Neural machine translation for low-resource and variation-rich languages

Yves Scherrer, University of Helsinki (+ soon University of Oslo)
yves.scherrer@helsinki.fi

University of Salzburg, 25 May 2023

Outline

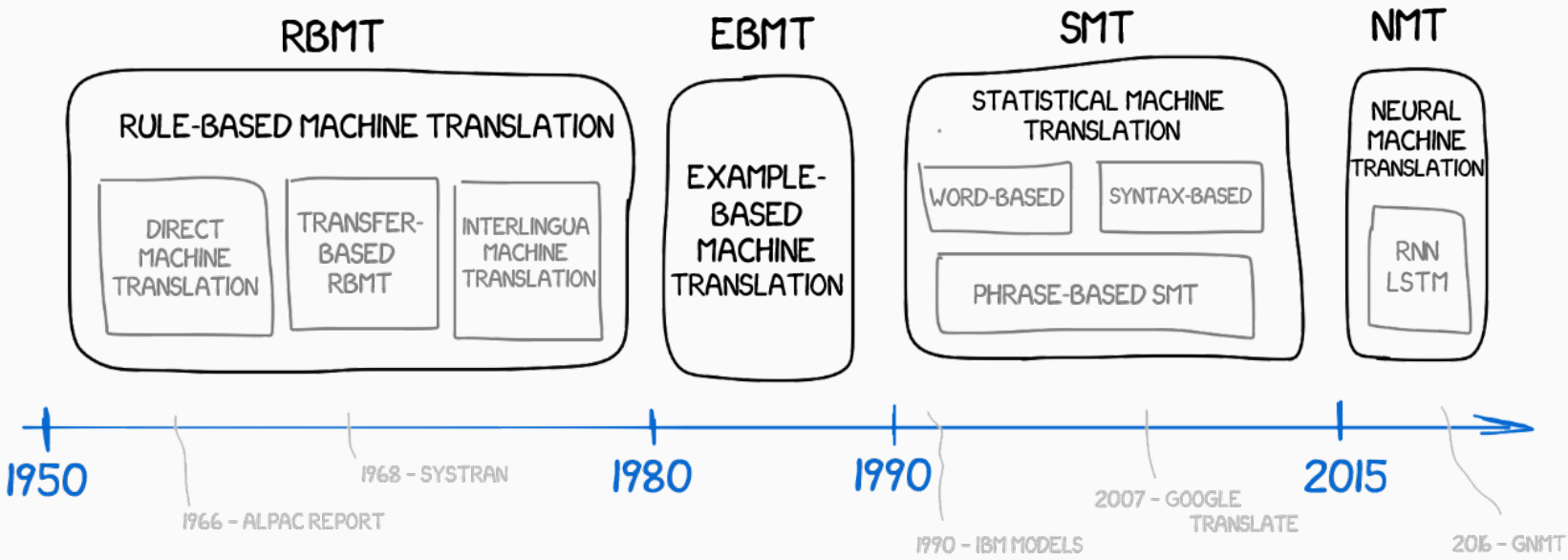
1. Neural machine translation, sequence-to-sequence models and encoder-decoder architectures

2. NMT for low-resource languages:
   The AmericasNLP shared tasks

3. NMT for variation-rich languages:
   Normalization of Finnish and Norwegian dialects

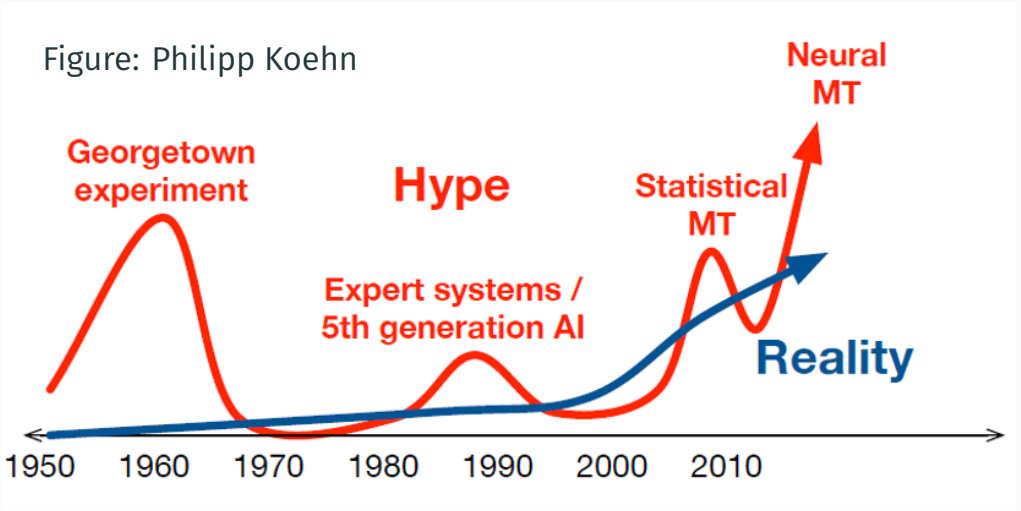
Neural machine translation, sequence-to-sequence models and encoder-decoder architectures

A brief history of machine translation



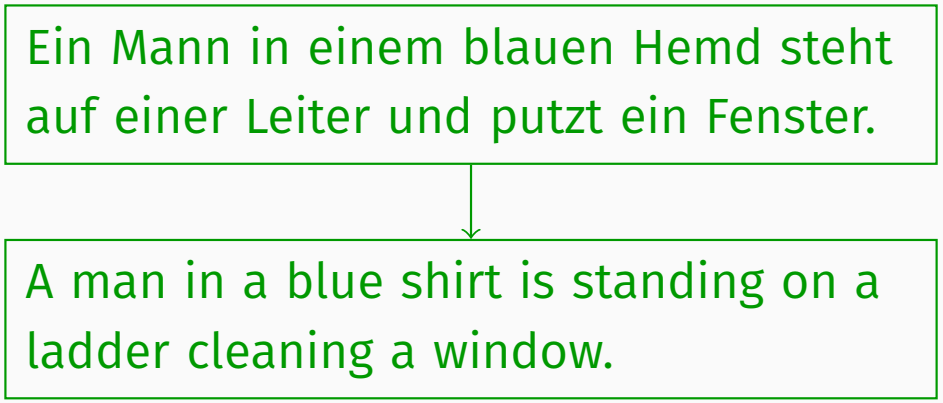
https://vas3k.com/blog/machine_translation/

Figure: Philipp Koehn



Machine translation

Machine translation is a **sequence-to-sequence transformation** (seq2seq) problem:



Input: a sequence of words in the **source language**

Output: a sequence of words in the **target language**

- No 1-to-1 mapping:
  - input and output may have different lengths
  - the word order may vary
- Successful translation if:
  - the output is grammatically correct
  - the output conveys the same meaning as the input

Data-driven machine translation

Statistical and neural machine translation systems are trained using large amounts of data:

- **Parallel corpora** or **bitexts** are sets of sentence pairs with the same meaning.

Popular sources of parallel corpora:

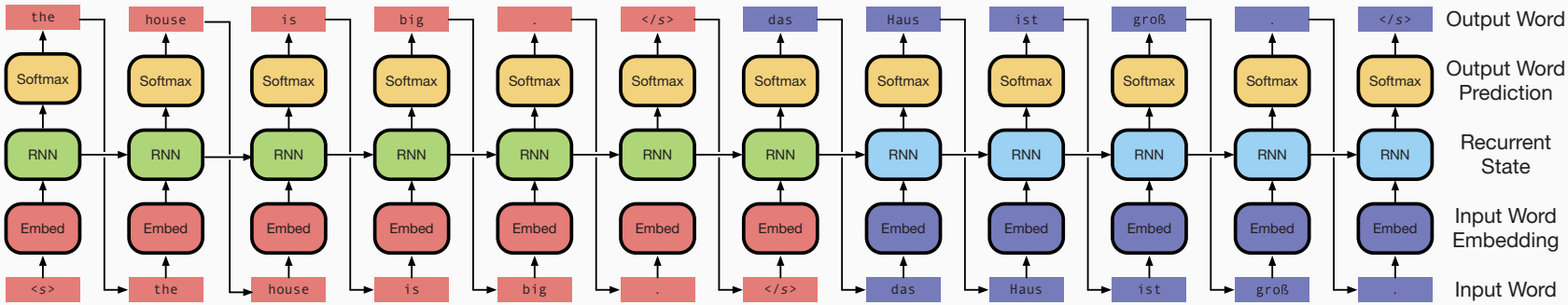
- Proceedings of the European Parliament
- Movie subtitles
- Bibles (and other translated literature)
- Multilingual websites

Encoder-decoder architectures

Neural machine translation systems are based on the so-called **encoder-decoder architecture**.

The **encoder** produces abstract numeric representations of the input sequence.

The **decoder** converts the abstract numeric representations into the output sequence.



the house is big. ⟶ *das Haus ist groß.*

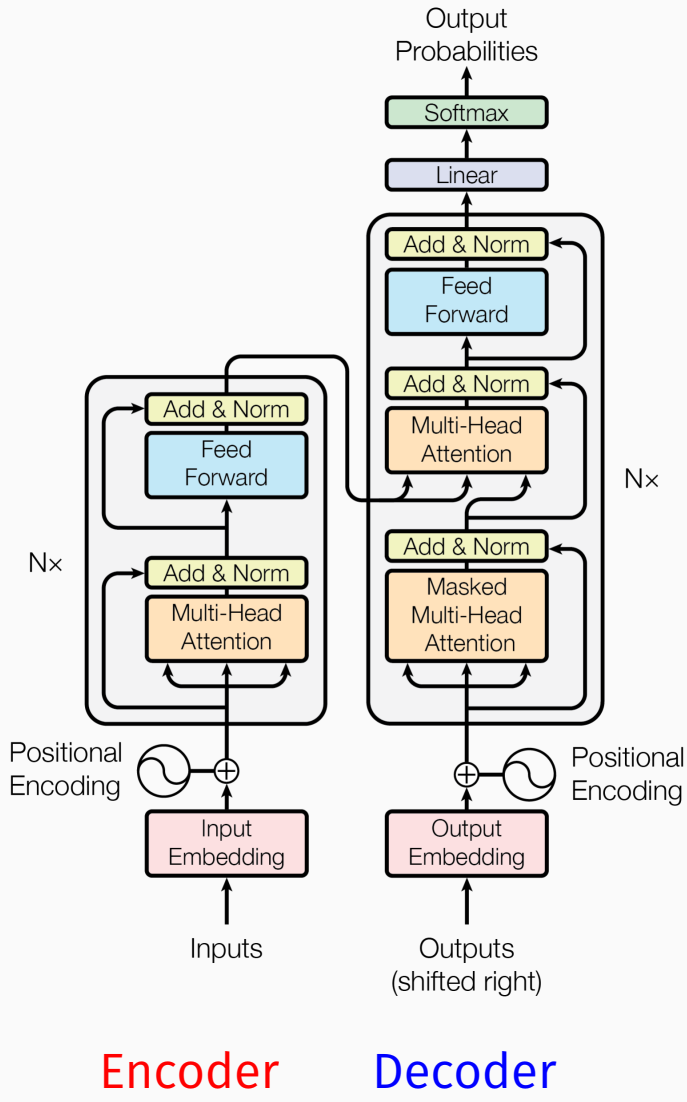
- Decoding only starts when the source sentence is fully encoded.

Encoder-decoder architectures

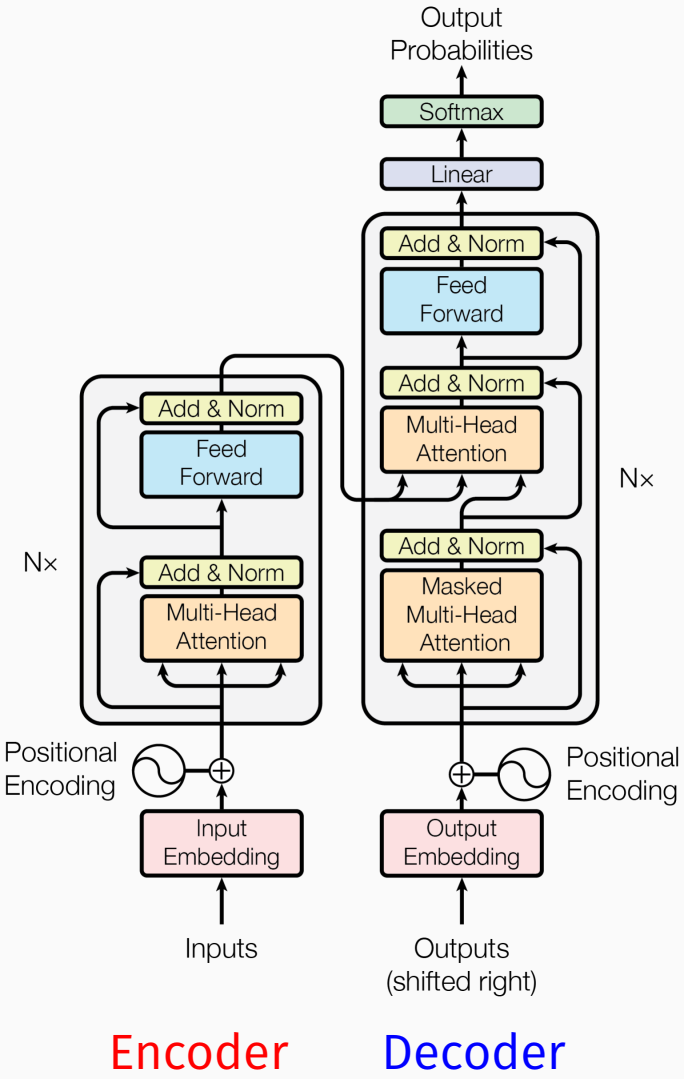
Encoder-decoder architectures vary according to:

- encoder and decoder types (RNN, CNN, bidirectionality, self-attention)
- connections between encoder and decoder (attention mechanism)

The most popular architecture nowadays is called **Transformer** (Vaswani et al. 2017).



The Transformer architecture



Although the Transformer was initially developed for sequence-to-sequence transformations, its building blocks are used in most current pre-trained models:

- **BERT models** only contain a Transformer encoder
- **GPT models** only contain a Transformer decoder

Sequence-to-sequence transformation problems

- Machine translation
  Ein Mann in einem blauen Hemd steht auf einer Leiter und putzt ein Fenster. → A man in a blue shirt is standing on a ladder cleaning a window.
- Paraphrasing
  I will not let you down. → I won't disappoint you.
- Grammatical error correction
  We had enjoy time. → We had a great time.
- Historical text normalization
  Ce ſeroit une marque de la force de voſtre merite pluſtoſt que de ma facilité. → Ce serait une marque de la force de votre mérite plutôt que de ma facilité.
- Dialect-to-standard normalization
  ich ha das ales inere kasette won ich de schlüssel nüme ha dezue → ich habe das alles in einer kassette wo ich den schlüssel nicht mehr habe dazu

"Monolingual translation tasks"

Sequence-to-sequence transformation problems

- Lemmatization
- Grapheme-to-phoneme conversion
- Transliteration
- Speech recognition (speech-to-text)
- Sign language transcription

Character-level transduction tasks

Multimodal settings

Machine translation is one of several seq2seq problems.

Seq2seq problems are implemented by neural encoder-decoder architectures such as the Transformer.

One model architecture (and software toolkit) is sufficient for all tasks. The only thing that changes is the training data.

Multilingual machine translation

One model can learn to translate between multiple language pairs. Two “special words”, called **language labels**, are appended to each sentence to tell the model about the languages used:

<FROM_ES> <TO_FR> Visitaré a los niños.	Je viendrai voir les enfants.
<FROM_EN> <TO_ES> You did well, you did very well.	Bien hecho. Genial.
<FROM_ES> <TO_EN> Llegaremos enseguida.	We will be arriving soon.
<FROM_FR> <TO_ES> C’est la voix de notre âme qui parle.	Es la voz del alma que habla.

The model automatically learns to make use of the language labels when deciding which target words to generate.

Experimentation with sequence-to-sequence models

What can be done to improve performance on a sequence-to-sequence transformation task?

- Data optimization
  - Collection of additional training data
  - Generation of synthetic training data
  - Cleaning and filtering of noisy data

  Easy
- Hyperparameter optimization
  - Number of encoder/decoder layers
  - Size of vectors
  - Learning rate
  - Segmentation of long words

  Easy
- Architectural changes
  - Changes to attention mechanism
  - Shared modules for related languages

  Difficult

NMT for low-resource languages: The AmericasNLP shared tasks

The AmericasNLP shared tasks

Machine translation from Spanish into 11 indigenous languages of the Americas. Two editions, 2021 and 2023:

- Spanish – Chatino (only 2023)
- Spanish – Hñähñu (Otomi)
- Spanish – Nahuatl
- Spanish – Rarámuri (Tarahumara)
- Spanish – Wixarika (Huichol)

Mexico

- Spanish – Bribri

Costa Rica

- Spanish – Shipibo-Konibo
- Spanish – Asháninka
- Spanish – Aymara
- Spanish – Quechua
- Spanish – Guaraní

Peru, Bolivia, Paraguay

The AmericasNLP shared tasks

Results – chrF scores:

	aym	bzd	cni	czn	gn	hch	nah	oto	quy	shp	tar
Baseline 2021	18.8	7.7	10.4	–	22.0	12.6	18.2	5.9	33.1	13.9	4.6
Helsinki 2021	28.3	16.5	25.8	–	33.6	30.4	26.6	14.7	34.3	32.9	18.4
Helsinki 2023	33.4	22.5	28.4	32.1	**40.4**	**32.3**	26.9	**15.3**	33.3	33.4	**19.2**
Best contender 2023	**36.2**	**26.1**	**30.0**	**40.0**	39.3	32.3	**27.3**	14.8	**39.5**	**33.4**	18.7

For comparison – chrF scores:

Spanish–English	60
Chinese–English	49
English–Spanish	56
English–German	66

How did we get there?

Our strategies

- Collection of additional parallel data
- Data augmentation: back-translation and pivoting
- Curriculum and transfer learning
- Knowledge distillation
- Addressing spelling variation and dialectal variation

Additional sources of parallel data

- Constitutions, laws, declaration of human rights
  La Nación Mexicana es única e indivisible. → Mekiku kwieyari kaniyuxewini. (Wixarika)
- Educational material from websites and PDFs
  Dicen los mayores que la danta permanecía allá abajo donde tenía su casa, pues ella es la madre de esta tierra. → Naĩ' che kẃkëpa tö cha, ie' bitsȍ ke cha kȏ dià ẽ, e' ẽ ie' ù mérkĩ. (Bribri)
- Translated news
  La Secretaría de Politicas Lingüísticas premió a 16 promotores de lenguas indigenas. → Paraguái Ñe'ẽnguéra Sambyhyha ombojopói 16 Ypykuéra ñe´ẽ rayhuhárape. (Guaraní)
- Bibles (source and quality sometimes unclear)
  Esta es la lista de los antepasados de Jesucristo, descendiente de David y de Abrahán: → Meeca nosanquenajeitempiro ivajiropee icharinepeeni Jesoquirishito. Tempa irio ishanincani Iravirini aisati Avaramani. (Asháninka)

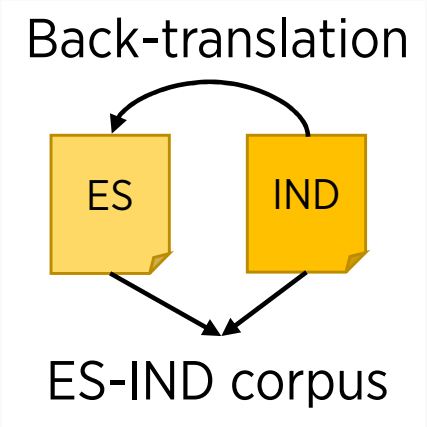
Data augmentation strategies

Back-translation (Sennrich et al. 2016)

If monolingual data is available (e.g. from Wikipedia), translate it to Spanish using a translation system trained in the opposite direction:

Nozan cateh chicuēyimpōhualcaxtōloncē tōnalli (176) xiuhpan. (Nahuatl)
⇒ Las plazas serán públicas el año 176.

This has been shown to be beneficial even if the translation system is of bad quality.

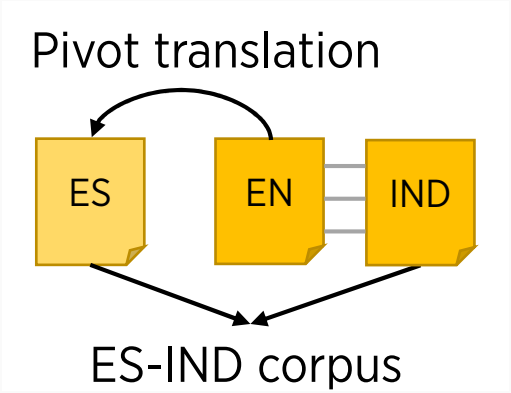


Data augmentation strategies

Pivot translation (e.g. Xia et al. 2019)

Some datasets are parallel with English instead of Spanish. We use an existing English-to-Spanish MT system to create Spanish versions:

How's farm life treating you? I hear its a simple, peaceful existence. It sounds nice. $\rightarrow$
Kunjamakiskis ranja ukanxa? Jasakiw satwa ist'txa. Ukax wali askiwa. (Aymara)
$\Rightarrow$ ¿Cómo te trata la vida en la granja? Escucho que es una existencia simple y pacífica.



Data collection and augmentation

Language	Code	Organizer-provided	Additional non-Bible	Additional Bible
Ashaninka	cni	3,878	8,593	23,321
Aymara	aym	6,039	27,265	92,082
Bribri	bzd	7,490	588	23,103
Chatino	czn	354	4,798	47,570
Guarani	gn	26,012	72,597	23,687
Hñähñu	oto	4,888	8,593	23,849
Nahuatl	nah	15,863	22,558	47,674
Quechua	quy	109,372	209,814	123,829
Raramuri	tar	14,495	2,194	23,678
Shipibo-Konibo	shp	14,553	36,029	47,638
Wixarika	hch	8,960	2,932	23,867

Impact of additional data

Ablation study from 2021:

- Backtranslations help despite their questionable quality (green vs blue)
- Bibles don’t seem to help (blue vs orange)

Modelling choices

Our best model is trained in two phases with different proportions of data:

1. 90% Spanish–English + 10% Spanish–Indigenous (~1% per indigenous language)
2. 37% Spanish–English + 63% Spanish–Indigenous (~6% per indigenous language)

We take advantage of abundant Spanish–English data to train a good Spanish encoder. This is a case of **transfer learning**.

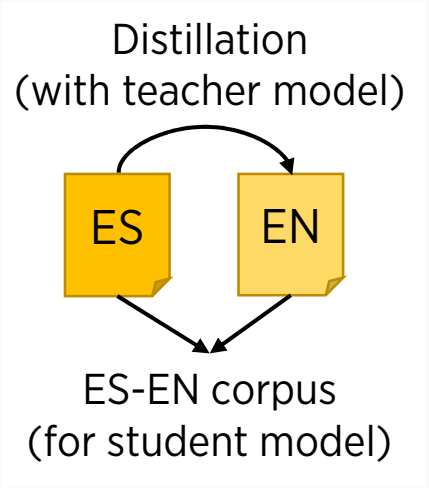
Using different proportions of data in different training phases is known as **curriculum learning**.

Keeping Spanish–English in phase 2 prevents **catastrophic forgetting**.

Another approach: knowledge distillation

The general approach:

1. Find an existing high-quality MT model (the **teacher model**)
2. Find large source language texts and translate them to the target language
3. Use this parallel dataset to train a new **student model**



Why does this work?

- Parallel datasets contain alignment errors and free translations, but model-generated translations are generally quite literal.
- Model-generated translations are easy to produce, and therefore also easy to learn.

Tackling spelling variation

- **Guaraní** uses different diacritics to mark nasal vowels. We normalize all to circumflex: ãäāâ → â
- **Bribri** has two spelling norms, *Constenla* and *Jara*. The task organizers provide a conversion table to an "intermediate form". Our additional datasets come in both norms and had to be converted.
- For **Hñähñu**, the organizers expect translation output in a dialect spoken by 100 elderly people. The training data comes from a different dialect and/or spelling norm. The variants are too different to infer systematic conversions. We remove all diacritics that occur in the training data, but not in the expected output format: öǒòø → o

Tackling spelling and dialect variation

- For **Chatino**, the organizers expect translation output with tone markers as superscripts:
  ntqo^{E} lo^{J} ran^{f} neq^{c} sqen^{E} no^{A} ngwa^{c} junta^{K} wa^{c} 26 qo^{E} koq^{f} sayu^{K} yjan^{A} 2022, tqa^{J} ka^{E} nten^{K} skwen^{B} kwan^{K} yaq^{c}.
  95% of the training data does not contain such markers. We add two **variant labels**, *<default>* and *<plain>*, to help the model distinguish the two variants.
- Most of the **Quechua** data is in Ayacucho Quechua (the standard variant of Southern Quechua), but some datasets are in Cuzco Quechua or in a Bolivian dialect. The exact relationships between the dialects is unknown. We add three variant labels, *<default>*, *<quz>* and *<quh>*, to help the model distinguish the dialects.

Discussion

	aym	bzd	cni	czn	gn	hch	nah	oto	quy	shp	tar
Baseline 2021	18.8	7.7	10.4	–	22.0	12.6	18.2	5.9	33.1	13.9	4.6
Helsinki 2021	28.3	16.5	25.8	–	33.6	30.4	26.6	14.7	34.3	32.9	18.4
Helsinki 2023	33.4	22.5	28.4	32.1	**40.4**	**32.3**	26.9	**15.3**	33.3	33.4	**19.2**
Best contender 2023	**36.2**	**26.1**	**30.0**	**40.0**	39.3	32.3	**27.3**	14.8	**39.5**	**33.4**	18.7

- Plateau reached for the lowest-resource languages
- Considerable improvements for high-resource languages
- Spelling normalization and variant handling helped
- Additional data collection efforts did not always pay off
- New approaches (knowledge distillation, use of other toolkits) did not surpass our 2021 approach

NMT for variation-rich languages: Normalization of Finnish and Norwegian dialects

Dialect-to-standard normalization

The task: automatically convert phonetic transcriptions of dialectal utterances to standard orthography.

Finnish – SKN corpus:

mä oon syänys seittemän silakkaa, aiva niin, häntä erellä.
→ minä olen syönyt seitsemän silakkaa, aivan niin, häntä edellä.
'I have eaten seven herrings, that's right, tail first.'

Norwegian – NDC corpus:

å får eg sje sjøra vår bil før te påske
→ og får jeg ikke kjøre vår bil før til påske
'and I don't get to drive our car until Easter'

- A seq2seq task
- Mostly "small" changes at character level, monotonic
- Useful for syntactic annotation, lexical search

Dialect-to-standard normalization

The datasets:

Language	Corpus	Texts	Speakers	Locations	Sentences	Words
Finnish	SKN	99	99	50	41,407	630,665
Norwegian	NDC	684	438	111	126,460	1,684,059

- SKN has two levels of transcriptions. We use the simplified ones.
- For NDC, each speaker takes part in two texts, an interview and a conversation with another speaker.
- NDC is normalized towards Bokmål.
- Each speaker has a unique ID that includes their place of origin.

The normalization model

- We use a standard Transformer architecture for normalization.
- We train one model per language. This corresponds to a “multilingual” many-to-one setup.
  - Many different source dialects, one standardized target variety
- We append the speaker ID as labels to each sentence pair.
  - This is not strictly needed in a many-to-one setting, but we’re interested in the labels themselves.
- We break up the words into small segments using the BPE (byte-pair encoding) method.

<SKN34-Markkova> _ mi e _ po i ka in _ kan s _ ol en _ ka h en _ te ä l _
→ _ minä _ po i ka ni _ kan ssa _ ol en _ ka hd en _ tä ä llä _
‘Me and my son are alone here.’

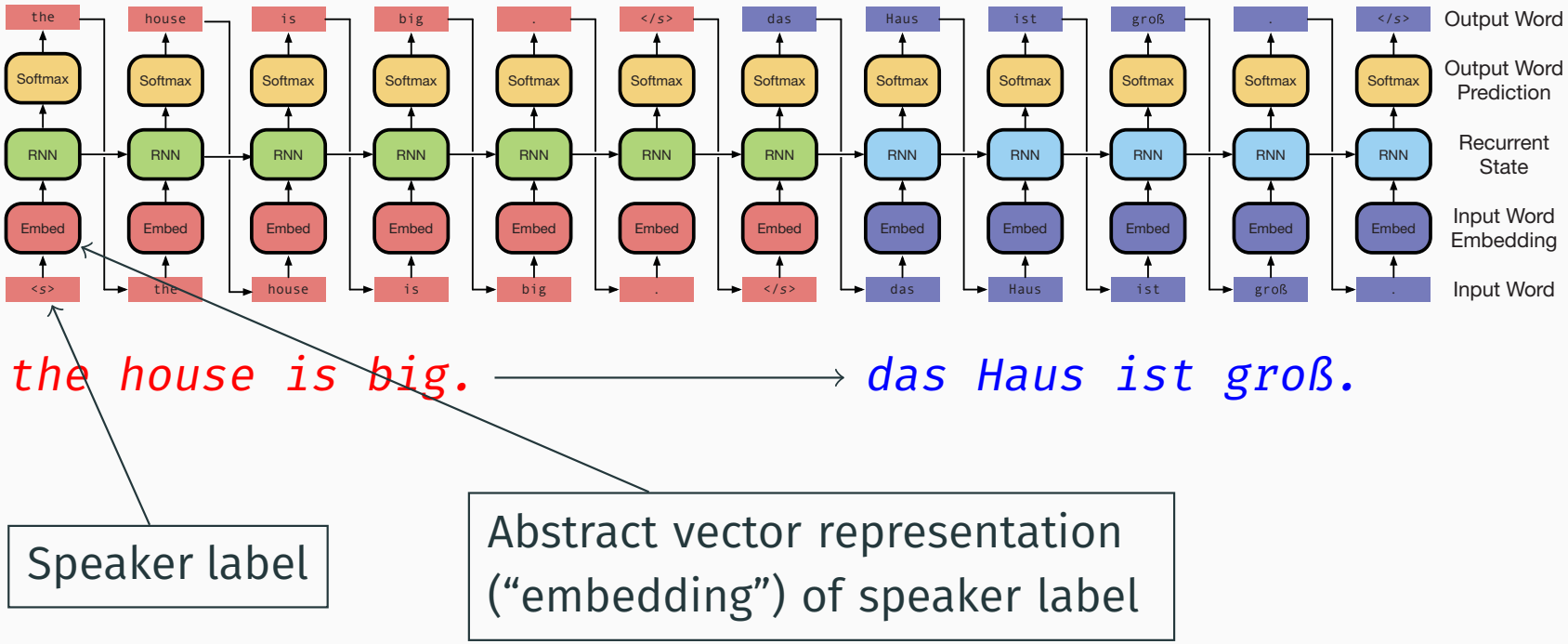
Normalization results

Not quite as good as previous work, but close enough:

Word error rates (↓)	SKN	NDC
Partanen et al. (2019)	5.73	—
This work	6.11	4.89

The main goal of this work is not to achieve optimal normalization performance, but to see what the model learns about the speaker labels.

Speaker label embeddings



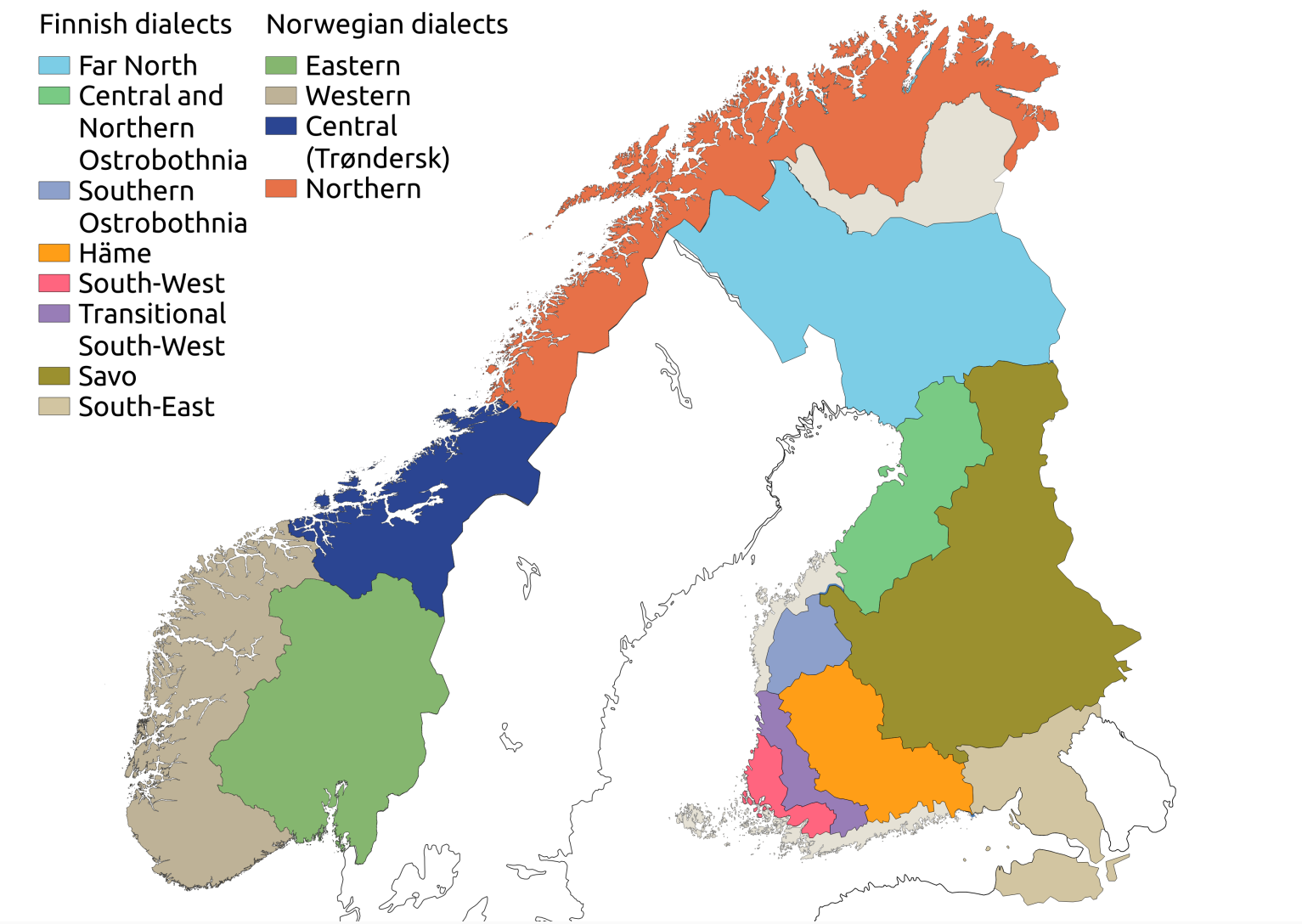
We look at the speaker label embeddings. In our setting, each speaker label embedding is a vector with 512 dimensions.

Some inspiration from dialectometry

A rough sketch of a dialectometrical experiment:

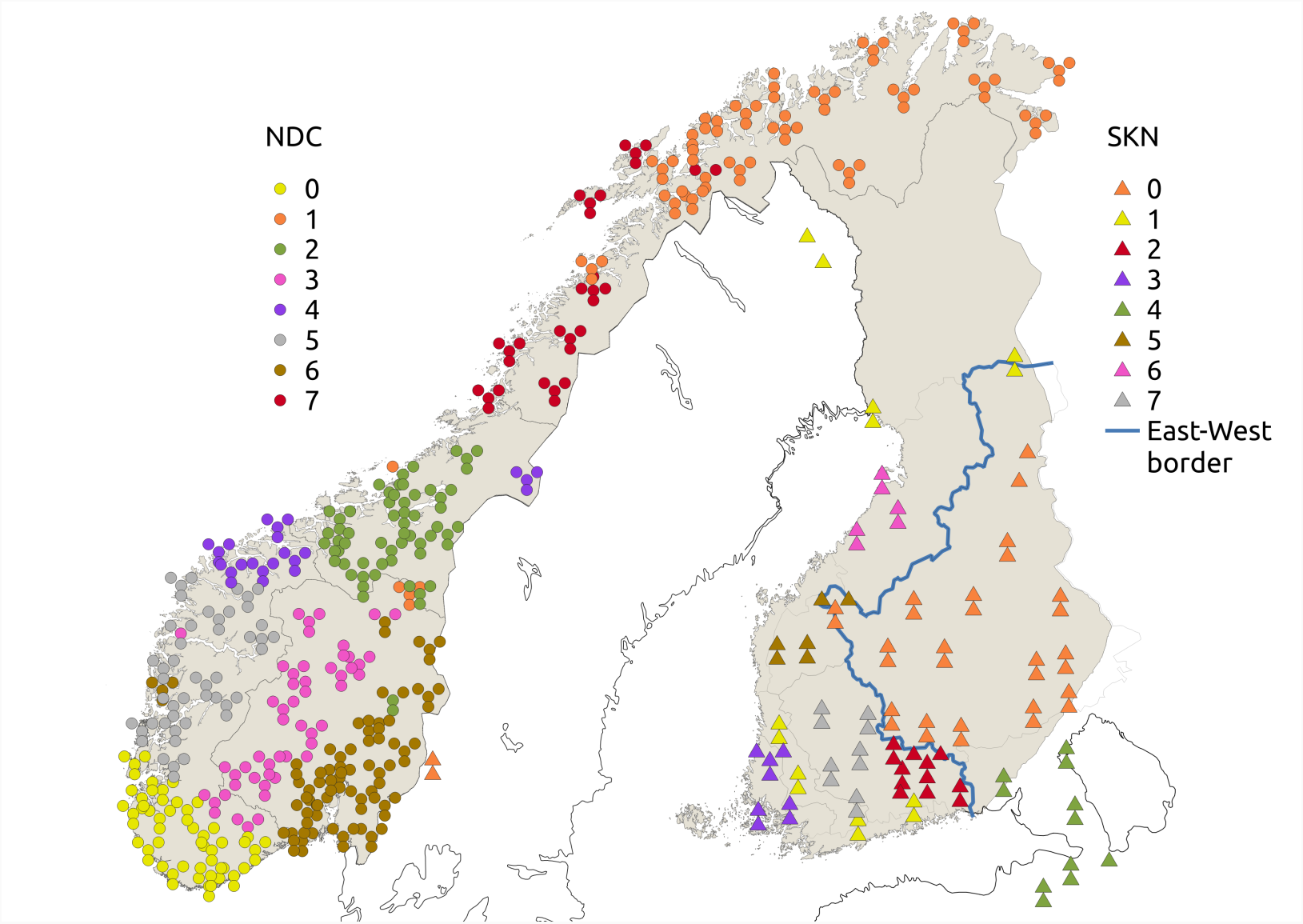
1. Build a vector characterizing each dialect
   - Data vector: each dimension represents a linguistic item, the value marks presence or absence
   - Distance/similarity vector: each dimension represents the distance/similarity to one other dialect
   - We just use our speaker embedding vectors here.
2. Project these high-dimensional vectors into a lower-dimensional space
   - Cluster analysis
   - Multidimensional scaling, principal component analysis, factor analysis, ...
3. Assign each value a color and plot on a map

Expected dialect classifications

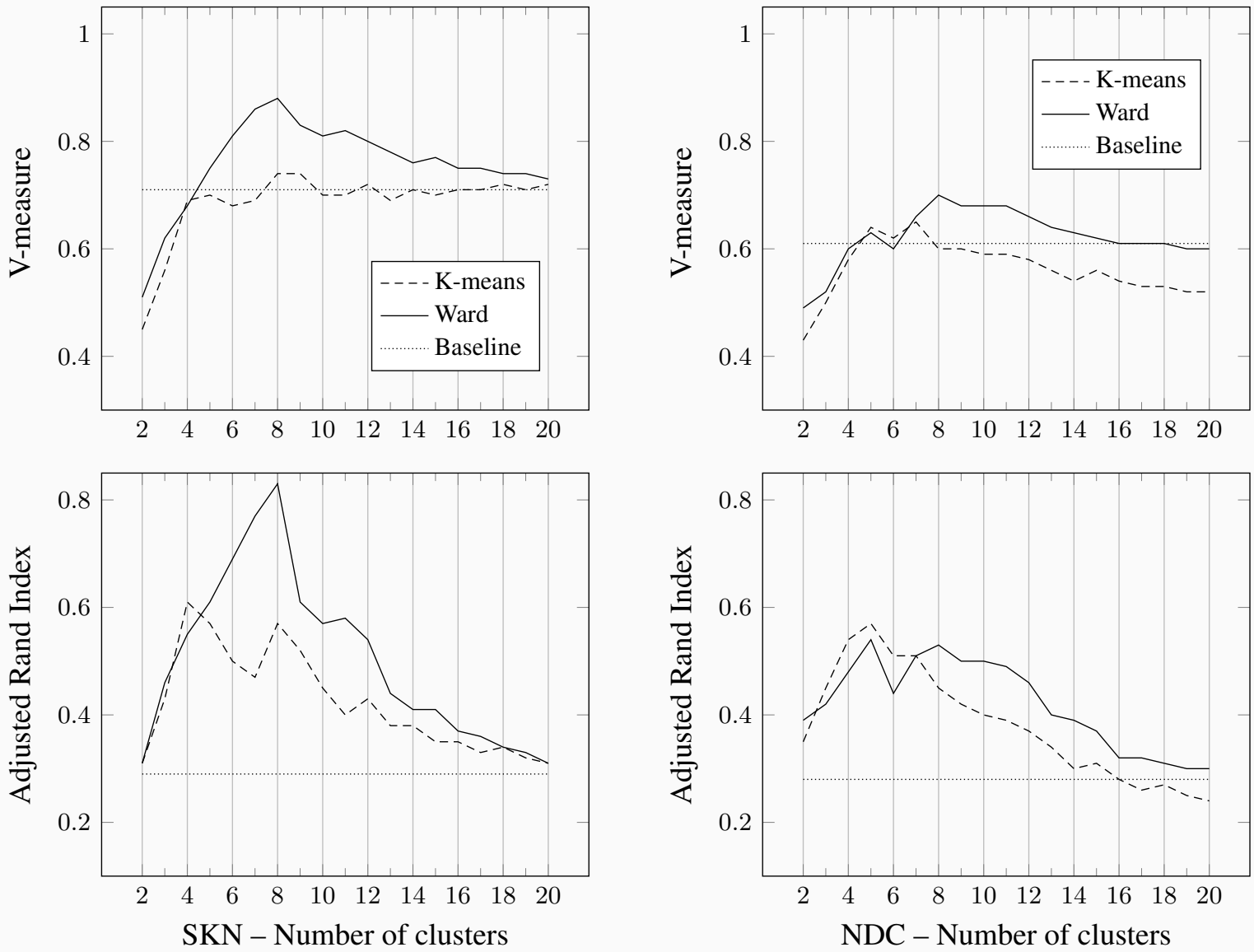


Norwegian division based on Hanssen (2010–2014).
Finnish division based on Itkonen (1989).

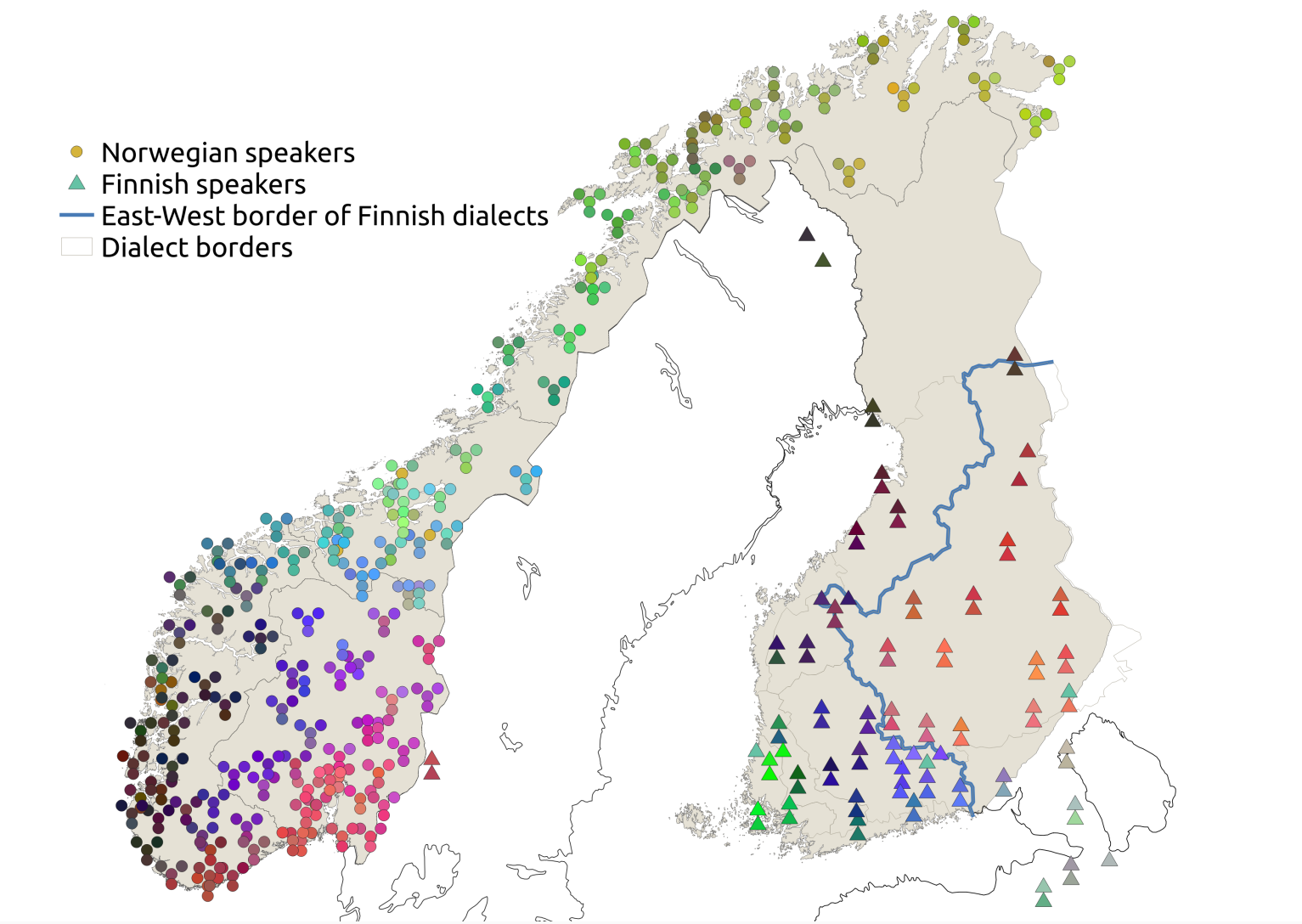
Hierarchical clustering (Ward, 8 clusters per language)



Clustering evaluation



Principal component analysis (3 dimensions → RGB)



Explained variance: 9% for Norwegian, 14% for Finnish.

Discussion

- The speaker labels learned for the normalization task reflect the dialectal (or geographic) origin of the speakers.
  - The model could also have ignored them completely.
  - The model could also have used them for something else.
- Speakers from the same place are almost always placed in the same cluster.
- The major dialect borders are visible in the embeddings.
- The explained variance of the PCA is low. It remains to be investigated if the embeddings contain other interesting variation patterns.
- It would be interesting to see **when** and **how** the normalization model makes most use of the labels. This could be achieved by analyzing the attention weights.

Outline

1. Neural machine translation, sequence-to-sequence models and encoder-decoder architectures

2. NMT for low-resource languages:
   The AmericasNLP shared tasks

3. NMT for variation-rich languages:
   Normalization of Finnish and Norwegian dialects

Credits

Thanks to **Ona de Gibert, Raúl Vázquez, Mikko Aulamo, Sami Virpioja and Jörg Tiedemann** for their work on the AmericasNLP shared tasks!

This work has been supported by the FoTran (ERC, No. 771113) and HPLT (Horizon Europe, No. 101070350) projects.

References

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann (2021). The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas*, pages 255–264, Online. *https://aclanthology.org/2021.americasnlp-1.29/*

Ona de Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann (to appear). Four Approaches to Low-Resource Multilingual NMT: The Helsinki Submission to the AmericasNLP 2023 Shared Task In *Proceedings of the Third Workshop on NLP for Indigenous Languages of the Americas*.

Credits

Thanks to **Olli Kuparinen** for the work on dialect normalization!

This work has been supported by the CorCoDial project (Academy of Finland, No. 342859).

References

Olli Kuparinen and Yves Scherrer (2023). Dialect Representation Learning with Neural Dialect-to-Standard Normalization. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 200–212, Dubrovnik, Croatia.
https://aclanthology.org/2023.vardial-1.20/