Findings of the SIGMORPHON 2023 Shared Task on Interlinear Glossing

Michael Ginn[♢] **Sarah Moeller**[♣] **Alexis Palmer**[♢] **Anna Stacey**[♠]
Garrett Nicolai[♠] **Mans Hulden**[♢] **Miikka Silfverberg**[♠]
[♢]University of Colorado Boulder [♣]University of Florida
[♠]University of British Columbia
michael.ginn@colorado.edu miikka.silfverberg@ubc.ca

Abstract

This paper presents the findings of the SIGMORPHON 2023 Shared Task on Interlinear Glossing. This first iteration of the shared task explores glossing of a set of six typologically diverse languages: Arapaho, Gitksan, Lezgi, Natügu, Tsez and Uspanteko. The shared task encompasses two tracks: a resource-scarce closed track and an open track, where participants are allowed to utilize external data resources. Five teams participated in the shared task. The winning team Tü-CL achieved a 23.99%-point improvement over a baseline RoBERTa system in the closed track and a 17.42%-point improvement in the open track.

1 Introduction

Roughly half of the world's languages are currently endangered (Seifart et al., 2018). As a result, language preservation and revitalization have become significant areas of focus in linguistic research. Both of these endeavors require thorough documentation of the language, which is crucial for creating grammatical descriptions, dictionaries, and educational materials that aid in language revitalization. However, traditional manual language documentation is a time-consuming and resource-intensive process due to the costs associated with collecting, transcribing, and annotating linguistic data. Therefore, there is a need to expedite the documentation process through the use of automated methods. While these methods can never fully replace the expertise of a dedicated documentary linguist, they have the potential to greatly facilitate and accelerate the annotation of linguistic data (Palmer et al., 2009).

Linguistic annotation involves several interconnected subtasks, including: (1) transcription of speech recordings, (2) morphological segmentation of transcribed speech, (3) glossing of segmented morphemes, and (4) translation of the transcriptions into a matrix language, such as English. These processes result in a semi-structured output known as an interlinear gloss, as demonstrated in the Natügu example below:

(1) ma yrkr-tx-o-kz-Ø .
house finish-INTS-GDIR.DOWN-also-3MINIS .
Houses were gone too.

This paper presents the findings of the SIGMORPHON 2023 Shared Task on Interlinear Glossing[1], which focuses on automating step (3) of the language documentation pipeline. Notably, this shared task represents the first initiative specifically dedicated to interlinear glossing. Despite the prevalence of interlinear glossed text as a data format in language documentation, the automatic generation of glossed text remains relatively underexplored in the field of natural language processing (NLP). We hope that this shared task can help stimulate further work in automated glossing.

2 Background

Existing work in data driven automated glossing has utilized both traditional feature-based approaches like maximum entropy classifiers (MEMM) (Ratnaparkhi, 1996) and conditional random fields (CRF) (Lafferty et al., 2001) as well as more recent neural models like LSTM encoder-decoders (Sutskever et al., 2014) and transformers (Vaswani et al., 2017). Palmer et al. (2009) investigate active learning for interlinear glossing using the MEMM architecture. McMillan-Major (2020) incorporated translations as auxiliary supervision in a CRF glossing model. Moeller and Hulden (2018) and Barriga Martínez et al. (2021) compare traditional feature-based models and LSTM encoder-decoder models. Zhao et al. (2020) present a modified multi-source transformer model which incorporates translations as auxiliary supervision.

The current literature on automatic glossing exhibits notable gaps, as several techniques that have

[1] https://github.com/sigmorphon/2023glossingST

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 186–201

proven valuable for other morphology tasks have yet to be explored for glossing. There are several intriguing directions for future research, including:

1. Crosslingual training (Çöltekin, 2019; Anastasopoulos and Neubig, 2019) has shown promise for morphological inflection and could be investigated for its potential in glossing.

2. Incorporating additional noisy training data (Wiemerslage et al., 2023) can improve accuracy for low-resource inflection and could help improve the performance of glossing models as well. In the context of interlinear glossing, this data could come from large multilingual databases like ODIN (Lewis and Xia, 2010) which is automatically created with the aid of web crawling and is known to be noisy.

3. Data augmentation techniques (Liu and Hulden, 2021; Anastasopoulos and Neubig, 2019; Silfverberg et al., 2017) are now a well-established technique in morphological inflection and could enhance the training process for glossing models.

4. Hard attention models (Aharoni and Goldberg, 2017; Makarov et al., 2017) have delivered strong performance for several morphology tasks in low-resource settings and could also be applied to interlinear glossing.

5. Multitask training (Rama and Çöltekin, 2018) and meta-learning (Kann et al., 2020) techniques could be leveraged to enhance glossing performance.

6. Finally, pretrained language models like ByT5 (Xue et al., 2022) have demonstrated strong performance in various morphology tasks, yet their potential for interlinear glossing remains unexplored.

The submissions in this shared task explore several of these techniques, including the use of pretrained language models, data augmentation, utilization of external data, and the application of hard attention models.

3 Tasks and Evaluation

3.1 Interlinear Glossed Text

Interlinear Glossed Text (IGT) serves as a means to capture the syntactic and morphological characteristics of words within a corpus. It is a semi-structured format which lacks strict annotation standards, leading to variations in annotation practices among different annotators. These variations can be influenced by documentation requirements, adopted theoretical frameworks, and other factors (Palmer et al., 2009).

For this shared task, the data adheres to the Leipzig glossing conventions (Lehmann, 1982). The Leipzig format follows a three-line documentation style, including morphological segmentation of the input tokens, glosses of individual morphemes, and translations. Below is an example from Arapaho, one of the languages used in the shared task:

(2) nih-bii3ihi-noo nohkuseic
2S.PAST-eat-1S morning
I ate this morning.

In this example, the first line represents the morphological segmentation, the second line provides glosses for each morpheme, and the third line presents the corresponding translation.

The transcription line (*nih-bii3ihi-noo nohkuseic* in Example 2) gives the orthographic transcription of a sentence, phrase or utterance in the source language. The transcription may be segmented with dashes to indicate morpheme boundaries.

The gloss line ("2S.PAST-eat-1S morning" in Example 2) provides a linguistic gloss for each morpheme in the transcription line. For glossing, morphemes are grouped into two distinct categories:

1. *Functional morphemes* or *grams* include affixes and functional words which do not carry their own lexical meaning. Functional morphemes are glossed using uppercase labels like 1S (first-person singular affix) which indicate grammatical category and/or syntactic function. Portmanteau morphs, which denote multiple functions, can be glossed using compound labels like 2S.PAST. Gloss labels typically come from a fixed inventory like UniMorph (Sylak-Glassman, 2016; Kirov et al., 2018; Batsuren et al., 2022b), although conventions are not standardized and are often varied to fit the needs of the language.

2. In contrast to functional morphemes, *lexical morphemes* or *stems* are open-class words and stems which carry semantic meaning. These are glossed in lowercase using their translation

in a matrix language like English or Spanish; thus, for example, *bii3ihi* is glossed as *eat*.

The translation line ('I ate this morning.' in Example 2) of an IGT entry provides a translation in a high-resource language such as English. The tokens in the translation are not necessarily aligned with specific words in the source language, as languages often express equivalent concepts in differing numbers of words.

3.2 The Interlinear Glossing Task

The objective of the shared task is to develop automated systems capable of predicting the gloss of a given input utterance, using its orthographic transcription and translation as input. The glossing task presents several key challenges, such as disambiguation of ambiguous morphemes and accurate translation of word stems. The shared task explores two distinct resource settings, referred to as tracks, which differ in terms of the supervision provided during model training and at test-time.

The Closed Track (Track 1) In the closed track, the input consists of the orthographic transcription of the target utterance, for example, *nihbii3ihinoo nohkuseic* (Arapaho), and its translation to a matrix language like English: 'I ate this morning' (note the lack of morpheme boundaries in the transcription). The aim is to generate a gloss `2S.PAST-eat-1S morning`. This setting poses a significant challenge since the glossing model does not have access to a morphological segmentation of the input utterance. Therefore, it must infer the number of morphemes and the identity of the component morphemes for each input word without any supervision. The closed setting draws inspiration from the work of Zhao et al. (2020), which utilizes a similar setup.

The Open Track (Track 2) In a practical language documentation setting, various types of resources can be available as auxiliary supervision when training glossing systems. These resources may include manually glossed text, morphological segmentations, dictionaries, raw text in the target language, and more. The open track aims to explore the extent of glossing performance achievable when participants are allowed to utilize auxiliary resources. In addition to the data provided in the closed track, morphological segmentations are provided in the open track. For instance, for the Arapaho example mentioned earlier, a morphological segmentation *nih-bii3ihi-noo nohkuseic* would be included. Gold standard segmentations are provided both for model training and at test-time. Moreover, participants are encouraged to make use of external data resources except for additional glossed text in the target language.

3.3 Evaluation of Glossing Performance

We evaluate glossing performance with regard to two metrics: word-level and morpheme-level glossing accuracy. Word-level glossing accuracy is defined as the fraction of words in the test data which received a fully correct gloss like `2S.PAST-eat-1S`:

$$\mathrm{w}_{acc} = \frac{\text{Count(correctly glossed tokens)}}{\text{Count(all tokens)}} \quad (1)$$

Note that all the individual morphemes in the word have to be correctly glossed. In contrast, morpheme-level glossing accuracy is defined as the fraction of morphemes in the test data which received the correct gloss:

$$\mathrm{m}_{acc} = \frac{\text{Count(correctly glossed morphemes)}}{\text{Count(all morphemes)}} \quad (2)$$

In the closed track, where gold standard morphological segmentations are not provided, it may happen that the system predicts too few or too many glosses for an input word. This complicates computation of morpheme-level glossing accuracy. When too few morphemes are predicted, we pad the predictions with `NULL` morphemes until the number of morphemes corresponds to the gold standard gloss (e.g. `2S.PAST-eat` → `2S.PAST-eat-NULL`). When too many morphemes are predicted, we discard extra morphemes at the end of the output (e.g. `2S.PAST-eat-1S-PL` → `2S.PAST-eat-1S`).

For the official shared task results, we compute accuracy over multiple languages. We then report micro average glossing accuracy across the different languages. Micro average word-level glossing accuracy is used for the official ranking of the participating submissions.

3.4 Comparison to Other NLP Tasks

While interlinear glossing forms a distinct and interesting NLP task in its own right, it has connections to many commonly explored NLP tasks, particularly part-of-speech (POS) tagging, lemmatization, morphological tagging[2], and morphological segmentation (McCarthy et al., 2019; Cotterell and

[2]Also known as morphological analysis in context.

Heigold, 2017; Müller et al., 2015; Batsuren et al., 2022a). All of these tasks involve varying degrees of grammatical analysis.

Interlinear glossing is particularly strongly connected to morphological tagging as both involve morphological annotation in context. However, there are two major differences between the tasks:

1. In interlinear glossing, a morpheme-level annotation of the input sentence is generated. The output of a glossing model provides the order of various morphological elements in the input tokens, indicating the position of different affixal elements. In contrast, morphological tagging provides a more abstracted representation where the order of morphemes is lost.

2. Another difference between morphological tagging and interlinear glossing is related to the treatment of lexical elements. In morphological tagging, it is common to return the lemma of input words along with the associated grammatical information of the inflected input word. In glossing, on the other hand, it is common to annotate word forms with a translation of the input lexeme in a matrix language like English. This substantial difference between the tasks introduces elements of machine translation into the morphology task.

Following the approaches of McMillan-Major (2020) and Zhao et al. (2020), the shared task datasets provide gold standard translations of the input sentences as additional supervision during both training and test time. Thus, the task of lexeme translation involves retrieving the lemma of the correct lexemes from the provided translation.

4 Data

4.1 Languages and Glossed Data

Arapaho [arp] is an Algonquian language with a few hundred speakers in Wyoming, USA. It is highly agglutinating and polysynthetic, with the verb carrying the heaviest morphological load (Cowell and Moss, 2008). Polysynthesis in Arapaho includes noun incorporation, where special forms of certain nouns become part of the verb. The corpus used in this shared task contains narratives and conversation that have been documented starting in the 1880s until the present day, including a few religious texts that are translations from English. It is written in the popular Arapaho orthography. Much of the data is available through the Endangered Languages Archive[3] or the Center for the Study of Indigenous Languages of the West[4].

Gitksan [git] The Gitksan are one of the Indigenous peoples of the northern interior region of British Columbia, Canada. Today, Gitksan is the most vital Tsimshianic language, but is still critically endangered with an estimated 300-850 speakers (Dunlop et al., 2018). The language has an "analytic to synthetic" morphology (Rigsby, 1986, 1989) and, unlike many Canadian Indigenous languages, it is not polysynthetic. It has a rich assortment of derivational morphemes and substantial capacity for compounding; consequently, its degree of word-complexity has been described as similar to German (Tarpent, 1987). The data used for the shared task were extracted from a paper containing three stories by the Gitksan elders Barbara Sennot, Hector Hill and Vincent Gogag (Forbes et al., 2017).

Lezgi [lez] (aka Lezgian) is a Nakh-Daghestanian (Northeast Caucasian) language spoken by over 500,000 speakers in Russia and Azerbaijan (Eberhard et al., 2023). The corpus used is from the Qusar dialect in Azerbaijan (Donet, 2014). It is a highly agglutinative language with overwhelmingly suffixing morphology (Haspelmath, 1993). Noun cases are formed by case-stacking which is a unique characteristic of Nakh-Daghestanian languages. Instead of a unique morpheme for each case, case-stacking composes case inflections by "stacking" sequences of case suffixes as illustrated in Table 1.

Natügu [ntu] belongs to the Reefs-Santa Cruz group in the Austronesian family. It is spoken by about 4,000 people in the Temotu Province of the Solomon Islands. It has primarily agglutinative morphology with complex verb structures (Åshild Næss and Boerger, 2008). The corpus used for the shared task contains transcribed narratives and a large written text.[5]

Tsez [ddo] (aka Dido) belongs to the Tsez-Hinukh branch of the Nakh-Daghestanian family.

[3] https://elar.soas.ac.uk/Collection/MPI189644

[4] https://www.colorado.edu/center/csilw/arapaho-language-archives

[5] Natqgu grammar and large text available at https://www.langlxmelanesia.com/tilp

```
\t heetne'ii'P woowooyoo'ohk heet-ne'ii'cencei'soo'
\m heet-ne'ii'-P woo-wooyoo'-ohk heet-ne'ii'-cen-cei'soo-'
\g FUT-that's.when-pause REDUP-new-SUBJ FUT-that's.when-very-different-0S
\l It will be , pretty soon it will all be different  [ from how it is now ].
```

Figure 1: A glossed Arapaho sentence in the official shared task format for the open track (i.e. track 2).

WORD FORM	GLOSS
itim	SG.ABS 'man'
itim-ar	PL.ABS 'men'
itim-ar-di	PL-ERG 'men'
itim-di-k	OBL-AD.ESS 'near a man'
itim-di-k-di	OBL-AD-DIR 'toward a man'
itim-ar-di-k-ay	PL-OBL-AD-EL 'from men'

Table 1: A simplified example of Lezgi case-stacking on the noun root *itim* 'man'. Absolutive (ABS) and essive (ESS) cases and singular number (SG) are marked by null morphemes. The plural suffix (PL) attaches directly to the noun stem. The ergative (ERG) and the oblique (OBL) suffixes attach after the number. The adessive case (AD.ESS) attaches to the oblique suffix. The elative (EL) and directive (DIR) cases are added in the fourth slot after the root.

It has about 14,000 speakers in Daghestan, Russia. It has a rich agglutinative, suffixing morphology. The corpus is part of the Tsez Annotated Corpus Project (Comrie et al., 2022; Abdulaev and Abdullaev, 2010).[6]

Tutrugbu [nyb] (aka Nyagbo, Nyangbo) is a Niger-Congo language with a few thousand estimated speakers in Ghana (Eberhard et al., 2023). It is a highly agglutinative language that features some reduplication (Essegbey, 2019). The corpus from which the shared task data was extracted contains a variety of spontaneous data supplemented with elicited data collected with a range of documentary techniques.[7]

Uspanteko [usp] (aka Uspantek) belongs to the K'ichean branch of the Mayan language family spoken by as many as 6000 speakers in the Guatemalan highlands and in diaspora communities (Bennett et al., 2016). Uspanteko is a lightly agglutinative language with complex verbal morphology and ergative-absolutive alignment (Coon, 2016). Uspanteko is unusual among Mayan languages for its use of contrastive lexical tone (Bennett et al., 2022).[8] The texts were collected, transcribed, translated and annotated as part of an OKMA Mayan language documentation project (Pixabaj et al., 2007) and are currently accessible via the Archive of Indigenous Languages of Latin America.[9] The corpus includes oral histories, personal experience texts, and stories; preprocessing of the corpus is described in Palmer et al. (2010).

4.2 Shared Task Data

Shared task datasets were generated from original glossed source data in various formats (LaTeX, CLDF[10] and Flex[11]) using dedicated conversion scripts. We aimed to make minimal changes to the original glossed data while ensuring consistent annotation practices across languages. All morpheme boundaries were converted to a unified format using hyphens ("-"), all glossed word stems were lowercased (or titlecased in the case of proper nouns) and all affix glosses were uppercased. Apart from potential changes to casing, gloss symbols were not modified. Portmanteau morphs, where morpheme-boundaries cannot be identified, were glossed using a period syntax (".") as in the examples here.it.is and 2S.PAST.

An example of a glossed Arapaho sentence in the official shared task format is given in Figure 1. This entry comes from the open track (track 2), where morphological segmentations are provided. The following lines are included in the gloss:

`\t` the orthographic representation,

`\m` the morphological segmentation of the orthographic representation,

`\g` the gloss of the orthographic representation and

`\t` the English or Spanish translation.

[6] https://tsezacp.clld.org/

[7] 'Unpublished Nyangbo (Tutrugbu) texts' compiled by Dr. James Essegby

[8] Tone is not, however, marked in the shared task dataset.

[9] https://ailla.utexas.edu

[10] https://cldf.clld.org/

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

The token counts in the transcription, segmentation and gloss of a given example have to match. However, the token count in the translation line is allowed to differ. Examples in the source data which did not follow this restriction were filtered out.

We split the datsets into non-overlapping training, development and test data. For languages where there was a clear division into separate texts, we aimed to use one complete text for development and testing, respectively, and the rest of the data for training. This was the case for Gitksan and Arapaho.[12] For the rest of the languages, we used 80% of the sentences for training, and 10% for development and testing, respectively. Statistics on data sizes are provided in Table 2. Note that the table gives token counts, not sentence counts, and the counts do not, therefore, exactly correspond to an 80-10-10 split.

Data characteristics The shared task datasets encompass a range of diverse data conditions. The training data size, as shown in Table 2, varies from approximately 140k tokens for Arapaho to a mere 261 tokens for Gitksan, with most languages having between 2k and 15k tokens of training data. With the potential exception of Arapaho and Uspanteko, all the languages qualify as low-resourced datasets. Additional characteristics of the datasets are presented in Table 3:

1. Type-token-ratio (TTR) for most languages falls within the 20-30% range with the notable exception of Gitksan where TTR is 61.3% which is likely to be related to the very small size of the training set.

2. We compute out-of-vocabulary (OOV) rates on the test set. For most languages, OOV rates are below 30% with Gitksan once again being a notable exception with OOV rate of 79.9%. In general, these rates are high compared to typical OOV rates for English text.

3. As a further analysis, we also report morpheme-level OOV rates on the test set, which can be more illuminating for morphologically complex languages. These fall below 10% for most languages with the exception of Gitksan, where morpheme-level OOV is 41.2%, again due to the very small training set.

	TRAIN	DEV	TEST
ARAPAHO	139714	17573	17597
GITKSAN	261	388	384
LEZGI	7029	992	886
NATÜGU	10140	1280	1076
NYANGBO	8669	1093	1057
TSEZ	37458	4761	4701
USPANTEKO	41923	928	2405

Table 2: Token counts for shared task train, development and test data. The counts are the same for both the open and closed track.

In Table 3, we also report statistics related to the morphological characteristics of the languages:

1. The average number of morphemes per word can be computed based on the morphological segmentations provided for track 2. For training data, this ranges from 1.4, for Uspanteko, to 2.0 for Tsez, meaning that many multimorphemic words can be found in all of the datasets.

2. Finally, we also compute the gloss ambiguity, that is, the average number of distinct glosses that a morpheme receives in the training data. For example, the English suffix *-s* is ambiguous between two readings because it can be both a number and tense marker. Glossing ambiguity can be seen as one indicator of the difficulty of a glossing task. For most of the shared task languages, it is very close to 1. The only exceptions are Gitksan (1.3) and Uspanteko (1.2), both of which contain frequent and ambiguous affixes.

The shared task datasets also provide English or Spanish translations, which can be valuable when glossing word stems. Table 4 presents statistics on how often the correct stem translation can be found in the utterance translation.[13]. We present separate statistics for in-vocabulary tokens, which have been observed in the training set, and for out-of-vocabulary (OOV) tokens, which are absent from the training set. The coverage ranges from 37% for Uspanteko (40% for OOV tokens) to 71% for

[12]For Arapaho, text 56 is used for development and text 63 for testing. For Gitksan, we used Hector Hill's story for development and Vincent Gogag's story for testing.

[13]To compute these statistics, the translations in the test set were first lemmatized using the Stanza toolkit (Qi et al., 2020).

Tsez (72% for OOV tokens). This demonstrates that translations are likely to contain valuable information for the glossing task, particularly for OOV tokens, which can be challenging to gloss without access to stem translations.

5 Glossing Systems

5.1 The Baseline System

The baseline system utilizes the RoBERTa architecture with default hyperparameters (Liu et al., 2019). The glossing task is treated as a token classification task, where words or morphemes form the input, and the IGT gloss (or gloss compound) forms the output label. In the closed track, we train word-level models; in the open track, where morphological segmentations are provided, morphemes form the input units to the glossing model. The baseline model is trained on the shared task training data without pretraining. We train one model for each language. For a detailed presentation of the baseline system, please see Ginn (2023).[14]

A transformer-based architecture is an effective choice for this task, as interlinear glossing often involves disambiguating homonymous morphemes based on context. For example, the English plural morpheme *-s* is spelled the same as the present-tense third-person singular verb morpheme, and the correct label must be determined from the lexical and sentence context. We decided to use a masked architecture rather than a sequence-to-sequence setup. During initial development, we also experimented with a sequence-to-sequence architecture, but this required more data to converge, and delievered inferior performance. Error analysis revealed this to be due to isolated insertions and deletions of morphemes. This is difficult to fix because there exists no a priori restriction on the morpheme count generated by the model.

The baseline system includes a number of known limitations which leave room for improvement; particularly, it can not effectively handle out-of-vocabulary words or morphemes, does not perform any segmentation in the closed track, and does not make use of part-of-speech tags or other resources in the open track. The system also does not utilize translations.

[14]Code for the baeline system can be found in the shared task repository https://github.com/sigmorphon/2023glossingST/tree/main

5.2 Participant Systems

Here we describe the participating systems. Table 5 provides an overview of the strategies employed by the different teams.

COATES (Coates, 2023) This system is based on the LSTM encoder-decoder architecture (Sutskever et al., 2014) and participated in the closed track of the shared task. The input to the glossing system consists of short context windows centered at the target word. Windows of width 1 and 2 are used to generate candidate predictions and the final output prediction of the model is generated using weighted voting among the output candidates.

LISNTEAM (Okabe and Yvon, 2023) This submission is a hybrid CRF-neural system and participated in the open track of the shared task. The system is a combination of two components: (1) An unsupervised neural alignment system SimAlign (Sabet et al., 2020) originally intended for machine translation, and (2) A CRF sequence labeling system Lost (Lavergne et al., 2011). The alignment system is used during training to associate word stems with lexemes in the translations. It uses cosine similarity of BERT representations (Devlin et al., 2019) to score the association between lexemes in the translation and the word stems in the gloss. Alignment allows the system to learn to pick lexemes from the translation line for stems which do not occur in the training data and thus to gloss unseen word forms. The CRF model is used to gloss the morphemes in the input sentence. The team submitted two systems LISNTEAM$_1$ and LISNTEAM$_2$ which differ with regard to the featurization of the CRF model.

SIGMOREFUN (He et al., 2023) This team submitted transformer-based systems and participated in the open track of the shared task. The authors experiment with the pretrained byte-level transformer model ByT5 (Xue et al., 2022) and the multilingual pretrained transformer XLM-RoBERTa (Conneau et al., 2020) fine-tuned for glossing. Interestingly, the ByT5 model falls behind the XLM-RoBERTa model in terms of glossing accuracy. To boost performance, the team incorporate additional glossed data from the ODIN database and, for Gitksan, lexemes from a Gitksan morphological analyzer (Forbes et al., 2021). The team also experiment augmenting the gold standard training sets with artificially generated glossing data. This team incorporated both translations and segmentations into

	ARAPAHO	GITKSAN	LEZGI	NATÜGU	NYANGBO	TSEZ	USPANTEKO
(1) TTR	31.9%	61.3%	27.0%	27.5%	22.3%	29.4%	21.9%
(2) OOV	25.8%	79.9%	15.2%	21.4%	8.4%	18.1%	20.5%
(3) MORPH OOV	3.6%	41.2%	4.9%	2.8%	1.1%	0.5%	5.3%
(4) MORPHS PER WORD	1.8	1.6	1.5	1.6	1.6	2.0	1.4
(5) GLOSSES PER MORPH	1.0	1.3	1.0	1.0	1.0	1.0	1.2

Table 3: Statistics concerning the shared task datasets: (1) TTR type-token-ratio in training data, (2) Amount of OOV, or out-of-vocabulary, tokens in the test set, (3) Amount of OOV morphemes in the test set, (4) average number of morphemes per word in the training data, and (5) Average number of possible glosses per morpheme in the training data.

	TOK. RECALL	OOV TOK. RECALL
ARAPAHO	51.32	49.87
GITKSAN	44.29	44.13
LEZGI	42.98	44.89
NATÜGU	58.72	58.21
TSEZ	71.17	71.66
USPANTEKO	36.49	40.14

Table 4: Amount of stem glosses which are found in the translation of the sentence. We present figures separately for all tokens and OOV tokens which are not found in the training data. Nyangbo is missing from this table because translations are not provided.

the model input using specialized prompts. The team made four submissions to the shared task SIGMOREFUN$_1$ – SIGMORFUN$_4$ displaying different combinations of model and data augmentation strategy.

TEAMSIGGYMORPH (Cross et al., 2023) This team participate both in the open and closed track. They investigate the performance of different input and output representations: character-based, byte-based and subword-based. For the closed track, the team used a vanilla transformer model. For the open track, they applied a BiLSTM encoder-decoder model and the ByT5 byte-level transformer model. The team accomplished stem-translation using a heuristic approach which combines translation statistics computed from the training set and copying of unseen stems, which often represent proper names. Like team SIGMOREFUN, this team also found that ByT5 underperformed compared to other model architectures.

TÜ-CL (Girrbach, 2023) This team participated both in the open and closed track of the shared task (in fact, the team also participated in this year's SIGMORPHON inflection shared task using the same model). The system uses straight-through gradient estimation (Bengio et al., 2013) to train a hard-attentional neural glossing model. For the closed track submission, the system induces a shallow morphological segmentation of the input text. This happens without any segmented training data which is not available in the closed track. Morpheme boundaries are assigned using the hard attention weights learned by the model. For the open track, gold standard segmentations are used. For both tracks, gloss tags and stems are then predicted for each morpheme using an MLP. This model delivers very strong performance while, surprisingly, not utilizing translations in any way.

6 Results and Discussion

6.1 Closed track (track 1)

The official shared task results for the closed track are presented in Table 6. Three teams participated in the closed track and two of these teams presented a complete submission for all shared task languages and beat the baseline system. Only teams with a complete submission (TÜ-CL and COATES) were eligible to participate in the official shared task evaluation. Of these two teams, TÜ-CL achieved the best micro average word-level glossing accuracy 71.30% with their second submission TÜ-CL$_2$. Team TÜ-CL also delivering the best performance for all individual languages in track 1.

It is noteworthy that both teams TÜ-CL and COATES beat the shared task baseline by wide margins: 23.99%-points for TÜ-CL and 12.24%-points for COATES. This demonstrates that even in the resource-scarce closed track setting, large improvements in glossing accuracy are possible over a baseline transformer system. All track 1 submissions strongly outperform the baseline for Nyangbo. Likewise, we see great improvements over the baseline for Lezgi and Natügu.

Results for morpheme-level glossing accuracy

	HA	TRANSFORMER	BYT5	LSTM	CRF-HYBRID	USE TRANSL.	EXT. DATA	DATA AUG.
COATES				X				
LISNTEAM$_1$					X	X		
LISNTEAM$_2$					X	X		
SIGMOREFUN$_1$		X	X			X	X	X
SIGMOREFUN$_2$		X	X			X	X	X
SIGMOREFUN$_3$		X	X			X	X	X
TEAMSIGGYMORPH$_1$		X						
TEAMSIGGYMORPH$_2$		X	X	X				X
TÜ-CL$_1$	X							
TÜ-CL$_2$	X							

Table 5: Summary of design features in the shared task systems: Hard attention (HA), use of transformer architecture TRANSFORMER, use of the BYT5 pretrained model, use of LSTM encoder-decoder architecture, use of a hybrid CRF and neural model (CRF-HYBRID), use of the provided translations (USE TRANSL.), use of external data (EXT. DATA), and use of data augmentation techniques (DATA AUG.).

WORD-LEVEL ACCURACY

Submission	Arp	Ddo	Git	Lez	Ntu	Nyb	Usp	AVG	Complete?
TÜ-CL$_2$	**78.79**	80.94	**21.09**	**78.78**	**81.04**	85.05	**73.39**	71.30	**YES**
TÜ-CL$_1$	77.90	**80.96**	4.69	78.10	80.20	**85.34**	68.86	68.01	**YES**
COATES$_1$	55.56	74.45	6.51	65.69	70.63	77.01	66.99	59.55	**YES**
BASELINE	71.14	73.41	16.93	49.66	42.01	5.96	72.06	47.31	**YES**
TEAMSIGGYMORPH$_1$	-	52.46	-	22.91	41.82	59.22	57.26	46.73	

MORPHEME-LEVEL ACCURACY

Submission	Arp	Ddo	Git	Lez	Ntu	Nyb	Usp	AVG	Complete?
TÜ-CL$_2$	**78.47**	**73.95**	**11.72**	**62.10**	56.32	85.24	**70.05**	62.55	**YES**
TÜ-CL$_1$	76.56	70.29	9.26	62.03	**56.38**	**86.74**	60.42	60.24	**YES**
TEAMSIGGYMORPH$_1$	-	53.19	-	28.13	31.86	66.25	59.73	47.83	
COATES$_1$	45.42	64.43	9.84	40.74	37.55	72.82	56.02	46.69	**YES**
BASELINE	44.19	51.23	8.54	41.62	18.17	14.22	57.24	33.60	**YES**

Table 6: Word-level accuracy (above) and morpheme-level accuracy (below) for track 1. The AVG column gives the micro average accuracy across languages. Averages are not comparable for partial submissions, where results for some languages are missing.

largely mirror those of word-level accuracy. Again TÜ-CL delivers the best performance for all languages. A general observation is that morpheme-level accuracies in track 1 are lower than word-level accuracies. This can be attributed to the fact that multi-morphemic words are often difficult to gloss correctly when the morphological segmentation is not give. A single incorrectly identified morpheme boundary will often result in several incorrectly glossed morphemes. To see why this is the case, consider the English past tense verb form *walked*. If the word is incorrectly analyzed as a monolithic adjective, both the stem *walk* and past tense marker *-ed* will be incorrectly glossed. This effect weighs down morpheme-level accuracy for the closed track.

6.2 Open track (track 2)

The official shared task results for the open track are presented in Table 7. In the open track, we got submissions from four teams, two of which presented complete submissions for all shared task languages. Both of these teams beat the baseline with regard to micro averaged word-level glossing accuracy. Similarly as in the closed track, TÜ-CL achieved the best overall performance and the best performance for most languages. For Arapaho, the SIGMORFUN team achieved the best performance and, for Natügu and Gitksan, LISNTEAM achieved the best performance. TÜ-CL beat the baseline system with regard to micro average word-level glossing accuracy by a wide margin of 17.42%-points.

Overall performance in the open track is, un-

WORD-LEVEL ACCURACY

Submission	Arp	Ddo	Git	Lez	Ntu	Nyb	Usp	AVG	Complete?
TÜ-CL$_2$	85.80	**85.79**	26.56	83.41	87.92	**87.98**	**78.46**	76.56	**YES**
TÜ-CL$_1$	85.12	85.68	13.80	**85.44**	87.83	85.90	77.21	74.43	**YES**
SIGMOREFUN$_2$	82.92	80.07	31.25	77.77	78.72	85.53	77.51	73.39	**YES**
LISNTEAM$_1$	-	84.85	28.39	83.41	88.85	-	76.30	72.36	
SIGMOREFUN$_1$	**85.87**	73.77	27.86	74.15	82.99	80.61	73.47	71.25	**YES**
TEAMSIGGYMORPH$_2$	-	79.28	26.56	81.72	87.73	76.25	75.84	71.23	
SIGMOREFUN$_4$	80.56	82.79	20.57	63.77	77.97	82.59	75.72	69.14	**YES**
LISNTEAM$_2$	-	-	**31.51**	82.73	**89.31**	-	-	67.85	
BASELINE	85.44	75.71	16.41	34.54	41.08	84.30	76.55	59.14	**YES**
SIGMOREFUN$_3$	73.27	62.37	4.17	38.60	55.11	69.25	70.85	53.38	**YES**

MORPHEME-LEVEL ACCURACY

Submission	Arp	Ddo	Git	Lez	Ntu	Nyb	Usp	AVG	Complete?
TÜ-CL$_2$	**91.37**	**92.01**	50.22	**87.61**	92.32	**91.40**	**84.51**	84.21	**YES**
SIGMOREFUN$_2$	89.34	88.15	**52.39**	82.36	85.53	89.49	83.08	81.48	**YES**
LISNTEAM$_1$	-	91.39	50.80	87.17	92.60	-	82.42	80.88	
TEAMSIGGYMORPH$_2$	-	88.36	47.76	86.59	92.10	82.74	82.22	79.96	
SIGMOREFUN$_1$	91.36	84.35	47.47	80.17	88.35	85.84	80.08	79.66	**YES**
TÜ-CL$_1$	90.93	91.16	17.08	83.45	90.17	89.96	83.45	78.03	**YES**
LISNTEAM$_2$	-	-	51.09	86.52	**92.77**	-	-	76.79	
BASELINE	91.11	85.34	25.33	51.82	49.03	88.71	82.48	67.69	**YES**
SIGMOREFUN$_4$	80.81	78.24	12.74	50.00	63.39	85.30	73.25	63.39	**YES**
SIGMOREFUN$_3$	72.10	57.93	2.60	26.24	35.62	70.01	67.73	47.46	**YES**

Table 7: Word-level accuracy (above) and morpheme-level accuracy (below) for track 2. The AVG column gives the micro average accuracy across languages. Averages are not comparable for partial submissions, where results for some languages are missing.

derstandably, higher than in the closed track due to the fact that gold standard morphological segmentations were provided during training and test time, and additional resources were allowed, which some of the participants utilized. However, absolute improvement over the baseline is lower in the open track than the closed track. This may be a consequence of the fact that the learning problem in the open track is easier. It is also noteworthy that morpheme-level performance is higher than word-level performance for the open track, whereas the opposite is true for the closed track. This is understandable because gold standard morphological segmentations are provided and a single isolated glossing error is less likely to ruin the gloss for the complete word form in the open track.

6.3 Analysis of performance

We now present a more detailed analysis of the shared task results. This analysis is related to Figure 2 which presents average performance of shared task systems on the different languages and their relationship with training data size, out-of-vocabulary (OOV) rate and type-token-ratio (TTR).

Impact of training data size The size of the training set is one of the most influential factors determining the performance of natural language processing systems. This observation also holds true for the shared task results. The training data sizes vary from 261 tokens for Gitksan (git) to 139,714 tokens for Arapaho (arp). It is evident that the highest micro average word-level glossing performance in the open track is achieved for Arapaho, which benefits from the largest training set. In the closed track, Arapaho stands among the top three languages in terms of glossing accuracy, but the best performance is observed for Tsez (ddo), which has approximately 37,000 training tokens. This places it among the higher-resourced languages in the shared task. Conversely, Gitksan, with the smallest training set, consistently exhibits the lowest glossing performance. Overall, a clear trend emerges, demonstrating an improvement in glossing performance as the training data size increases.

Impact of OOV rate While out-of-vocabulary (OOV) rate computed on the test set is an important predictor of performance in tasks like morphological tagging (Müller et al., 2015), it does not seem to have a clear impact on system performance in this shared task. While the highest OOV rate and lowest performance are attained for Gitksan, this is largely an artefact of its very small training data size. If we disregard Gitksan, the impact of OOV rate both for the open and closed track is unclear. In fact, in the open track, the best average glossing performance is attained for Arapaho which has the second-highest OOV rate. Nevertheless, Arapaho also has the largest training set. This might seem like a surprising coincidence but we must remember that Arapaho is highly morphologically complex which tends to lead to higher OOV rates.

Impact of TTR Similarly to OOV, Type-token-ratio in the training set can be seen a measure of the diversity of the training data. We would expect a higher TTR to improve glossing performance. However, according to the statistics presented in Figure 2, the trend is not very clear. While the best performance in the closed track is attained for Tsez, which has moderately high TTR, the second-best performance is attained for Uspanteko with the lowest TTR.

7 Future Directions

The submissions in this shared task have explored several novel techniques that have not been previously applied to automatic interlinear glossing. Surprisingly, pretrained language models like ByT5 did not perform as well as one might expect based on their strong performance on other morphology tasks. This unexpected outcome raises the need for further investigation.

One interesting observation is that the winning submission, TÜ-CL, completely disregards the provided translations. While this could suggest that translations may not be as useful for the glossing task, we believe there is still room for improvement in this area. Incorporating large pretrained English models as a reliable source of translated text could potentially lead to additional enhancements.

Considering the availability of extensive morphological resources for many languages, such as those provided by UniMorph and similar projects, multi-task learning holds promise for interlinear glossing. Additionally, we encourage further exploration of crosslingual approaches, leveraging the ODIN database of interlinear glossed text, which despite being noisy, offers a highly multilingual resource for research purposes.

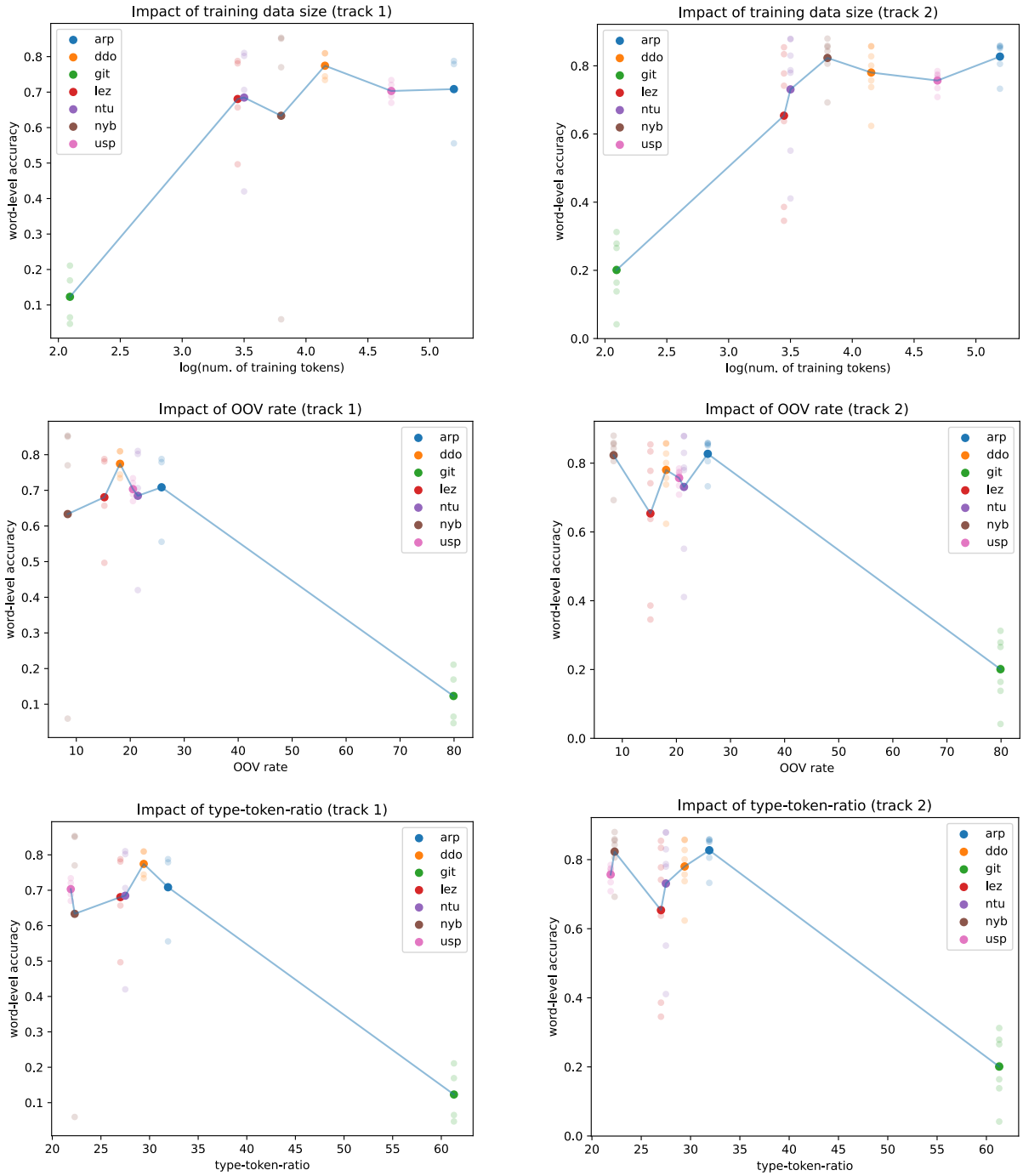

Figure 2: Impact of different data characteristics (training data size, out-of-vocabulary rate and type-token-ratio) on average word-level glossing accuracy. In addition to the average performance, we also plot the performance of each individual system. Only complete complete submissions, for all shared task languages, are included in these plots. Abbreviations refer to languages: Arapaho (arp), Tsez (ddo), Gitksan (git), Lezgi (lez), Natügu (ntu), Nyangbo (nyb) and Uspanteko (usp).

8 Conclusion

The 2023 SIGMORPHON Shared Task on Interlinear Glossing received submissions from five teams which presented a wealth of interesting techniques greatly expanding the field of automated interlinear glossing. The submissions achieved substantial improvements over a baseline RoBERTa system. The winning team Tü-CL achieved a 23.99%-point improvement over the baseline in the closed track and a 17.42%-point improvement in the open track using a hard attention model.

Acknowledgements

We would like to express our gratitude to the organizers of the SIGMORPHON workshop and all the participants of the shared task for their valuable contributions. We would also like to extend our sincerest thanks to the speakers and linguists who have dedicated their efforts to the development of the corpora used in this shared task. Lastly, Miikka Silfverberg wants to acknowledge the assistance provided by ChatGPT during the editing process of this manuscript.

References

A. K. Abdulaev and I. K. Abdullaev, editors. 2010. *Cezyas folklor/Dido (Tsez) folklore/Didojskij (cezskij) fol´klor*. "Lotos", Leipzig–Makhachkala.

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996.

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.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtskỳ, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, et al. 2022a. The sigmorphon 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022b. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Ryan Bennett, Jessica Coon, and Robert Henderson. 2016. Introduction to Mayan Linguistics. *Lang. Linguistics Compass*, 10:455–468.

Ryan Bennett, Meg Harvey, Robert Henderson, and Tomás Alberto Méndez López. 2022. The phonetics and phonology of uspanteko (mayan). *Language and Linguistics Compass*.

Edith Coates. 2023. An ensembled encoder-decoder system for interlinear glossed text. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics.

Çağrı Çöltekin. 2019. Cross-lingual morphological inflection with explicit alignment. In *Proceedings of the 16th workshop on computational research in phonetics, phonology, and morphology*, pages 71–79.

Bernard Comrie, A. K. Abdulaev, and I. K. Abdullaev, editors. 2022. *The Tsez Annotated Corpus Project (v1.0)*. Zenodo, Leipzig.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jessica Coon. 2016. Mayan morphosyntax. *Lang. Linguistics Compass*, 10:515–550.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759.

Andrew Cowell and Alonzo Moss. 2008. *The Arapaho Language*. University Press of Colorado.

Ziggy Cross, Michelle Yun, Ananya Apparaju, Jata MacCabe, Garrett Nicolai, and Miikka Silfverberg. 2023. Glossy bytes: Neural glossing using subword encoding. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Charles Donet. 2014. The Importance of Verb Salience in the Followability of Lezgi Oral Narratives. Master's thesis, Graduate Institute of Applied Linguistics, Dallas, TX.

Britt Dunlop, Suzanne Gessner, Tracey Herbert, and Aliana Parker. 2018. Report on the status of BC First Nations languages. Report of the First People's Cultural Council. Retrieved March 24, 2019.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, twenty-sixth edition. SIL International, Dallas, Texas.

James Essegbey. 2019. *Tutrugbu (Nyangbo) Language and Culture*. Brill, Leiden/Boston.

Clarissa Forbes, Henry Davis, Michael Schwan, and the UBC Gitksan Research Laboratory. 2017. Three Gitksan texts. In *Papers for the 52nd International Conference on Salish and Neighbouring Languages*, pages 47–89. UBC Working Papers in Linguistics.

Clarissa Forbes, Garrett Nicolai, and Miikka Silfverberg. 2021. An fst morphological analyzer for the gitksan language. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 188–197.

Michael Ginn. 2023. Sigmorphon 2023 shared task of interlinear glossing: Baseline model. *arXiv preprint arXiv:2303.14234*.

Leander Girrbach. 2023. Tü-CL at SIGMORPHON 2023: Straight-Through gradient estimation for hard attention. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics.

Martin Haspelmath. 1993. *A grammar of Lezgian*. Mouton de Gruyter, Berlin; New York.

Taiqi He, Lindia Tjuatja, Nathaniel Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig, and Lori Levin. 2023. SigMoreFun submission to the SIGMORPHON shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics.

Katharina Kann, Samuel R Bowman, and Kyunghyun Cho. 2020. Learning to learn morphological inflection for resource-poor languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 05 (34), pages 8058–8065.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. Unimorph 3.0: Universal morphology. *ArXiv*, abs/1810.11101.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Thomas Lavergne, Alexandre Allauzen, Josep Maria Crego, and François Yvon. 2011. From n-gram-based to CRF-based translation models. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 542–553, Edinburgh, Scotland. Association for Computational Linguistics.

Christian Lehmann. 1982. Directions for interlinear morphemic translations. *Folia Linguistica - FOLIA LINGUIST*, 16:199–224.

William D Lewis and Fei Xia. 2010. Developing odin: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*, 25(3):303–319.

Ling Liu and Mans Hulden. 2021. Backtranslation in neural morphological inflection. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 81–88.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Peter Makarov, Tatyana Ruzsics, and Simon Clematide. 2017. Align and copy: Uzh at sigmorphon 2017 shared task for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 49–57.

Arya D McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J Mielke, Jeffrey Heinz, et al. 2019. The sigmorphon 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 355–366, New York, New York. Association for Computational Linguistics.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 84–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2268–2274.

Shu Okabe and François Yvon. 2023. LISN @ SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics.

Alexis Palmer, 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.

Alexis Palmer, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology*, 3.

Telma Can Pixabaj, Miguel Angel Vicente Méndez, Mar ía Vicente Méndez, and Oswaldo Ajcot Damián. 2007. *Text Collections in Four Mayan Languages*. Archived in The Archive of the Indigenous Languages of Latin America (AILLA).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Taraka Rama and Çağrı Çöltekin. 2018. Tübingen-oslo system at sigmorphon shared task on morphological inflection. a multi-tasking multilingual sequence to sequence model. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 112–115.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on empirical methods in natural language processing*.

Bruce Rigsby. 1986. Gitxsan grammar. Ms., University of Queensland, Australia.

Bruce Rigsby. 1989. A later view of Gitksan syntax. In M. Key and H. Hoenigswald, editors, *General and Amerindian Ethnolinguistics: In remembrance of Stanley Newman*. Mouton de Gruyter, Berlin.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*.

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

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data augmentation for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (UniMorph Schema).

Marie-Lucie Tarpent. 1987. *A Grammar of the Nisgha Language*. Ph.D. thesis, University of Victoria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Adam Wiemerslage, Changbing Yang, Garrett Nicolai, Miikka Silfverberg, and Katharina Kann. 2023. An investigation of noise in morphological inflection. *arXiv preprint arXiv:2305.16581*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free

future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Åshild Næss and Brenda H. Boerger. 2008. Reefs–santa Cruz as Oceanic: Evidence from the verb complex. *Oceanic Linguistics*, 47:185–212.