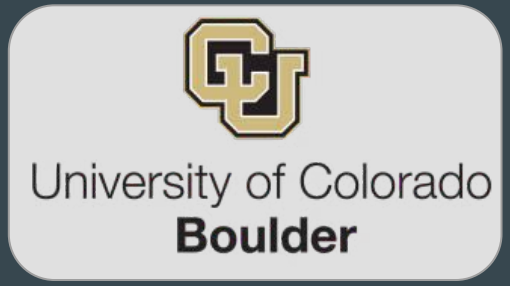
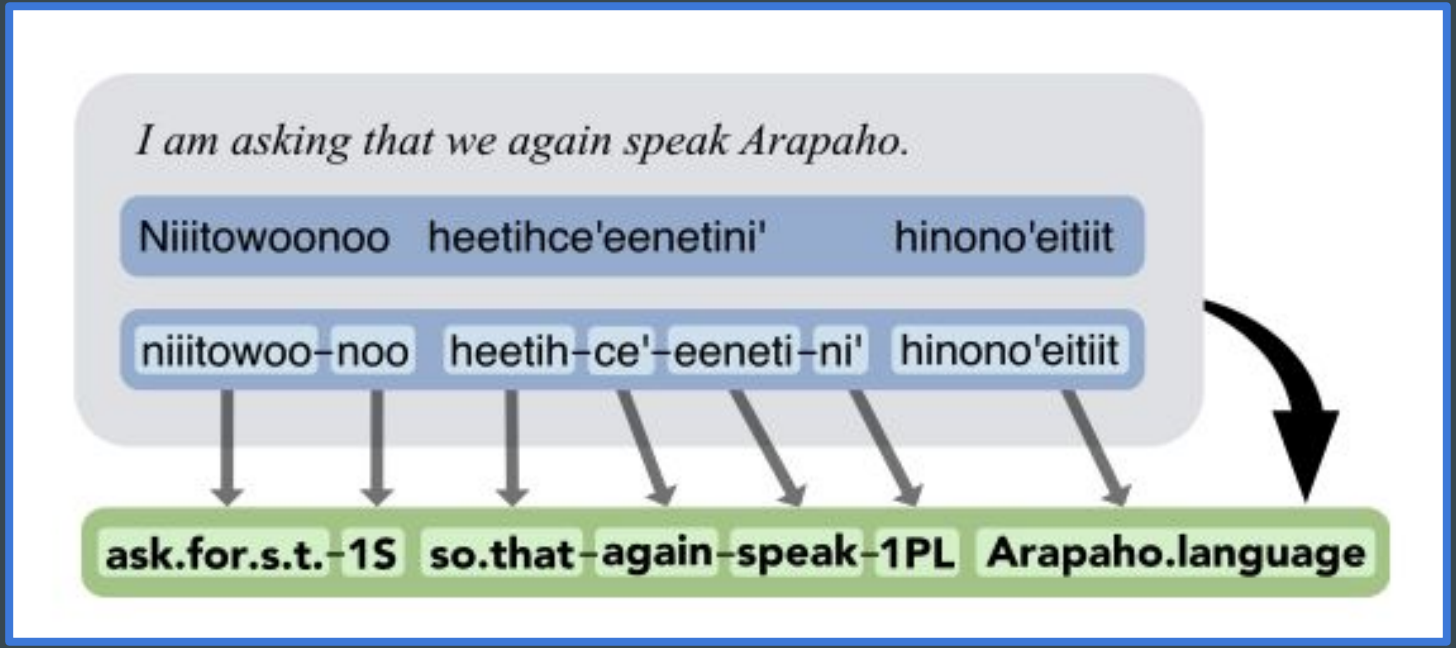


Automating Interlinear Glossing

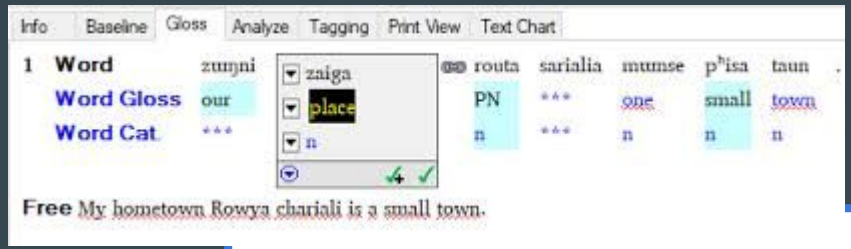
...

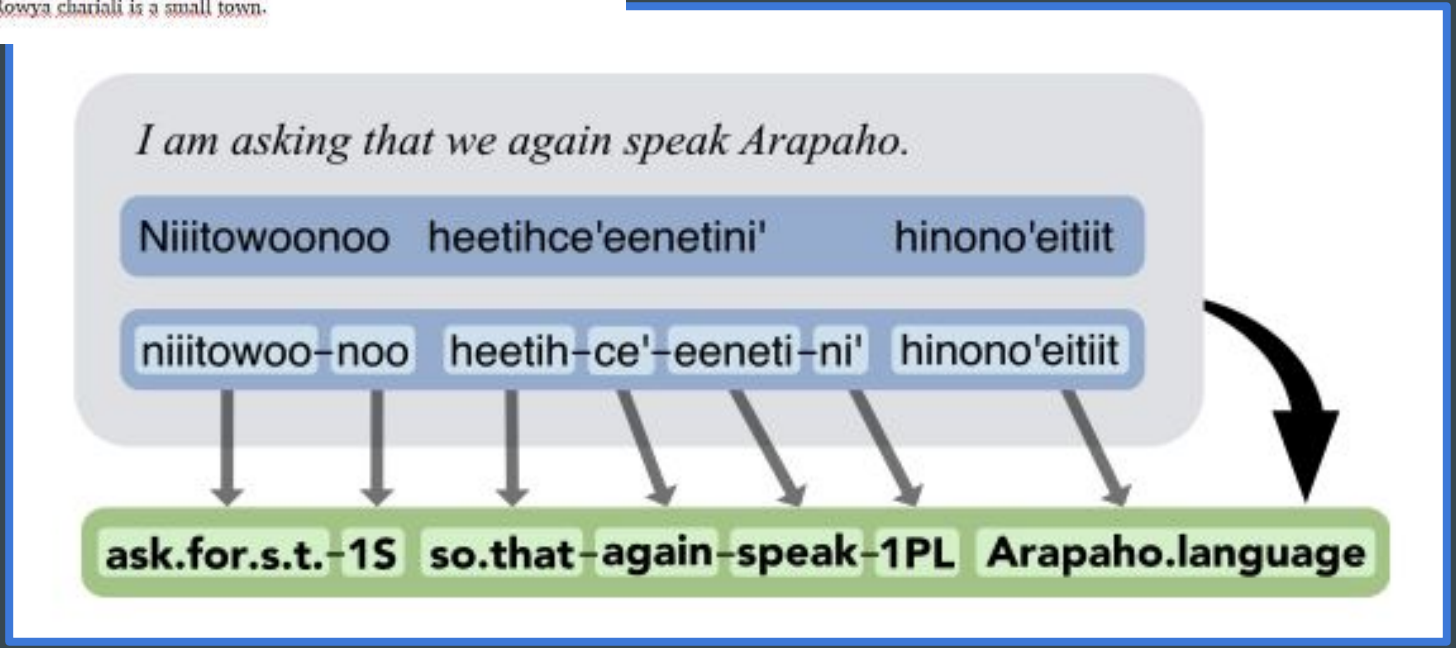
Alexis Palmer, FieldMatters, March 29, 2026

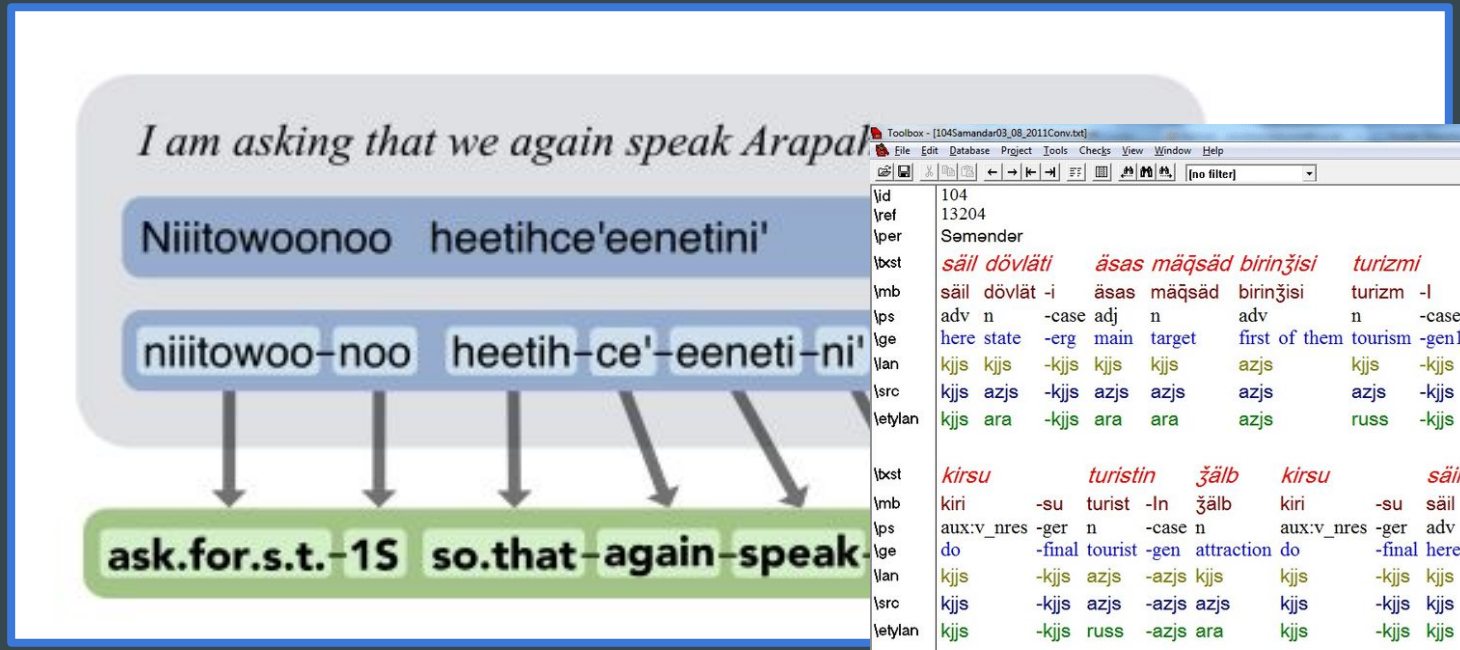


Arapaho: data from Cowell 2020

Fieldworks Language Explorer (FLEx)







Toolbox - [104Samandar03_08_2011Conv.txt]

File Edit Database Project Tools Checks View Window Help

[no filter]

\id	104								
\ref	13204								
\per	Səməndər								
\txst	*säil*	*dövläti*		*äsas*	*mäqsäd*	*birinǯisi*	*turizmi*		*inkišaf*
\mb	säil	dövlät	-i	äsas	mäqsäd	birinǯisi	turizm	-I	inkišaf
\ps	adv	n	-case	adj	n	adv	n	-case	n
\ge	here	state	-erg	main	target	first of them	tourism	-gen1	development
\lan	kjjs	kjjs	-kjjs	kjjs	kjjs	azjs	kjjs	-kjjs	kjjs
\src	kjjs	azjs	-kjjs	azjs	azjs	azjs	azjs	-kjjs	azjs
\etylan	kjjs	ara	-kjjs	ara	ara	azjs	russ	-kjjs	ara

\txst	*kirsu*		*turistin*		*ǯälb*	*kirsu*		*säil*
\mb	kiri	-su	turist	-In	ǯälb	kiri	-su	säil
\ps	aux:v_nres	-ger	n	-case	n	aux:v_nres	-ger	adv
\ge	do	-final	tourist	-gen	attraction	do	-final	here
\lan	kjjs	-kjjs	azjs	-azjs	kjjs	kjjs	-kjjs	kjjs
\src	kjjs	-kjjs	azjs	-azjs	azjs	kjjs	-kjjs	kjjs
\etylan	kjjs	-kjjs	russ	-azjs	ara	kjjs	-kjjs	kjjs

\fta Burada dövlətin əsas məqsədi birincisi turizmin inkişaf etdirməsi,

\fte **Here, the main target of the state (the government) is to develop tourism**

turistlərin cəlb etməsi buraya.

and to attract tourists here.

\id 104 \ref 104:004 4/13 Khinalug.prj

Toolbox

ELAN

BESEMAH SENTENCE:

ngibal-ngibal di pinggir ghimbe ni tadi, adinge ini tadi tekinak ngaghi jambu, katah besake.

INTONATION UNIT:	ngibal-ngibal	di	pinggir	ghimbe	ni	tadi,
MORPHEME:	ngibal-ng-ibal	di	pinggir	ghimbe	ni	tadi,
GLOSS:	RED-AV-walk	at	edge	forest	this	earlier

INTONATION UNIT:	adinge	ini tadi	tekinak	ngaghi	jambu,	
MORPHEME:	ading-e	ini tadi	te-kinak	ngaghi	jambu,	
GLOSS:	younger.brother-3	this earlier	INTR-see	with	guava	

INTONATION UNIT:	katah	besake.
MORPHEME:	katah	besak-e.
GLOSS:	INTENS	big-3

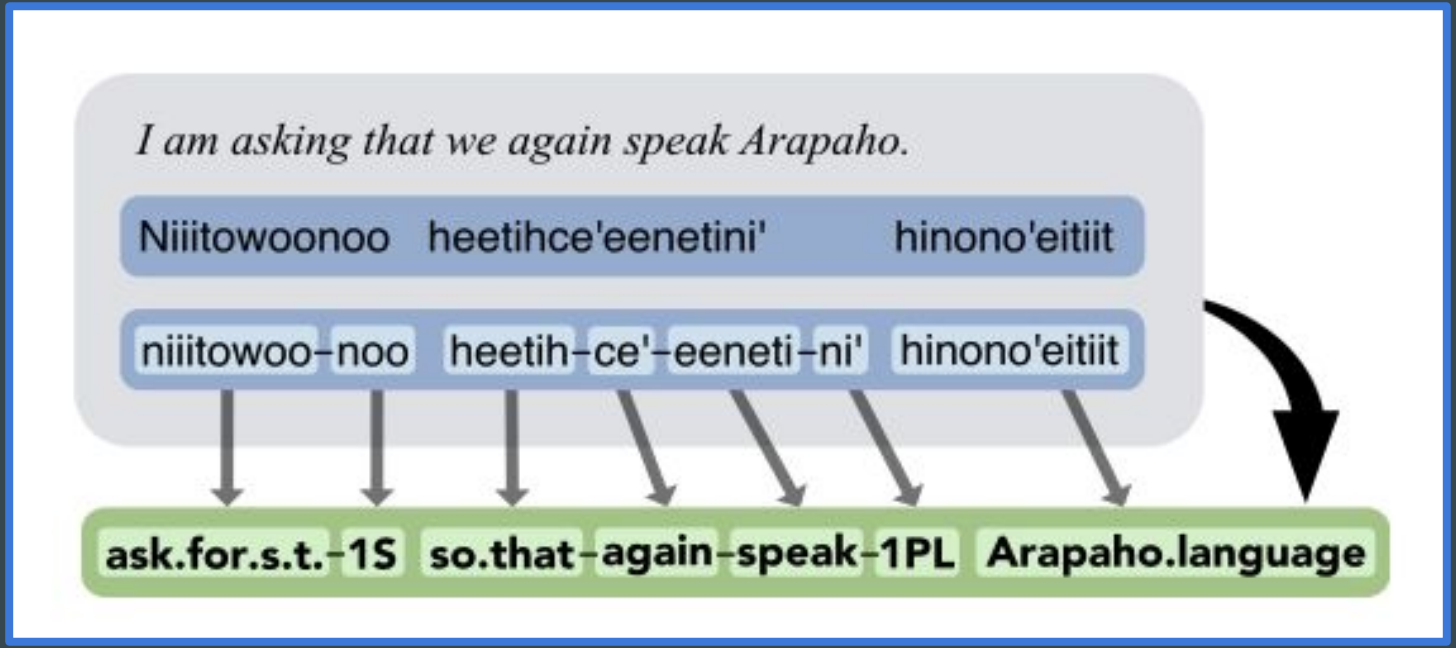
FREE TRANSLATION:

'Playing at the edge of the forest, the younger brother saw a really big guava.'

Fourth, language documentation projects typically manifest a yawning gap between the amount of material recorded and archived and the amount of data that is minimally annotated (transcribed and translated), let alone more thoroughly analyzed (e.g. morphologically segmented and glossed). As a result, the corpus size resulting from most language documentation programs is much smaller than what is required to answer many questions, severely diminishing its utility for future generations. This gap reflects the 'transcription bottleneck': an hour of recorded material can take between 40 and 100 hours to transcribe. A concerted attack on this problem needs to involve three steps:

Seifert et al. 2018, Language documentation twenty-five years on, *Language,* 94:4


ANNOTATE
TAG OR BUST
DATA
MORE DATA
Phew...
ENUNTHING.
TODO
TAG DATA!
FUEL
NOUN
NOUN
VERB
ADJ
ENTITY
NOUN
VERB
ADJ
TAGGING...



Arapaho: data from Andy Cowell's text collection


PRATA
TRAINING
DSTATE
FML
NLP
RULI
ANNOTATION
MODEL
DATA
DATA
B
BATA
NOUN
VERB
ADJ
VERB


I'M SO SMART
ANNOTATION MODEL
DATA
TRAINING
NLP
NOUN
VERB
ADJ
VERB


DATA
SOVEREIGNTY
DICTIONARY
AUDIO RECORDINGS
BIG
TECH
LANGUAGE
DATA
LANGUAGE
DATA
LANGUAGE
DATA

Historical Perspective

Recent Approaches

Centering the User

Preliminaries for automating IGT

- Data formats for IGT
  - Bow Hughes Bird 2003 - Towards a general model of interlinear text
  - Schroeter & Thieberger 2006 - EOPAS format
- EMELD - Electronic Metadata for Endangered Language Documentation
  - 2001-2006, NSF-funded project - workshops, best practices, resources, publications
  - Goal of harmonizing workflows and bringing tech on board
- TLSX - Computational Linguistics for Less-Studied Languages (2006)
- GOLD - General Ontology for Linguistic Description
  - Farrar & Langendoen 2003, 2010; Farrar & Lewis 2007
  - Goal: interoperability of IGT datasets
  - Proposal: atheoretic set of linguistic concepts
  - Idea: linguist maps into the ontology

Morpheme[C]

IRI	`http://purl.org/linguistics/gold/Morpheme`
Is Defined By	http://purl.org/linguistics/gold
Description	The smallest functioning unit in the composition of words, and the minimal distinctive unit of grammar. Morphemes are commonly classified into free forms (morphemes which can occur as separate words) adn bound forms (morphemes which cannot so occur - mainly affixes). A further distinction may be made between lexical and grammatical morphemes; the former are morphemes used for the construction of new words in a language; the latter are morphemes used to express grammatical relationships between a word and its context. [Crystal 2008: 300]
Sub Class Of	GrammarUnit[C]
Super Class Of	BoundMorpheme[C] Clitic[C] Compound[C] FreeMorpheme[C] Root[C] Stem[C] Affix[C]

AspectProperty [C]

IRI http://purl.org/linguistics/gold/AspectProperty

Is Defined By http://purl.org/linguistics/gold

Description

The term 'aspect'; designates the perspective taken on the internal temporal organization of the event, and different values of the Aspect Feature distinguish different ways of viewing the internal temporal constituency of the same event [Comrie 1976: 3ff], after [Holt 1943: 6; Bybee 2003: 157]. The 'event' is understood here as a general term covering any situation type (a state, activity, accomplishment, achievement, etc.) as expressed by the verb phrase of the construction. Unlike Tense Feature, which expresses event-external time and is deictic, Aspect Feature is event-internal and non-deictic, as it is not concerned with relating the time of the event to any other time point. [Kibort 2008e]

Sub Class Of MorphosemanticProperty [C]

Super Class Of

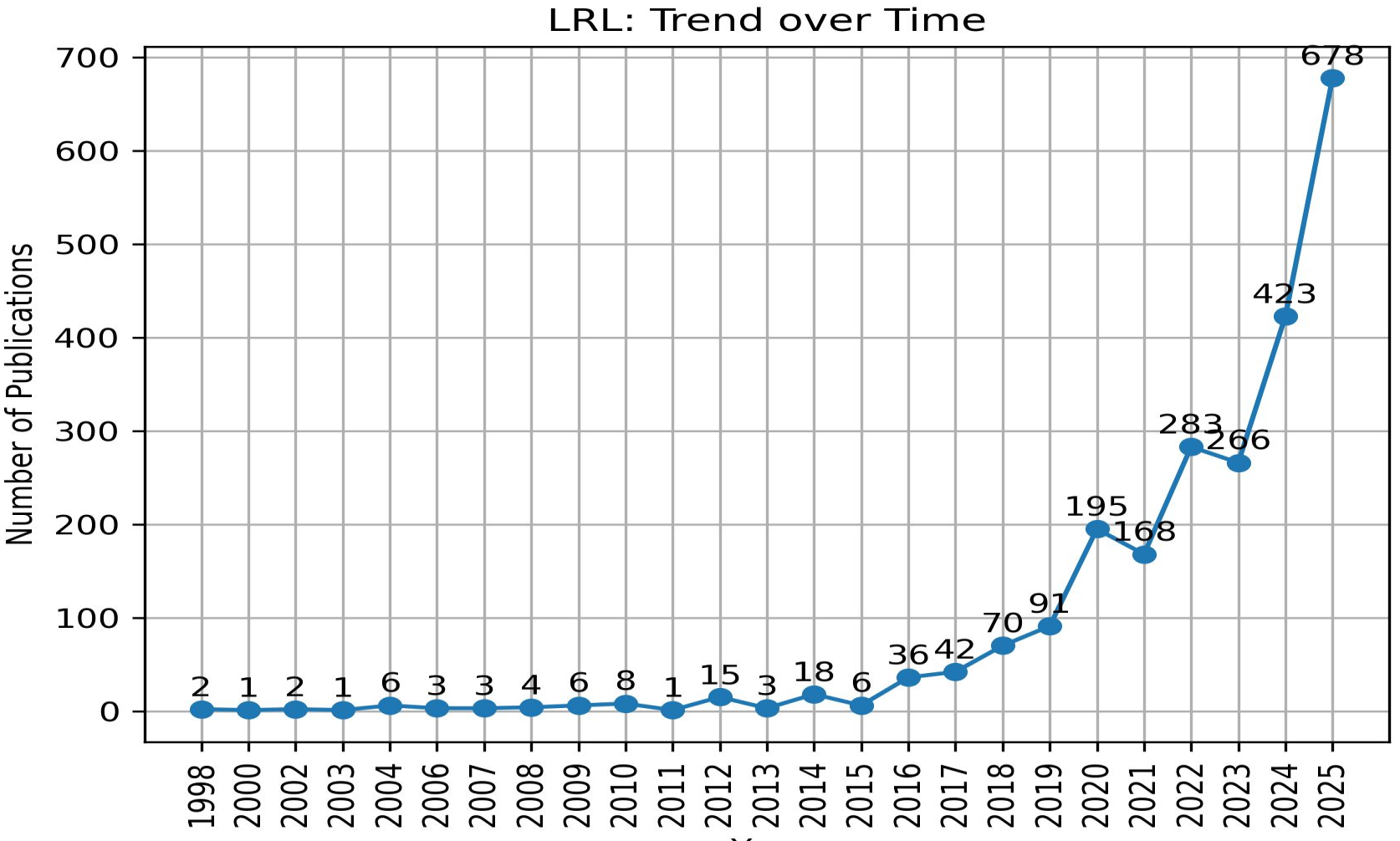
CompletiveAspect [C]
ContinuousAspect [C]
DistributiveAspect [C]
DurativeAspect [C]
FrequentiveAspect [C]
HabitualAspect [C]
ImperfectiveAspect [C]
InceptiveAspect [C]
IterativeAspect [C]
NonProgressiveAspect [C]
PerfectiveAspect [C]
PhasalAspect [C]
ProgressiveAspect [C]
SemelfactiveAspect [C]
SimultaneousAspect [C]
TerminativeAspect [C]
QuantificationalAspect [C]

Early glossing models (~2009)

- Morphological segmentation and morpheme labeling treated as entirely separate tasks
- Moon et al. 2009 investigate unsupervised segmentation, following earlier work
- Treat morpheme labeling as fine-grained POS tagging
  - Maximum entropy classifier
  - Feature-based supervised model
  - Assuming gold segmentation, performance around 79% (unigram probs: ~77.5%)
  - Palmer et al. 2009, Baldridge & Palmer 2009, Palmer 2009, Palmer et al. 2010

Publication trends over time

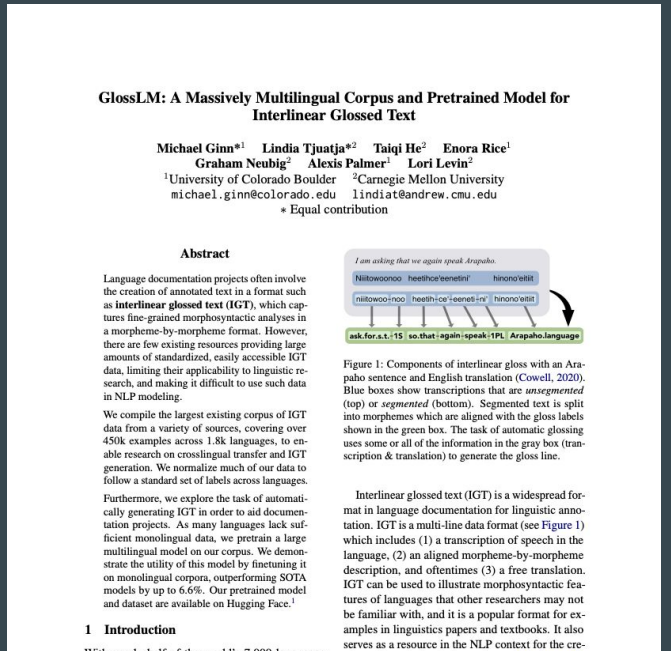
- Before 2016: few works on less-resourced or endangered languages
- 2014: first iteration of ComputEL workshop
- 2021/21: founding of SIGEL
- 2022: first iteration of AmericasNLP workshop
- 2023: SIGMORPHON shared task on glossing
- Many other research areas related to low-resource languages


LRL: Trend over Time
Number of Publications
0
100
200
300
400
500
600
700
2 1 2 1 6 3 3 4 6 8 1 15 3 18 6 36 42 70 91 195 168 283 266 423 678
1998
2000
2002
2003
2004
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025

(Ginn et al. 2024a)

GlossLM

- Large dataset of IGT (250k unique examples) spanning 2k languages
- Continued pretraining on ByT5 models to create multilingual IGT generation system
- Demonstrated benefit from transfer learning
- New SOTA on eval languages



GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text

Michael Ginn[*1] **Lindia Tjuatja**[*2] **Taiqi He**[2] **Enora Rice**[1]
Graham Neubig[2] **Alexis Palmer**[1] **Lori Levin**[2]
[1]University of Colorado Boulder [2]Carnegie Mellon University
michael.ginn@colorado.edu lindiat@andrew.cmu.edu
∗ Equal contribution

Abstract

Language documentation projects often involve the creation of annotated text in a format such as **interlinear glossed text (IGT)**, which captures fine-grained morphosyntactic analyses in a morpheme-by-morpheme format. However, there are few existing resources providing large amounts of standardized, easily accessible IGT data, limiting their applicability to linguistic research, and making it difficult to use such data in NLP modeling.

We compile the largest existing corpus of IGT data from a variety of sources, covering over 450k examples across 1.8k languages, to enable research on crosslingual transfer and IGT generation. We normalize much of our data to follow a standard set of labels across languages.

Furthermore, we explore the task of automatically generating IGT in order to aid documentation projects. As many languages lack sufficient monolingual data, we pretrain a large multilingual model on our corpus. We demonstrate the utility of this model by finetuning it on monolingual corpora, outperforming SOTA models by up to 6.6%. Our pretrained model and dataset are available on Hugging Face.[1]

Figure 1: Components of interlinear gloss with an Arapaho sentence and English translation (Cowell, 2020). Blue boxes show transcriptions that are *unsegmented* (top) or *segmented* (bottom). Segmented text is split into morphemes which are aligned with the gloss labels shown in the green box. The task of automatic glossing uses some or all of the information in the gray box (transcription & translation) to generate the gloss line.

1 Introduction

With nearly half of the world's 7,000 languages considered endangered, communities of minoritized language speakers are working to preserve and revitalize their languages (Seifart et al., 2018). These efforts often involve collection, analysis, and annotation of linguistic data. Annotated text can be used in the creation of reference materials (such as dictionaries and grammars) as well as to develop language technologies including search-

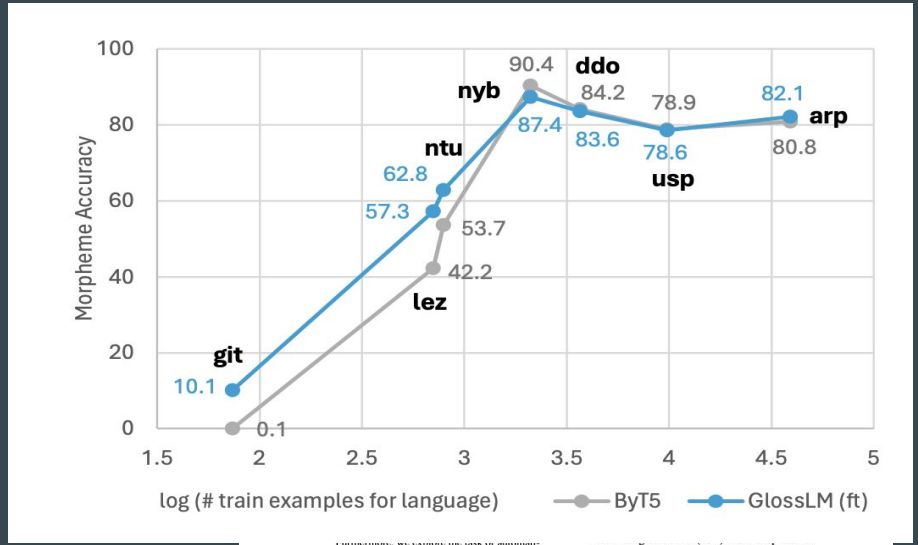
Interlinear glossed text (IGT) is a widespread format in language documentation for linguistic annotation. IGT is a multi-line data format (see Figure 1) which includes (1) a transcription of speech in the language, (2) an aligned morpheme-by-morpheme description, and oftentimes (3) a free translation. IGT can be used to illustrate morphosyntactic features of languages that other researchers may not be familiar with, and it is a popular format for examples in linguistics papers and textbooks. It also serves as a resource in the NLP context for the creation of morphological paradigms (Moeller et al., 2020), machine translation (Zhou et al., 2019), generating precision grammars (Bender et al., 2013), and other tools including POS taggers and dependency parsers (Georgi, 2016).

Compiling IGT Data Though IGT often follows a common glossing format, gloss conventions vary wildly. Furthermore, IGT data is rarely compiled into lar d corpora, often existing as scat rch papers. To address t corpus of digitize es, with over 4 We explore r

(Ginn et al. 2024a)

GlossLM

- Large dataset of IGT (250k unique examples) spanning 2k languages
- Continued pretraining on ByT5 models to create multilingual IGT generation system
- Demonstrated benefit from transfer learning
- New SOTA on eval languages



(Rice et al. 2025)

User-Centered Design for Language Documentation

- We asked 3 linguists to incorporate GlossLM into workflow (they didn't like it)
- All agreed: annotating texts from scratch would be faster and easier than correcting GlossLM outputs
- Major sticking point: lack of segmentation

Interdisciplinary Research in Conversation: A Case Study in Computational Morphology for Language Documentation

Enora Rice[1] **Katharina von der Wense**[1,2] **Alexis Palmer**[1]
[1]University of Colorado Boulder [2] Johannes Gutenberg University Mainz
enora.rice@colorado.edu

Abstract

Computational morphology has the potential to support language documentation through tasks like morphological segmentation and the generation of Interlinear Glossed Text (IGT). However, our research outputs have seen limited use in real-world language documentation settings. This position paper situates the disconnect between computational morphology and language documentation within a broader misalignment between research and practice in NLP and argues that the field risks becoming decontextualized and ineffectual without systematic integration of User-Centered Design (UCD). To demonstrate how principles from UCD can reshape the research agenda, we present a case study of GlossLM, a state-of-the-art multilingual IGT generation model. Through a small-scale user study with three documentary linguists, we find that, despite strong metric-based performance, the system fails to meet core usability needs in real documentation contexts. These insights raise new research questions around model constraints, label standardization, segmentation, and personalization. We argue that centering users not only produces more effective tools, but surfaces richer, more relevant research directions.

1 Introduction

Morphological analysis plays a central role in language documentation, and computational morphology is well-positioned to support this work through tasks such as morphological segmentation and the generation of Interlinear Glossed Text (IGT), a key linguistic annotation format. Yet, despite over two decades of interest—including early calls for NLP to engage more deeply with endangered languages (Bird, 2009)– we still lack broadly usable tools that support documentation workflows. This disconnect has been described as the "NLP gap" in language documentation (Gessler, 2022), and it presents not only a technical challenge but also a deeper disciplinary mismatch. We add to recent work that has highlighted the importance of incorporating user perspectives and rethinking evaluation practices (Ganesh et al., 2023; Liao and Xiao, 2025), and suggest that we need deep structural changes in how interdisciplinary systems are designed and assessed. These changes are especially urgent when research focuses on very low-resource or endangered languages, where care in collaboration is critical: otherwise, we risk building systems that extract data or prestige without meaningfully serving the communities involved (Schwartz, 2022; Bird, 2024).

We argue that User-Centered Design (UCD)—an iterative development approach from Human-Computer Interaction that emphasizes early and sustained engagement with end users—offers not only a path to more usable tools for morphological analysis, but also to a richer research process. We illustrate this through a case study of GlossLM (Ginn et al., 2024b), a state-of-the-art multilingual model for generating IGT. Since the stated aim of Ginn et al. (2024b) is to "explore the task of automatically generating IGT in order to aid documentation projects," we recruit 3 linguists to complete a small glossing task with GlossLM and share their perspectives on how it might fit into their documentation workflow. Our findings reveal that, despite strong performance on standard metrics, GlossLM falls short for real-world use: it lacks segmentation, enforces prescriptive glossing conventions, and produces out-of-domain labels. This feedback enables us to articulate new directions for research that are more accurately grounded in documentation workflows. Our findings raise the following research

Proceedings of the 2025 Conference on Empirical
November 4-9, 2025 ©2025 A

(Ginn et al. 2026, Ginn et al. 2024b)

PolyGloss

- New model based on user studies
- Expanded dataset (91k new examples)
- **Goal: multilingual model that predicts aligned morpheme segmentation and gloss labels**
- Introduced new alignment score

```
the   cat-s   ru-n       ⟶  x x-x x-x  ⟶  0.78
DET   cat-Pl  run.3P.Pl      x x-x x
```

Massively Multilingual Joint Segmentation and Glossing

Michael Ginn[1] Lindia Tjuatja[2] Enora Rice[1] Ali Marashian[1]
Maria Valentini[1] Jasmine Xu[2] Graham Neubig[2] Alexis Palmer[1]
[1]University of Colorado Boulder [2]Carnegie Mellon University
michael.ginn@colorado.edu

Abstract

Automated interlinear gloss prediction with neural networks is a promising approach to accelerate language documentation efforts. However, while state-of-the-art models like GLOSSLM (Ginn et al., 2024b) achieve high scores on glossing benchmarks, user studies with linguists have found critical barriers to the usefulness of such models in real-world scenarios (Rice et al., 2025). In particular, existing models typically generate morpheme-level glosses but assign them to whole words without predicting the actual morpheme boundaries, making the predictions less interpretable and thus untrustworthy to human annotators.

We conduct the first study on neural models that **jointly predict interlinear glosses and the corresponding morphological segmentation** from raw text. We run experiments to determine the optimal way to train models that balance segmentation and glossing accuracy, as well as the alignment between the two tasks. We extend the training corpus of GLOSSLM and pretrain POLYGLOSS, a family of seq2seq multilingual models for joint segmentation and glossing that outperforms GLOSSLM on glossing and beats various open-source LLMs on segmentation, glossing, and alignment. In addition, we demonstrate that POLYGLOSS can be quickly adapted to a new dataset via low-rank adaptation.

1 Introduction

Nearly half of the world's 7,000 languages face extinction. For many speakers and linguists of these languages, **language documentation** has become an urgent goal. Documentation projects commonly involve the creation of interlinear glossed text (IGT), a dense annotation format combining morphological segmentation, tagging, and translation (Figure 1). Due to its structured format and common usage among linguists, IGT has proven useful for linguistic analysis (Bender et al., 2013; Zamaraeva, 2016; Moeller et al., 2020), language pedagogy (Alast and Baleghizadeh, 2024; Bonilla Carvajal, 2025), and development of language technology such as taggers (Georgi, 2016), searchable text databases (Blokland et al., 2019; Rijhwani et al., 2023), educational tools (Uibo et al., 2017; Chaudhary et al., 2023), and machine translation systems (Zhou et al., 2020; Ramos et al., 2025).

Transcription: kooheesnee Andy
Segmentation: koo- he-esnee Andy
Glosses: INTERR-2S-hungry Andy

Figure 1: An interlinear glossed text example, showing the Arapaho for *"Are you hungry, Andy?"*. Our model predicts the segmentation and gloss line from the transcribed text.

Creating IGT is expensive, and a number of studies have proposed methods to automate IGT production with statistical and neural methods (McMillan-Major, 2020a; Zhao et al., 2020; Ginn et al., 2024a). In all of these studies, including the 2023 SIGMORPHON shared task (Ginn et al., 2023), the task is formulated as predicting the gloss line from the transcription or segmentation line. The former is more difficult (but also more useful), as it requires the model to infer morphological segmentation in addition to predicting glosses, and has been the primary focus of recent work.

Though state-of-the-art glossing models such as GLOSSLM (Ginn et al., 2024b) have achieved high accuracy across many languages, Rice et al. (2025) discovered several issues when using these models in a realistic documentation scenario:

1

PolyGloss IGT corpus

- Available on hugging face
- 9 eval languages

Language	Train	Eval	Test
Arapaho (arp)	36776	4687	4499
Tsez (ddo)	3626	444	442
Gitksan (git)	89	42	37
Uspanteko (usp)	8338	170	566
Ainu (ain)	6726	218	590
Lezgi (lez)	646	51	53
Natugu (ntu)	786	99	99
Nyangbo (nyb)	1221	225	248
Ruuli (ruc)	2158	212	333

Table 2: Number of examples for each evaluation language across train, eval, and test splits.

Statistic	Count
Total examples	353,266
Unique languages	2,077
Train examples	340,251
Eval examples	6,148
Test examples	6,867
No glottocode	13,428
No metalang. glottocode	10,894
No segmentation	93,648
No translation	5,921
Misaligned	34,894

Table 1: PolyGloss corpus statistics

	arp	ddo	git	usp	ain	lez	ntu	nyb	ruc	*Avg.*
Qwen 3 0.6B (ICL)	0.868	0.904	0.919	0.730	0.773	0.895	0.877	0.706	0.883	0.839
Gemma 3 4B (ICL)	0.489	0.597	0.826	0.476	0.351	0.668	0.473	0.430	0.723	0.559
Aya Expanse 8B (ICL)	0.545	0.749	0.871	0.514	0.464	0.740	0.591	0.492	0.802	0.641
GLOSSLM	0.161	0.095	0.870*	0.163	0.909*	0.940*	0.893*	0.990*	0.731*	0.639*
POLYGLOSS (ByT5, interleaved)	**0.152**	**0.072**	**0.597**	**0.160**	**0.095**	**0.357**	**0.142**	**0.222**	**0.306**	**0.234**

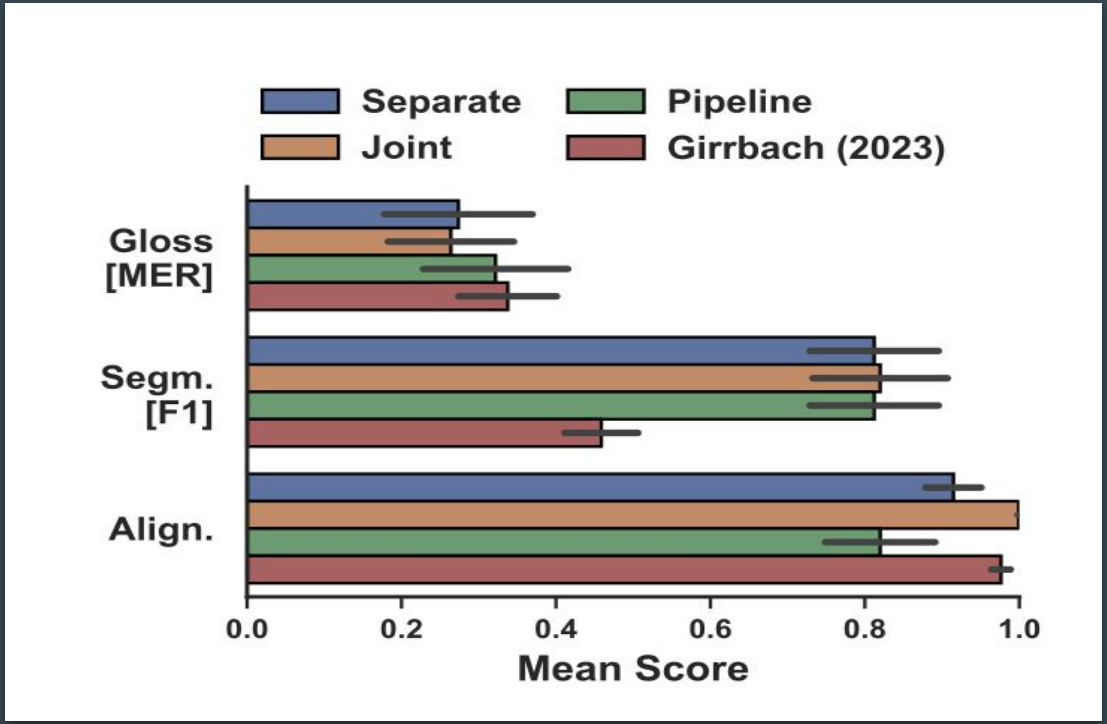
Table 3: Morpheme error rate (↓) for **glossing** on the held-out test set with multilingual models. For GLOSSLM, the only eval languages explicitly included in the pretraining corpus are `arp`, `ddo`, and `git`, so scores on other languages (marked with *) are very poor.

	arp	ddo	git	usp	ain	lez	ntu	nyb	ruc	*Avg.*
Qwen 3 0.6B (ICL)	0.868	0.904	0.919	0.730	0.773	0.895	0.877	0.706	0.883	0.839
Gemma 3 4B (ICL)	0.489	0.597	0.826	0.476	0.351	0.668	0.473	0.430	0.723	0.559
Aya Expanse 8B (ICL)	0.545	0.749	0.871	0.514	0.464	0.740	0.591	0.492	0.802	0.641
GLOSSLM	0.161	0.095	0.870*	0.163	0.909*	0.940*	0.893*	0.990*	0.731*	0.639*
POLYGLOSS (ByT5, interleaved)	**0.152**	**0.072**	**0.597**	**0.160**	**0.095**	**0.357**	**0.142**	**0.222**	**0.306**	**0.234**

Table 3: Morpheme error rate (↓) for **glossing** on the held-out test set with multilingual models. For GLOSSLM, the only eval languages explicitly included in the pretraining corpus are `arp`, `ddo`, and `git`, so scores on other languages (marked with *) are very poor.

PolyGloss

- Joint modeling outperforms separate models & pipeline
- New SOTA for the 9 eval languages
- Adapt quickly to a new language with low-rank adapter



PolyGloss

- Joint modeling outperforms separate models & pipeline
- New SOTA for the 9 eval languages
- Adapt quickly to a new language with low-rank adapter

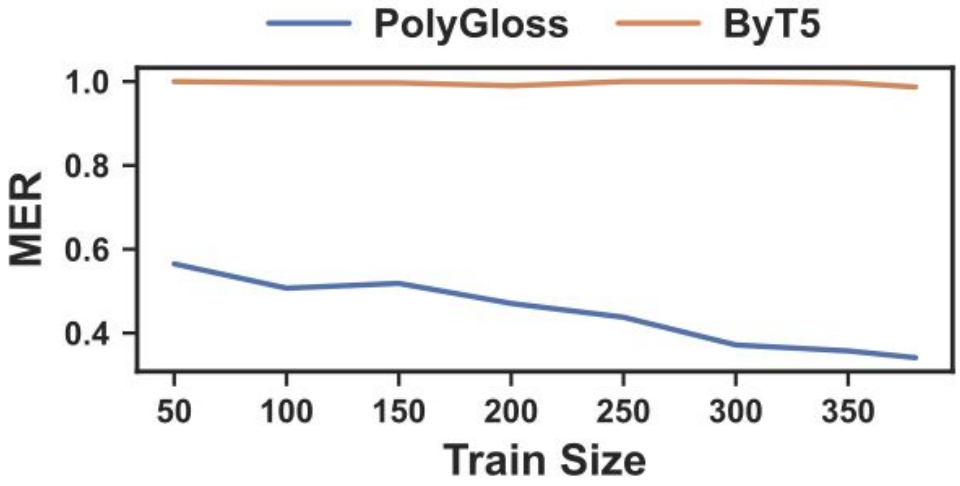


Figure 6: Morpheme error rate for Vamale when training LoRAs on the PolyGloss interleaved model and ByT5 with different size training sets

Historical Perspective

Recent Approaches

Centering the User

Very few linguists use NLP systems

The NLP Gap

- Poor interoperability between language documentation applications and NLP tools (Gessler 2022, Gessler and von der Wense 2024, Gessler et al. 2025)
- Institutional and disciplinary barriers (Flavelle and Lachler 2023)
  - Conflicting incentives
  - Limited interdisciplinary training
- A survey of "NLP applications" track for two 2020 NLP conferences (Ganesh et al., 2023) found that **nearly half lacked evaluations that reflected realistic deployment settings**
- Intrinsic evaluations often fail to predict real-world effectiveness and/or to align with human preferences (Ethayarajh & Jurafsky 2020; Kunz et al. 2022; Callison-Burch et al., 2006, among others)

Also... language documentation is not just any application

- Margaret Speas 2009: “Someone Else’s Language”

“IF WE TRULY WANT TO BE HELPFUL TO SOMEONE WITH A GOAL OF STABILIZING THEIR LANGUAGE, WE CANNOT ASSUME THAT WE KNOW BEST WHAT IS NEEDED BY A COMMUNITY THAT IS NOT OUR OWN”

Also... language documentation is not just any application

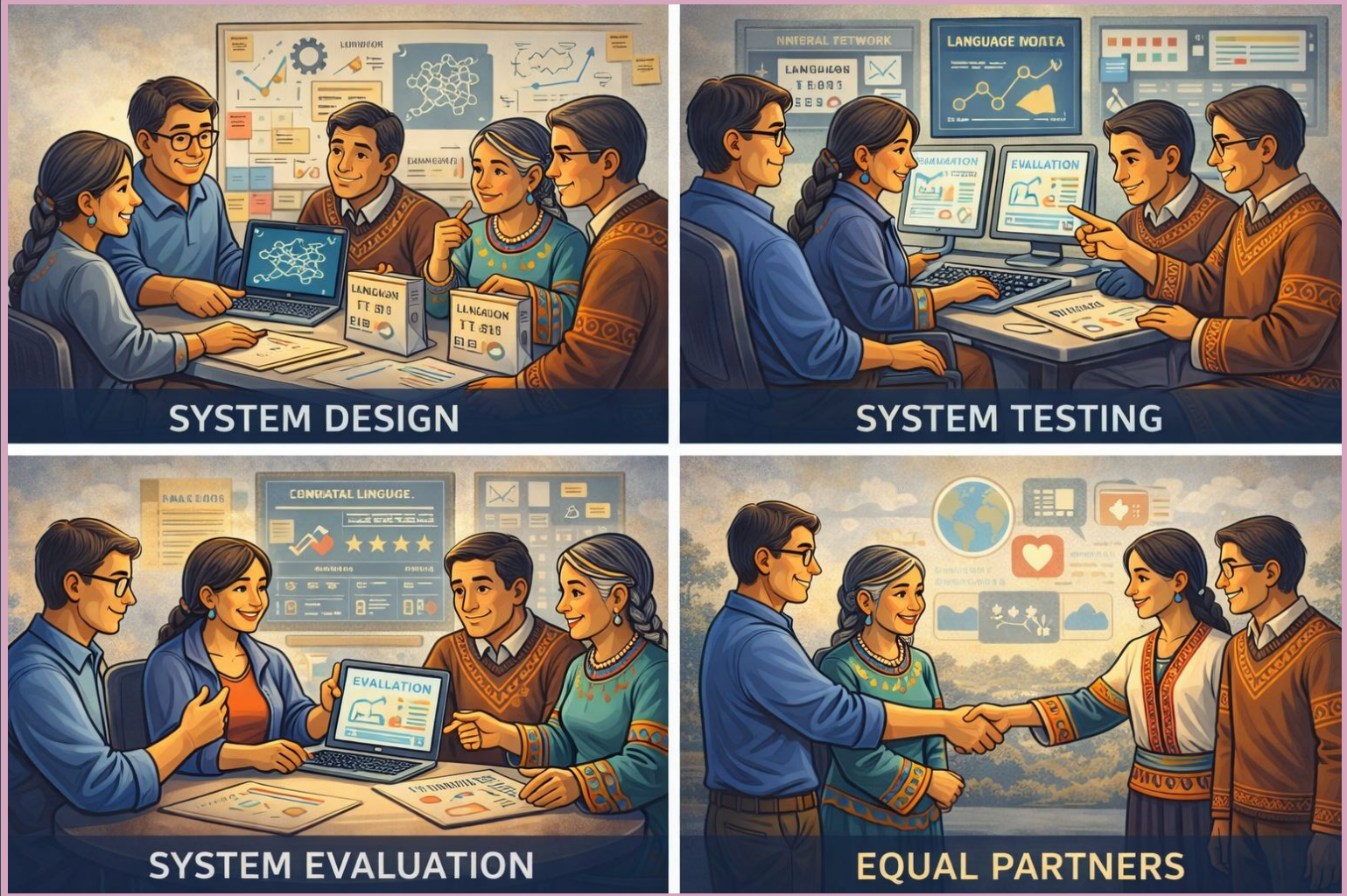
- Ken Hale: "... languages belong to those who speak them, not to those who study them"
- Interdisciplinary domain where researchers navigate the varied perspectives of diverse stakeholders
- (Hale: Linguistics is for those who like linguistic analysis - it does not help in furthering transmission of human languages)
- Several recent works call on NLP to decolonize practices when working with Indigenous languages in particular (Bird 2024, Schwartz 2022, Bird 2020, and others)

Back to user-centered design (UCD)

- Places user experience at the core of system development through iterative cycles of design, prototyping, and feedback
- Facilitates effective cross-disciplinary conversation
- Good for system design & development
- Underexplored for discovering novel research directions

UCD-inspired research questions for automating IGT

- Can we extract latent segmentation from glossing models? Do we need to?
  → PolyGloss
- Can (and should) we constrain glossing model outputs to pre-defined language specific labels? Or should we instead standardize gloss labels across languages?
- Can (and should) we tune glossing model outputs to fit the personal glossing conventions of individual linguists?
- Can we do accurate glossing without incorporating existing language-specific information, in a rule-guided fashion?


SYSTEM DESIGN
SYSTEM TESTING
SYSTEM EVALUATION
EQUAL PARTNERS

Thank you!

References - 1

Baldridge, Jason and Alexis Palmer. 2009. How well does active learning actually work? Time-based evaluation of cost-reduction strategies for language documentation.. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. ACL.

Bird, Steven. 2024. Must NLP be Extractive?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14915–14929, Bangkok, Thailand. ACL.

Bird, Steven. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bow, Catherine, Baden Hughes, and Steven Bird. 2003. Towards a general model of interlinear text. In *Proceedings of EMELD Workshoph 2003: Digitizing and Annotating Texts and Field Recordings.* LSA Institute: Lansing, MI, USA.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. ACL.

Cowell, Andrew. 2020. The Arapaho lexical and text database. Department of Linguistics, University of Colorado. Boulder, CO, USA.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. ACL.

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

References - 2

Farrar, Scott and William D. Lewis. 2007. “The GOLD Community of Practice: an infrastructure for linguistic data on the Web.” *Language Resources and Evaluation* 41 (2007): 45-60.

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

Flavelle, Darren and Jordan Lachler. 2023. Strengthening Relationships Between Indigenous Communities, Documentary Linguists, and Computational Linguists in the Era of NLP-Assisted Language Revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. ACL.

Ganesh, Ananya, Jie Cao, E. Margaret Perkoff, Rosy Southwell, Martha Palmer, and Katharina Kann. 2023. Mind the Gap between the Application Track and the Real World. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1833–1842, Toronto, Canada. ACL.

Gessler, Luke. 2022. Closing the NLP Gap: Documentary Linguistics and NLP Need a Shared Software Infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. ACL.

Gessler, Luke and Katharina von der Wense. 2024. NLP for Language Documentation: Two Reasons for the Gap between Theory and Practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 1–6, Mexico City, Mexico. ACL.

References - 3

Gessler, Luke, Alexis Palmer, and Katharina Von Der Wense. 2025. Understanding the Gap: an Analysis of Research Collaborations in NLP and Language Documentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 867–877, Vienna, Austria. ACL.

Ginn, Michael, 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, Toronto, Canada. ACL.

Ginn, Michael, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024a. 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. ACL.

Ginn, Michael, Mans Hulden, and Alexis Palmer. 2024b. Can we teach language models to gloss endangered languages?. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. ACL.

Ginn, Michael, Lindia Tjuatja, Enora Rice, Ali Marashian, Maria Valentini, Jasmine Xu, Graham Neubig, and Alexis Palmer. 2026. Massively Multilingual Joint Segmentation and Glossing. arxiv:2601.10925.

References - 4

Kunz, Jenny, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann. 2022. Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164–177, Abu Dhabi, United Arab Emirates (Hybrid). ACL.

Moon, Taesun, Katrin Erk, and Jason Baldridge. 2009. Unsupervised morphological segmentation and clustering with document boundaries. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 668–677, Singapore. ACL.

Palmer, Alexis. 2009. *Semi-automated annotation and active learning for language documentation*. Doctoral dissertation. University of Texas at Austin.

Palmer, Alexis, Taesun Moon, and Jason Baldridge. 2009. Evaluating Automation Strategies in Language Documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44, Boulder, Colorado. ACL.

Palmer, Alexis, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation: A case study in creating interlinear texts for Uspanteko. *Linguistic Issues in Language Technology*, *volume 3*.

References - 5

Rice, Enora, Katharina von der Wense, and Alexis Palmer. 2025. Interdisciplinary Research in Conversation: A Case Study in Computational Morphology for Language Documentation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11273–11285, Suzhou, China. ACL.

Schroeter, Ronald and Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In *Proceedings of Sustainable Data from Digital Fieldwork.* University of Sydney: Sydney University Press.

Schwartz, Lane. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. ACL.

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

Speas, Margaret. 2009. Someone Else's Language: On the Role of Linguists in Language Revitalization. Indigenous Language Revitalization: Encouragement, Guidance, and Lessons Learned. pp. 23-36.