c" state="normalized">
22      <item id="n1" alignment="c1" line="91" tag="L+CR">    ¼ ¯ VWLJD Mukel&lt;, [nW P OHVN¯</item>
23      <item id="n2" alignment="c2" line="94" tag="G">open the gate dog QP 1SG.NOM to.him</item>
24      <item id="n3" alignment="c3" line="95" tag="T">&gt;open the gate, dog&lt;, cried I to him.</item>
25    </tier>
```

Figure 8: A randomly drawn XML file from the ODIN corpus. This file states that it is about Welsh Romani. Even in the block called "normalized", there are clear encoding errors (line 22)"

```
−<S id="S1">
   <AUDIO start="0.0" end="56.88"/>
   <FORM>amo bari maira mutime di bio</FORM>
   <TRANSL xml:lang="pm">'Ai be dina duahia korea'</TRANSL>
   <TRANSL xml:lang="en"> 'We count the time with coconut leaves' </TRANSL>
  −<W>
    −<M>
       <FORM>amo</FORM>
       <TRANSL xml:lang="en">1PL</TRANSL>
     </M>
   </W>
  −<W>
    −<M>
       <FORM>bari</FORM>
       <TRANSL xml:lang="en">day.ABST</TRANSL>
     </M>
   </W>
  −<W>
    −<M>
       <FORM>maira</FORM>
       <TRANSL xml:lang="en">time</TRANSL>
     </M>
   </W>
  −<W>
    −<M>
       <FORM>muti</FORM>
       <TRANSL xml:lang="en">count</TRANSL>
     </M>
    −<M>
       <FORM>-me</FORM>
     </M>
   </W>
  −<W>
    −<M>
       <FORM>di</FORM>
       <TRANSL xml:lang="en">coconut.ABST</TRANSL>
     </M>
   </W>
 </S>
```

Figure 9: XML format used by Pangloss. In this file, interlinear morpheme translation is available

6.4 VLO

The CLARIN Virtual Language Observatory (VLO, https://vlo.clarin.eu/search) lists pointers to 532k resources, which are varied in nature and include text, corpora, audio. A resource here is a document, that typically consists of various utterances/sentences. It is possible to filter on language, data type, availability and more. VLO claims to hold data for over 5k languages. Just as with Pangloss, the language information is encoded as a string, not as an ISO-639-3 code or glottocode.[7] There are records on "Russian" (1002), "Russisch" (21), "Russian Language" (2) as well as "Old Russian" (25) and "OldRussian" (24). It is not possible to filter on geographical area or language family, which makes the amount of resources rather overwhelming.

VLO uses Lucene in the backend, allowing for advanced complex queries. The following query excludes all major European languages, as well as the Europeana collections, which mainly hold newspaper articles and the like of European languages

(8) https://vlo.clarin.eu/search/?15&fq=licenseType:PUB&fq=resourceClass:text&fqType=licenseType:or&fqType=resourceClass:or&q=NOT+language:English+AND+NOT+language:German+AND+NOT+language:Unspeficied+AND+NOT+language:Bulgarian++AND+NOT+language:Slovenian+AND+NOT+language:Latin+AND+NOT+language:French++AND+NOT+language:Italian++AND+NOT+collection:Europeana*

This query returns 10 114 hits. Of these, 5413 are from the "OAI frontend" and have no metadata about language.

Drilling down, it turns out that 2 750 of the 4 701 remaining hits relating to lesser-resourced languages actually come from the *COllections de COrpus Oraux Numériques* ('Collection of digital oral corpora'), which in turn sources it from Pangloss (see Section 6.3).

7 Legal aspects

IMTVault is hosted in Germany, part of the EU.[8] We only use data which are available under a CC-licence, but the legal status of IGT data such as 'The man is eating the fish.' in Example (1) above or Its high tide; the canoes float. in (3) is interesting. In continental Europe a text is a copyrightable work only if some creativity is involved. A sentence like 'The man is eating the fish' is clearly not creative, so the question arises to what extent it would be copyrightable in the first place. Furthermore, it is the expression which is copyrighted, not the facts contained therein. Mapping the morpheme *-x̂* to the meaning 'singular' is a factual assertion, and as such not copyrightable.

Even if we assume for the sake of argument that 'The man is eating the fish' fell under copy-

[7] At least in the interface. The backend seems to store ISO 639-3 where available

[8] A reviewer wonders whether examples would fall under Fair Use. Fair Use is a US concept which has no clear counterpart in Europe.

right law, the question is who the copyright holder would be.

The CC-BY licence points to

- Ellen Woolford. 2017. Mainland Scandinavian object shift and the puzzling ergative pattern in Aleut. In Laura R. Bailey and Michelle Sheehan, editors, *Order and structure in syntax I: Word order and syntactic structure*, page 117133. Language Science Press, Berlin

which gives the sentence, but adds the source "(Boyle 2000: 3 (6a) from Bergsland 1969: 27)". "Bergsland (1969)" resolves to

- Knut Bergsland. 1969. A problem of transformation in Aleut. *Word*, 25(1–3):2438

In that article, the author states

> In 1952 at Atka in the central Aleutians I had a number of English sentences translated into Aleut by one of my informants, a former G.I. who was a perfect bilingual (the Aleut material obtained in this way was checked with his 70-year-old father).

The journal WORD is now owned by Taylor and Francis, who assert their copyright. It may be the case that Knut Bergsland transferred his copyright to T&F, but since he is not the original creator of the sentence *Tayaĝu-x̂ qa-x̂ qa-ku-x̂.*, he does not own any rights to this sentence in the first place, and hence, they could not be transferred to T&F (to the extent that it is copyrightable at all).

This is only one example of 200k, but it shows that the aggregation of examples comes with its own kinds of legal problems. These problems may, however, be less serious than could be feared, at least for trivial examples.

There are of course creative narratives, myths, and songs which clearly meet the threshold required by continental copyright law. This is the case for the examples provided by INEL and OTC for instance, and these contain clear creator information.

This being said, the existence of a copyright framework in the EU jurisdiction will never exempt the individual researcher from making their own ethical evaluation of the circumstances under which a particular example can or cannot be used or distributed.

8 Outlook

IMTVault has grown from 40k examples to 177k examples. Further data providers have been added and more facets for querying have been provided. Given the legal analysis presented above, an inclusion of the examples found for instance in the 10 000 books included in the DReaM corpus (Virk et al., 2020) should be possible. It can also not be excluded that publishers will actually be happy to find their data in IMTVault, as a way to channel readers towards their publication and generate traffic and revenue.

Named Entity Extraction and linking do currently work, but a comparison of different algorithms and ontologies might lead to significant improvements here, both in terms of precision/recall as well as in terms of cultural appropriateness. A systematic evaluation is out-of-scope for this paper, but will be covered in future research.

Limitations

IMTVault is an aggregation project and relies on the data providers for accuracy. It cannot be assumed that all examples tagged for a given language use the same orthography, the same morphosyntactic abbreviations, or even the same morphosyntactic analysis for that matter.

IMTVault has written represtations as its stated scope. For many research questions, access to audio (e.g. for intonation) or video (e.g. for interaction, gesture, gaze) is necessary. These questions cannot be addressed with the data made available via IMTVault.

While it is possible to provide links at the document level, deep links to the exact position where a given example is found are currently not possible. For corpora/longer texts, this would allow to check e.g. for information structure effects. For typological treatises, it would allow for the appreciation of the argumentative context in which a given example is used, and what pecularities have to be observed for examples of this kind. While mistakes should probably cancel each other out in quantitative analyses, qualitative analyses should not be based on examples retrieved from IMTVault. Rather, researchers should go back to the original publications (which are all freely available) and familiarize themselves with the surrounding context.

IMTVault is not a treebank. The current interface allows for the combination of various facets, but these are all on the level of sentence/utterance.

It is possible to ask for examples featuring "animal" and "plural", but it is not possible to require that it has to be the animals which have to be plural. 'The dog eats bones' would for instance meet the former criterion but fail the latter.

The current provenance information provide metadata on the bibliographical level, i.e. the authors of scientific books and articles. There is currently no principled way to signal the authorship of particular speakers.

References

Knut Bergsland. 1969. A problem of transformation in Aleut. *Word*, 25(1–3):2438.

John Boyle. 2000. The Aleut effect: Competition at TP. In Mary Andronis, Christopher Ball, Heidi Elston, and Sylvain Neuvel, editors, *Proceedings of CLS 37*, page 221238. Chicago Linguistics Society, Chicago.

Christian Chiarcos and Maxim Ionov. 2019. Ligt: An LLOD-native vocabulary for representing interlinear glossed text as RDF. In *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.

Christian Chiarcos, Maxim Ionov, Monika Rind-Pawlowski, Christian Fäth, Jesse Wichers Schreur, and Irina Nevskaya. 2017. LLODifying linguistic glosses. In Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Proceedings of Language, Data and Knowledge (LDK-2017), Galway, Ireland, June 2017*, number 10318 in Lecture Notes in Artificial Intelligence. Springer, Cham.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.

Sebastian Drude. 2003. *Advanced Glossing: A language documentation format and its implementation with Shoebox*. International Workshop on Resources and Tools in Field Linguistics.

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.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201.

Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024. GlossLM: A massively multilingual corpus and pretrained model for interlinear glossed text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12267–12286, Miami, Florida, USA. Association for Computational Linguistics.

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.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. *Glottolog 5.1*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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 62826293. ACL.

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

Sebastian Nordhoff. 2020a. 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. European Language Resources Association.

Sebastian Nordhoff. 2020b. Modelling and annotating interlinear glossed text from 280 different endangered languages as Linked Data with LIGT. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 93–104, Barcelona, Spain. Association for Computational Linguistics.

Sebastian Nordhoff, Christian Döhler, and Mandana Seyfeddinipur. 2024. Open Text Collections as a resource for doing NLP with Eurasian languages. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 18–23, Torino, Italia. ELRA and ICCL.

Sebastian Nordhoff and Thomas Krämer. 2022. IMT-Vault: Extracting and enriching low-resource language interlinear glossed text from grammatical descriptions and typological survey articles. In *Proceedings of The 13th Language Resources and Evaluation Conference*, Marseille, France.

Shu Okabe and François Yvon. 2023. Towards multilingual interlinear morphological glossing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5958–5971, Singapore. Association for Computational Linguistics.

Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg, and Søren Wichmann. 2020. The DReaM corpus: A multilingual annotated corpus of grammars for the world's languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 878–884, Marseille, France. European Language Resources Association.

Eline Visser. 2022. *A grammar of Kalamang*. Number 4 in Comprehensive Grammar Library. Language Science Press, Berlin.

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.

Ellen Woolford. 2017. Mainland Scandinavian object shift and the puzzling ergative pattern in Aleut. In Laura R. Bailey and Michelle Sheehan, editors, *Order and structure in syntax I: Word order and syntactic structure*, page 117133. Language Science Press, Berlin.