Are Modern Neural ASR Architectures Robust for Polysynthetic Languages?

Éric Le Ferrand[1], **Zoey Liu**[2], **Antti Arppe**[3], **Emily Prud'hommeaux** [1]
[1]Boston College, USA
[2]University of Florida, USA
[3]University of Alberta, Canada

Abstract

Automatic speech recognition (ASR) technology is frequently proposed as a means of preservation and documentation of endangered languages, with promising results thus far. Among the endangered languages spoken today, a large number exhibit complex morphology. The models employed in contemporary language documentation pipelines that utilize ASR, however, are predominantly based on isolating or inflectional languages, often from the Indo-European family. This raises a critical concern: building models exclusively on such languages may introduce a bias, resulting in better performance on languages with simpler morphological structures. In this paper, we investigate the performance of modern ASR architectures on morphologically complex languages. Results indicate that modern ASR architectures appear less robust in managing high OOV rates for morphologically complex languages in terms of word error rate, while character error rates are consistently higher for isolating languages.

1 Introduction

Traditional morphological typology (Von Humboldt, 1822; Brown, 2011) recognizes a range of morphological complexity in the world's languages. At one end are isolating languages, like Mandarin, where grammatical information is not encoded in affixes and words typically consist of one or more lexical morphemes. Languages that do use affixes to indicate grammatical information, such as person and tense on a verb, or case and number on a noun, are typically described as either fusional – where a single morpheme encodes multiple features, as in Russian, Sanskrit, French – or as agglutinative – where each morpheme has a single function, as in Turkish. Languages with an extreme degree of morphological complexity, including phenomena like noun incorporation and "sentence words", are often described as polysynthetic. Many languages indigenous to the Americas, northern Australia, New Guinea, and Siberia are polysynthetic. English, while with rich derivational morphology, is quite limited in its inflectional morphology with only a handful of regular grammatical suffixes, making it closer in many ways to an isolating language.

Historically, NLP has often operated under the assumption that techniques effective for English will work well for other languages. However, this is gradually changing with the emergence of highly multilingual models that account for a significant diversity in language morphology (Radford, 2018; Radford et al., 2023). In areas such as machine translation, text generation, and speech recognition, the broader adoption of methods like Byte Pair Encoding (Sennrich et al., 2016) has facilitated the inclusion of languages with more complex morphology. Despite these advances, however, polysynthetic languages remain largely absent from large language models and automatic speech recognition systems.

In this paper, we select a variety of languages with differing degrees of morphological complexity to examine the effectiveness of state-of-the-art automatic speech recognition (ASR) models for transcription. We focus specifically on one key aspect: the potential for these languages to have a high number of out-of-vocabulary (OOV) words due to their extensive word construction possibilities and the models' ability to handle this challenge. On one hand, these models are trained on a large number of languages, including some agglutinative ones, and they operate on subwords, which might be sufficient to manage the production of long words with multiple morphemes. On the other hand, no polysynthetic languages were included in the original pretraining sets, leaving the ability of these models to handle such complex morphology unknown.

Using corpora for ten languages ranging typologically from isolating to polysynthetic, we extract

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 2953–2963

several features that capture aspects of morphological complexity. We then fine-tune ASR models for different splits of each corpus using two state-of-the-art multilingual ASR models. Comparing performance for each language in terms of word error rate (WER) and character error rate (CER) across models and splits, we find that both ASR architectures are less robust for polysynthetic languages and that certain measures of morphological complexity are associated with degradation of both WER and CER.

2 Related Work

Much of the NLP research on polysynthetic languages has focused primarily on computational morphology with the goal of developing methods for identifying the component parts of complex word structures. Early work included approaches such as diphones (Daland and Pierrehumbert, 2011), Bayesian methods (Goldwater et al., 2009), and transitional probabilities (Perruchet and Desaulty, 2008). Hybrid methods such as adaptor grammars were also recognized as powerful tools for uncovering subword units, aiding in language documentation (Sirts and Goldwater, 2013; Botha and Blunsom, 2013; Godard et al., 2018).

Recently, the trend has shifted towards combining traditional methods like finite-state transducers (FSTs) with neural networks to segment words and analyze their glosses. These approaches leverage human expertise and employ technology to scale up to larger datasets (Lane and Bird, 2019, 2020; Chen et al., 2020). Methods like BPE (Gutierrez-Vasques et al., 2021) and SentencePiece (Kudo and Richardson, 2018) are currently in wide use as a substitute or proxy for linguistically-informed morphological segmentation because of their efficiency, simplicity, and scalability. They have also been widely adopted into modern language processing pipelines, including contemporary text generation models (e.g. Touvron et al., 2023).

The idea of incorporating ASR into language documentation pipelines started to be explored with HMM-based architectures and has resulted in reductions in both speed of transcription and transcription error rates. (Prud'hommeaux et al., 2021; Shi et al., 2021). On similar architectures, attempts to incorporate morphological information in ASR architectures have flourished. These methods included not only BPE-like segmentation or syllable tokenization (Tachbelie et al., 2010; Manohar and Rajan, 2023) but also more complex methods that rely on WFSTs (Sak et al., 2009).

	Iso	Fus	Agglu	Poly	Other
WHISPER	17%	53%	27%	0%	2%
XLSR	22%	32%	42%	0%	4%

Table 1: Distribution of the morphological type of the languages used in ASR pretraining.

The expansion of transformer-based architectures has significantly enhanced the potential of ASR in documentary linguistics. Pretrained models like the WHISPER suite (Radford et al., 2023), wav2vec models (Baevski et al., 2020), and MMS (Pratap et al., 2024) have drastically reduced the amount of data needed to achieve reasonable results. These models have demonstrated efficiency in real-world documentary pipelines (Le Ferrand et al., 2023, 2024). However, their effectiveness for morphologically complex languages remains uncertain. An examination of the data used to build these models reveals that the vast majority of the included languages have relatively simple morphology, and none are polysynthetic (see Table 1). These architectures typically handle morphology by relying on subword units extracted from a pretrained SentencePiece model or by using segmentation based on a CTC loss function (Graves et al., 2006).

3 Data

For our study, we will compare ASR performances on three mostly isolating languages: Namakura, East Uvean, and Bambara, and seven languages with different levels of morphological complexity: Kanakanavu, Rukai, Hupa, Enenlhet, Kunwok, Plains Cree, and St Lawrence Yupik.

Namakura (ISO-nmk) and East Uvean (ISO-wls) are Austronesian languages spoken respectively on Wallis Island and Vanuatu in the Pacific Ocean. They are both isolating languages with simple morphology. Both corpora are spontaneous fieldwork recordings extracted from the Pangloss database[1]. The total durations of the collections are 2h9m for East Uvean and 2h18m for Namakura.

Bambara (ISO-bam) is one of the primary spoken languages in Mali and is part of the Mande language family. The language is isolating with simple morphology. The audio data consists of fieldwork recording of spontaneous speech. The

[1]https://pangloss.cnrs.fr/?lang=en

total duration of the collection is 7h11m. While the corpus is not currently publicly available, it is expected to be released this year (Tapo et al., 2024).

Kanakanavu (ISO-xnb) and Rukai (ISO-dru) are members of Formosan language family, the indigenous languages of Taiwan. Although not polysynthetic, these languages prominently feature complex morphological features such as distinct voice and case markings. The audio data consists of a mixture of fieldwork recording through the NTU corpus[2] (wen Su et al., 2008) and pedagogical resources from ePark [3]. Both collections are part of the FormosanBank project (Hartshorne et al., 2024). The total duration of the collections are 4h44m for Kanakanavu and 3h35m for Rukai.

Hupa (ISO-hup) is a critically endangered Native American language spoken in the Hoopa Valley in Northern California in the United States. It is part of the Dene/Athabaskan language family and has polysynthetic morphology. The audio data consists of narratives from one female elder native speaker of the language. The total duration of the collection is 9h12m. This corpus was shared for research purposes by a linguist who has been working with the Hupa community for a decade, with consent from the elder speaker. Due to ethical considerations of data sovereignty the data is not publicly available.

Toba-Maskoy or Enenlhet (ISO-tmf) is an Indigenous language spoken in Paraguay. It is part of the Maskoy language family. The language is polysynthetic but has very little documentation. The audio data consists of spontaneous stories about life and cultural topics. The total duration of the collection is 5h17m. The corpus was extracted from the Archive of the Indigenous Languages of Latin America[4].

Bininj Kunwok (ISO-gup) is an Aboriginal language spoken in Northern Australia. It is part of the Macro-Gunwinyguan language family and is polysynthetic. The audio data consists of guided tours by seven men and elicited sentences from three women speakers. The total duration of the collection is 1h3min. While the community has granted permission to use this data for research purposes, the corpus is not publicly available.

Plains Cree (ISO-crk) is an Indigenous language spoken in the plains of western Canada. It is part of the Algonquian language family and is highly polysynthetic but relatively agglutinative. The audio consists of a reading-out-loud of the book of Psalms by a fluent native speaker of Plains Cree. While various forms of audio for the entire Bible are available online, members of our research team have previously received access to high-quality recordings of Psalms for the purpose of developing a speech-synthesizer (Harrigan et al., 2019) and other research, and we make use of the same audio in this work. The total duration of this selection is 2h4min.

St Lawrence Yupik (ISO-ess) is an Indigenous language spoken on St Lawrence Island in the Bering straight in Alaska. It is part of the Eskimo-Aluit language family and is highly polysynthetic. The audio data consists of readings of the entire Bible. The total duration of the collection is 33h32m.

As noted above, all the languages used for this project are either extracted from open-source data collections constructed for research purposes or were made available to us through explicit agreements from the language community or associated scholars. All the languages here are Indigenous and most of them are endangered. None of these languages is included in the original training set of either of the state-of-the-art ASR models described later. To our knowledge, none of these languages display unusual phonetic or phonological features that could bias the experiments described below.

4 Methods

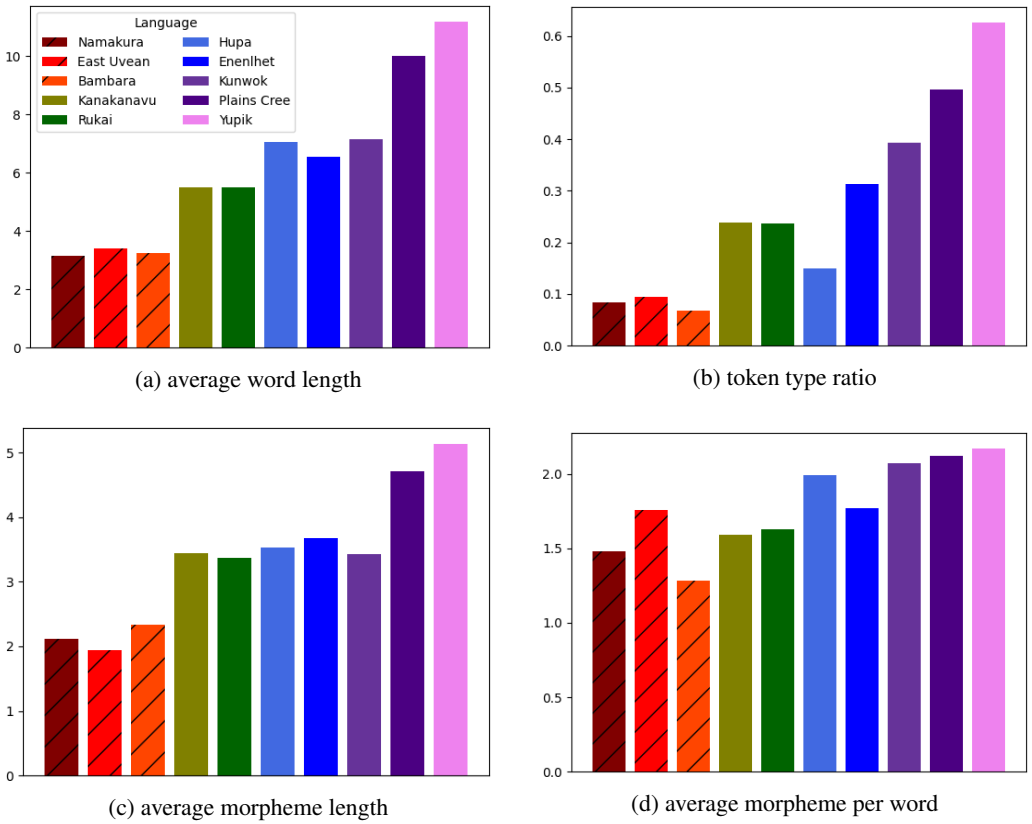
4.1 Morphological complexity

As noted above, morphological typology is a spectrum. On one end we can find isolating languages with simple morphology, such as Mandarin where information related to tense, person, or number is encoded in separate words. On the other end, we find polysynthetic languages where very complex word construction is possible, often resulting in very long words that would be translated into entire sentence with multiple words in English. Consider this exanple from Yupik:(1)

(1) **Mangteghaghllangllaghyugtukut**
Mangtegha-ghlla-ngllagh-yug-tu-kut
house-big-to.make-to.want.to-IND.INTR-1PL
'We want to make a big house.'
(Jacobson, 2001)

Among languages categorized as polysynthetic, different levels of morphological complexity can be observed. The morphological complexity of a language can be defined by the ability to add multiple

[2]https://corpus.linguistics.ntu.edu.tw/#/language/
[3]https://web.klokah.tw/
[4]https://ailla.utexas.org/collections/844/)



(a) average word length

(b) token type ratio

(c) average morpheme length

(d) average morpheme per word

Figure 1: Morphological complexity values for all languages according to the four criteria. The barred lines are for isolating languages, the plain bars are for morphologically complex languages

morphemes into a single word construction and how much information a single morpheme can hold. The more morphemes a language allows us to add in a single word construction, the more morphologically complex we can consider the language. Logically, a more complex language will end up with a much larger vocabulary because of all the possible combinations of morphemes that can be added to a single root or stem.

Based on this, we select four criteria to assess the morphological complexity of all our languages: The average word length (wrd len), the token type ratio (ttr), the average number of morphemes per word (mrp p/w), and the average morpheme length (mrp len). The first two criteria are easy to compute from the raw transcription. However, for the other two, we need to have access to some morphological segmentation. While some of the languages have existing morphological parsers for segmentation (Lane and Bird, 2019; Chen et al., 2020; Harrigan et al., 2017), this method cannot be generalized to all the languages here.

An alternative is to use unsupervised morphological segmentation. Morfessor is a suite of probabilistic methods that propose morphological segmentations for words from raw text (Virpioja et al., 2013). For each language, we train a segmentation model on all the audio transcripts and use that model to segment the data. We then compute the average morpheme length and number of morphemes per word on the resulting segmentation. The resulting values can be found in Figure 1.

For all four criteria, the segmentations proposed by Morfessor align with the typological descriptions of the languages. The five polysynthetic languages display more morphological complexity. They are followed by the two Formosan languages and the three isolating languages. Hupa displays a substantially lower token-to-type ratio. This is likely due to the fact that, in contrast to the audio data for the other languages, the audio recordings for Hupa consist of speech from just one elder female speaker, one of only a few living speakers of the language, whose individual preferences may limit the size of her vocabulary.

4.2 Testing word making ability

One key component of morphologically complex languages with repsect to ASR is their potential to have a very large vocabulary caused by the range

of possible word constructions. This phenomenon also increases risk of having a very large number of words out of vocabulary (OOV) compared to isolating languages. Because the chance of having a higher OOV rate is increased for polysynthetic languages, we want to see how much the performance of a system will degrade as OOV rates increase. To do so, we create two splits for each language where we vary the OOV rate based on types. The OOV rate is defined as the number of types (i.e., unique words) in the test set that do not appear in the training set, divided by the total number of types. To do this, for each language we create first a `random` split. Then from this split, we swap utterances that have a high rate of rare words in the training data with utterances that have a high rate of frequent words in the test data. We refer to this split as the `max` split. We swapped these utterances iteratively 700 times; we selected the split that had the highest OOV rate[5]. The OOV rates for each language can be found Figure 2.

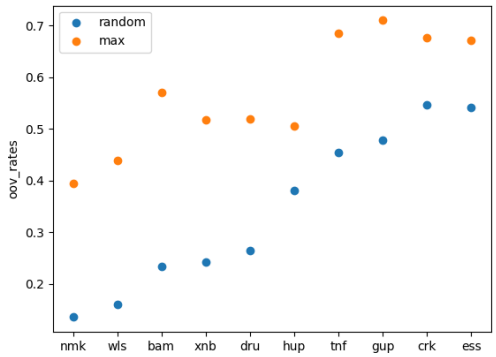


Figure 2: OOV rates for each language.

4.3 Frameworks

To build the ASR systems for each language, we fine-tune two pretrained models: WHISPER (Radford et al., 2023)) and XLSR-53 (Conneau et al., 2021).

WHISPER is an encoder-decoder architecture based on transformers. It takes as input raw spectrograms, passes them through encoder blocks, and then uses an autoregressive decoder to generate text from the encoded representation. For our task, we chose WHISPER medium which offered a good compromise between performance and resource requirements. We fine-tuned the model for 30 epochs using the standard hyperparameters described in the main WHISPER tutorial[6] and kept the last model for evaluation.

XLSR-53 (XLSR) is a pre-trained model based on the WAV2VEC architecture (Baevski et al., 2020). WAV2VEC is an encoder-decoder speech feature extractor based on transformers. Typical use for ASR consists of fine-tuning an extra transformer head with a CTC loss function on the top of the feature extractor. We fine-tuned XLSR for ASR for 30 epochs using the hyperparameters described in the main XLSR tutorial[7] and kept the last model for evaluation. The decoding of the model is done with a trigram language model trained on the transcripts of the audio training set for each language.

In an ideal scenario, a validation set would be extracted and the best model would be chosen; however, the amount of data for some languages is insufficient for this kind of development. For consistency, we chose to apply the same setup for all the languages. All training was conducted on a Nvidia A100 GPU for at most 5h for each model.

4.4 Analysis

The main challenge here is to be able to compare performance across languages. A simple comparison of the WER scores would not be fair as we are working with corpora that are very different. A very low WER might be observed in Yupik, which has the most data, while a very high WER might be observed for Kunwok, which has the least data. Instead, we compute the relative impact of the performances of two models. Concretely, for each model, we compute the WER, the CER, and the OOV robustness. The latter measure is defined as the number of OOVs that have been correctly transcribed out of all the OOVs in the test set. Then for each score, we compute the relative improvement or degradation as $(r - m)/s$ where r is the score of the model trained from the `random` split, and m is the score of the model trained from the `max` split. We compute this score for WER, CER, and OOV robustness for each language.

The Pearson correlation coefficient measures the linear relationship between two datasets. A correlation coefficient of +1.0 or -1.0 implies a perfect linear relationship. We use the Pearson coefficient to assess the impact of morphological complexity on ASR performances using the scores described in Section 1. Additionally, to ensure that the partitioning described in Section 4.2 does not bias our

[5]the partition-making script can be found in https://github.com/eleferrand/ASR_poly/tree/main

[6]https://huggingface.co/blog/fine-tune-whisper

[7]https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

results, we compute the correlation between the ASR scores and the change of OOV based on tokens, types and subwords and subtypes (i.e., unique subwords) based on the tokenization module of WHISPER. The degree of change is computed exactly as we computed the degradation score of the ASR models.

5 Results

All the figures are sorted from the least to the most morphologically complex language. The language order is the following: Namakura (nmk), East Uvean (wls), Bambara (bam), Kanakanavu (xnb), Rukai (dru), Hupa (hup), Enenlhet (tnf), Kunwok (gup), Plains Cree (crk) and Yupik (ess).

5.1 General observations

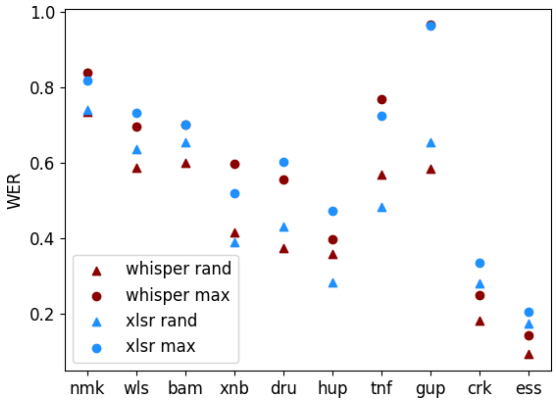


Figure 3: WERs for all the models.

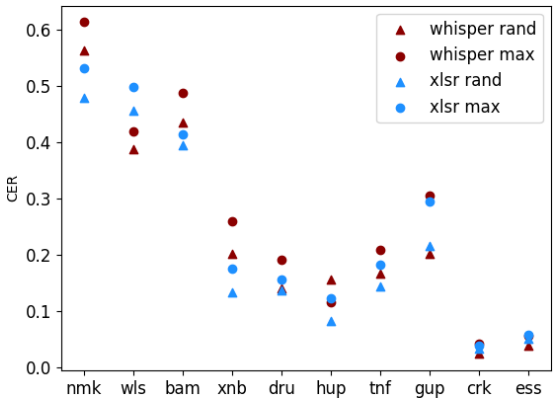


Figure 4: CERs for all the models.

The WERs for all the models trained can be found in Figure 3. The performance of the models across languages cannot be directly compared due to influencing factors beyond the language's typology, such as the quantity of data and the quality of the recordings. For example, Yupik (ess), which has the most data and cleaner recordings, achieves the best scores, while Kunwok (gup), with only one hour of spontaneous speech, has a much higher WER. No clear trend can be observed when comparing isolating languages with more morphologically complex languages. There is also no clear trend indicating whether WHISPER or XLSR is better.

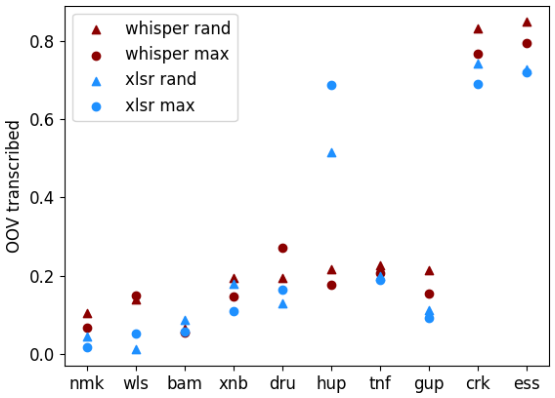


Figure 5: The amount of OOV correctly transcribed.

	WHISPER rand	WHISPER max	XLSR rand	XLSR max
Correlation	-0.9	-0.8	-0.9	-0.9
p-value	0.0002	0.001	0.000	0.0003

Table 2: The effect of language type on CERs.

The CERs for all models can be found in Figure 4. XLSR seems to provide generally better CER than WHISPER. A clear observation that can be made from the results is that the three isolating languages yield much higher scores than all the other languages across both architectures. Looking at the correlation between morphological type (isolating or non-isolating), we see that the CER is significantly lower for non-isolating languages for all four configurations (see Table 2).

With regards to the robustness of OOV (Figure 5), generally WHISPER seems to be more efficient in inferring the orthography of OOVs except for Hupa (hup) and Bambara (bam). In this aspect, all the models seem to be slightly more robust for languages with complex morphology. Interestingly, it is not always the case that models with higher OOV rates correctly transcribe a proportionally higher number of words. While both architectures seem to more accurately predict unseen words for morphologically complex languages, the difference is not pronounced.

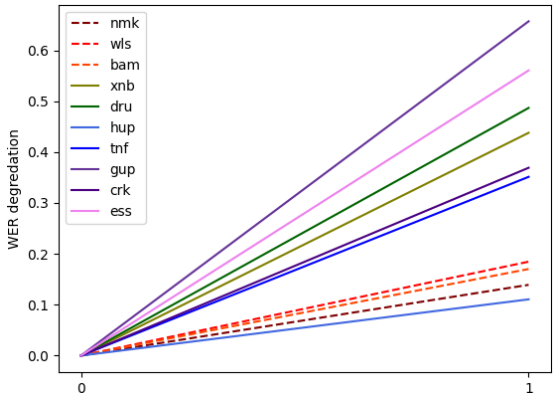


Figure 6: Degradation of WER with WHISPER. Dotted lines are isolating languages and plains line are languages with complex morphology

5.2 Results for WHISPER

The relative degradation of WER between the random model and the max model can be found in Figure 6. The first observation that can be made is that all of the isolating languages have a smaller increase in WER than the other languages except for Hupa. Other than this gap, the increase in WER does not seem to follow the ranking of the morphological diversity that can be observed in Figure 1.

Looking at the correlation between the increases in WER and the morphological factors (Table 3), we can see that none of the factors are significantly correlated other than token-type ratio. Hupa, with one of the longest average word lengths and the lowest WER degradation, renders this criterion insignificant. This reinforces the token-type ratio as the most relevant criterion for this study.

	wrd len	ttr	mrp len	mrp p/w
Correlation	0.56	**0.75**	0.59	0.48
p-value	0.08	**0.01**	0.06	0.15

Table 3: The effect of morphological factors on WERs with WHISPER. Significant values are in bold.

The correlation between the increases in CER and the morphological factors can be found in Table 4. Here again, the only significant factor is the token-type ratio. Regarding OOV robustness, none of the morphological factors show a significant correlation.

Looking at the impact of the augmentation of OOVs, none of the factors were significant for WER and OOV robustness. The evolution of subtypes and to a smaller extent subtokens seem to have a negative impact on CER (see Table 5).

	wrd len	ttr	mrp len	mrp p/w
Correlation	0.58	**0.77**	0.59	0.42
p-value	0.08	**0.009**	0.07	0.22

Table 4: The effect of morphological factors on CER with WHISPER. Significant values are in bold.

These scores mean that as the OOV rate computed on subtokens and subtypes increases, CER decreases. This sounds counterintuitive as a higher OOV rate usually correlates with higher error rates.

	tokens	types	subwords	subtypes
Correlation	-0.53	-0.36	**-0.67**	**-0.78**
p-value	0.11	0.29	**0.03**	**0.007**

Table 5: The effect of OOV rate factors on CER with WHISPER. Significant values are in bold.

5.3 Results for XLSR

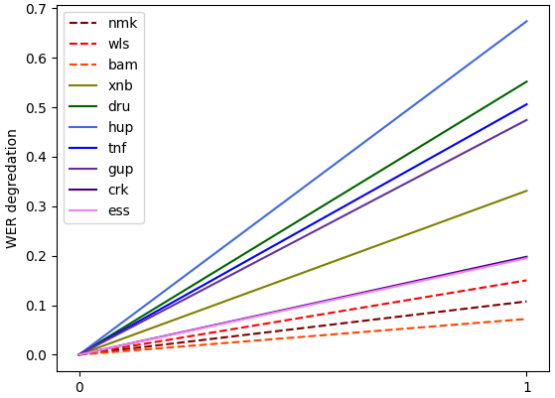


Figure 7: Evolution of the WER with XLSR. Dotted lines are isolating languages and plains line are languages with complex morphology

Examining the degradation in WER for XLSR (see Figure 7), the same disparity between the two language groups is evident. The three isolating languages still have the smallest increases but they are closely followed by Cree and Yupik. The other languages have much more dramatic increases. Interestingly, Hupa also behave radically differently under XLSR compared to WHISPER and has now the largest increase in WER among all languages.

Regarding the impact of the morphological factors on WER, CER, and OOV robustness, none of the criteria were found to be significantly correlated. However, we mentioned before that Yupik and Cree do behave differently than the other languages. We can see in Figure 8 the distribution of the data points of the correlation between WER and

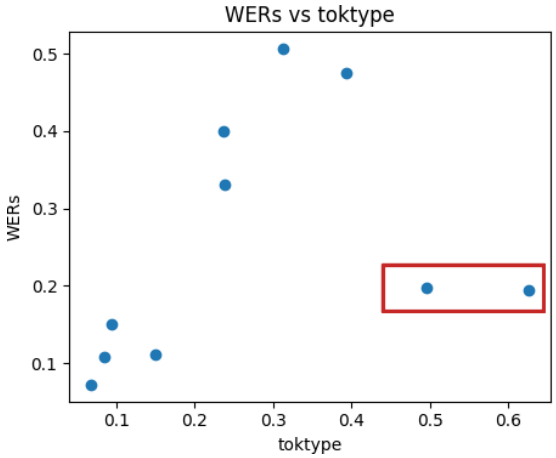


Figure 8: Correlation between WER and token type ratio for XLSR. The squared dots correspond to Cree and Yupik.

the token-type ratio. We can see a clear trend between all the languages other than Cree and Yupik, which have almost the same WER degradation. The data from both languages were extracted from the Bible which seems to produce a substantial bias in the analysis. We attempted to automatically identify the outliers using a linear regression, but the number of data points was probably too small. Arbitrarily removing these two languages led to a strong correlation between WER and the token-type ratio and to a lesser extent the average morpheme length (see Table 6), and between CER and the average word length and to a lesser extent the average morpheme length and the average number of morpheme per words (see Table 7).

	wrd len	ttr	mrp len	mrp p/w
Correlation	0.66	**0.94**	**0.71**	0.43
p-value	0.07	**0.0004**	**0.05**	0.28

Table 6: The effect of morphological factors on WER with XLSR without Cree and Yupik.

	wrd len	ttr	mrp len	mrp p/w
Correlation	**0.89**	0.55	**0.77**	**0.78**
p-value	**0.003**	0.15	**0.02**	**0.02**

Table 7: The effect of morphological factors on CER with XLSR without Cree and Yupik.

In terms of increasing OOV rates, none of the factors were significant for WER and OOV robustness. However, the increase of the tokens and subtokens out of vocabulary is significantly correlated to CER (see Table 8). This means that as the number of unknown tokens increases, CER increases.

	tokens	types	subwords	subtypes
Correlation	**0.86**	0.03	**0.68**	0.59
p-value	**0.001**	0.91	**0.03**	0.07

Table 8: The effect of OOV rate evolution factors on CER with XLSR.

5.4 Discussion

Conducting this study on two architectures allowed us to confirm performance trends related to morphological factors, but the results did not allow us to determine whether one architecture was uniformly better than the other. Generally, however, the differences in performance between `random` and `max` are less pronounced for XLSR.

Morphological complexity does have an influence on ASR performance. The degradation of WER is much more pronounced for morphologically complex languages than for isolating languages. Additionally, for both architectures, WER significantly increases when the token-type ratio increases. Both architectures, however, seem to be more robust in terms of CER.

Regarding the word-making ability of the architecture, while we couldn't pinpoint any one influencing factor, we were able to rule out the increase of OOV words and subwords as a significant factor, as well as morphological factors. Further study is necessary to understand this phenomenon.

While analyzing the potential biases introduced by partitioning, an interesting phenomenon emerged regarding the changes in CER. We anticipated that CER would increase with the rise in OOV rate, which was indeed the case with XLSR. The opposite trend was observed with WHISPER, where a higher number of OOVs resulted in better detection of individual characters.

Three languages stood out: Hupa, Cree, and Yupik. These languages probably had the lowest language diversity among all the languages in this study. The Hupa corpus includes a single speaker, and the diversity of her vocabulary might now be limited. The data extracted for Cree and Yupik is exclusively religious content, and the Cree data come from a single speaker. These differences created challenges during the analysis, as the performance of the ASR models varied significantly between these languages, irrespective of their morphological complexity.

6 Conclusion

In this paper, we investigated how morphological complexity affects the performance of state-of-the-art ASR models. We focused on one key aspect: the potential for polysynthetic languages to generate a vast number of morpheme combinations, leading to a high number of unknown words during model testing. We selected a set of 10 typologically diverse languages, created datasets with varying OOV rates, and attempted to correlate the degradation of WER with factors related to morphological complexity. We found that modern ASR models are generally less robust for polysynthetic languages. Among the various morphological factors we examined, a higher token-type ratio is associated with greater degradation of both WER and CER for morphologically complex languages. However, these models tend to produce higher CER for isolating languages.

The primary aim of this study was to clarify the relationship between complex morphology and ASR performance. Along the way, several research questions emerged, such as the impact of elements like the diversity of data collections and the combined effect of various influences. Additionally, the elements influencing the word formation capability of these models remain unclear.

Morphological complexity is a topic that triggers interest from a small but committed research community, but their interest is often limited to textual data. Extending research contributions to speech data could lead to improved understanding of the current biases within modern ASR systems and new directions in using ASR for language documentation purposes.

7 Limitations

We acknowledge two main limitations of our work: the number of languages included and the constraints of the correlation analysis. Ethically, we chose to work only with languages for which we had explicit permission to use available data for research purposes. Furthermore, we selected languages not included in the original training sets of the models to avoid biases, which restricted the number of languages we could study. Nevertheless, the selected languages are typologically quite diverse for such a small set, in terms of their phylogenic and geographic provenance.

Various factors could influence the performance of ASR models beyond morphological complexity. We only explore a few of these factors. Additionally, correlation analysis provides only a partial understanding of these impacts; more robust analyses could yield stronger trends.

Acknowledgements

We have obtained prior approval for the use of the various datasets. For the Bininj Kunwok data, we extend our gratitude to Steven Bird, Maïa Ponsonnet, and the Bininj community members who agreed to be recorded by Éric Le Ferrand. For the Yupik data, we thank Lane Schwartz and SIL Americas. Our thanks go to Raina Heaton for the Enenlhet data, Mrs. Verdena Parker and Dr. Justin Spence for the Hupa language data, Valentin Vydrin for Bambara. We are also grateful to Li-May Song and Yuyang Liu for facilitating access to NTU and ePark, respectively, for the Rukai and Kanakanavu data. For the Plains Cree data, we are grateful for Dolores Sand and the Canadian Bible Society for agreeing to our use of their data. This material is based upon work supported by the National Science Foundation under Grant #2319296. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the Computing Research Association.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jan A Botha and Phil Blunsom. 2013. Adaptor grammars for learning non-concatenative morphology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 345–356.

D Brown. 2011. Morphological typology. in: Jae jung song (ed.) the handbook of linguistic typology, 487-503. In *Handbook of Linguistic Typology*, 22, pages 487–503. Oxford University Press.

Emily Chen, Hyunji Hayley Park, and Lane Schwartz. 2020. Improved finite-state morphological analysis for St. lawrence Island Yupik using paradigm function morphology. In *12th International Conference on Language Resources and Evaluation, LREC 2020*, pages 2676–2684. European Language Resources Association (ELRA).

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Robert Daland and Janet B Pierrehumbert. 2011. Learning diphone-based segmentation. *Cognitive science*, 35(1):119–155.

Pierre Godard, Laurent Besacier, François Yvon, Martine Adda-Decker, Gilles Adda, Hélène Maynard, and Annie Rialland. 2018. Adaptor grammars for the linguist: Word segmentation experiments for very low-resource languages. In *Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–42.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468.

Atticus Harrigan, Antti Arppe, and Timothy Mills. 2019. A preliminary Plains Cree speech synthesizer. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 64–73.

Atticus G Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27:565–598.

Joshua K. Hartshorne, Éric Le Ferrand, Li-May Sung, and Emily Prud'hommeaux. 2024. Formosanbank and why you should use it. In *Architectures and Mechanisms in Language Processing (AMLaP) Poster*.

Steven A Jacobson. 2001. *A Practical Grammar of the St. Lawrence Island/Siberian Yupik Eskimo Language*. ERIC.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

William Lane and Steven Bird. 2019. Towards a robust morphological analyzer for Kunwinjku. *ALTA 2019*, page 1.

William Lane and Steven Bird. 2020. Bootstrapping techniques for polysynthetic morphological analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6652–6661.

Éric Le Ferrand, Raina Heaton, and Emily Prud'hommeaux. 2024. Enenlhet as a case-study to investigate asr model generalizability for language documentation. In *The Fourth Workshop on NLP for Indigenous Languages of the Americas*.

Éric Le Ferrand, Fabiola Henri, Benjamin Lecouteux, and Emmanuel Schang. 2023. Application of speech processes for the documentation of Kréyòl Gwadloupéyen. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 17–22.

Kavya Manohar and Rajeev Rajan. 2023. Improving speech recognition systems for the morphologically complex malayalam language using subword tokens for language modeling. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):47.

Pierre Perruchet and Stéphane Desaulty. 2008. A role for backward transitional probabilities in word segmentation? *Memory & cognition*, 36(7):1299–1305.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation and Conservation*, 15:491–513.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Haşim Sak, Murat Saraşlar, and Tunga Güngör. 2009. Integrating morphology into automatic speech recognition. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 354–358. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yoloxóchitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Wolfgang Menzel. 2010. Morpheme-based automatic speech recognition for a morphologically rich language-Amharic. In *SLTU*, pages 68–73.

Allahsera Tapo, Éric Le Ferrand, Zoey Liu, Christopher Homan, and Emily Prud'hommeaux. 2024. Leveraging speech data diversity to document indigenous heritage and culture. In *Interspeech 2024*, pages 5088–5092.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

S Virpioja, P Smit, SA Grönroos, and M Morfessor Kurimo. 2013. 2.0: Python implementation and extensions for morfessor baseline. *Helsinki: Department of Signal Processing and Acoustics, Aalto University*.

Wilhelm Von Humboldt. 1822. *Über das Entstehen der grammatischen Formen, und ihren Einfluss auf die Ideenentwicklung: Gelesen in der Akademie der Wissenschaften am 17. Januar 1822*.

Lily I wen Su, Li-May Sung, Shuping Huang, Fuhui Hsieh, and Zhemin Lin. 2008. NTU corpus of Formosan languages: A state-of-the-art report. *Corpus Linguistics and Linguistic Theory*, 4(2):291–294.