Open Text Collections as a Resource for Doing NLP with Eurasian Languages

Sebastian Nordhoff, Christian Döhler, Mandana Seyfeddinipur
Berlin-Brandenburg Academy of Sciences and Humanities
nordhoff@bbaw.de, doehler@bbaw.de, seyfeddinipur@bbaw.de

Abstract

The Open Text Collections project establishes a high-quality publication channel for interlinear glossed text from endangered languages. Text collections are made available in an open interoperable format and as a more traditional book publication. The project addresses a variety of audiences, eg. community members, typological linguists, anthropologists, and NLP practitioners.

Keywords: text collection, interlinear glossed text, language resources

1. Introduction

Franz Boas established the "Boasian Trilogy" in language documentation and description (Himmelmann, 1998), consisting of a grammatical description, a dictionary, and a text collection. All three levels of description are necessary to get a comprehensive overview of a language, and more importantly they complement each other. Linguists working in any field will often find themselves going back and forth between all three components. While we have good outlets for grammars (eg. Comprehensive Grammar Library[1]) and dictionaries (eg. Dictionaria[2]), such is not the case for text collections. This means that only few of them are published, and even fewer follow the FAIR principles of findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016).

The project Open Text Collections (henceforth OTC)[3] remedies this by making high quality text collections from endangered languages available in an open interoperable format. Next to providing pdfs and/or printed books to researchers and to the language communities themselves, this setup makes the data available in CLDF format (Forkel et al., 2018) for downstream use in NLP applications.

Most reference grammars published today are the result of a language documentation project, often part of authors' dissertation projects. These grammars should be data-driven and accompanied by a corpus in order to facilitate the verification or falsification of the analysis (Mosel, 2012). While countless hours are invested into the structuring and glossing of texts, in many cases, however, these texts are not made available in a reusable way. Linguists tend to have them somewhere on their hard drive, or uploaded to an archive, but there is no generally established way of publishing them, at least not in a format which would feed further research downstream (eg. linguistic typology, corpus-based language description, or NLP). This means that these valuable results of language documentation often fail to be discovered.

OTC establishes a quality venue for publishing text collections, following the setup created by Language Science Press. The platform is community-driven and aims at being attractive to both data producers (ie. language documenters) as well as data users (ie. language communities, typologists, NLP practitioners). For data producers, the platform sets up guidelines for quality control, rigorous peer review, and top-notch publishing (pdf and print-on-demand), making sure that the time invested in a text collection will not harm job prospects. For data consumers, different output format are available to suit different needs: printed books without interlinearization for the language communities; pdfs/books with interlinearization and a search interface for typologists (prototype available at `https://imtvault.org`), and all the data in CLDF format for NLP practitioners. By making reuse easy, the research output will spread more widely, which in turn is very attractive for the data producers.

As of today, there are 5 regional boards and 45 proposed text collections. This paper showcases the platform, its motivations, and its benefits for data producers and consumers.

2. Content coverage

Text collections are an old publication format, which has its origin in history, human geography, and social anthropology. In modern linguistics, the study of texts has given rise to entire subfields, for example corpus linguistics, and it is now standard prac-

[1] `https://langsci-press.org/catalog/series/cogl`

[2] `https://dictionaria.clld.org`

[3] `https://opentextcollections.github.io/`

2nd EURALI Workshop, pages 18–23
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

tice to add a few sample texts to grammatical descriptions. In some cases, grammar authors have published collections of texts as separate monographs in book form. For example, Jeffrey Heath's descriptive trilogy of the Australian language Nunggubuyu consists of a text collection (1980), a dictionary (1982), and a grammar (1984).

But what is the difference between a text corpus and a text collection? What is the difference between an archive deposit and a text collection? A language corpus of one of the major languages is technologically way more advanced than what is feasible for low-resource languages, where, very often, there is only one researcher working on a language. Moreover, corpus linguistics aims for representativeness, for a broad coverage of different criteria: genre, spoken or written style, topic, speaker background. This sets the bar too high for a language documentation project. On the other hand, an archive collection from a documentation project generally has a focus on natural, unedited, spoken language. It includes audio-visual recordings of speech events of various genres. For the OTC project, we endorse a notion of text as "written oral literature".

Moreover, archives tend to have a kind of "Russian doll" structure (Evans and Dench, 2006, 25) with a small core of well-analysed material, a medium number of translated texts in the middle and a huge amount of raw data with no significant transcription or translation at the outside. This small core of well-analysed texts potentially falls within the scope of the OTC project, but the archives in their entirety have a much larger scope.

The OTC project is located between corpora and archive collections, and the intended output differs from both in various ways. Therefore, the project has to find its own definition of "text collection". To this end, we have defined the following criteria to gauge submissions:

Curation: The submission has made a careful selection of texts from a language (eg. from a documentation project) and provides them as a coherent whole. A text collection may be structured by variety, topic or genre. This is different from a full corpus or a deposit in a language archive, in that selectivity and content coherence are ranked higher than quantity and representativeness.

Contextualization: The submission has a prose introduction, which gives geographical, anthropological, historical and linguistic context. This includes an introduction to the speech community, the language, the recording methods, the individual narrators, etc. Contextualization should go beyond the metadata as can be found in a language archive. Such contextualisation gives full credit to the original authors (narrators/speakers) because, after all, these texts are much more than just data points. Moreover, contextualization is demanded by researchers from many fields, for example anthropology, oral history, sociolinguistics or comparative narratology.

Ethics: The submission ensures that as much input is collected from key stakeholders as possible, especially on the topics of cultural sensitivities, access control, publishing licenses, and intellectual property. In most cases, the researcher submitting a text collection to OTC will consult the language community and/or the individual speakers on these points, but in cases of legacy material this can include the heirs of the speakers, or the heirs of the collector.

Editing: The submission has adapted the source material to be understandable outside of the immediate context (time and place) of narration, and the changes applied to the original source are documented and justified. Contributors may choose to edit out false starts, pauses, self-corrections, etc., but the criteria for doing so should be stated explicitly. OTC endorses a notion of "text" that is closer to "written oral literature" than to the close transcriptions that are useful for detailed analysis of speech phenomena.

Transparency: The submission has good provenance, which includes well-structured metadata, but also links to the original recordings deposited in an archive or scans in the case of legacy material. Furthermore, all decisions and steps in the editing process are documented.

Accessibility: The text collection will be available under an open and interoperable format following the FAIR standards of findability, accessibility, interoperability, and reusability.

Glossed: The submission has been fully interlinearized and glossed, following the Leipzig Glossing Rules.

3. Social setup

OTC is based on a bottom-up, scholar-led, community-driven structure. The platform is provided by the Berlin-Brandenburg Academy of Sciences and Humanities in co-operation with the publishing structure of Language Science Press.

Interested researchers can form a regional board to cover a given area. Currently, there are 5 such areas (Africa, Caucasus, Eurasia, Papunesia, South

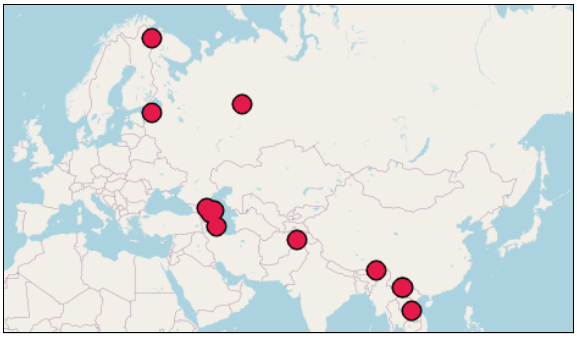
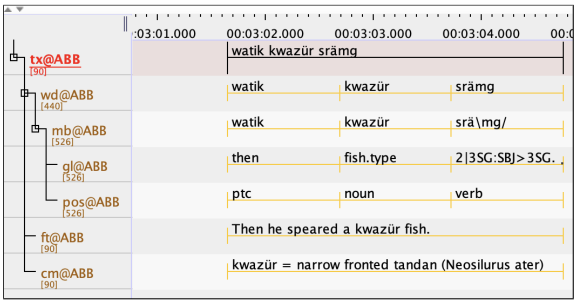


Figure 1: Expressions of interest for languages of Eurasia in OTC

Phylum	# Languages
Burmo-Qiangic	1
Indo-Aryan	2
Iranian	1
Macro-Tani	1
Nakh-Dagestanian	7
Tai-Kadai	1
Uralic	3
total	16

Table 1: Expressions of interest per phylum, Eurasia only.



	:03:01.000	00:03:02.000	00:03:03.000	00:03:04.000
tx@ABB [90]	watik kwazür srämg			
wd@ABB [440]	watik	kwazür	srämg	
mb@ABB [526]	watik	kwazür	srä\mg/	
gl@ABB [526]	then	fish.type	2|3SG:SBJ>3SG.	
pos@ABB [526]	ptc	noun	verb	
ft@ABB [90]	Then he speared a kwazür fish.			
cm@ABB [90]	kwazür = narrow fronted tandan (Neosilurus ater)			

Figure 2: The sample text in ELAN

America). The regional boards organise a rigorous peer review process that ensures high-quality results. Peer review is organised as a two-step process. An initial proposal will contain the linguistic, anthropological and philological context, accompanied by one sample file. The proposal is peer evaluated by the regional editors. If the proposed is judged positively on merits of focus, coherence, adequacy, ethics and technical quality of the sample file, the compiler is invited to submit the full collection. The full collection will undergo peer review, with one text being selected for in-depth review, while from the remaining texts, only a subset of randomly drawn sentences will be highlighted for review. This ensures both depth and breadth of reviewing without overburdening the reviewers. Text collections can number several hundred pages, which would be very time-consuming to review one by one. Consistency and adherence to guidelines will be checked computationally.

More areas, or regional boards, than the initial five can be added, but have to undergo vetting by the existing regional boards. It is envisioned, for instance, to split the rather large area of "Eurasia" into several subareas of a more manageable size. Figure 1 gives an overview of the collections which have been proposed to OTC for the languages of Eurasia.

4. Geographical coverage

For languages of Eurasia, the relevant regional boards are OTC Caucasus and OTC Eurasia. At the time of writing, there are 16 collections which have been proposed to OTC, whose affiliation is given in Table 1.

5. Technology

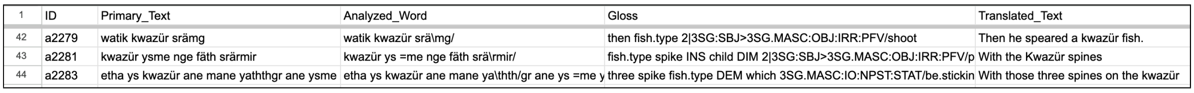
OTC can ingest several types of file formats commonly used in language documentation formats. These are converted to a common backend in CLDF format, from which a variety of output formats can be generated.

5.1. Ingestion

There are a number of different language documentation projects, which typically submit their work to one of the DELAMAN[4] archives, eg. AILLA,[5] ELAR,[6] PARADISEC[7] or TLA.[8] The most commonly used programs to produce interlinear glossed text (IGT) are ELAN and FLEx.

ELAN is a program, shown in Figure 2, which allows users to annotate multimedia on different "tiers" (Wittenburg et al., 2006). Different speakers will have different tiers, and tiers can be of different types, eg. transcription, translation, and glosses. Relations between tiers are explicit. Users have a lot of freedom about which tiers to define and what features to assign to them, leading to a vast heterogeneity of tier types (von Prince and Nordhoff, 2020; Nordhoff, 2020). ELAN uses an XML format as its backend. The library *eldpy* reads ELAN files and applies a number of heuristics to find the most probable tiers for transcription, translations, glosses. Criteria evaluated are: the name of the

[4] https://www.delaman.org
[5] https://ailla.utexas.org/
[6] https://www.elararchive.org
[7] https://www.paradisec.org.au
[8] https://archive.mpi.nl/tla



1	ID	Primary_Text	Analyzed_Word	Gloss	Translated_Text
42	a2279	watik kwazür srämg	watik kwazür srä\mg/	then fish.type 2|3SG:SBJ>3SG.MASC:OBJ:IRR:PFV/shoot	Then he speared a kwazür fish.
43	a2281	kwazür ysme nge fäth srärmir	kwazür ys =me nge fäth srä\rmir/	fish.type spike INS child DIM 2|3SG:SBJ>3SG.MASC:OBJ:IRR:PFV/p	With the Kwazür spines
44	a2283	etha ys kwazür ane mane yaththgr ane ysme	etha ys kwazür ane mane ya\thth/gr ane ys =me y	three spike fish.type DEM which 3SG.MASC:IO:NPST:STAT/be.stickin	With those three spines on the kwazür

Figure 3: The sample text in CLDF (.csv) format.

tier ('ft' is typically indicative of "free translation", 'ge' is "gloss english" etc), the relation to other tiers ("symbolic association" is either a translation or a gloss), and the language of the tier (translation tiers should pass a language detection test for English; transcription tiers should fail such a test). Based on these criteria, content is extracted and stored as the CLDF fields "Primary_Data", "Analyzed_Text", "Glosses", and "Translation" (see Figure 3).

FLEx is another program which is often used in language documentation projects. It allows the linguist to tokenize and gloss a transcribed text with the help of a lexicon. The lexicon grows as more and more texts are ingested. FLEx also uses an XML backend. The CLDF library *cldflex* (Matter, 2024) can be used to extract the relevant content and store it as CLDF. By and large, FLEx shows a lot less heterogeneity than ELAN.

tex and xlsx are other formats which are structured enough to provide import routines. The *langsci-gb4e* package for the LaTeX typesetting language is commonly used in grammar writing, and the content can easily be extracted with *linglit*, as has been shown for IMTVault (`https://imtvault.org`). These two latter formats are less prevalent than ELAN or FLEx, but still frequent enough to warrant import routines.

5.2. Backend

OTC stores the interlinear glossed text in the Cross-Linguistic Data format (CLDF, [9] (Forkel et al., 2018), Figure 3), a format which is an emerging standard for research data in linguistic typology and beyond and which can easily be ingested into CLLD (cross-linguistic linked data) applications. CLDF provides several components, of which the component "examples"[10] is the most pertinent for OTC. The relevant columns are Primary_Data, Analyzed_Text, Glosses, and Translation, complemented by a column for Glottocode,[11] and a column for comments. The CLDF format is extensible, meaning that additional columns can easily be added, but no promise is made that the content therein can be consumed.

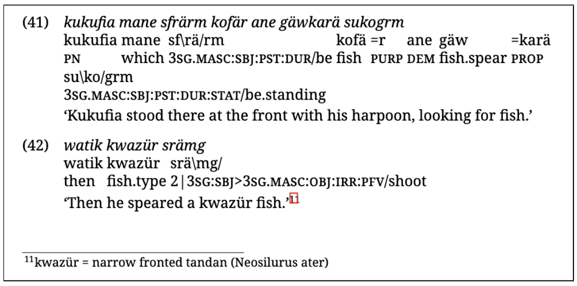
The creation and refinement of the text collection is done on GitHub, with releases being automatically archived on Zenodo[12] using the GitHub-Zendodo bridge.

5.3. Output formats

There are three main target groups for OTC content: NLP practitioners, linguists, and speaker communities. For NLP practitioners, a csv dump is made available (cf. Figure 3), next to a rendering in JSON-LD. Linguists can use the csv dump for quantitative research or an ElasticSearch HTML frontend for qualitative explorations, based on work done for IMTVault (Nordhoff and Krämer, 2022). The text is also made available as a pdf with interlinearized examples (Figure 4). Language communities finally can use the pdfs generated from the backend with a two-column layout with vernacular on the left and translation on the right (Figure 5). Both pdf formats are fed into the print-on-demand pipelines established by Language Science Press. These printed books are then available world wide via the usual distribution channels (eg Amazon, local bookstores, Verzeichnis lieferbarer Bücher etc.)

6. Use downstream

A number of recent studies have shown the usefulness of well-structured textual data for NLP approaches. Most of them focus on ways to overcome bottlenecks in the production of IGT, for example segmentation and glossing (McMillan-Major 2020, Barriga Martínez et al. 2021, Liu et al. 2021, Moeller and Hulden 2021). Two example studies of NLP approaches are explained in more detail here.



(41) *kukufia mane sfrärm kofär ane gäwkarä sukogrm*
kukufia mane sf\rä/rm kofä =r ane gäw =karä
PN which 3SG.MASC:SBJ:PST:DUR/be fish PURP DEM fish.spear PROP
su\ko/grm
3SG.MASC:SBJ:PST:DUR:STAT/be.standing
'Kukufia stood there at the front with his harpoon, looking for fish.'

(42) *watik kwazür srämg*
watik kwazür srä\mg/
then fish.type 2|3SG:SBJ>3SG.MASC:OBJ:IRR:PFV/shoot
'Then he speared a kwazür fish.'[11]

[11] kwazür = narrow fronted tandan (Neosilurus ater)

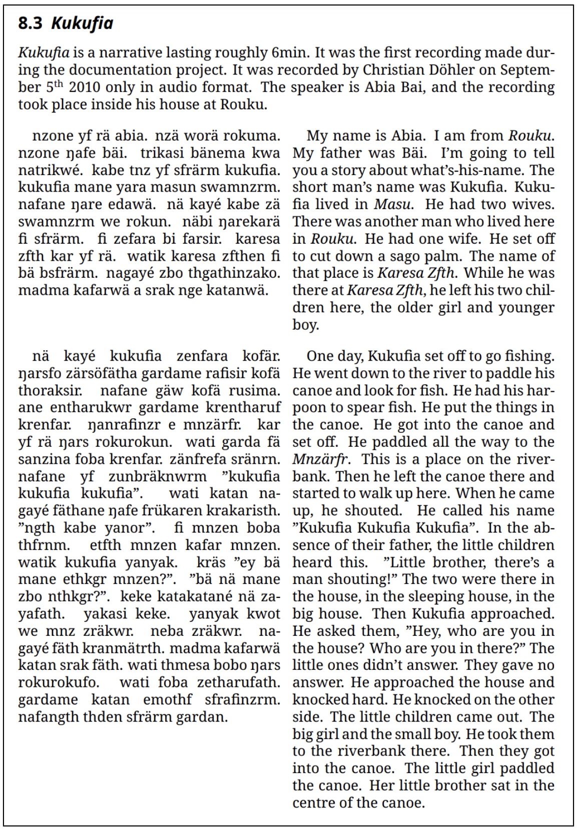
Figure 4: The sample text in "scientific" format

[9] `https://cldf.clld.org`

[10] `https://github.com/cldf/cldf/tree/master/components/examples`

[11] `https://glottolog.org`

[12] `https://zenodo.org/communities/otc/records?q=&l=list&p=1&s=10&sort=newest`



8.3 *Kukufia*

Kukufia is a narrative lasting roughly 6min. It was the first recording made during the documentation project. It was recorded by Christian Döhler on September 5th 2010 only in audio format. The speaker is Abia Bai, and the recording took place inside his house at Rouku.

nzone yf rä abia. nzä worä rokuma. nzone ŋafe bäi. trikasi bänema kwa natrikwé. kabe tnz yf sfrärm kukufia. kukufia mane yara masun swamnzrm. nafane ŋare edawä. nä kayé kabe zä swamnzrm we rokun. näbi ŋarekarä fi sfrärm. fi zefara bi farsir. karesa zfth kar yf rä. watik karesa zfthen fi bä bsfrärm. nagayé zbo thgathinzako. madma kafarwä a srak nge katanwä.

My name is Abia. I am from *Rouku*. My father was Bäi. I'm going to tell you a story about what's-his-name. The short man's name was Kukufia. Kukufia lived in *Masu*. He had two wives. There was another man who lived here in *Rouku*. He had one wife. He set off to cut down a sago palm. The name of that place is *Karesa Zfth*. While he was there at *Karesa Zfth*, he left his two children here, the older girl and younger boy.

nä kayé kukufia zenfara kofär. ŋarsfo zärsöfätha gardame rafisir kofä thoraksir. nafane gäw kofä rusima. ane entharukwr gardame krentharuf krenfar. ŋanrafinzr e mnzärfr. kar yf rä ŋars rokurokun. wati garda fä sanzina foba krenfar. zänfrefa sränrn. nafane yf zunbräknwrm ”kukufia kukufia kukufia”. wati katan nagayé fäthane ŋafe frükaren krakaristh. ”ngth kabe yanor”. fi mnzen boba thfrnm. etfth mnzen kafar mnzen. watik kukufia yanyak. kräs ”ey bä mane ethkgr mnzen?”. ”bä nä mane zbo nthkgr?”. keke katakatané nä zayafath. yakasi keke. yanyak kwot we mnz zräkwr. neba zräkwr. nagayé fäth kranmätrth. madma kafarwä katan srak fäth. wati thmesa bobo ŋars rokurokufo. wati foba zetharufath. gardame katan emothf sfrafinzrm. nafangth thden sfrärm gardan.

One day, Kukufia set off to go fishing. He went down to the river to paddle his canoe and look for fish. He had his harpoon to spear fish. He put the things in the canoe. He got into the canoe and set off. He paddled all the way to the *Mnzärfr*. This is a place on the riverbank. Then he left the canoe there and started to walk up here. When he came up, he shouted. He called his name ”Kukufia Kukufia Kukufia”. In the absence of their father, the little children heard this. ”Little brother, there's a man shouting!” The two were there in the house, in the sleeping house, in the big house. Then Kukufia approached. He asked them, ”Hey, who are you in the house? Who are you in there?” The little ones didn't answer. They gave no answer. He approached the house and knocked hard. He knocked on the other side. The little children came out. The big girl and the small boy. He took them to the riverbank there. Then they got into the canoe. The little girl paddled the canoe. Her little brother sat in the centre of the canoe.

Figure 5: The sample text in “community” format

For Kalamang, an endangered Papuan language, Tanzer et al. (2024) have tested the translation capabilities of language models versus humans by feeding them the grammatical description, an approach they call MTOB (Machine Translation from One Book), and then comparing their translations from Kalamang to English and vice versa. Their study shows that humans are more successful at present, but they also show several points for improving the models.

For Japhug, an small language of Southern China, and Tsez, a small language spoken in the Caucasus, Okabe and Yvon (2023) have experimented with Bayesian models for simultaneously segmenting utterances into words and morphemes. They have tested two models to simultaneously segment into words and morphemes: one segmenting in parallel and the other in a hierarchical manner. They show that in the unsupervised condition the hierarchical model produces higher accuracy. What's more is that the study makes a number of suggestions to improve the results, eg. by incorporating contextual word models or adding further levels of supervision like phonology.

Such examples show that NLP research, however preliminary, when applied to low-resource languages, can help both the linguists working in language documentation and description and the language communities in participating in the development of large language models, thereby, increasing the relevance of small languages and overcoming the digital divide.

7. Conclusion

The Open Text Collections project remedies the lack of recognized publication venues for text collections of under-resourced languages and thereby pushes further the efforts to make lesser-resourced language content available in digital formats. In order to overcome the digital divide, the project wants to provide existing structured data to speaker communities and academics alike, in a formats suitable for the respective groups. Furthermore, the project provides researchers the prestige they deserve (ie. a peer-reviewed book publication) for creating interlinear glossed texts. Finally, the project provides a source of data for NLP research and facilitates further typological research. There are currently 16 text collections being prepared for languages of Eurasia, and as the project grows, more data from the less-resourced languages of the continent will become available as data sources for NLP research and community purposes alike.

8. Bibliographical References

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Nicholas Evans and Alan Charles Dench. 2006. Introduction: Catching language. In Felix K Ameka, Alan Charles Dench, and Nicholas Evans, editors, *Catching Language: The Standing Challenge of Grammar Writing*, pages 1–40. Mouton de Gruyter, Berlin, New York.

Robert Forkel, Johann-Mattis List, Simon 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. *Sci Data*, 5.

Jeffrey Heath. 1980. *Nunggubuyu myths and ethnographic texts*. Australian Institute of Aboriginal Studies, Canberra.

Jeffrey Heath. 1982. *Nunggubuyu dictionary*. Australian Institute of Aboriginal Studies, Canberra.

Jeffrey Heath. 1984. *Functional grammar of Nunggubuyu*. Australian Institute of Aboriginal Studies, Canberra.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological Segmentation for Seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101, Online. Association for Computational Linguistics.

Florian Matter. 2024. Cldflex. URL: https://github.com/fmatter/cldflex.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. *Society for Computation in Linguistics*, 3:338–349.

Sarah Moeller and Mans Hulden. 2021. Integrating automated segmentation and glossing into documentary and descriptive linguistics. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 86–95. Association for Computational Linguistics.

Ulrike Mosel. 2012. Advances in the accountability of grammatical analysis and description by using regular expressions. *Language Documentation & Conservation Special Publication*, 4:235–250.

Sebastian Nordhoff. 2020. 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 and Thomas Krämer. 2022. IMTVault: 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. Joint word and morpheme segmentation with Bayesian non-parametric models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 640–654, Dubrovnik, Croatia. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. ArXiv preprint, arXiv:2309.16575.

Kilu von Prince and Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.