

# NLP beyond English: Do we need to think more about linguistics?

Marcel Bollmann

Linköping University, NLP Group



`marcel.bollmann.me`



`marcel.bollmann@liu.se`

# About me

 PhD in [Computational Linguistics](#) from Ruhr-Universität Bochum




 Postdoc in [CoAStAL NLP Group](#) @ DIKU



 Assistant Professor in [Jönköping AI Lab](#)

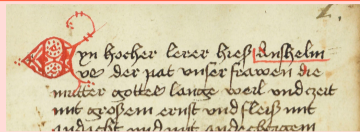


 Associate Professor at [LiU NLP](#)  
(since 16.01.2023)

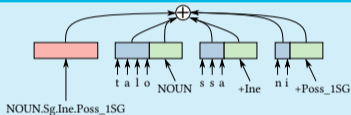


## Historical Documents

- Dealing with spelling variation
- Making old texts accessible for research



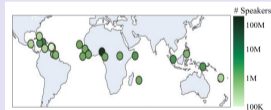
## Morphologically-Rich Languages



- Improving tokenization for MRLs
- “Bringing more linguistics into NLP”

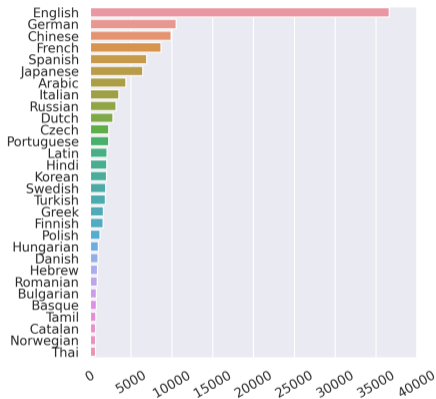
## Creoles

- Compiling datasets and models for creoles
- Data availability and quality issues



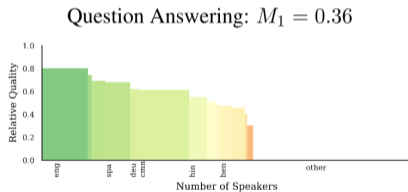
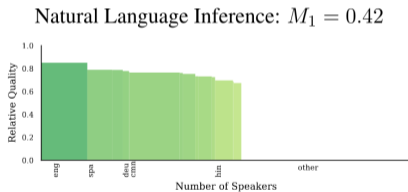
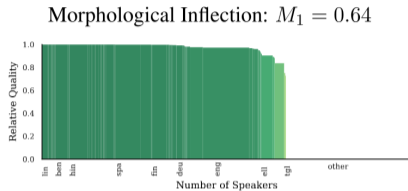
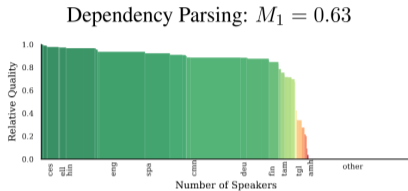
# Most NLP research is done on English

- How often are languages mentioned in papers in the [ACL Anthology](#)?
- 1,466 identified languages, but only 23% appear in more than 10 papers



■ Damian Blasi et al. (2022). "Systematic Inequalities in Language Technology Performance across the World's Languages".

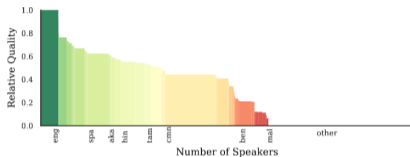
# Task quality vs. number of speakers



Damian Blasi et al. (2022). "Systematic Inequalities in Language Technology Performance across the World's Languages".

# Task quality vs. number of speakers

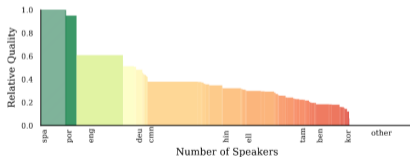
Speech Synthesis:  $M_1 = 0.32$



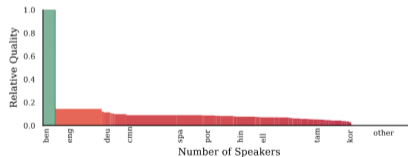
Machine Translation (X→English):  $M_1 = 0.49$



Machine Translation (X→Spanish):  $M_1 = 0.36$



Machine Translation (X→Bengali):  $M_1 = 0.10$



▮ Damian Blasi et al. (2022). "Systematic Inequalities in Language Technology Performance across the World's Languages".

- 1 Translating Indigenous American Languages
- 2 NLP for Creole Languages
- 3 NLP for Historical Documents
- 4 How do we improve on “NLP beyond English?”

- 1 Translating Indigenous American Languages
- 2 NLP for Creole Languages
- 3 NLP for Historical Documents
- 4 How do we improve on “NLP beyond English?”



# AmericasNLP 2021 Shared Task

- Translate from **Spanish** into one of ten **Indigenous American** languages.
- Low-resource setting
  - *Asháninka*: 4,000 sentences
  - *Quechua*: 125,000 sentences



## Monolingual Data

- **Wikipedia**
  - *Aymara, Guaraní, Nahuatl, Quechua*
- **Bible** translations
  - *Aymara, Guaraní, Quechua*
- **Individual books**

## Parallel Data

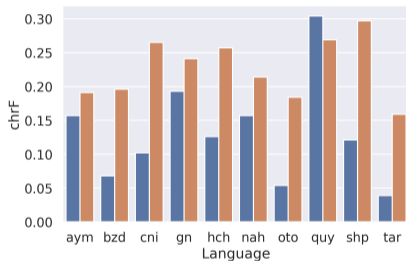
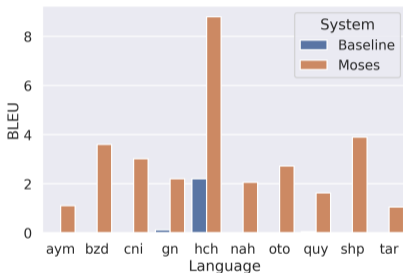
- **JW300** corpus
  - *Aymara*
- **Tatoeba** corpus
  - *Guarani*
- **Bible** corpus
  - *Nahuatl, Quechua*

---

▀ Marcel Bollmann et al. (2021). "Moses and the Character-Based Random Babbling Baseline: CoAStL at AmericasNLP 2021 Shared Task".

# Let's start with Moses!

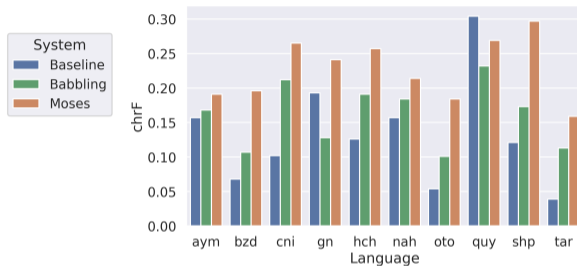
- **Statistical** machine translation ↔ naïve **neural baseline**



We did not manage to improve on this. 😞

# Are our results better than “random babbling”?

💡 What if we just  
randomly generated  
character  $n$ -grams?

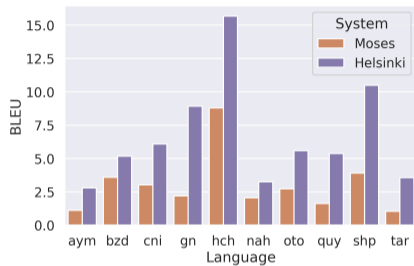


🤖 Almost **never ranked last** among all submissions

🤖 Scored **5th out of 12** on Asháninka

# Helsinki Team placed 1st on all languages

1. **Data quality:**  
Extensive cleaning & filtering
2. **Data quantity:** Backtranslation
3. **Transformer** model with carefully tweaked training regime



Overall, these BLEU scores are still very low!

▀ Raúl Vázquez et al. (2021). "The Helsinki submission to the AmericasNLP shared task".



## AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages

- [Mailing list for AmericasNLP 2023 shared task participants](#)
- [AmericasNLP 2023 shared task GitHub \(data+evaluation script\)](#)
- [Registration form](#)

### What?

The AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages is a competition aimed at encouraging the development of machine translation (MT) systems for Indigenous languages of the Americas. Participants will build systems that translate between Spanish and an Indigenous language.

### Why?

Many of the Indigenous languages of the Americas are under-resourced. This means that many approaches designed for translating between high-resource languages, such as neural machine translation, often struggle to perform well among languages frequently studied in natural language processing. One of the main goals of AmericasNLP is to motivate researchers to take on the challenge of developing MT systems for Indigenous languages.

**This year's results released on May 9th!**

ed. This means that many approaches designed for translating between high-resource languages exhibit linguistic properties uncommon among languages frequently studied in natural language processing. One of the main goals of AmericasNLP is to motivate researchers to

### How?

1 Translating Indigenous American Languages

2 NLP for Creole Languages

3 NLP for Historical Documents

4 How do we improve on “NLP beyond English?”

# Creoles: An example from Singlish

Tamil	Mandarin(我们)	Cantonese(拍拖)	English	Malay	Eng	Malay	Hokkien/ Hakka(店)	X
Dey	wǒ men	paktor	always	makan	at	kopi	tiam	one
Hey	,	we	date	always	eat	at	coffee shop	<INTJ>

Standard English: "Hey, when we date we always eat at the coffee shop"

Heather Lent et al. (2021). "On Language Models for Creoles".



# Should we even do this? Language technology needs

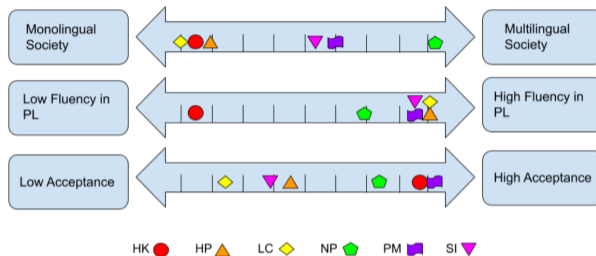


Figure 1: We map a sample of Creole languages to our proposed Creole continuum for language technology. PL here refers to “Prestige Language”. We map Haitian Kreyol (HK), Hawaiian Pidgin (HP), Louisiana Creole (LC), Nigerian Pidgin (NP), Papiamentu (PM), Singlish (SI).

Heather Lent et al. (2022). “What a Creole Wants, What a Creole Needs”.

# Haitian Creole: Where to obtain data?

- Not a lot of existing resources...

Haitian Kreyol	Haitian Disaster Response Corpus ( <a href="#">Munro, 2010</a> )	SMS	Verified; E-mail authors for access.
Haitian Kreyol	<a href="#">CMU Haitian Corpus</a>	Speech and Text Corpora	Verified
Haitian Kreyol	<a href="#">Corpus of Northern Haitian Creole</a>	Audio and Transcription	Not open source

 [creole-nlp.github.io](https://github.com/creole-nlp)

# What about web-crawled data?

- **OSCAR**: raw text in 166 languages
  - *Not a single creole* 😬
- Massively multilingual datasets often have **severe quality issues**
  - *7 out of 50 languages contained not a single correct sentence.*
  - *Poorest quality: African languages, minority languages “closely related to higher-resource languages”*



[oscar-project.org](https://oscar-project.org)

---

Julia Kreutzer et al. (2022). “Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets”. TACL 10.

# What about Wikipedia? (I)

- Often **template**-based

→ Not a lot of  
linguistic **variety**

## Hadley, Massachusetts

Depi Wikipedya, ansiklopedi lib

**Hadley, Massachusetts** se yon vil [Etazini](#). Li sitye nan leta [Massachusetts](#). Chèf-lye li se York.

## Albertville, Alabama

Depi Wikipedya, ansiklopedi lib

**Albertville, Alabama** se yon vil [Etazini](#). Li sitye nan leta [Alabama](#). Chèf-lye li se Marshall .

## Grover, Kolorado

Depi Wikipedya, ansiklopedi lib

**Grover, Kolorado** se yon vil [Etazini](#). Li sitye nan leta [Kolorado](#). Chèf-lye li se ? .

## Ball, Lwizyana

Depi Wikipedya, ansiklopedi lib

**Ball, Lwizyana** se yon vil [Etazini](#). Li sitye nan leta [Lwizyana](#). Chèf-lye li se ? .

## West Brookfield, Massachusetts

Depi Wikipedya, ansiklopedi lib

**West Brookfield, Massachusetts** se yon vil [Etazini](#). Li sitye nan leta [Massachusetts](#). Chèf-lye li se York.

# What about Wikipedia? (II)

- Largely **names** and **foreign-language titles**
- Word-/phrase-level language identification is **surprisingly hard!**

## Arrête-moi si tu peux

Depi Wikipedya, ansiklopedi lib

**Arrête-moi si tu peux** (nan angle : *Catch Me If You Can*) se yon **film ameriken** reyalize pa **Steven Spielberg** soti an **2002**. Fim sa a enspire pa lavi a **Frank Abagnale Jr.**

**Kontni** [kache]

- 1 Ekip teknik
- 2 Aktè
- 3 Referans
- 4 Lyen deyò

**Ekip teknik** [ modifye | modifye kòd ]

**Aktè** [ modifye | modifye kòd ]

- **Leonardo DiCaprio** : **Frank Abagnale Jr**
- **Tom Hanks** : Carl Hanratty, ajan **FBI**, anketè
- **Christopher Walken** : Frank Abagnale, Sr., papa Frank
- **Nathalie Baye** : Paula Abagnale, manman Frank
- **Amy Adams** : Brenda Strong
- **Martin Sheen** : Roger Strong, papa Brenda
- **James Brolin** : Jack Barnes, prezidan klib
- **Brian Howe** : ajan FBI Earl Amdursky

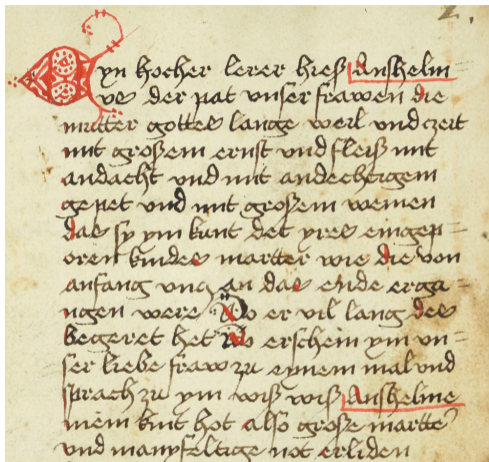
# A Creole benchmark dataset

Coming soon!

- 28 Creole languages, 6 NLP tasks
  - *Machine translation*
  - *Sentiment analysis*
  - *Named entity recognition*
  - ...
- Emphasis on **data quality**
  - *Parts were professionally translated*
- Hope: **encourage more research** into NLP for Creoles!

- 1 Translating Indigenous American Languages
- 2 NLP for Creole Languages
- 3 **NLP for Historical Documents**
- 4 How do we improve on “NLP beyond English?”

# NLP for historical documents



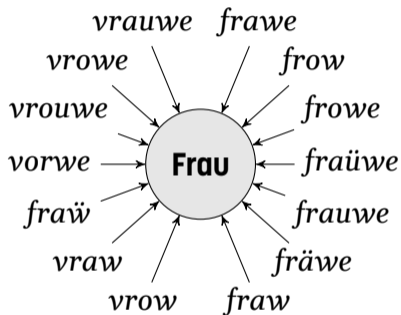
- Making historical documents **more widely accessible**
- Enabling **new research** directions
- Requires various NLP tasks
  - *Part-of-speech tagging*
  - *Named entity recognition*
  - *Entity linking*
  - ...

[linguistics.rub.de/comphist](https://linguistics.rub.de/comphist)

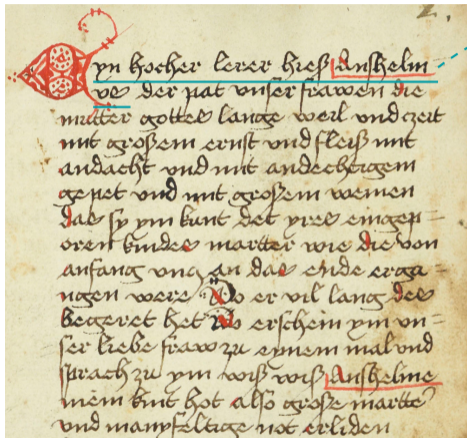


## Example: the Anselm corpus

- Early New High German  
(1350–1650)
- Semi-parallel texts,  
4,000–14,000 tokens each
- Lack of orthographic conventions:  
ca. 70% of all tokens are variants



# Historical text normalization pipeline



ayn hoher lere hieß anshelmve

ein hoher Lehrer hieß Anselm

ein hoher Lehrer hieß **Anselm<sub>PER</sub>**

A common theme?

## Quantitative Issues

- Few datasets to begin with
- Even **unlabelled data** can be hard to come by
  - *Many languages missing from OSCAR, Wikipedia, etc.*
  - *Historical documents can't be web-crawled*

- Hard to do reliable **language identification**

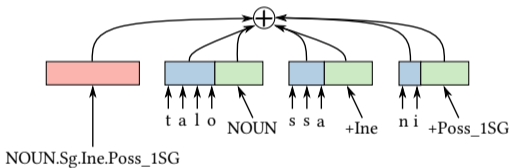
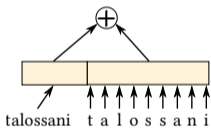
## Qualitative Issues

- Web-crawled data often **noisy**
- Even “curated” sources have **quality issues**
  - *Repetitive or “templatey” text*
  - *Proper nouns or foreign-language phrases*
- Surface-level **variation**

- 1 Translating Indigenous American Languages
- 2 NLP for Creole Languages
- 3 NLP for Historical Documents
- 4 How do we improve on "NLP beyond English?"

# My plan in 2018... 🔥 MorphIRe 🔥

- “Morphologically-Informed Representations for Natural Language Processing”



# Subword tokenization is statistically motivated

pappersindustriarbetareförbundet

*“paper industry worker’s union”*

- What we might hope to get:

pappers ##industri ##arbetare ##förbundet

- What mBERT’s tokenizer produces:

pa ##pper ##sin ##dust ##ria ##rb ##etar ##ef ##ör ##bundet

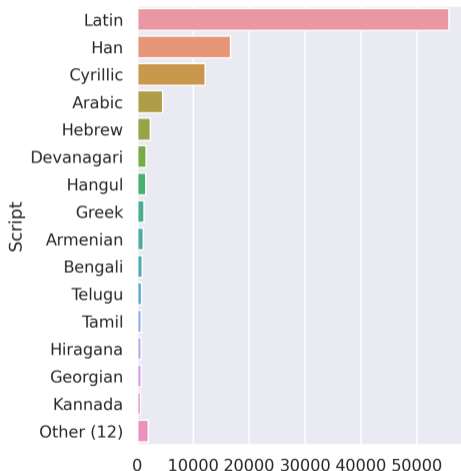
# A sample from mBERT's vocabulary

Энциклопедиясынын 完 738 sonucunda الجنسي #ènes chuyén legati ン  
poblíž filter #дала 完 Aragona prendre Rücktritt Robinson demografiala #だ  
ausgewiesen Brennan кхоллало #드 Messico #ξ Wereldoorlog 将 Spanien Сити  
##纵 ##郵 ##ミ— פרופסור ##있고 perlawanan Mujeres Pachina gospel বিপক্ষে  
excepto 남자 ##eve Вони Kandidaten 596 ##sivo 岬 ##ündən ##ійська Desire  
Sørensen inviato factores ば ##ينية 000 ##ла Breuning Patty ##胀 ##H му PA 釀  
ā ##b помоћ PK ##atus Play ##زن compringto ##준 ##ителями ##зная Lissabon  
##mione already belonging торган Москвы filosofo ند شدها do ##乞 шыв Bowl  
##sjonen ##i Catalogus ciudadanos ##hong ##žší ##zeka 1625 Wuppertal ##žiq юк

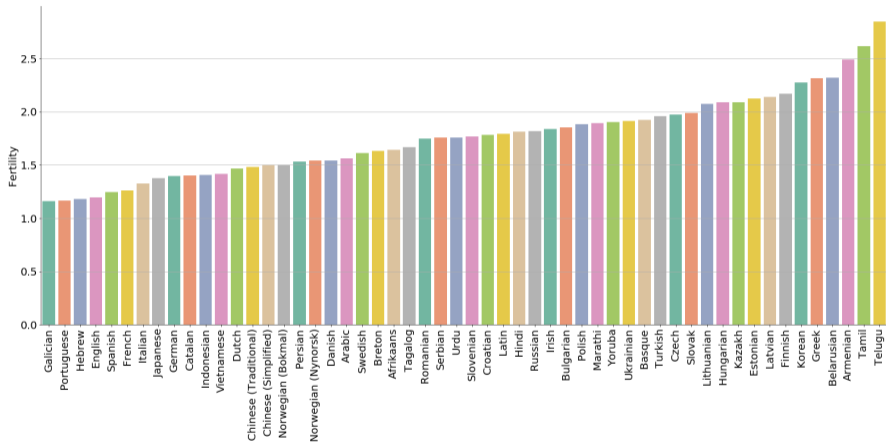


# Scripts represented in multilingual BERT

- mBERT can represent **27 different scripts**.
- Latin-script subwords make up **54%** of the total vocabulary.
- “Shared vocabulary”, but **it’s not shared equally!**



# Fertility: “How many subword tokens per word?”



Judit Ács (2019). [Exploring BERT's Vocabulary.](#)

## Fertility correlates with downstream performance

“ [T]he tokenizer’s ability of representing a language plays a crucial role [for downstream performance]; Consequently, choosing a *sub-optimal tokenizer* typically results in *deteriorated downstream performance*. ”

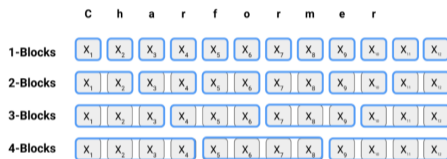
— Rust et al. (2021)

---

▀ Phillip Rust et al. (2021). “How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models”.

# Gradient-based subword tokenization

- **Charformer** learns tokenization “end-to-end.”



(a) Formation of subword blocks to be scored by  $F_R$ . Offsets and/or pre-GBST convolutions not shown.





(b) Block scores that have been expanded back to length  $L$ . Softmax is taken over block scores at each position  $i$  to form block weights for constructing latent subword representations.

Figure 2: Illustration of subword block formation and scoring.

Yi Tay et al. (2021). “Charformer: Fast Character Transformers via Gradient-based Subword Tokenization”. arXiv abs/2106.12672.

# Character-based tokenization

W e l c o m e \_ t o \_ L i n k ö p i n g

- **Canine**  uses Unicode codepoints.
- **ByT5**  uses raw bytes.

 **Efficiency? Characters/bytes carry much less information...**

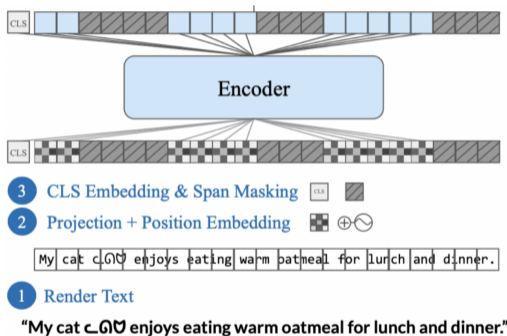
---

 Jonathan H. Clark et al. (2022). "Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation". In: TACL 10, pp. 73-91.

 Linting Xue et al. (2022). "ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models". In: TACL 10, pp. 291-306.

# Modelling language with pixels

- **PIXEL** encodes text rendered as an image.



Phillip Rust et al. (2022). "Language Modelling with Pixels". arXiv abs/2207.06991. ICLR 2023.



Some questions that interest me

# NLP beyond English

 What can we do to reduce the “language gap”?

- There’s a lot of room for improvement on under-explored languages and domains.

 Can we abstract away from different writing systems?

- It shouldn’t matter if e.g. Serbian is written in Cyrillic or Latin.

 How do we address data quantity/quality issues?

- We probably need to do more than just “collect more data.”



# Thank you!

 marcel.bollmann@liu.se  
 marcel.bollmann.me  
 @mbollmann@sigmoid.social