MUNI
FI

Evaluating Training Data Construction Strategies for Token-Level Language Identification

Outline

- Development of the Langtok model
- Strategies for creating the training dataset
- Evaluation datasets
- Results

Langtok model

- identification of the language for each token on the input

MUNI
FI

Why is this important?

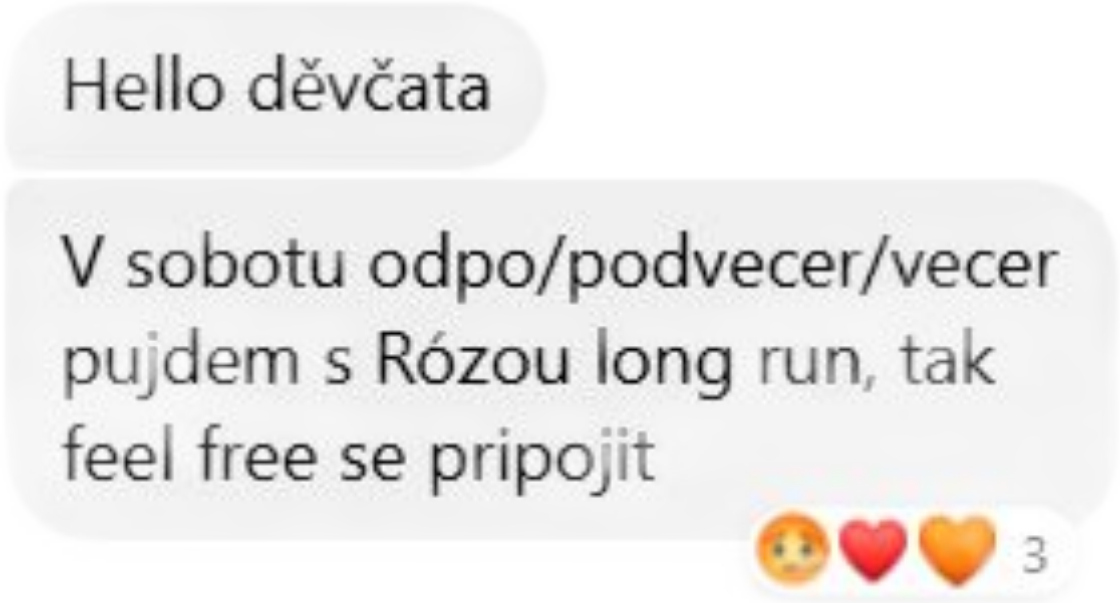
- Code-switching

> Process of shifting from one linguistic code (a language or dialect) to another, depending on the social context or conversational setting.
>
> Britannica

Examples of code-switching

REAL EXAMPLES OF WRITTEN INFORMAL CODE-SWITCHING

Hello



long run

feel free

MUNI FI

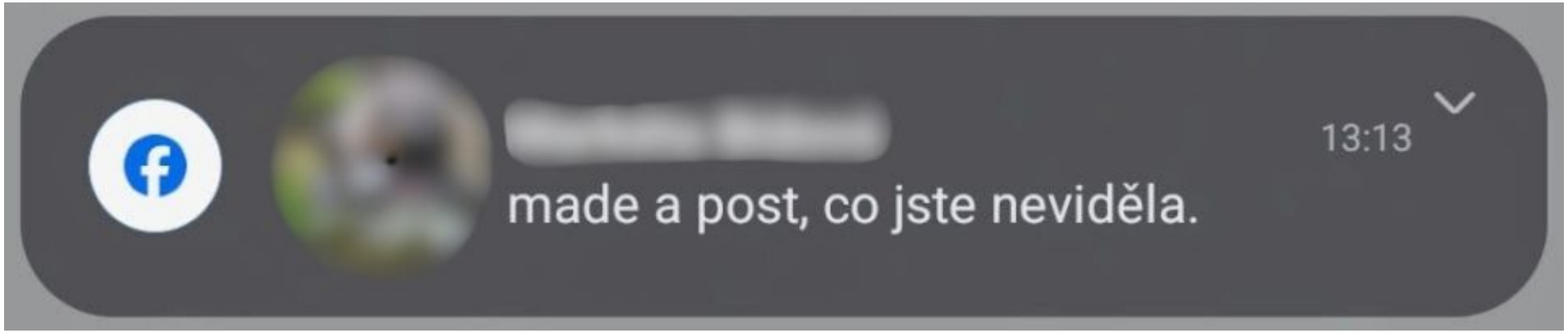
Examples of code-switching

EXAMPLES OF CODE-SWITCHING IN LITERATURE

- "**Jáva**. Borneo. Philippines. Fidji Islands. Solomon Islands. Carolines. Samoa. Damned Clipperton Island. A lot of damned islands, **chlapče**. **Proč?**"
- "Dvanáct tisíc **tons**, mládenče."
- Ale ty je dáš na loď, **see**? Ty koupíš tu loď, ty budeš ten **ship-owner** a možeš jet s sebou; ja, to možeš, abys viděl, jak tě na ní hospodařím. Ale ty peníze, co tam uděláme, ty budou **fifty-fifty**. To je poctivé **business**, ne?

(Karel Čapek, War with the Newts)

Examples of code-switching



MUNI
FI

Development

- Fine-tuning a pre-trained language model on synthetically created dataset

```
{
    'tokens': [
        'Stále', 'se', 'pokoušejí', 'určit', ',', 'jak',
        'velký', 'was', 'the', 'impact', 'and', 'jaký',
        'bude', 'mít', 'vliv', 'na', 'Zemi', '.'
    ],
    'labels': [
        'ces', 'ces', 'ces', 'ces', 'ces', 'ces',
        'ces', 'eng', 'eng', 'eng', 'eng', 'ces',
        'ces', 'ces', 'ces', 'ces', 'ces', 'ces'
    ]
}
```


MUNI FI

Training dataset strategies

- **random** - inserting random snippets on random positions
- **parallel** - inserting snippets from parallel sentences (translations of the original sentence)
- **translated** - replacing snippets by their translations

- raw data from **FLORES-200 dataset**

Training dataset strategies

- **random** - inserting random snippets on random positions

deu

Die Feuerwehr konnte das Feuer schließlich um 23:35 Uhr löschen.

slk

Ľudia každé ráno kvôli práci opúšťajú malé vidiecke mestá a prechádzajú okolo nich ďalší, ktorí pracujú tam, odkiaľ práve tí prví odišli.

Die Feuerwehr konnte das Feuer schließlich um 23:35 Uhr tam odkiaľ práve tí löschen.

Training dataset strategies

- **parallel** - inserting snippets from parallel sentences

eng **During his trip, Iwasaki ran into trouble on many occasions.**

pol **W trakcie swej wyprawy Iwasaki raz po raz wpadał w tarapaty.**

During wpadał w tarapaty his trip, Iwasaki ran into trouble on many occasions.

Training dataset strategies

- **translated** - replacing snippets by their translations

deu

Der Film mit Ryan Gosling und Emma Stone hat Nominierungen in allen Hauptkategorien erhalten.

deu

allen Hauptkategorien erhalten

ricevuto in tutte le categorie principali

ita

Der Film mit Ryan Gosling und Emma Stone hat Nominierungen in ricevuto in tutte le categorie principali.

MUNI FI

Training dataset strategies

VARIABLES

- **number of snippets** inserted into the original sentence
- **maximal length** of snippets inserted into the original sentence
- proportion of **snippets of length 1**

Training dataset strategies

VARIABLES

- **number of snippets** inserted into the original sentence

The Babylonians built each of their atrakcím$_{ces}$ gods
das was von Außenstehenden$_{deu}$ a primary weten dat het enorm$_{nld}$
temple that was considered the home of the god.$_{eng}$

Úkaz, ktorý tím očakával' by peut être placé$_{fra}$ spôsobili prílivové sily
medzi tmavou hmotou galaxie a tmavou hmotou Mliečnej dráhy.$_{slk}$

Training dataset strategies

VARIABLES

- **maximal length** of snippets inserted into the original sentence

Historycy krytykowali dotychczasową politykę FBI, skupiającą zasoby na sprawach łatwych do rozwiązania, zwłaszcza kradzieżach samochodów, w celu zwiększenia wskaźnika began it was to last for the next years and would be$_{eng}$ stopnia skuteczności agencji.$_{pol}$

Jelikož bylo k dispozici pouze osmnáct medailí denně, některým zemím se nepodařilo dostat como$_{spa}$ se na stupně vítězů.$_{ces}$

MUNI FI

Training dataset strategies

VARIABLES

- proportion of **snippets of length 1**
  - 30 % of islands in two code-switching datasets were of length 1
  - in training datasets marked with label *w* (= weighted), 30 % of the snippets have length 1
  - the proportion of snippets of the remaining lengths is distributed equally

Training dataset strategies - datasets

- random

dataset name	max snippets count	max snippet length	weighted
random-S1-L12-w	1	12	yes
random-S1-L12-e	1	12	no
random-S1-L06-w	1	6	yes
random-S1-L06-e	1	6	no
random-S1-L03-w	1	3	yes
random-S1-L03-e	1	3	no
random-S3-L06-w	3	6	yes
random-S3-L06-e	3	6	no
random-S3-L03-e	3	3	no

Training dataset strategies - datasets

- parallel
- translated

dataset name	max snippets count	max snippet length	weighted
parallel-s	1	6	yes
parallel-l	1	10	no
translated-s	1	6	yes
translated-l	1	10	no

Evaluation datasets

- **Denglisch dataset**
  - English and German code-switching
  - posts on Reddit
- **Bangor Miami corpus**
  - English and Spanish code-switching
  - transcripts of audio recordings

Denglish dataset examples

PS | : | Your | Deutsch | could | auch | use | some | improvements | .

Yes | ok | ja | I | have | to | gebe | zu | that | meine | geographiekenntnisse | now | not | the | beste | sind | , | especially | when | es | schon | so | late | ist | und | ich | gleich | sleepen | möchte | .

Bangor Miami corpus examples

Yeah | it’s | like | but | it’s | real | it’s | real | sand | .

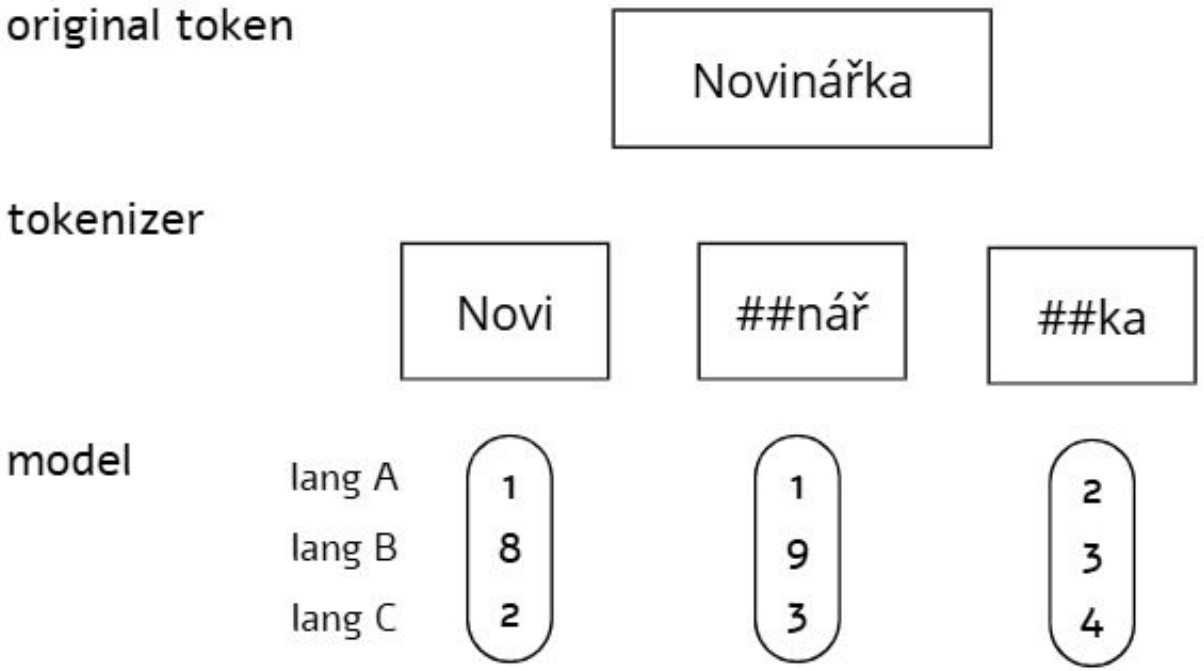
Ah | tú | solo | ?

And | bro | at | this | point | like | bro | the | most | ridiculous | pain | in | my | life | .

ay | dios | mío | how | embarrassing | !

Problem with token-level annotations

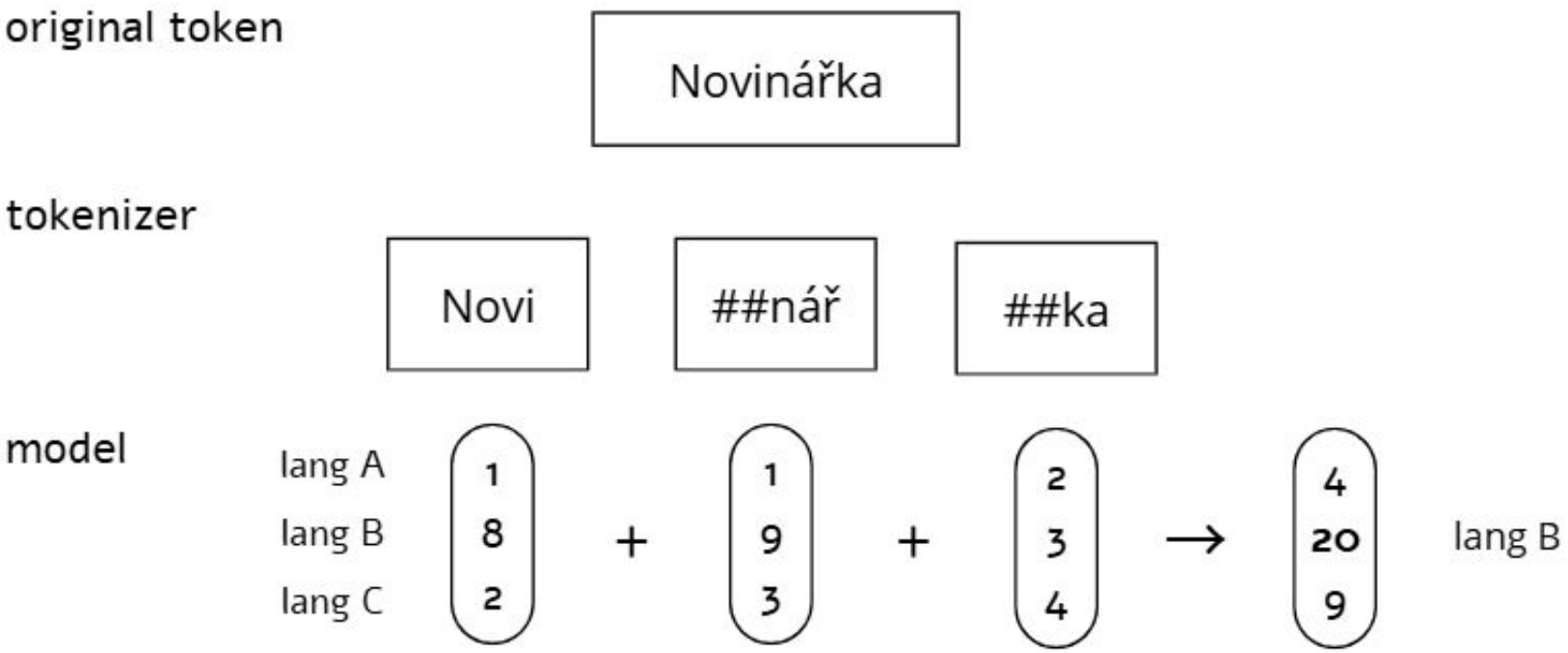
- Merging the model's outputs to match the original tokens



MUNI FI

Problem with token-level annotations

- Merging the model's outputs to match the original tokens

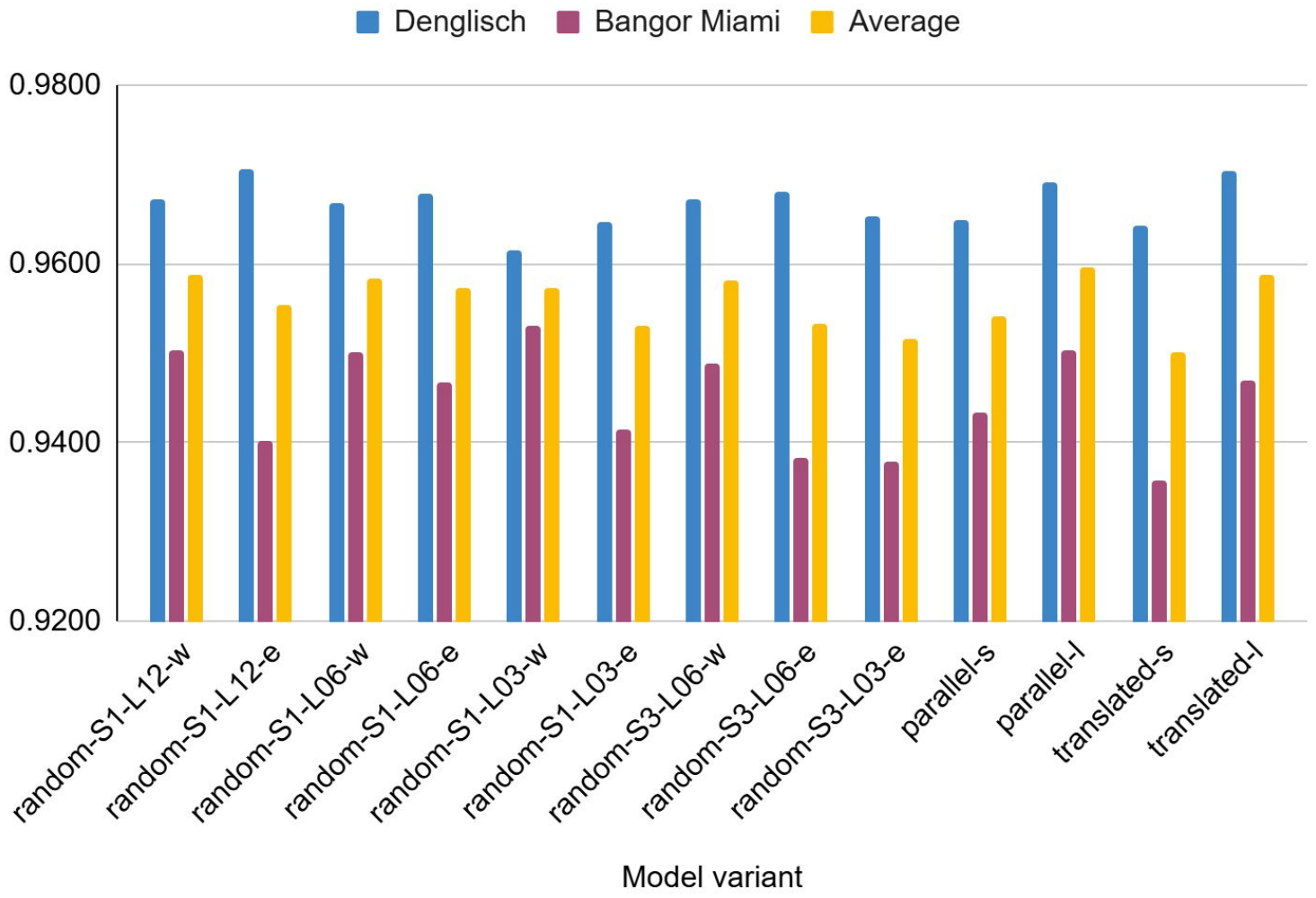


MUNI FI

Training dataset strategies - evaluation

model variant	Denglisch dataset	Bangor Miami corpus	Average
random-S1-L12-w	0.9672	0.9503	0.9588
random-S1-L12-e	**0.9707**	0.9403	0.9555
random-S1-L06-w	0.9667	0.9501	0.9584
random-S1-L06-e	0.9678	0.9467	0.9573
random-S1-L03-w	0.9616	**0.9530**	0.9573
random-S1-L03-e	0.9646	0.9414	0.9530
random-S3-L06-w	0.9673	0.9488	0.9581
random-S3-L06-e	0.9681	0.9384	0.9533
random-S3-L03-e	0.9654	0.9378	0.9516
parallel-s	0.9648	0.9433	0.9541
parallel-l	0.9691	0.9503	**0.9597**
translated-s	0.9643	0.9358	0.9501
translated-l	0.9704	0.9470	0.9587

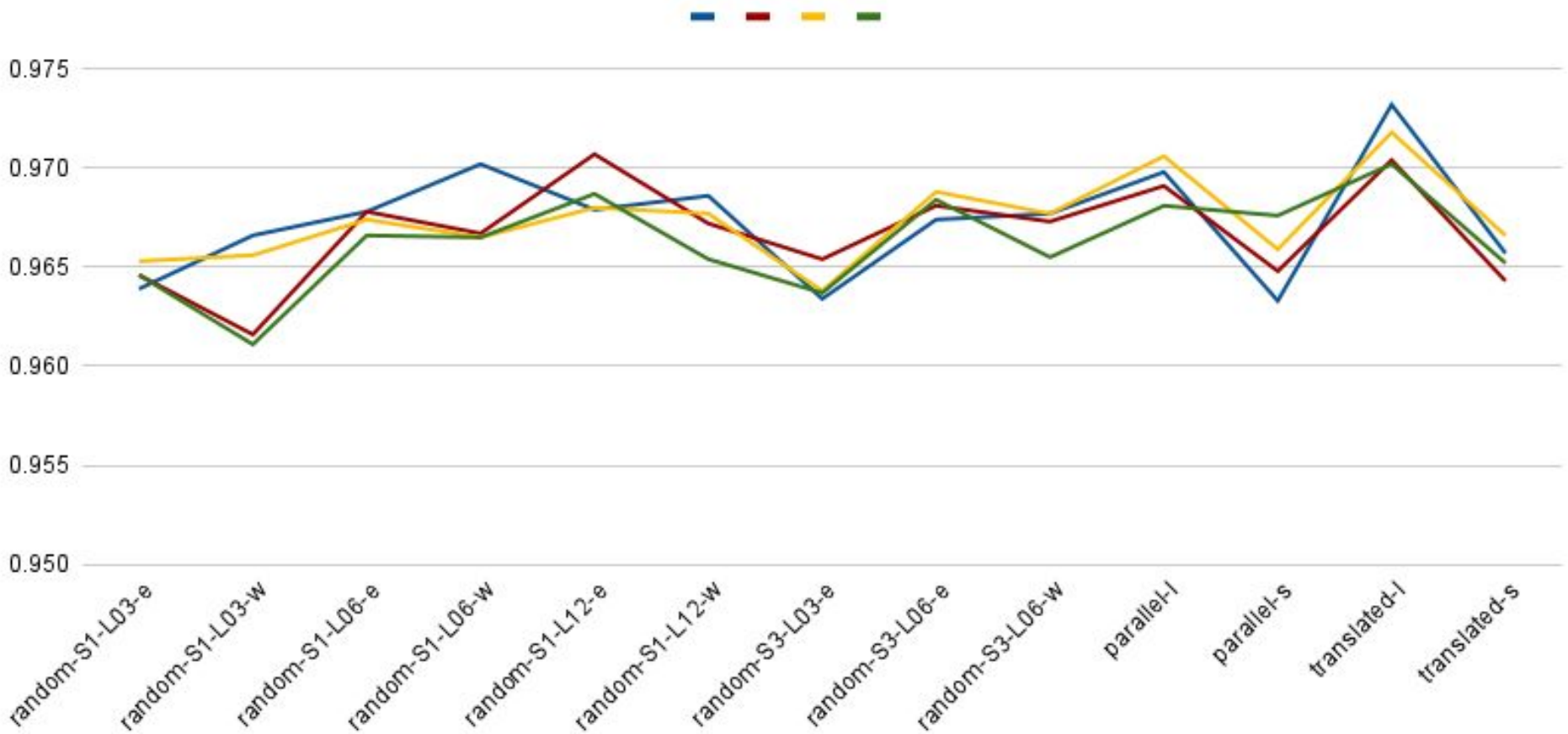
Training dataset strategies - evaluation



MUNI FI

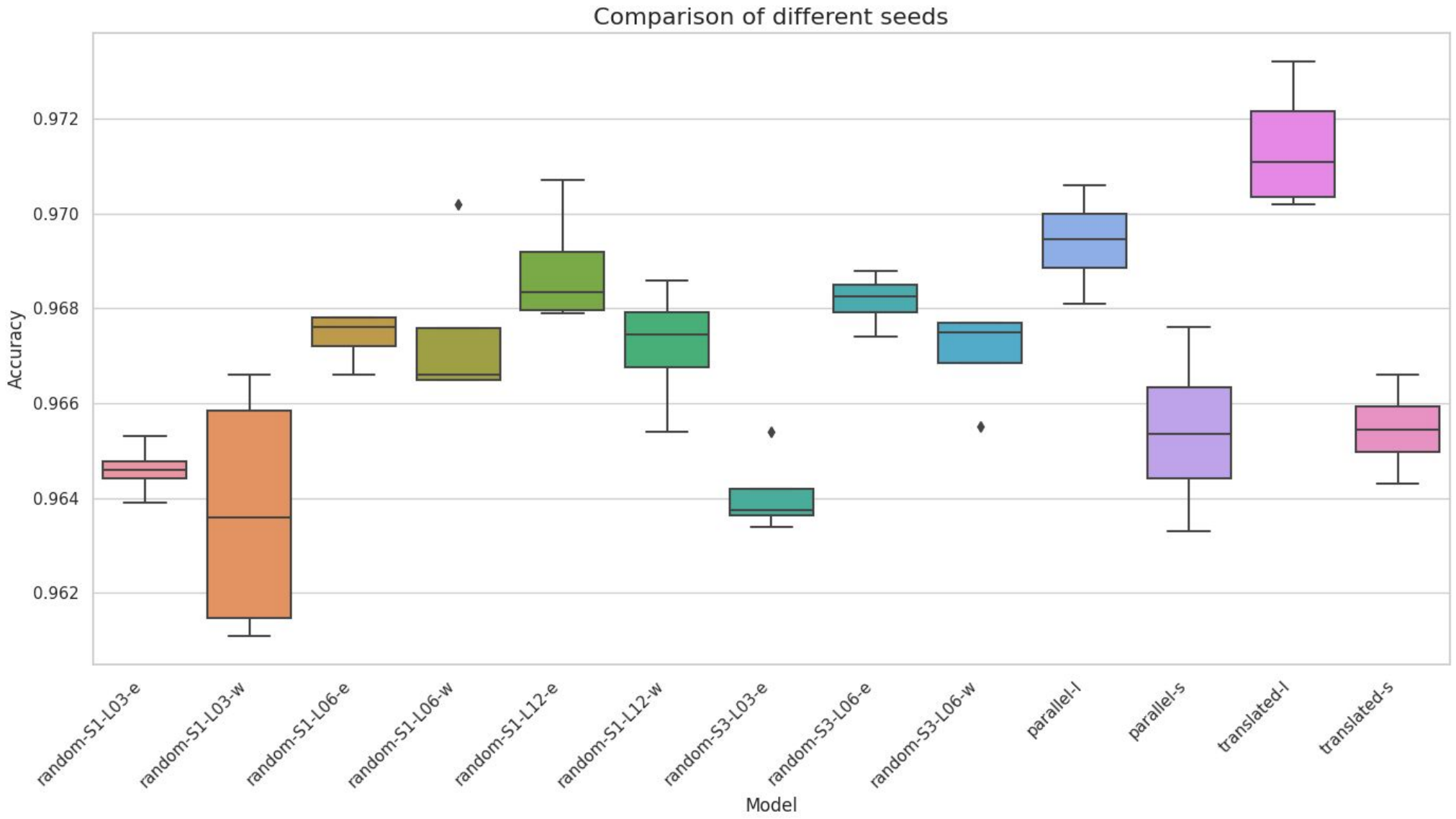
Comparison of different seeds

DENGLISCH DATASET



MUNI FI

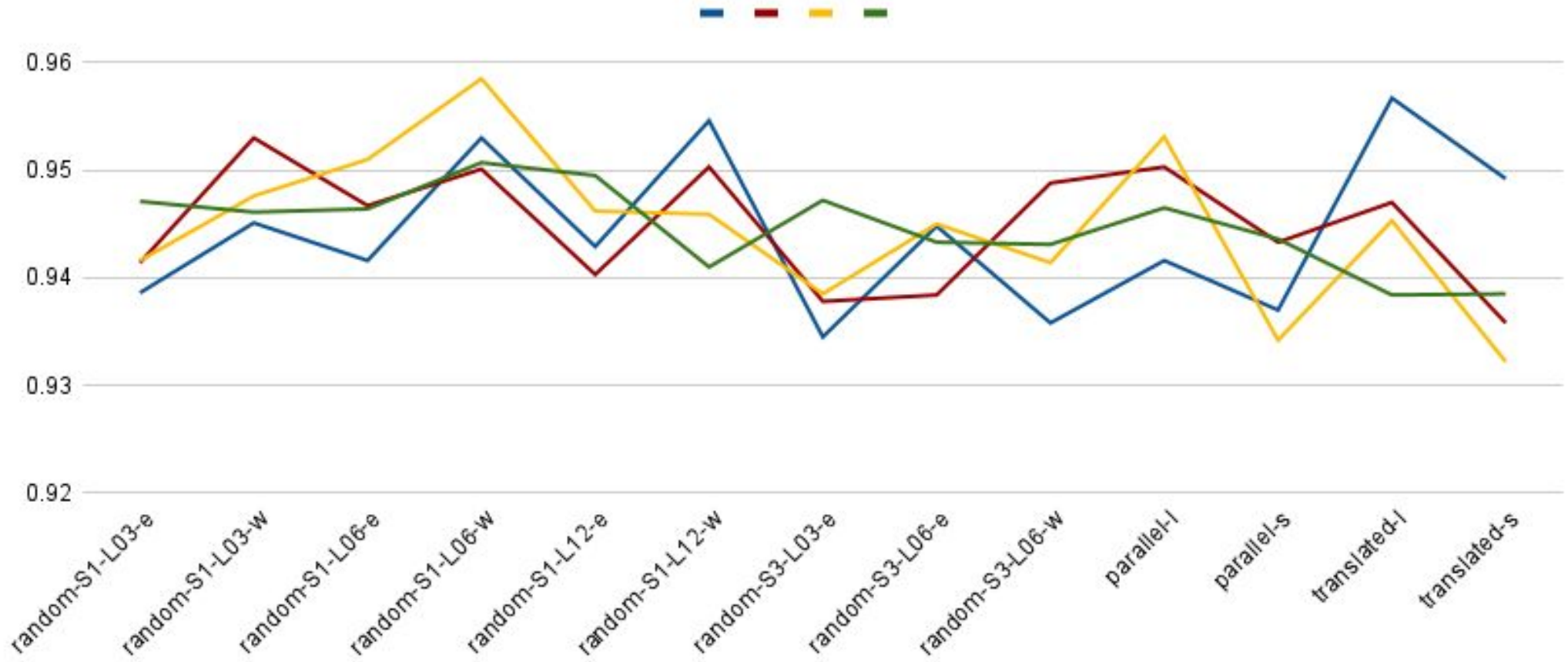
DENGLISCH DATASET



MUNI FI

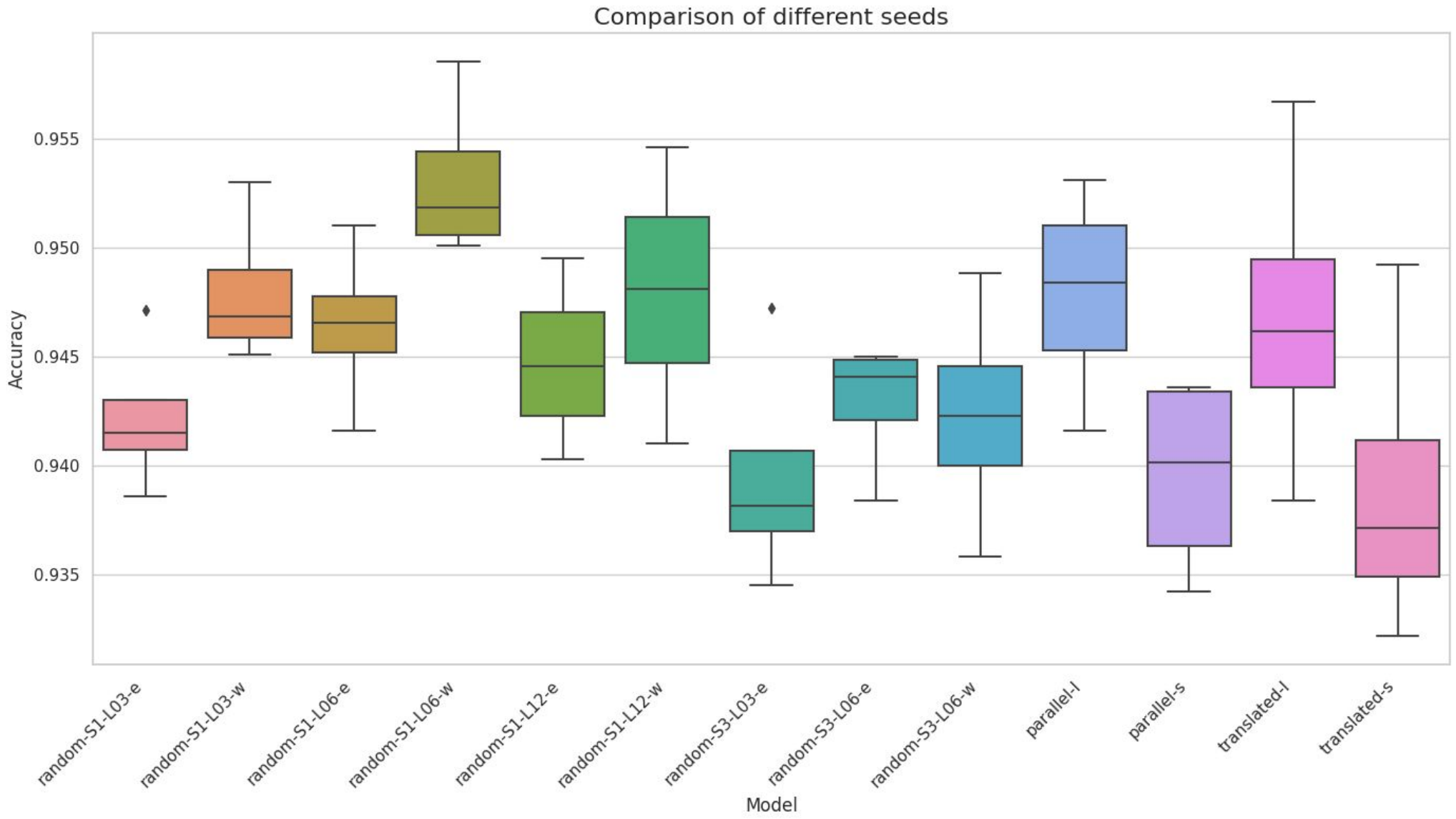
Comparison of different seeds

BANGOR MIAMI CORPUS



MUNI FI

BANGOR MIAMI CORPUS



MUNI FI

Conclusion

- Due to the small number of datasets, it is difficult to evaluate the results
- For parallel and translated strategies, longer snippets are preferable

Links and Resources

- FLORES-200 (Nature.com)
- Denglisch dataset (ACL Anthology)
- Bangor Miami corpus (TalkBank)

Training dataset strategies

- **translated**

fra Golsling et Stone ont respectivement été nominés pour le Meilleur Acteur et la Meilleure Actrice.

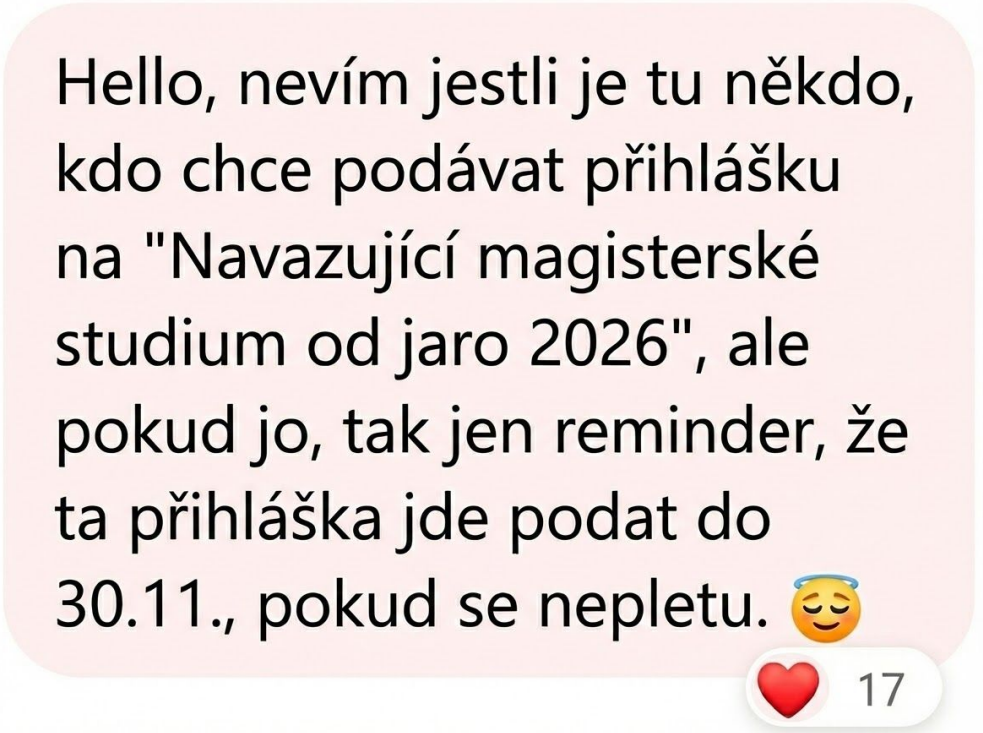
fra Stone ont → Kamién ma pol

Golsling et Kamień ma respectivement été nominés pour le Meilleur Acteur et la Meilleure Actrice.

Examples of code-switching

REAL EXAMPLES OF WRITTEN INFORMAL CODE-SWITCHING

Hello



reminder

MUNI FI

Examples of code-switching

REAL EXAMPLES OF WRITTEN INFORMAL CODE-SWITCHING

MUNI FI

Examples of code-switching

REAL EXAMPLES OF WRITTEN INFORMAL CODE-SWITCHING

- Jak moc ***hard*** jsou ty odpovedniky?
- A jak jste vybirali ***event***? Podle tech ***tickets***?
- Tvuj ***phone*** umi platit?
- Je to nejaky ***close cycle***
- Okej pohoda ***no stress***
- Ja jsem ***sick*** zase, ***no school today***

Thank you for your attention

MUNI
FI