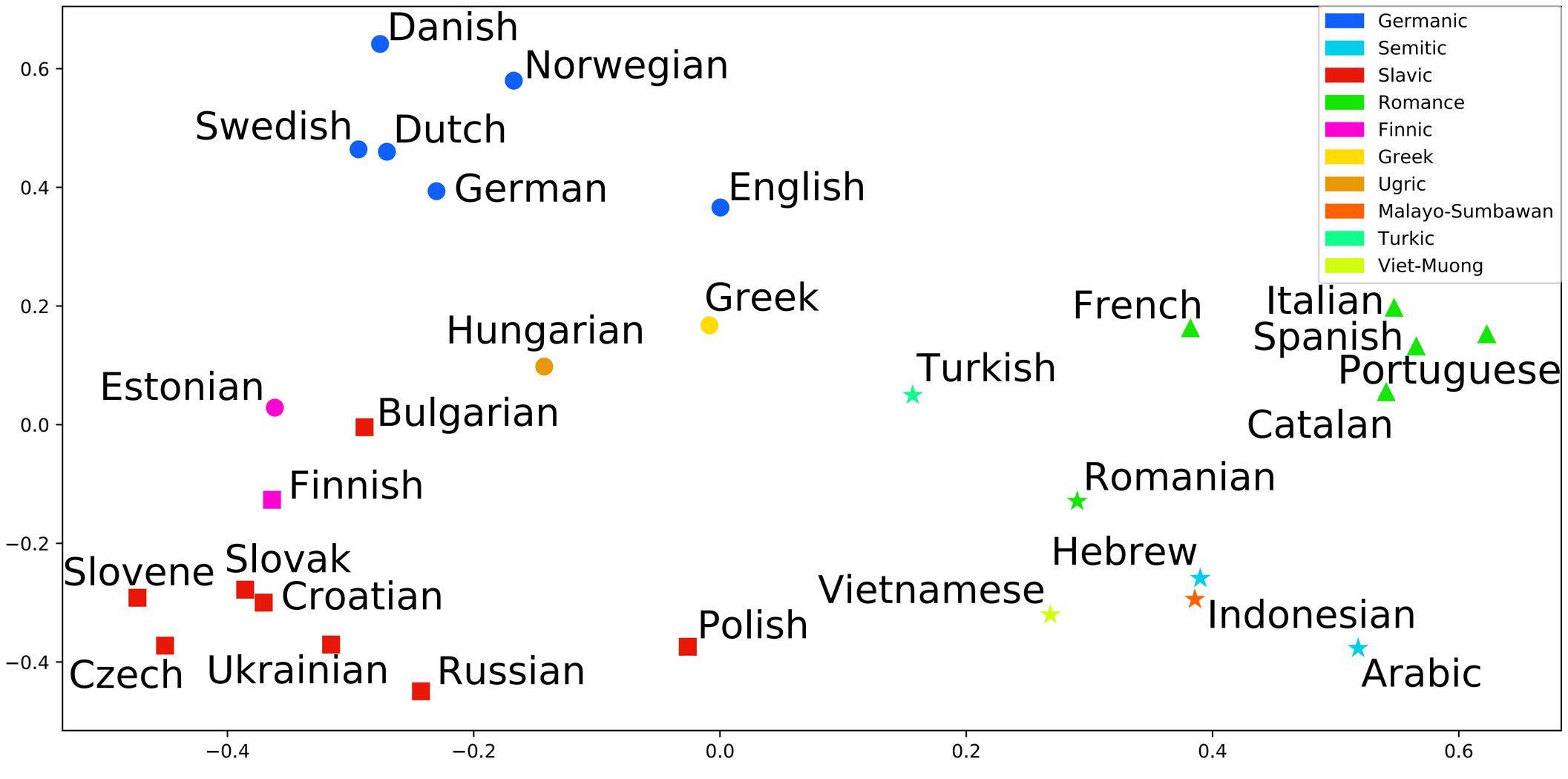
Syntactic Typology from Plain Text Using Language Embeddings

Taiqi He
tqhe@ucdavis.edu

Kenji Sagae
sagae@ucdavis.edu

University of California, Davis

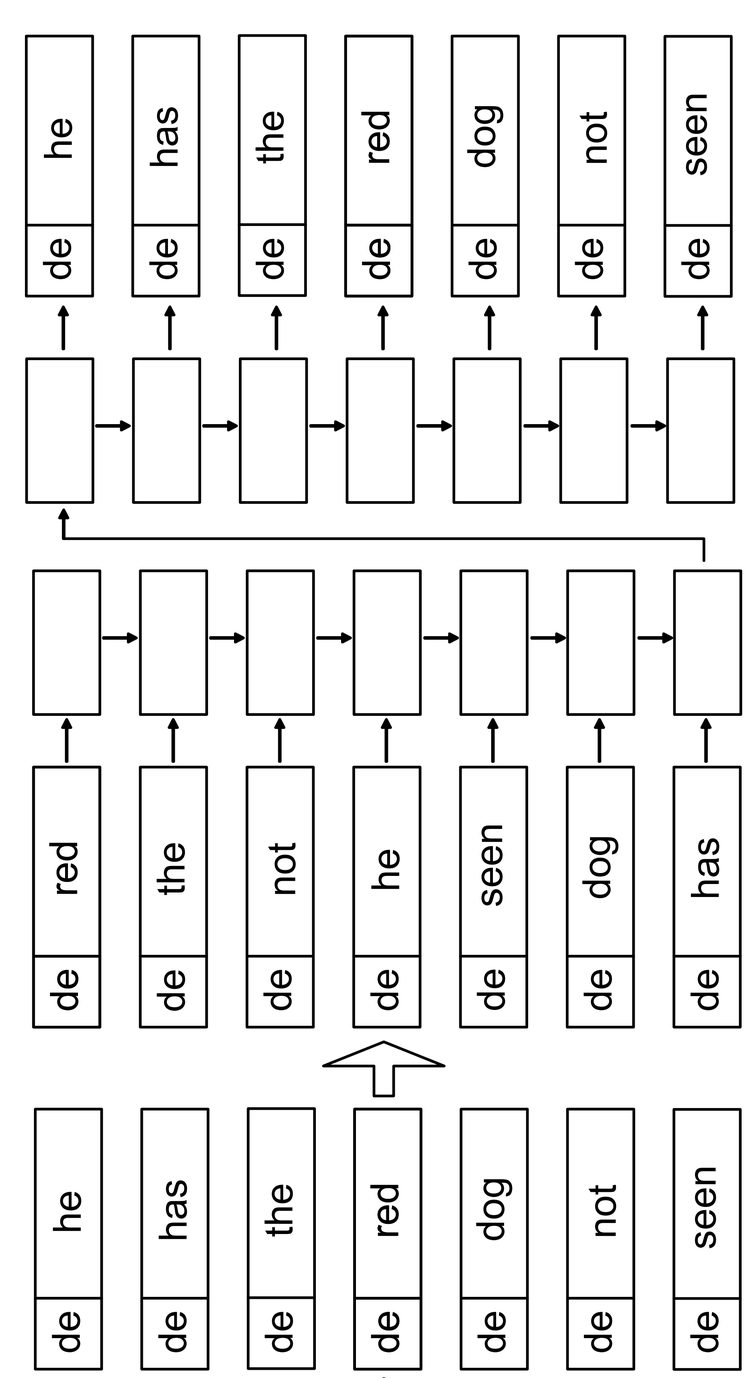


Motivation

- Typology from unannotated text corpora
- Continuous representation of languages

Methods

- Multilingual inputs to a shared encoder
- Words are mapped to English with MUSE
- Denoising autoencoder with language embeddings



Results

- Language embeddings outperform baseline on WALS and dependency prediction tasks
- Distribution of language embeddings captures genetic relationships

WALS Area	Lexicon	Morphology	Verbal Categories	Word Order
Baseline	0.68	**0.82**	0.66	0.81
Model	**0.90**	0.78	**0.69**	**0.86**