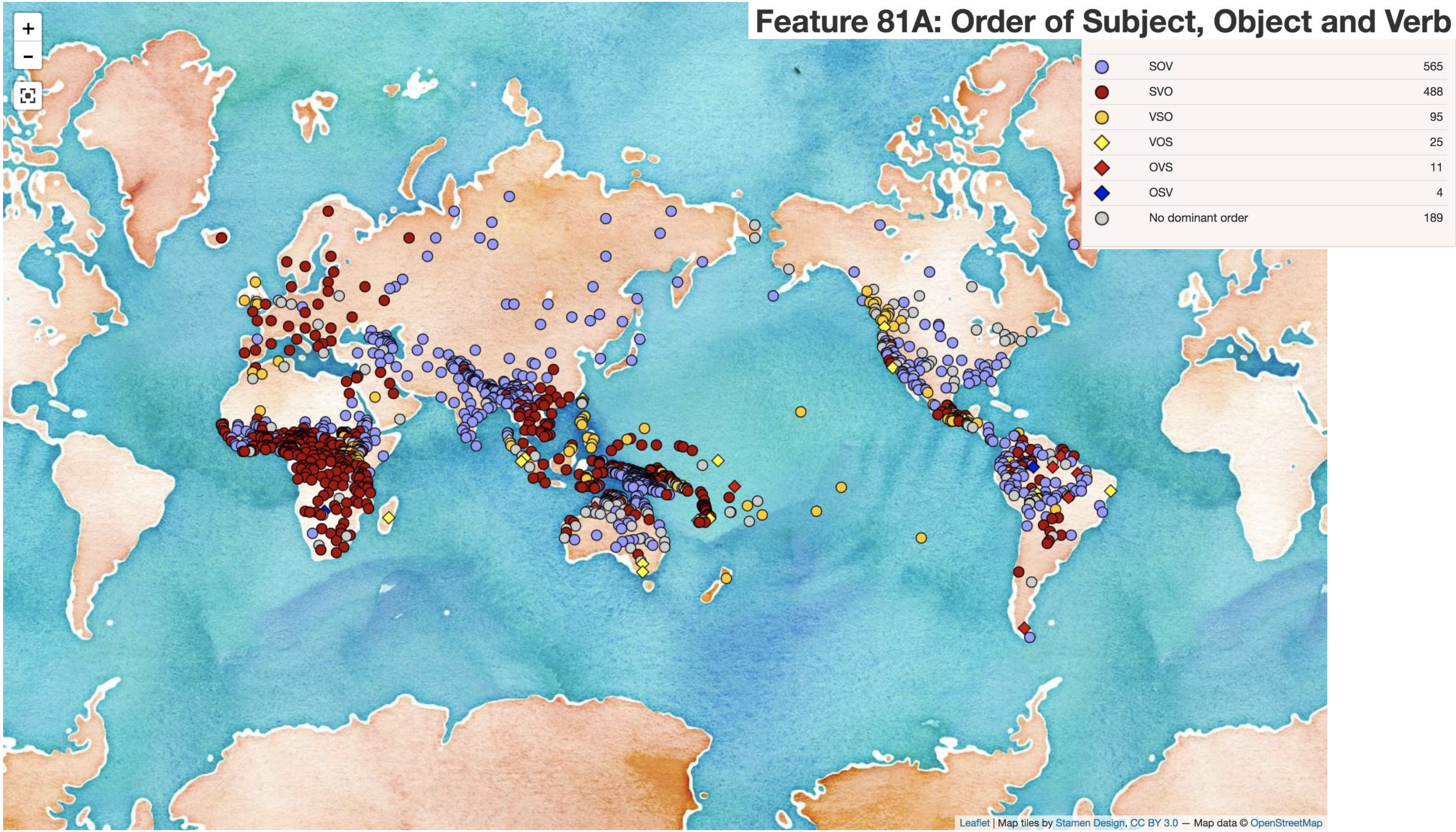
Tracking Typological Traits of Uralic Languages in Distributed Language Representations

Johannes Bjerva and Isabelle Augenstein
bjerva@di.ku.dk
http://bjerva.github.io

UNIVERSITY OF COPENHAGEN

Linguistic Typology

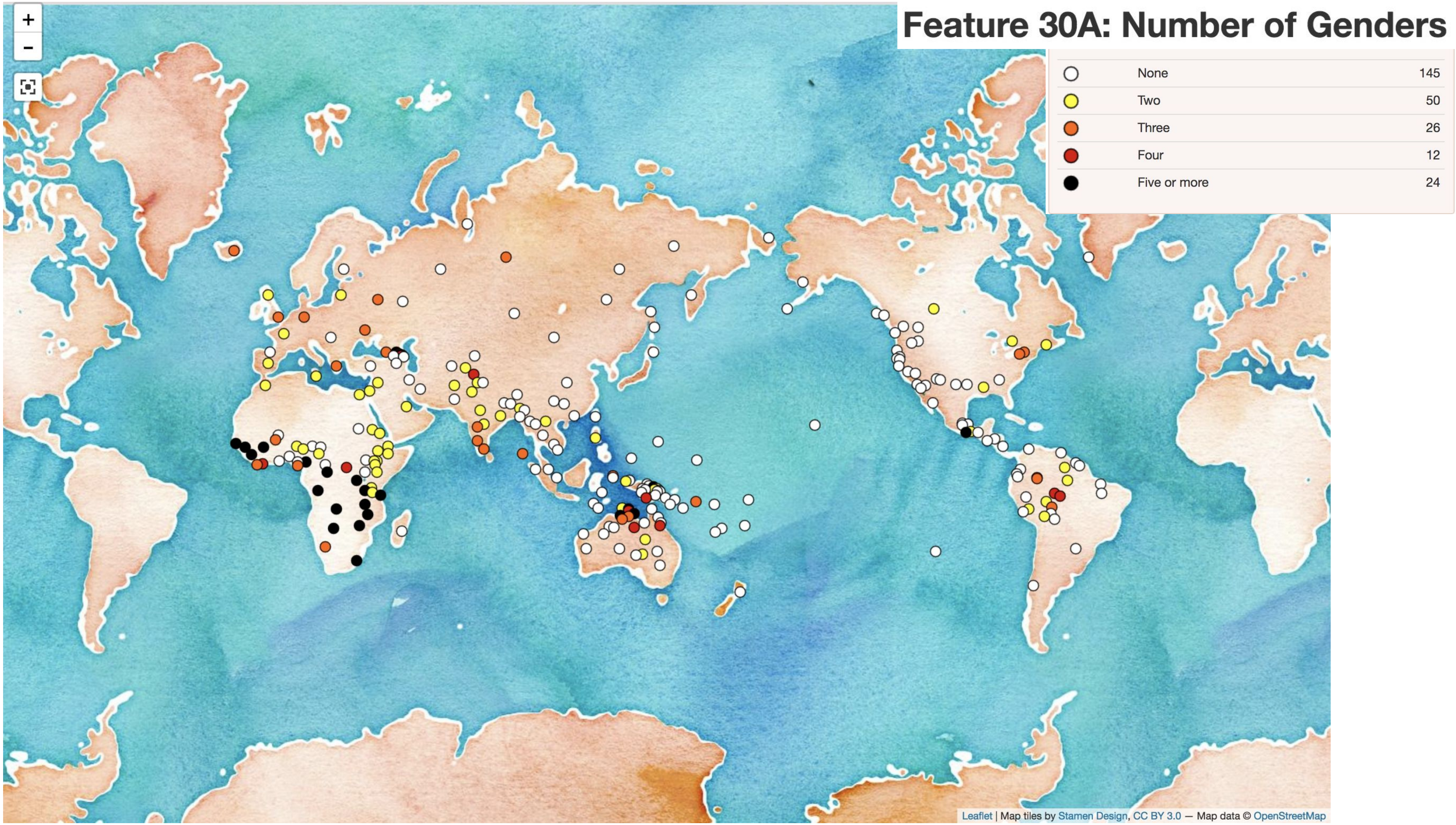
- ‘The systematic study and comparison of language structures’ (Velupillai, 2012)
- Long history (Herder, 1772; von der Gabelentz, 1891; …)
- Computational approaches (Dunn et al., 2011; Wälchli, 2014; Östling, 2015, ...)
- Potential to answer linguistic research questions on large scales
- This work:
  - Focus on features in the World Atlas of Language Structures (WALS)
  - Computational Typology via unsupervised modelling of languages in neural networks
  - Focussing on four Uralic languages (Finnish, Estonian, Hungarian, North Sami)



Feature 81A: Order of Subject, Object and Verb

●	SOV	565
●	SVO	488
●	VSO	95
◆	VOS	25
◆	OVS	11
◆	OSV	4
●	No dominant order	189

Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap



Feature 30A: Number of Genders

○	None	145
🟡	Two	50
🟠	Three	26
🔴	Four	12
●	Five or more	24

Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap

Resources exist for a lot of languages

- Universal Dependencies (>60 languages)
- UniMorph (>50 languages)
- New Testament translations (>1,000 languages)
- Automated Similarity Judgment Program (>4,500 languages)

Multilingual NLP and Language Representations

- No explicit representation
  - Multilingual Word Embeddings
- Google's "Enabling zero-shot learning" NMT trick
  - Language given explicitly in input
- One-hot encodings
  - Languages represented as a sparse vector
- **Language Embeddings**
  - Languages represented as a distributed vector

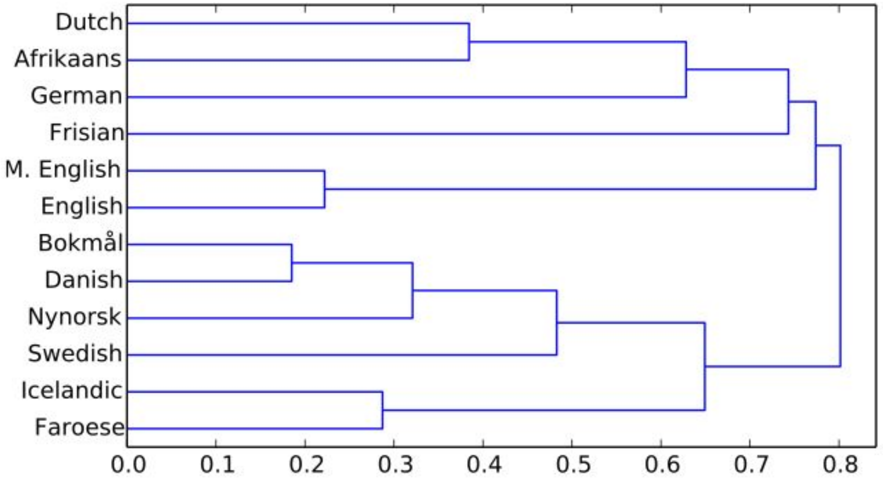


Figure 5: Hierarchical clustering of language vectors of Germanic languages.

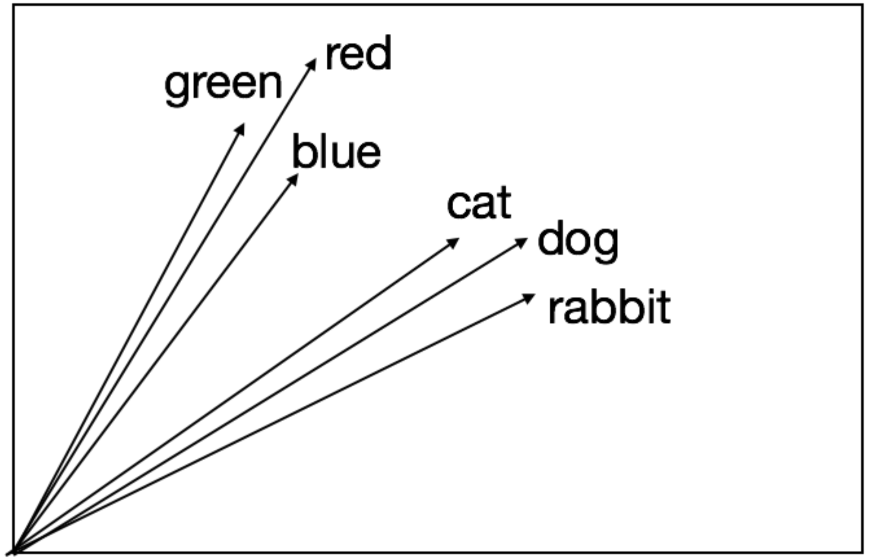
(Östling and Tiedemann, 2017)

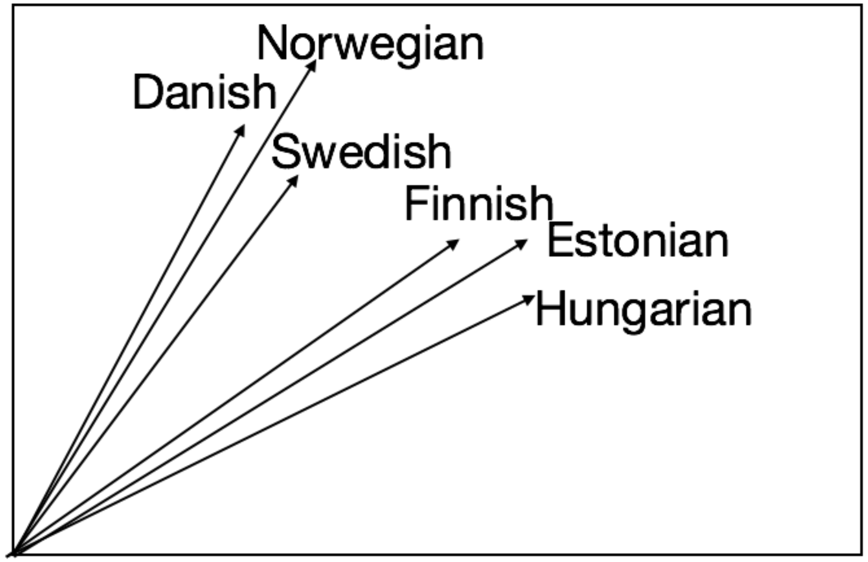
Data

- Pre-trained language embeddings from Östling and Tiedemann (2017)
  - Trained via Language Modelling on New Testament data
- PoS annotation from Universal Dependencies for
  - Finnish
  - Estonian
  - North Sami
  - Hungarian

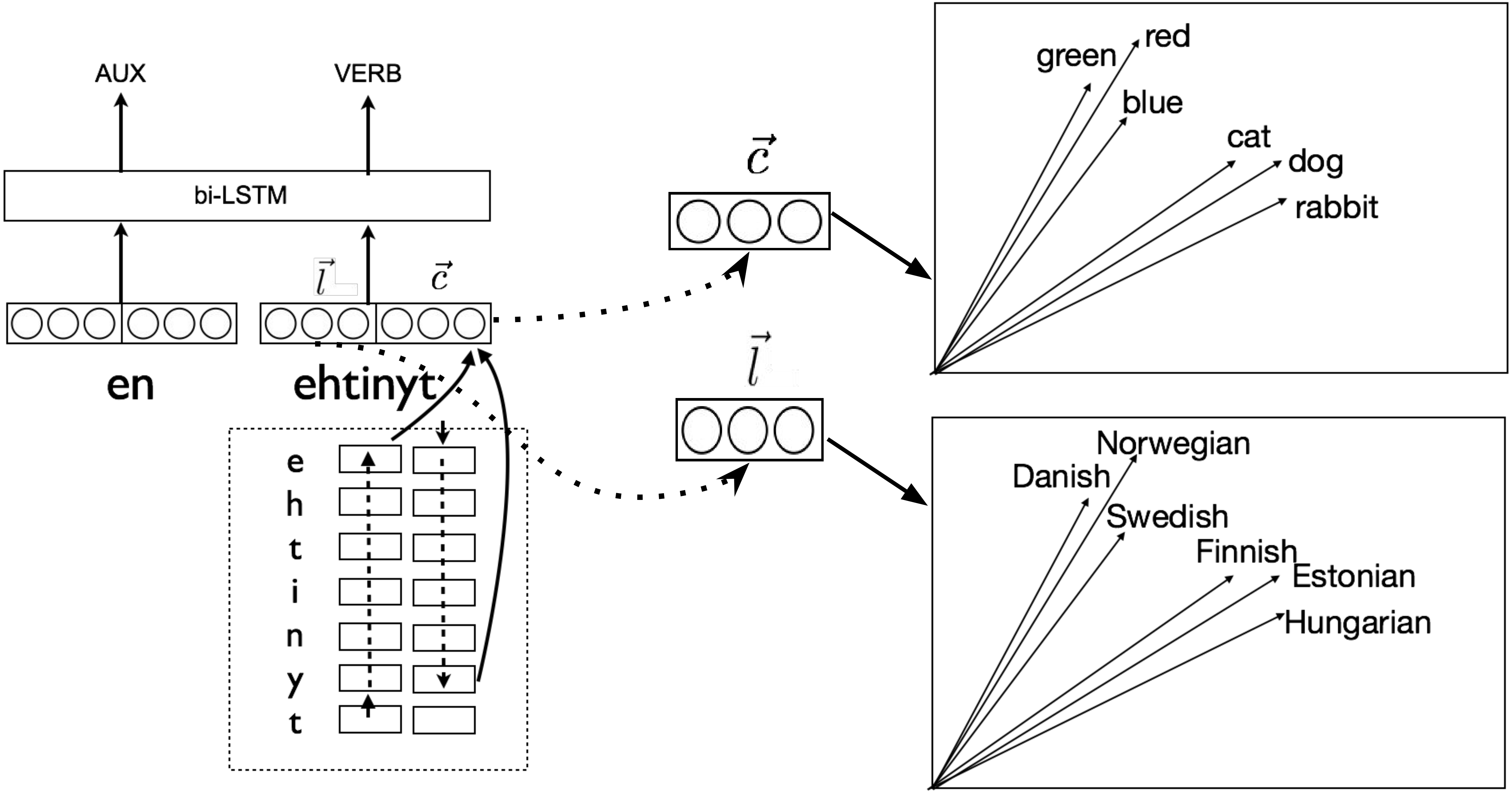
Distributed Language Representations

- Language Embeddings
- Analogous to Word Embeddings
- Can be learned in a neural network without supervision

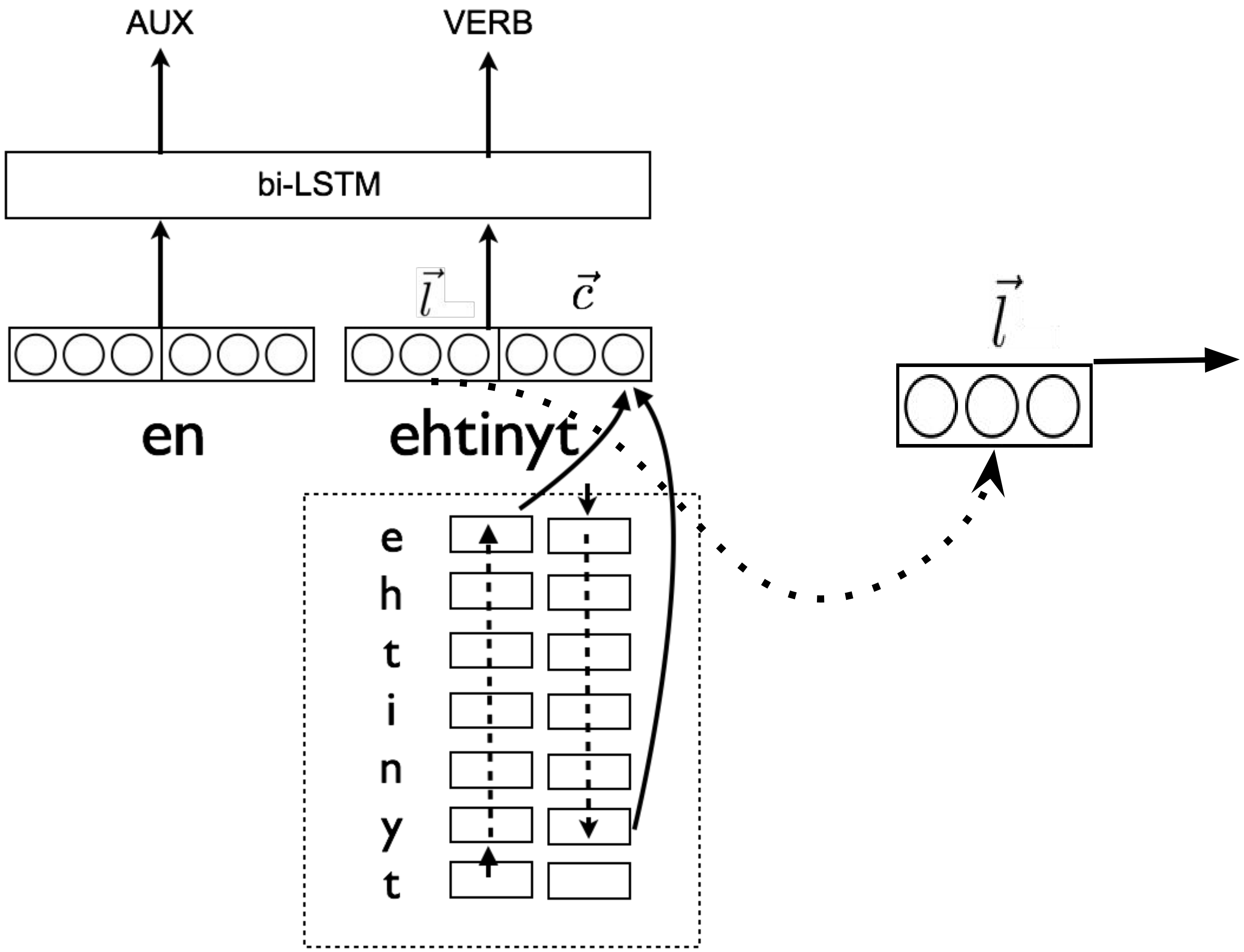




Language Embeddings in Deep Neural Networks



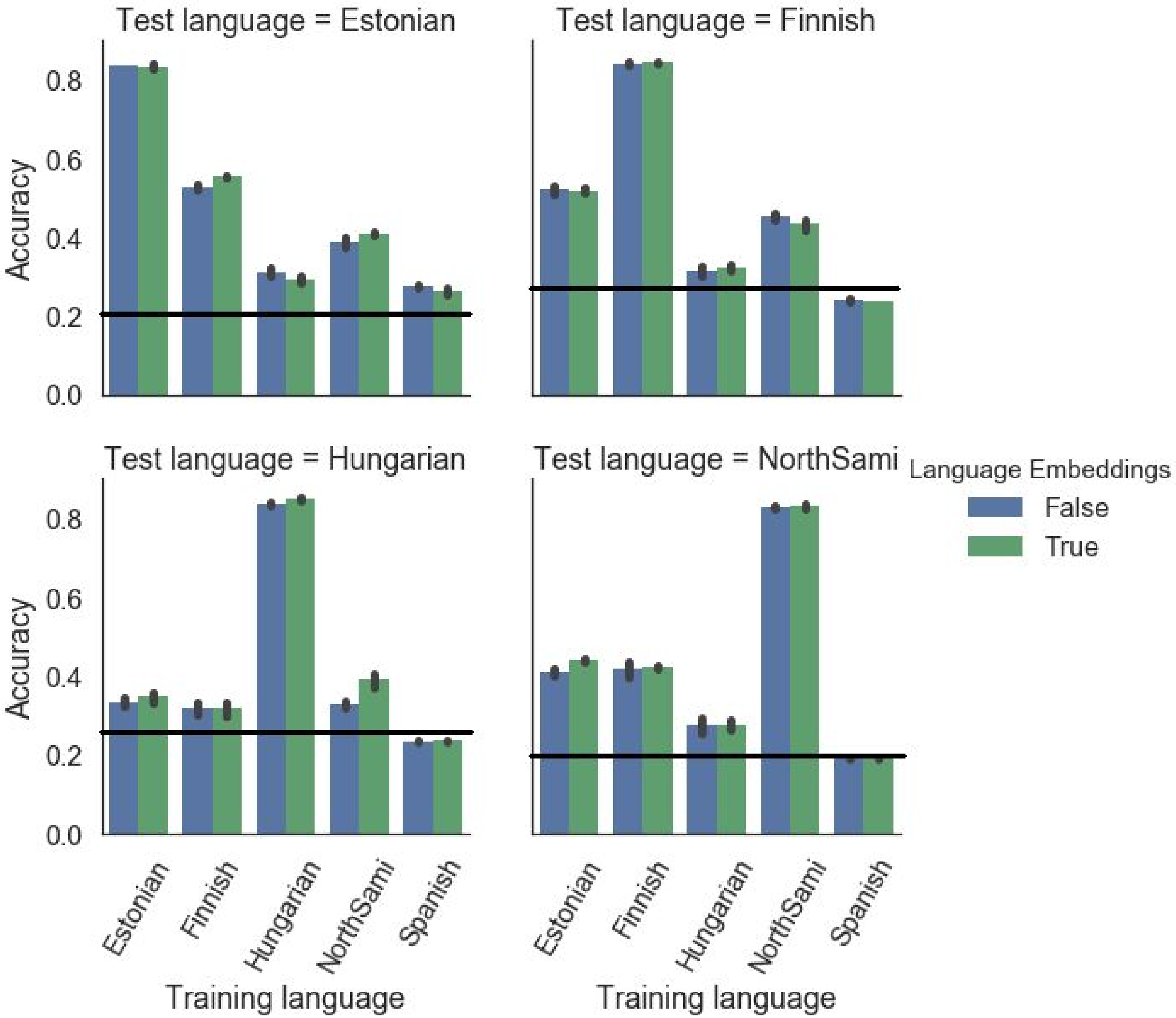
Language Embeddings in Deep Neural Networks



1. Do language embeddings aid multilingual modelling?
2. Do language embeddings contain typological information?

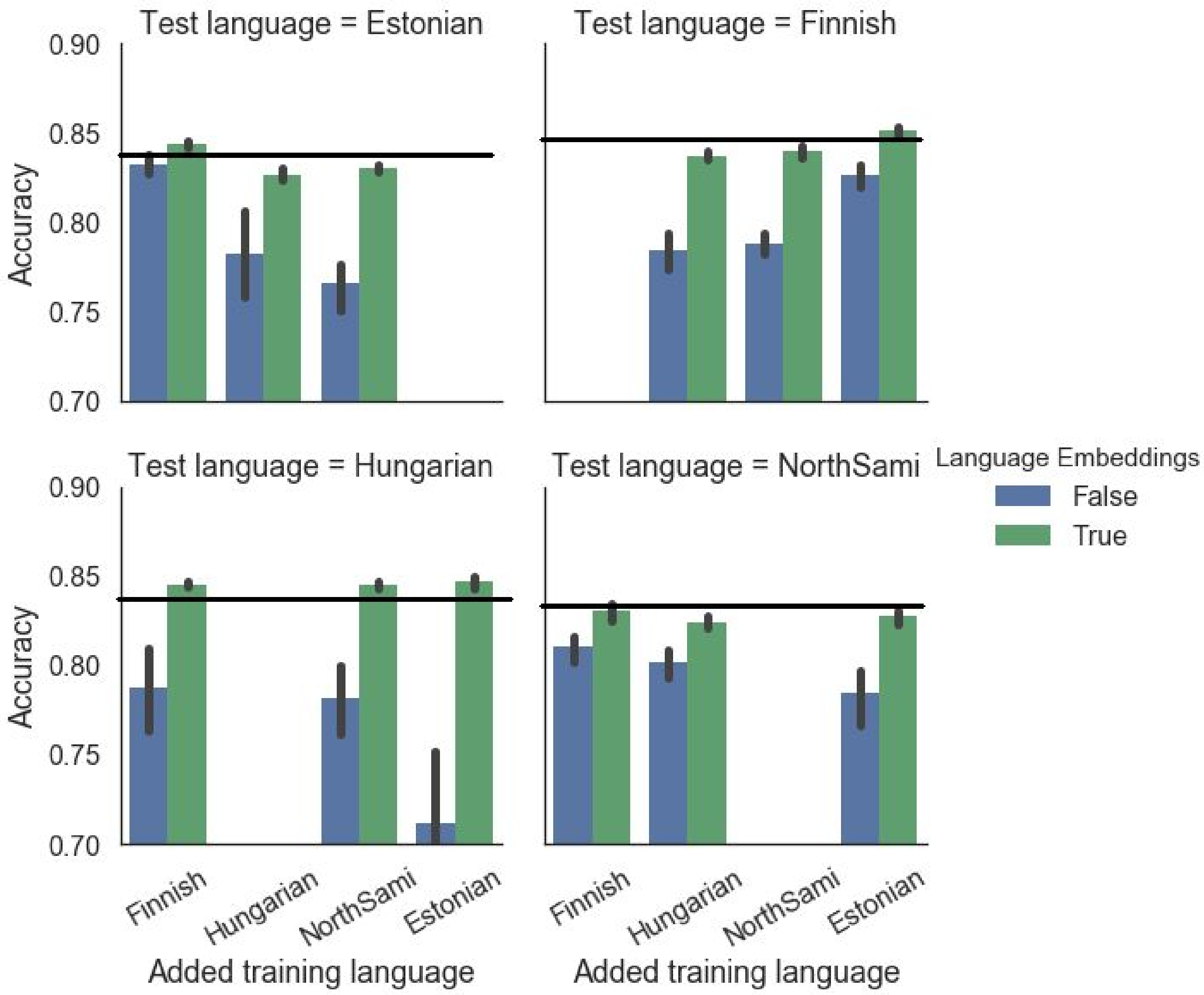
Model performance (Monolingual PoS tagging)

- Compared to most frequent class baseline (black line)
- Model transfer between Finnic languages relatively successful
- Little effect from language embeddings (to be expected)



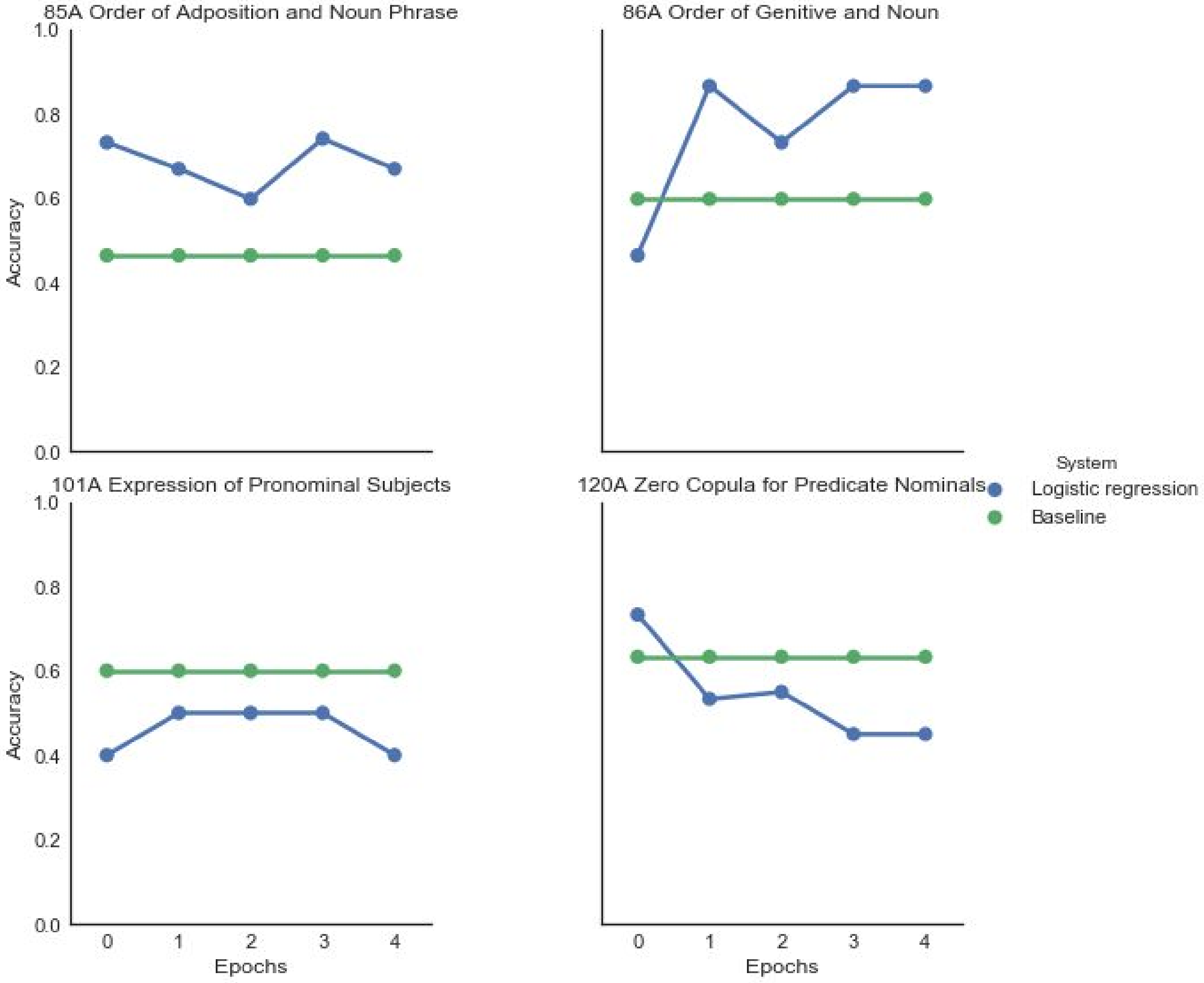
Model performance (Multilingual PoS tagging)

- Compared to monolingual baseline (black line)
- Model transfer between Finnic languages outperforms monolingual baseline
- *Language embeddings improve multilingual modelling*



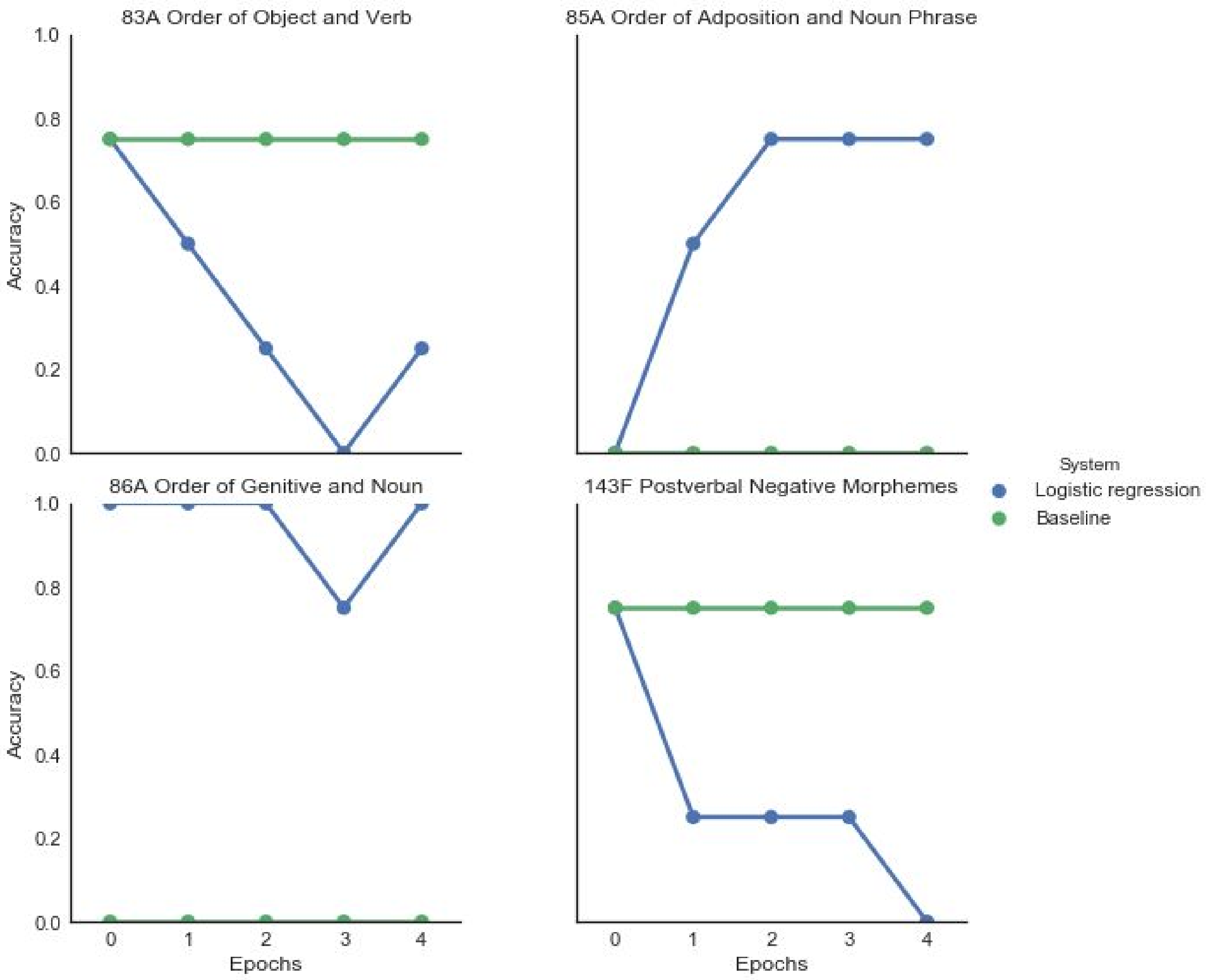
Tracking Typological Traits (full language sample)

- Baseline: Most frequent typological class in sample
- Language embeddings saved at each training epoch
- Separate Logistic Regression classifier trained for each feature and epoch
  - Input: Language embedding
  - Output: Typological class
- *Typological features encoded in language embeddings change during training*



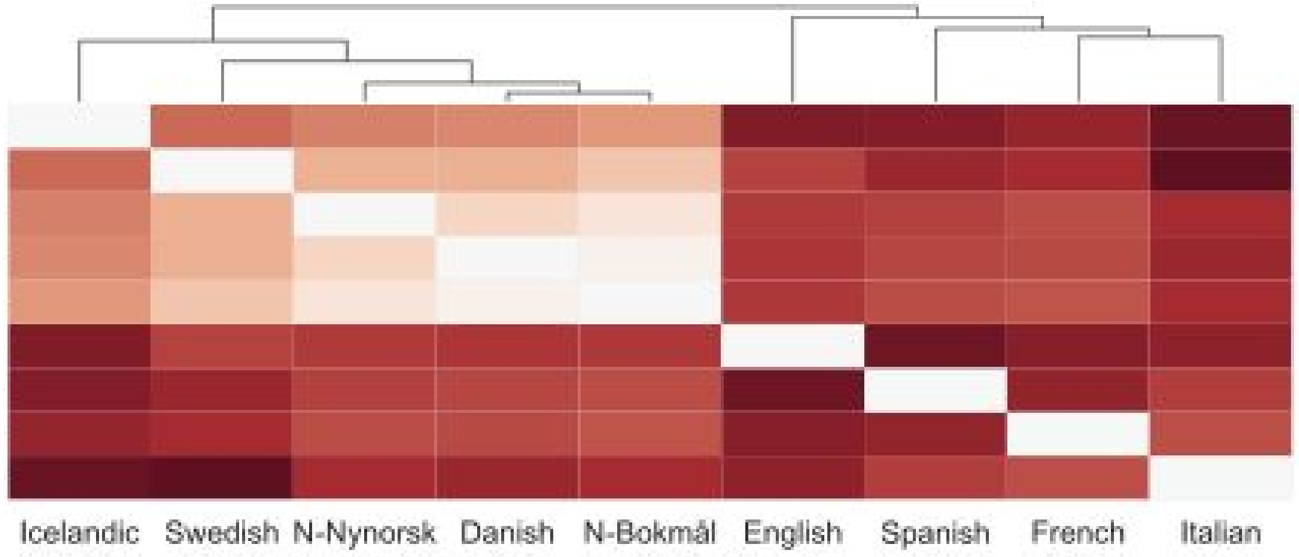
Tracking Typological Traits (Uralic languages held out)

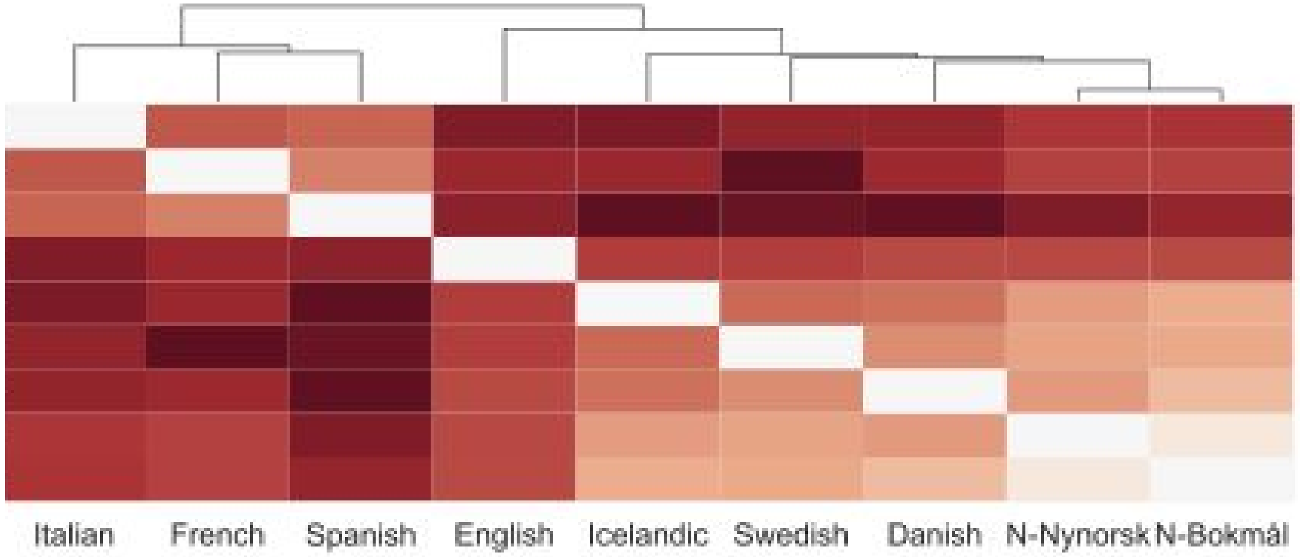
- Some typological features can be predicted with high accuracy for the unseen Uralic languages.



Beyond Uralic languages

- Hierarchical clustering of language embeddings
- Language modelling based language embeddings
  - English with Romance
  - Large amount of romance vocabulary
- PoS based language embeddings
  - English with Germanic
  - Morpho-syntactically more similar

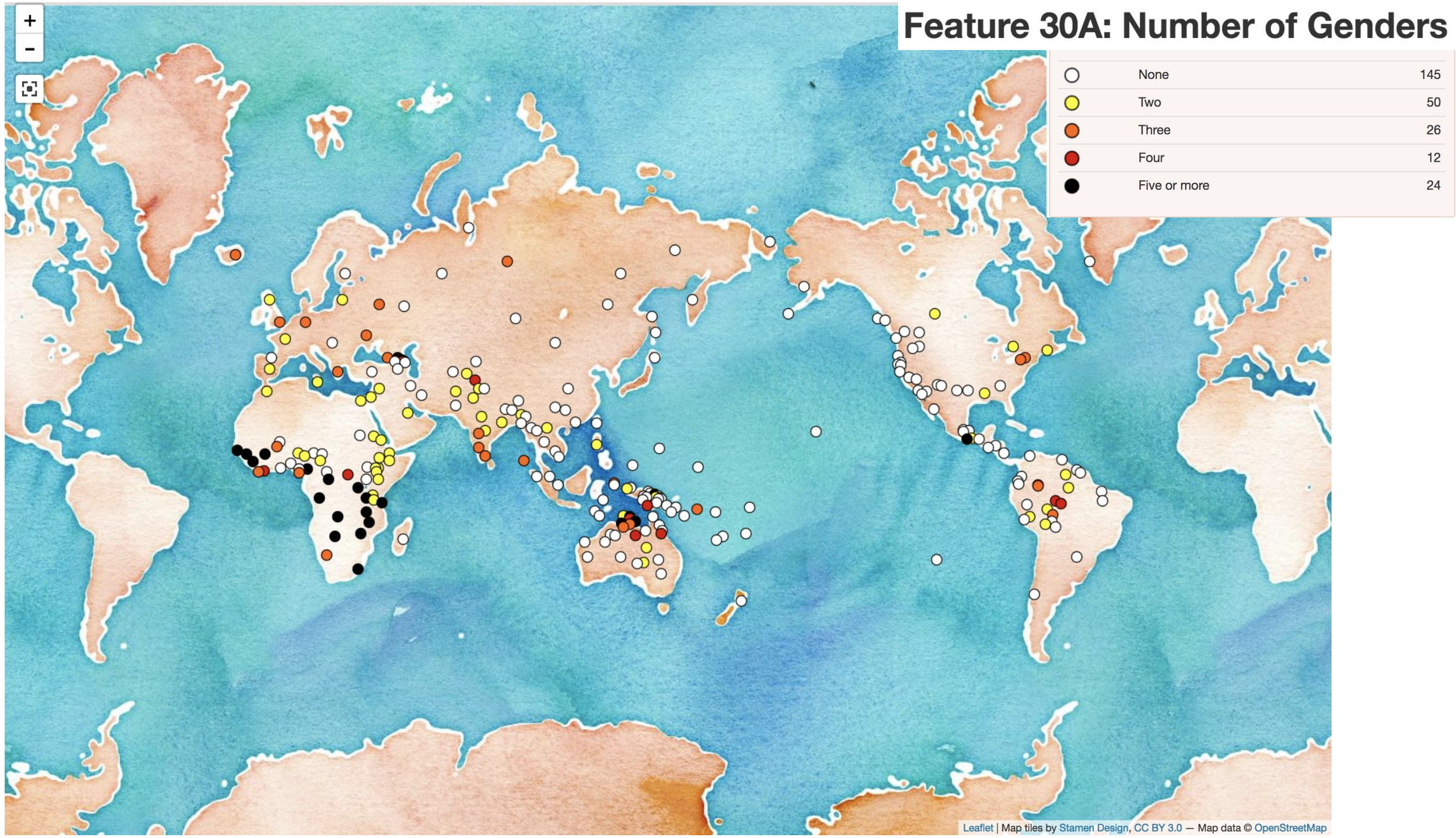




(Bjerva and Augenstein, Under review)

Future work

- Improve multilingual modelling
  - E.g., share morphologically relevant parameters for morphologically similar languages
- Automatically fill gaps in WALS by using Language Embedding predictions


Feature 30A: Number of Genders

+
-
○ None 145
● Two 50
● Three 26
● Four 12
● Five or more 24
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap

Thanks!

Questions?