Mapping Faroese in the Multilingual Representation Space: Insights for ASR Model Optimization

Dávid í Lág, Barbara Scalvini, Jón Guðnason
University of the Faroe Islands

[Motivation]

- Data-driven way to select close languages to Faroese
- Used for Cross-linguistic transfer to boost Faroese ASR

[Model]

- Wav2Vec 2.0 XLS-R 53
- Multilingual with 53 languages
- 24 hidden layers

[Data sets]

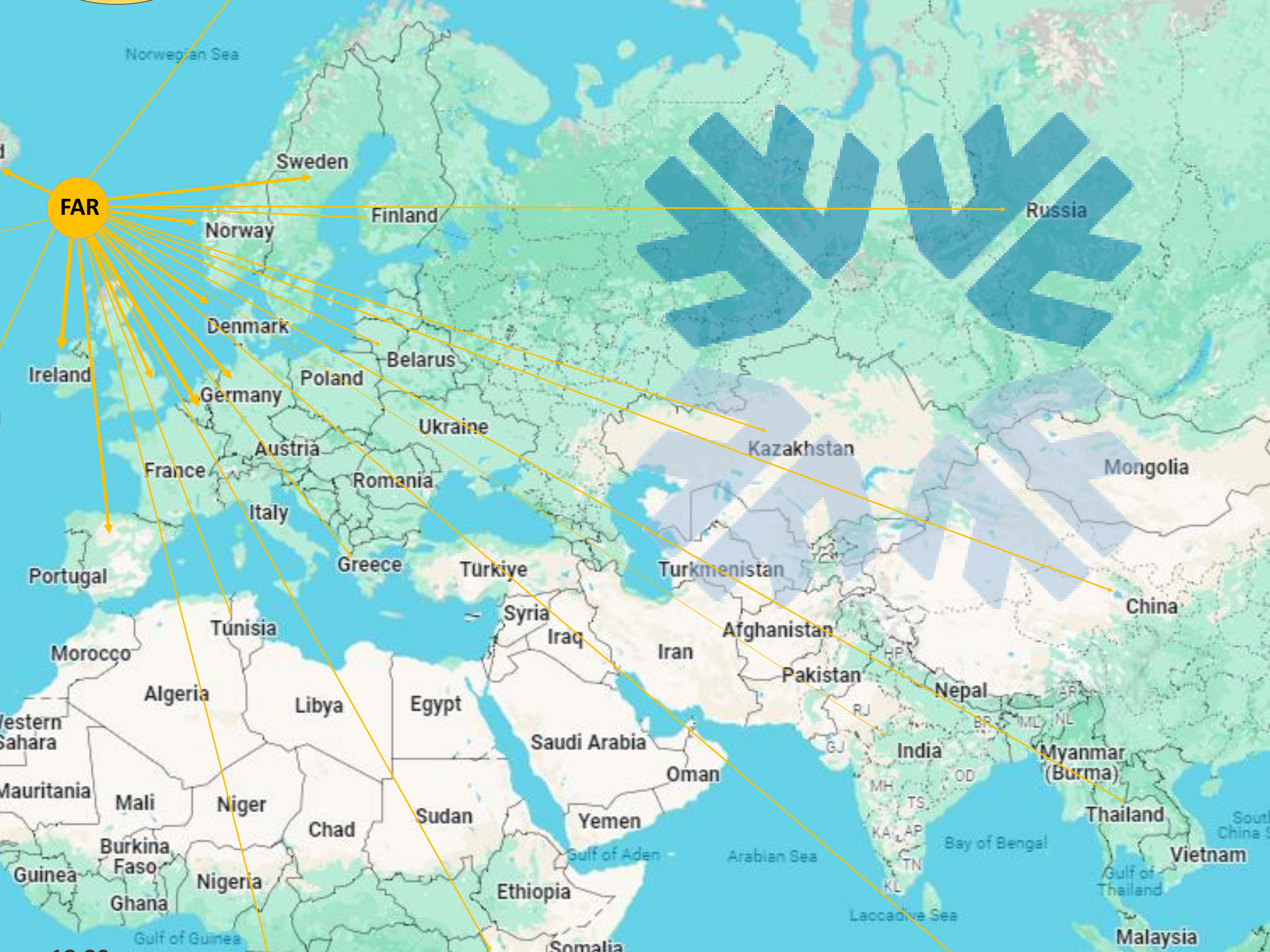
- Google Fleurs (all 102 languages)
- Ravnursson (Faroese)

[Method]

- Wav2Vec2 representation space
- Same 900 sentences for all 102 Fleurs languages
- Averaging of sentence-level representation per language per layer:

$$\mu_{l,j} = \frac{1}{N}\sum_{i=1}^{N} R_{l,i,j},$$

- Euclidean distance measured in the original representation vector space
- Clustering with PCA, t-SNE, UMAP



Clusters in layers 18-20

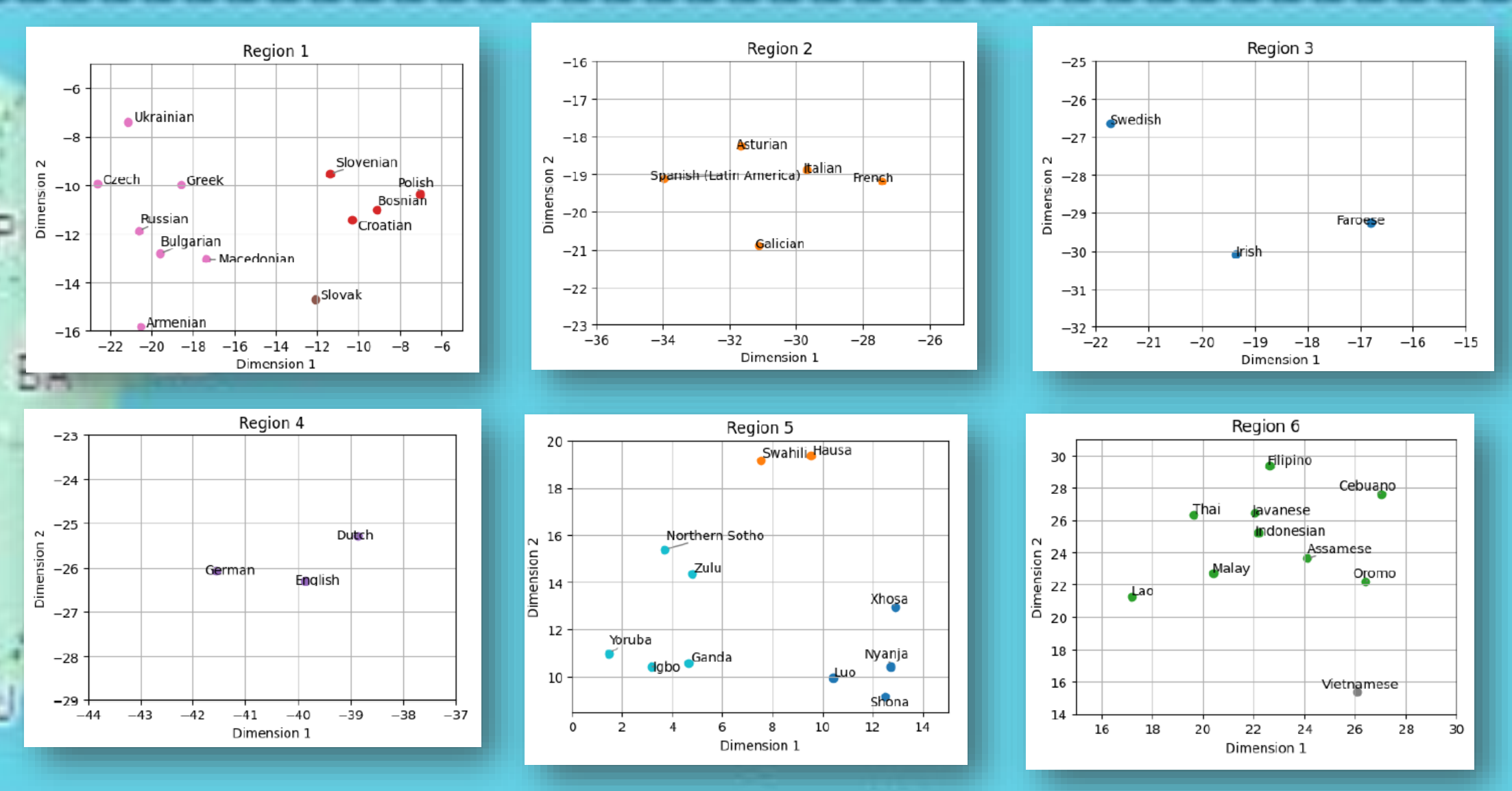


Figure 1: Clusters of closely related languages for layers 18-20 with t-SNE and K-Means with 18 clusters

Closest languages

Layers	0-4	5-9	10-14	15-19	20-24	0-24
1	Irish (10.8)	Irish (13.4)	Irish (13.4)	Irish (16.2)	Welsh (31.7)	Welsh (14.8)
2	German (11.3)	German (15.4)	Estonian (15.4)	Croatian (17.0)	Turkish (34.7)	Turkish (17.5)
3	Romanian (11.6)	Estonian (16.0)	Croatian (15.8)	Estonian (17.4)	Punjabi (47.4)	Punjabi (22.6)
4	Estonian (11.8)	Croatian (16.2)	Lithuanian (15.9)	Lithuanian (17.5)	Slovak (104.0)	Slovak (25.2)
5	Simplified Chinese (11.8)	Romanian (16.2)	Welsh (16.1)	Polish (17.7)	Georgian (110.1)	Georgian (25.8)
6	Catalan (12.0)	English (16.2)	Romanian (16.1)	Georgian (17.9)	Amharic (112.7)	Amharic (27.4)
7	Korean (12.1)	Welsh (16.4)	Polish (16.5)	Romanian (18.0)	Norwegian (126.4)	Norwegian (29.8)
8	Armenian (12.3)	Lithuanian (16.4)	Swedish (16.6)	Slovenian (18.0)	Vietnamese (145.8)	Armenian (32.5)

Table 1: Closest languages to Faroese measured in Euclidean distance in the original representation vector space

[Results]

- Proximity to Gaelic, Germanic and Nordic languages
- Irish and Welsh (Celtic) are consistently close to Faroese
- Layers 18-20 show close proximity to Irish and Swedish
- Clustering of language families is happening in Wav2Vec2 representation space

[Limitations]

- Only one model and one data set used
- Euclidian distance and the curse of dimensionality

HÁSKÓLINN Í REYKJAVÍK
REYKJAVIK UNIVERSITY
FRÓÐSKAPARSETUR FØROYA
UNIVERSITY OF THE FAROE ISLANDS