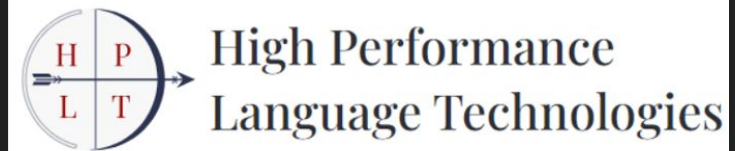


# Language identification for dataset building

Laurie Burchell ~ University of Edinburgh

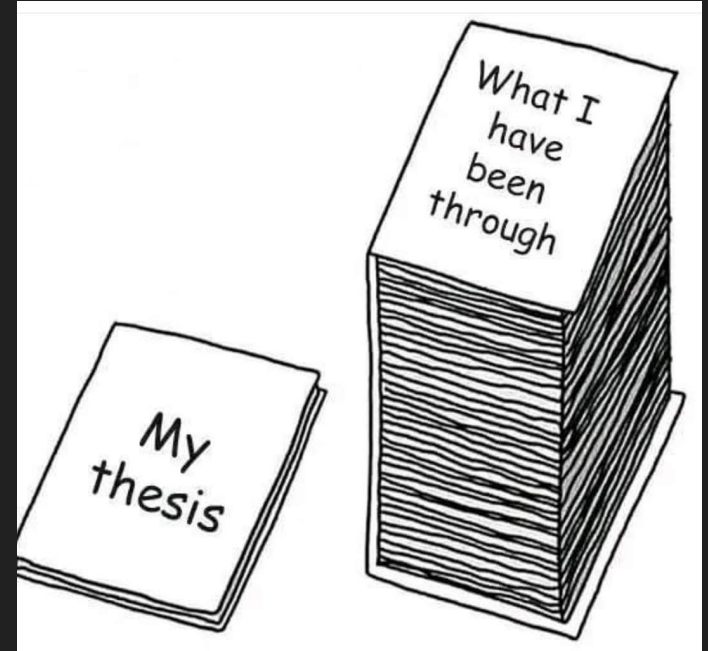


Slides?



# Overview

1. What is language identification?
2. Things to think about
3. Fail states (and what to do about it)
4. Recommendations



What is language identification?

# What is language identification?

English	All human beings are born free and equal in dignity and rights
MS Arabic	يولد جميع الناس أحراراً متساوين في الكرامة والحقوق
Scottish Gaelic	Tha gach uile dhuine air a bhreth saor agus co-ionnan ann an urram 's ann an còirichean
Armenian	Բոլոր մարդիկ ծնվում են ազատ ու հավասար՝ իրենց արժանապատվությամբ եւ իրավունքներով
Marathi	सर्व माणसे जन्मतः स्वतंत्र आहेत व प्रतिष्ठा आणि हक्कांच्या बाबतीत समान आहेत

Table 2.1: An extract from article 1 of the Universal Declaration of Human Rights (UDHR) in different languages. Marathi translation courtesy of Nikita Moghe; other translations from <https://www.ohchr.org>.



Pred. Language	Mined “Sentence” purporting to be in this language	Noise class
Manipuri		General noise
Twi (Akan)	me: why you lyyyn , why you always lyyyn	General noise
Varhadi	Òyáèè èè, áóðà- éyðèyðy ìàèèè ìèàí Éàãóá löyèèèì òyìy- yäyá-yðèàíú èèy áó íyñè [...]	Misrendered PDF
Aymara	Orilyzewuhubys ukagupixog axiqyh asozasuh uxilutidobyq osoqalelohan [...]	Non-Unicode font
Balinese	As of now  is verified profile on Instagram.	Boilerplate
Cherokee	“ALL mY IhΘRLs GREW bACK As fLOWERs ” ••• SWEET BΛBIES n DUGS	Creative use of Unicode
Oromo	My geology <b>essay</b> introduction <b>essay</b> on men authoring crosswords	Unlucky frequent n-gram
Pular	MEEOW	Repeated n-grams
Chechen	Жирновский ... Жирновский районный Фестиваль ТОСОВ	A N T S P E A K
Kashmiri	ਸ਼.	Short/ambiguous
Nigerian Pidgin	This new model features a stronger strap for a secure fit and increased comfort.	High-resource cousin
Uyghur	نۇرسۇلتان نازاربايەۋ قىتايدىڭ قازاقستانداغى ملشسىمەن	Out-of-model cousin
Dimli	The S</b><b class='b2'>urina</b><b class='b1'>m toa</b><b class='b3'>d is [...]	Deliberately Obfuscated

Table 2: Examples of several representative classes of noise in our initial web-crawl corpora.

Language identification is not solved (table from [Caswell et al. 2020](#))

42% languages had  
<50% useable data

[Kreutzer et al. \(2022\)](#)

Things to think about





## Things to think about (non-exhaustive)

1. Which languages?
2. Which model?
3. How to test?



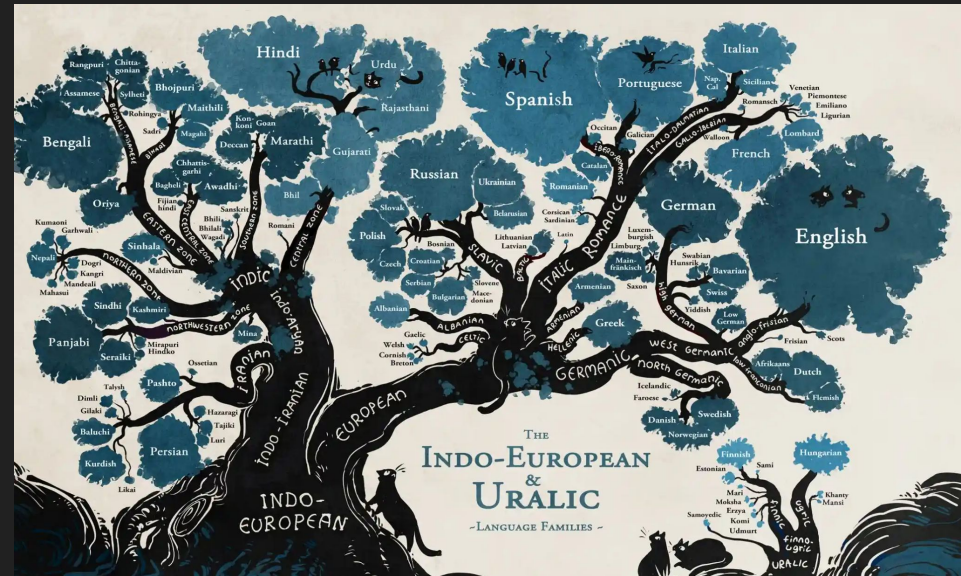
## Which languages?

- Trend towards more languages
-  More inclusion, more coverage
-  ...but I am sceptical
  - Checking quality in long tail?
  - Under-served  $\neq$  high-resource
  - More opportunities for confusion



# Which languages? My advice

- More languages not always better :(
- Think about downstream task (always)
- Check for close cousins
- Test properly!









## Models and architectures

- Web-scale data needs **fast** (and cheap) inference
- My standard pick: [fastText](#)
  - Alternative: [HeLI-OTS](#) (now [fast in Rust!](#))
- Two-stage models?
  - More expensive second stage



...at least not as default

## How to test? Metrics

- Recall 
- Precision  function of class balance 
- False positive rate  
- Confusion matrix clusters 

# How to test? Test sets



## Open Language Data Initiative

[Home] [Values] [Languages] [Guidelines]

**Latest news:** We are organising a [shared task at WMT24](#). Please consider participating!

## The FLORES+ evaluation benchmark for multilingual machine translation

We need better, more representative test sets!

Contributors 3



jeanm Jean



avidale David Dale

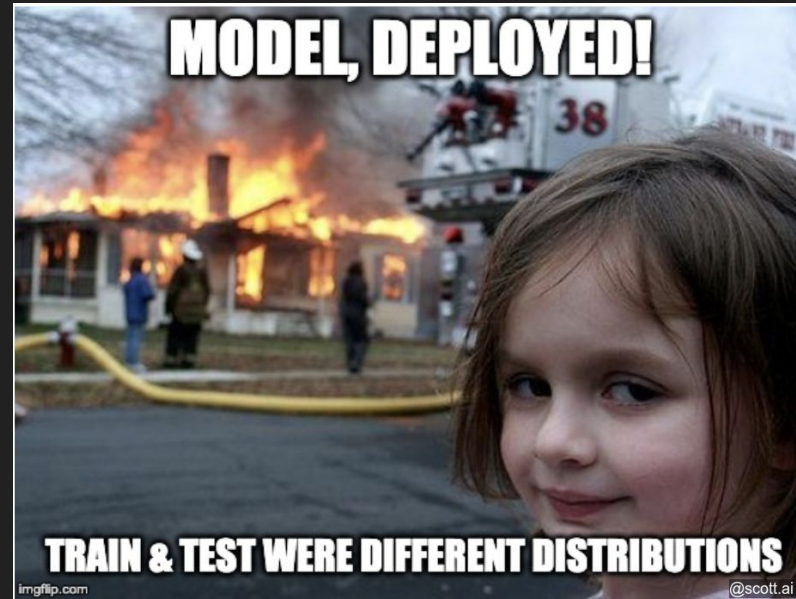


laurieburchell Laurie Burchell

omg who is she

# How to test? Test sets

- The dream: same domain for training, test and application 🥰
- The reality: unreliable test data 😞
- The ideal: check results with native speakers
  - Native speakers can only check native language reliably
  - ...but something is better than nothing ✌️

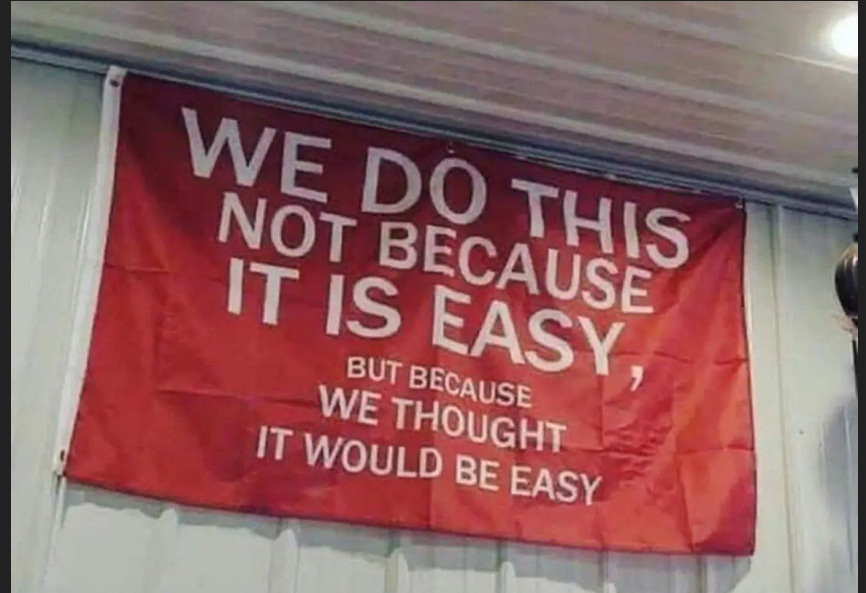


Fail states (and what to do about it)



## Fail states (a selection)

1. Short segments
2. Multilingual or code-switched input
3. Dialects/close languages



## Short segments

Problem: very short segments are often misclassified



andrey\_kutuzov 12:55

Actually, @laurie I am still interested why "not found nginx" is classified as Tagalog, etc. Just out of curiosity - is it really some ngram from "nginx" in the Tagalog training data? Can you check?



↩ 29 replies | Follow

Answer: extreme suspicion if <5-7 tokens

## Multilingual or code-switched input

- Breaks single-label assumption
- Multilingual = chunks in different languages
- Code-switched = multiple languages, same utterance

# Multilingual input

Text reading assistance: 昨日すき焼きを食べました。

*ENGLISH*

*JAPANESE*

El chico no tiene en la cabeza nada mas que el negocio. Der Junge hat ja nichts im Kopf als das Geschäft.

*SPANISH*

*GERMAN*

La signora lesse il messaggio e volse a Daisy uno sguardo di intesa. The lady read the message and looked up at Daisy in a knowing way.

*ITALIAN*

*ENGLISH*

Sliding window approach (e.g. [Kocmi and Bojar 2017](#))

## Code-switched input

➤  <USER> delete that tweet.. ya lo hize

➤  {eng\_Latn, spa\_Latn}

[Burchell et al. \(2024\)](#)

## Code-switched input: why is it hard?

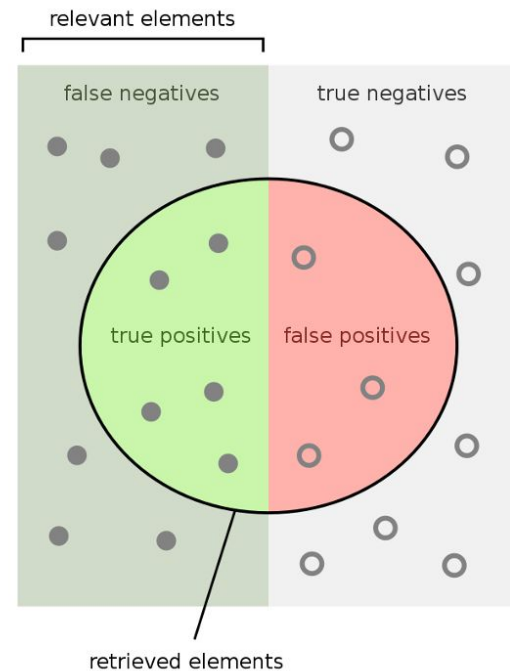
- Many labels  
= sparsity
- Short switches  
= not enough context

Example	Predictions	
	OpenLID	MultiLID
bir kahve dükkanında geçen film tadında güzel bir şarkıya ayrılısın gece <i>falling in love at a coffee shop</i>	Turkish	English & Turkish
<i>haters gon hate players gon play live a life man good luck mic drop</i> tam beklediğim gibi cikti çok efsane	English & Turkish	English
deri ceket sezonu acilsinnnnnnn <i>cool kids of bursaaaaa</i>	Standard Latvian	Latgalian & Wolof

Table 9: Examples from the Turkish–English test dataset where the gold labels are ‘English & Turkish’. English text is rendered *in italics* to distinguish it from Turkish.

# Code-switched input: why is it hard?

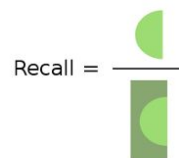
- Not enough test sets 🤔
- Multi-label tasks need different metrics
  - Big problem = irrelevant predictions
  - Precision & recall ignore this
  - Need metrics to reflect downstream (e.g. Hamming loss)



How many retrieved items are relevant?



How many relevant items are retrieved?



# Code-switched input: recommendations

If I did it again...

- Targeted crawls
- Two-stage classifier  
(maybe binary CS/not CS?)
- Pretrain on artificial code-switched data

Me giving  
advice to my  
PhD friends

Me working  
on my PhD





## Dialects/close languages



**Gretchen McCulloch** @gretchenmcc.bsky.social

@GretchenAMcC · [Follow](#)



tired: a language is a dialect with an army and a navy  
wired: a language is just some dialects in a trenchcoat

7:39 AM · Feb 20, 2022



713



Copy link

[Read 4 replies](#)

Dialects / close languages: what's the problem?

Big issue: single or multi-label?

- Some only valid in one dialect
- Some are valid in both
- Some depend on context

...plus a lack of data/test sets obviously



## Dialects/close languages: Arabic

- Arabic-speaking world = diglossia
  - Formal: Modern Standard Arabic
  - Spoken: many 'dialects'
- Dialect continuum
  - No hard borders
  - Intelligibility?
  - No standard orthography

**Arabic teacher: repeat after me, "ka-thī-ran"**

**Students: "kathīran".**

**Arabic teacher: Good! Now you know how to say "very much" in Arabic.**

**Students: Yaaay!**

***Arabic dialects:***



## Dialects/close languages: Arabic

- [OpenLID](#) did poorly on Arabic
  - Overall F1: 0.93 🥰
  - Arabic F1: 0.23 😓
- After data clean-up and two-stage models, still 😞
- ★ Suit labelling to task



Bonus: non-standard  
orthography?

Maybe PIXELS can help?

# Recommendations

## Recommendations

1. Check for reliable coverage of your language(s)
2. Think about speed/capacity trade-off
3. Remember softmax has to predict something
4. Check test set domain
5. False positive rate is key for dataset building

## Recommendations

6. Annotators are only gold for their native language  
(but something > nothing)
7. Be suspicious of short text (need >5-7 tokens)
8. Targeted collection for code-switching
9. Dialect labelling depends on downstream task
10. LOOK AT YOUR DATA



# Recommendations

1. Check for reliable coverage of your language(s)
2. Think about speed/capacity trade-off
3. Remember softmax has to predict something
4. Check test set domain
5. False positive rate is key for dataset building
6. Annotators are only gold for their native language (but something > nothing)
7. Be suspicious of short text (need >5-7 tokens)
8. Targeted collection for code-switching
9. Dialect labelling depends on downstream task

**10. LOOK AT YOUR DATA**

# Thank you!

laurie.burchell@ed.ac.uk



Please check out the Open Language Data Initiative! 