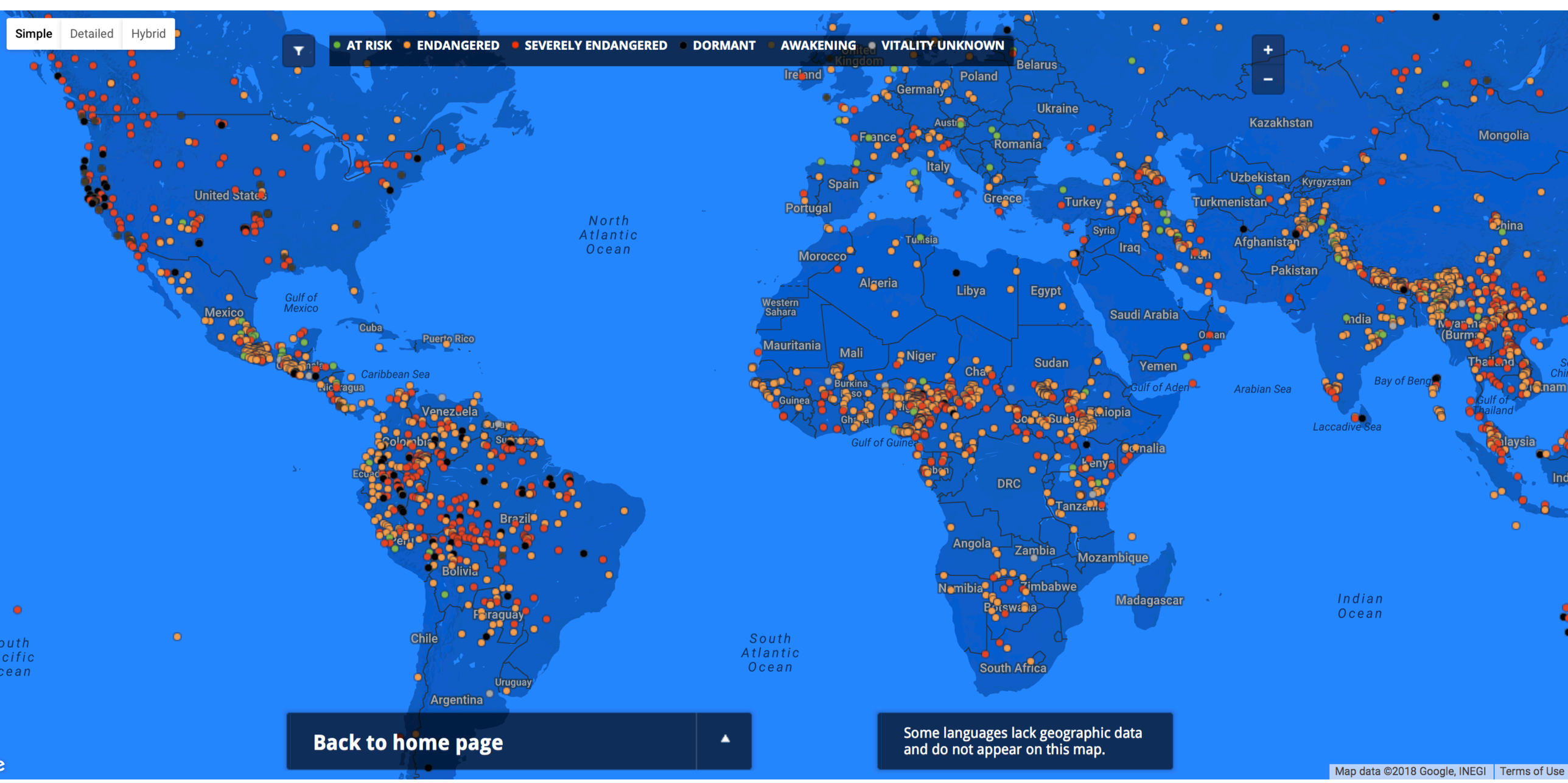
CS11-737: Multilingual Natural Language Processing

Graham Neubig, Alan Black, Shinji Watanabe

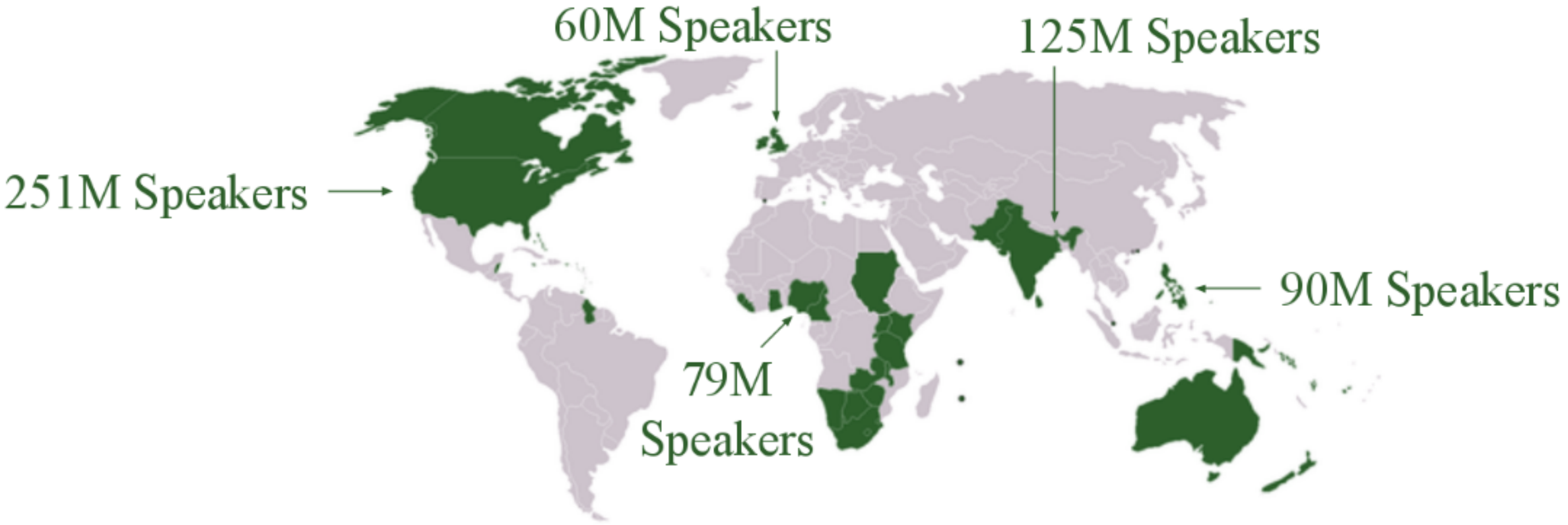
http://phontron.com/class/multiling2022/

Carnegie Mellon University
Language Technologies Institute



http://endangeredlanguages.com/

Language Varieties (e.g. English)



How do We Build NLP Systems?

- **Rule-based systems:** Work OK, but require lots of human effort for each language for where they're developed

- **Machine learning based systems:** Work really well when lots of data available, not at all in low-data scenarios

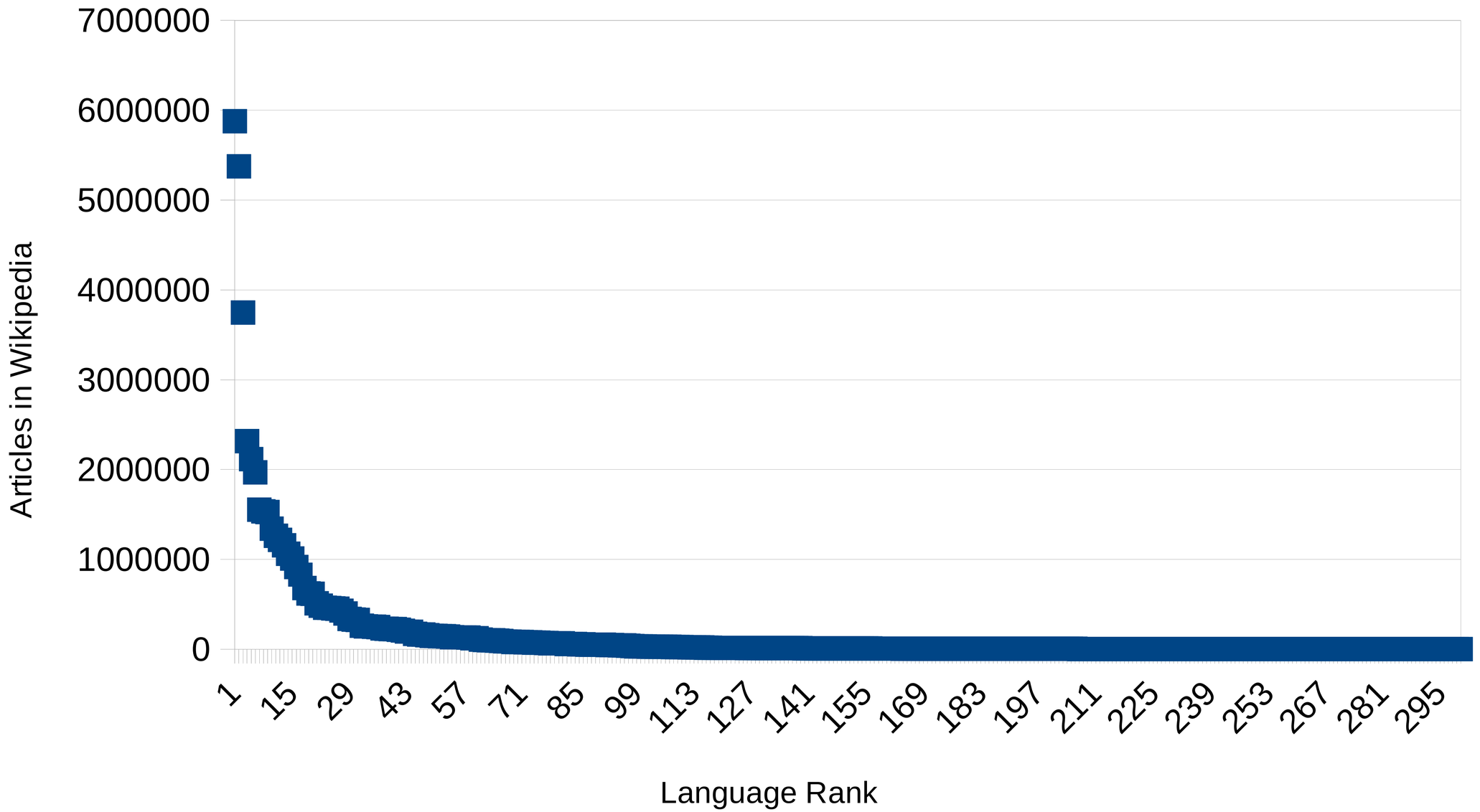
Machine Learning Models

- Formally, map an input *X* into an output *Y*. Examples:

Input *X*	Output *Y*	Task
Text	Text in Other Language	Translation
Text	Response	Dialog
Speech	Transcript	Speech Recognition
Text	Linguistic Structure	Language Analysis

- To learn, we can use
  - Paired data <*X*, *Y*>, source data *X*, target data *Y*
  - Paired/source/target data in *similar* languages

The Long Tail of Data



How to Cope?

- **Better Models or Algorithms:**
  - sophisticated modeling/training methods - know NLP/ML!
  - linguistically informed methods - know linguistics!
- **Better Data:**
  - every piece of relevant data can help - be resourceful!
  - make data if necessary - be connected!
- **Better Deployment:**
  - different situations require different solutions - be aware!

This Class Will Cover

- **Linguistics:** typology, orthography, morphology, syntax, language contact/change, code switching
- **Data:** annotated and unannotated sources, data annotation, linguistic databases, active learning
- **Tasks:** language ID, sequence labeling, translation, speech recognition/synthesis, syntactic parsing
- **Societal Considerations:** ethics, connection between language and society

All to:

Allow you to build a strong, functioning language system in a low-resource language that you do not know

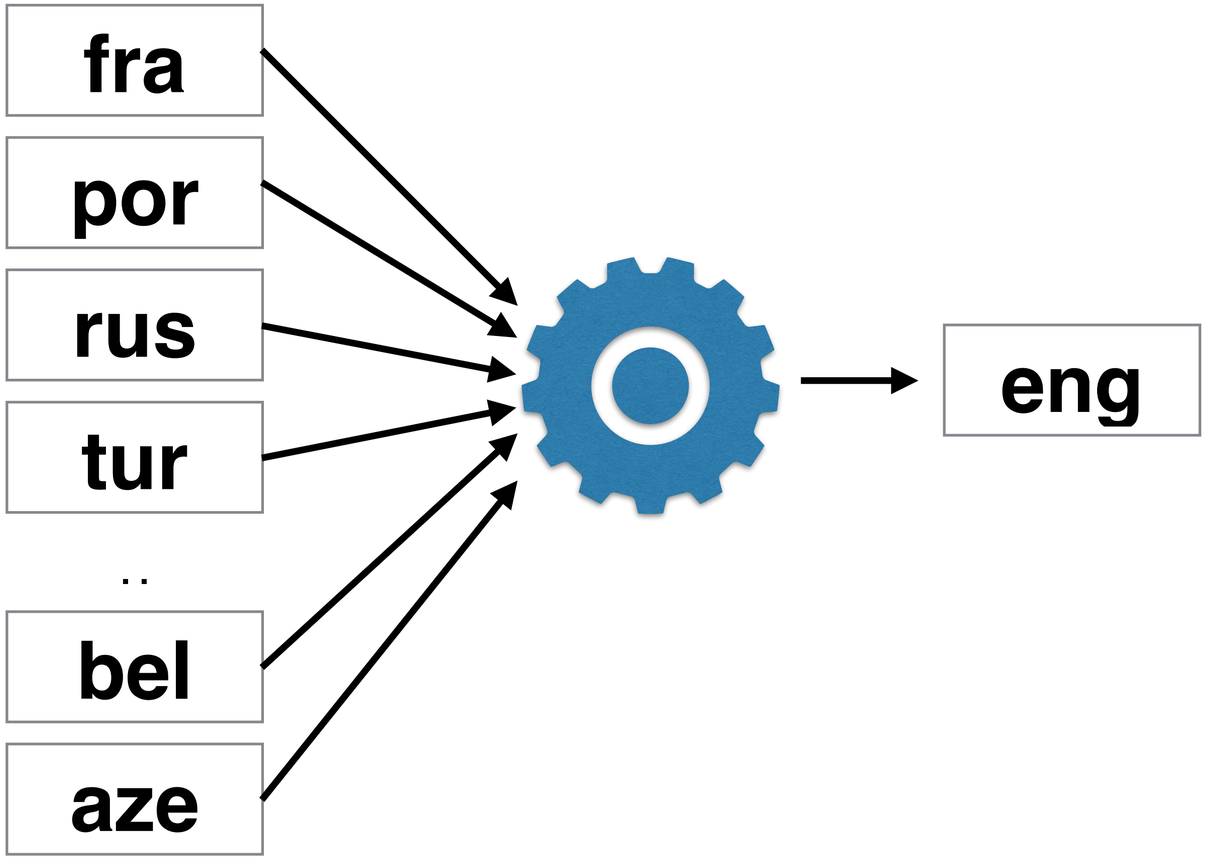
Training Multilingual NLP Systems

Data Creation/Curation

- First step is obtaining curated training data in your language
- What **types** of data? (monolingual? multilingual? annotated?)
- **Where** can we get it? (annotated data sources? curated text collections? scraping?)
- Can we **create** data? (efficient, high-quality creation strategies)
- How do we deal with the **ethical issues**? (working with communities, language ownership)

Multilingual Training

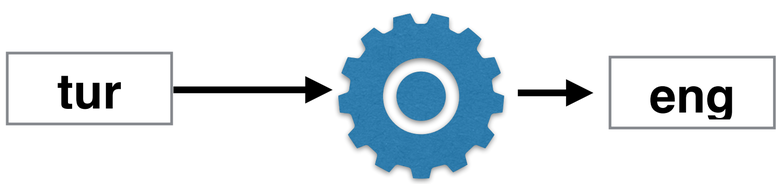
- Train a large multi-lingual NLP system

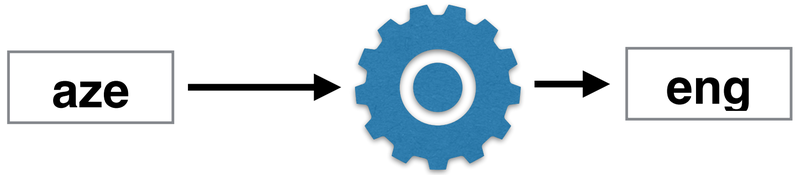


- **Challenges:** how to train effectively, how to ensure representation of low-resource languages

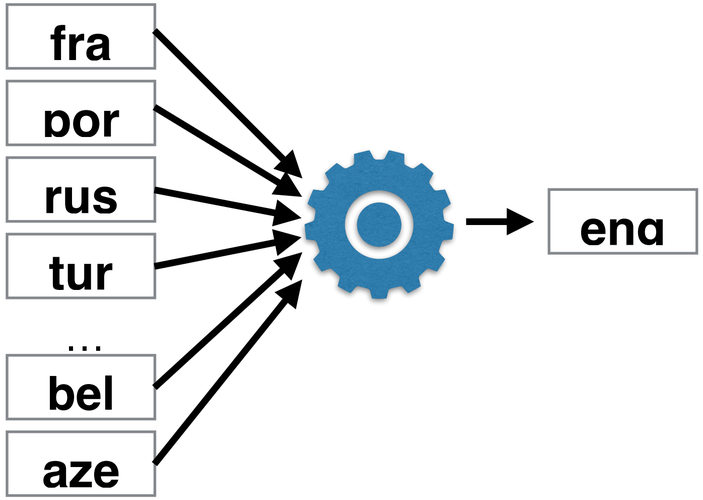
Transfer Learning

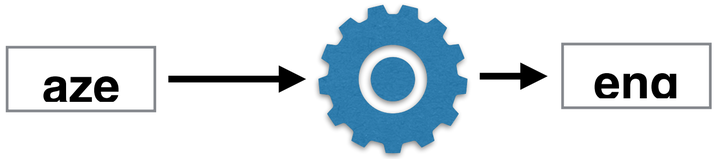
- Train on one language, transfer to another



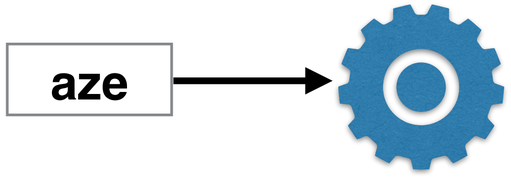


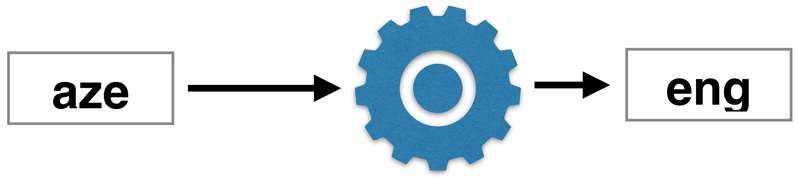
- Train on many languages, transfer to another





- Train on unannotated data, transfer to supervised tasks



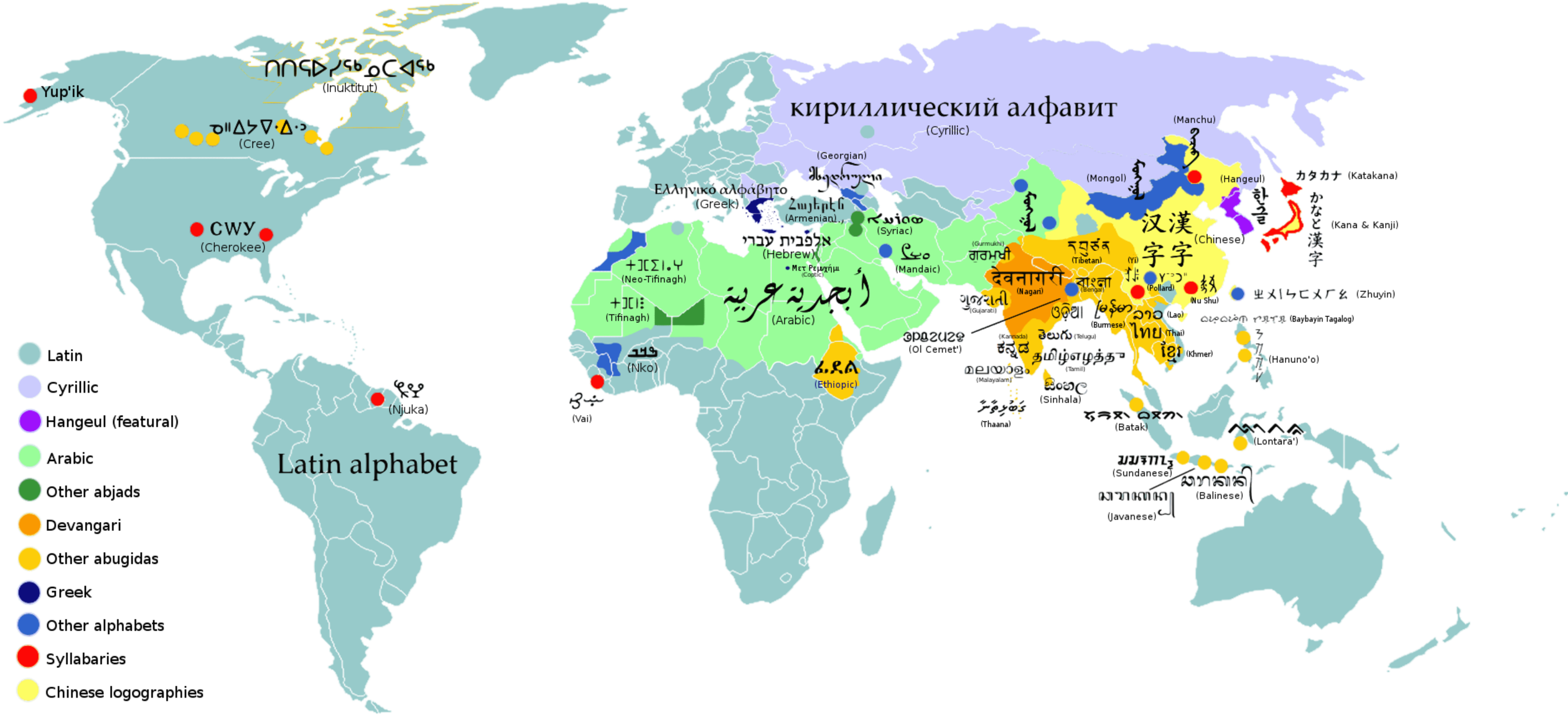


Multilingual Linguistics

Typology: The Space of Languages

- Languages across the world have similarities and differences
- **Typology** is the practice (and result) of organizing languages along axes

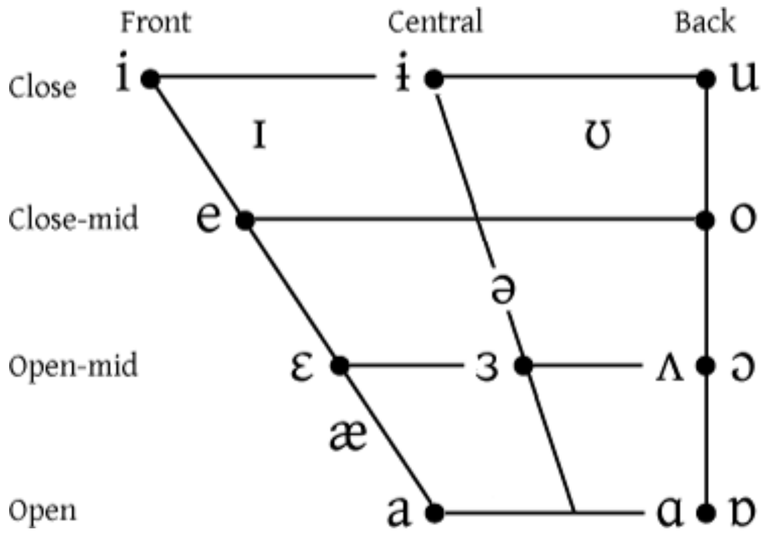
Scripts / Writing System



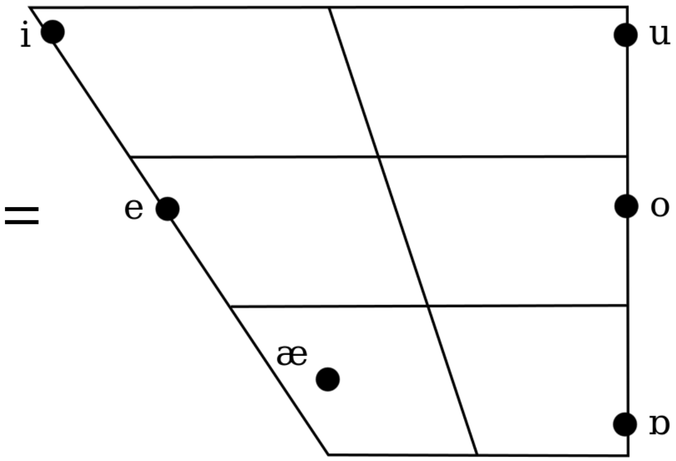
Phonology

- How is the language pronounced?
- e.g. what is the inventory of vowel sounds?

English =



Farsi =



Morphology, Syntax

- Morphology: what is the system of word formation?

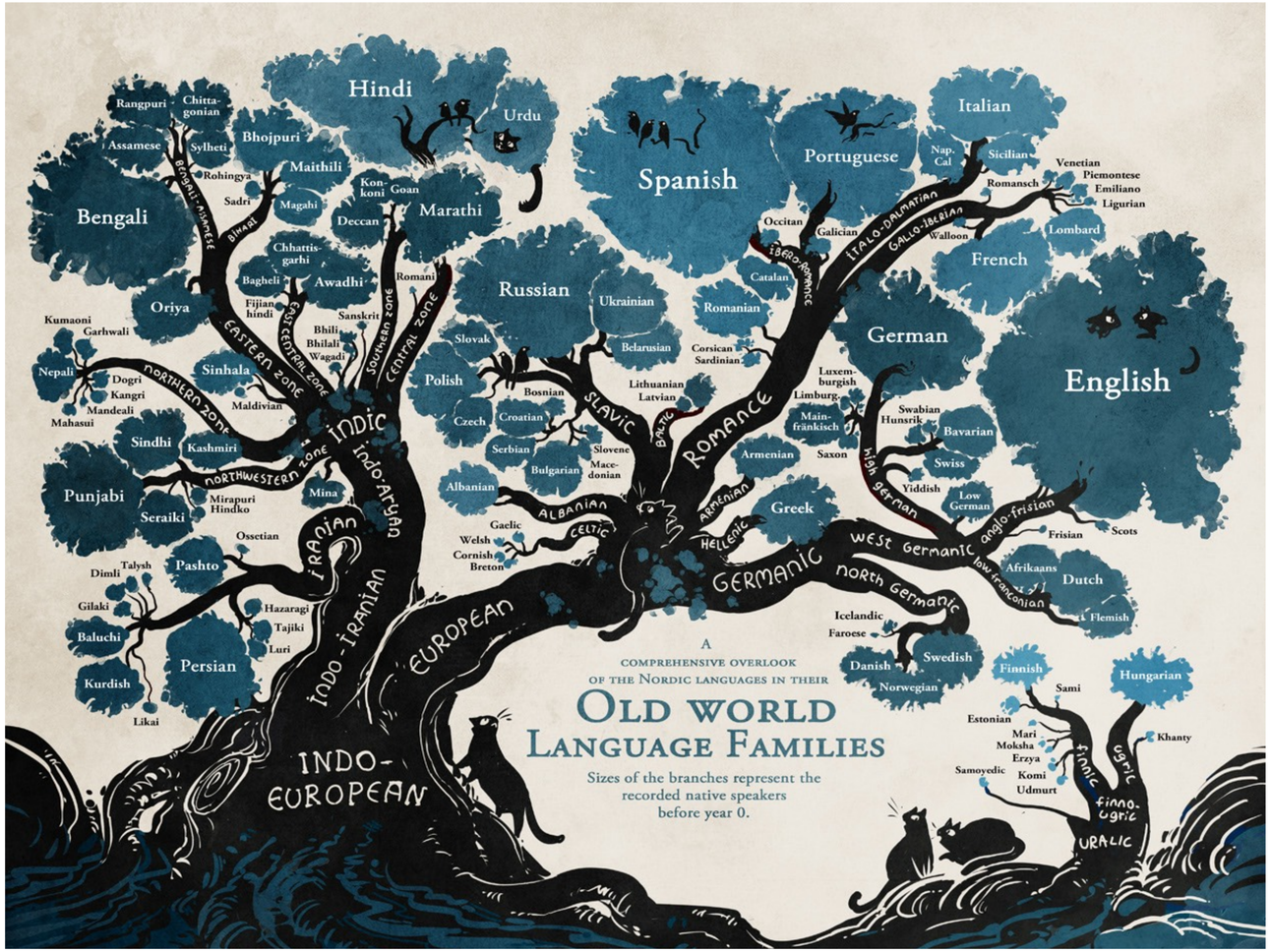
English = fusional: *she opened the door for him again*
Japanese = agglutinative: *kare ni mata doa wo aketeageta*
Mohawk = polysynthetic: *sahonwanhotónkwahse*

- Syntax: how are words brought together to make sentences?

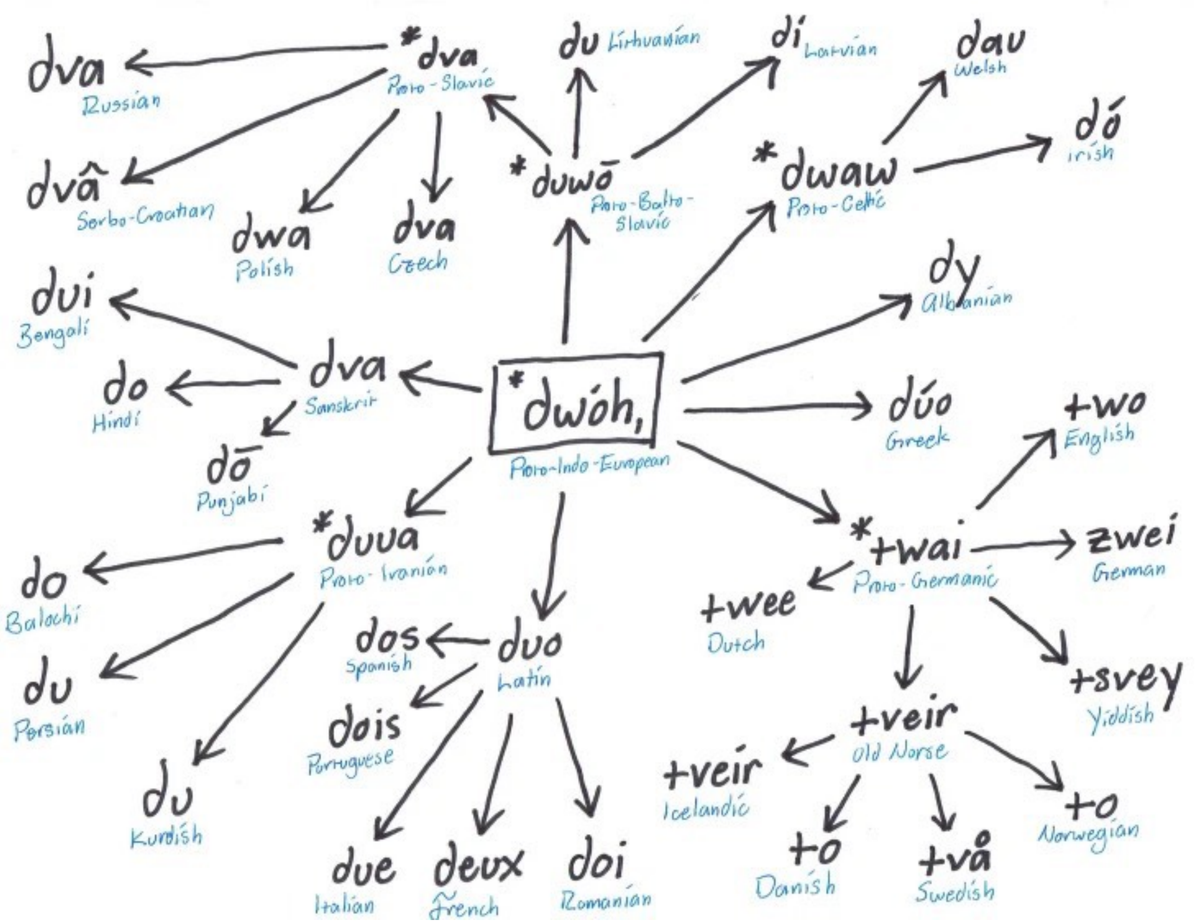
English = SVO: *he bought a car*
Irish = VSO: *cheannaigh sé carr*
Japanese = SOV: *kare wa kuruma wo katta*
Malagasy = VOS: *nividy fiara izy*

Language Varieties, Contact, and Change

- Languages contact from one-another, and gradually evolve
- Similarity in structure, but also words



https://www.sssscomic.com/comic.php?page=196

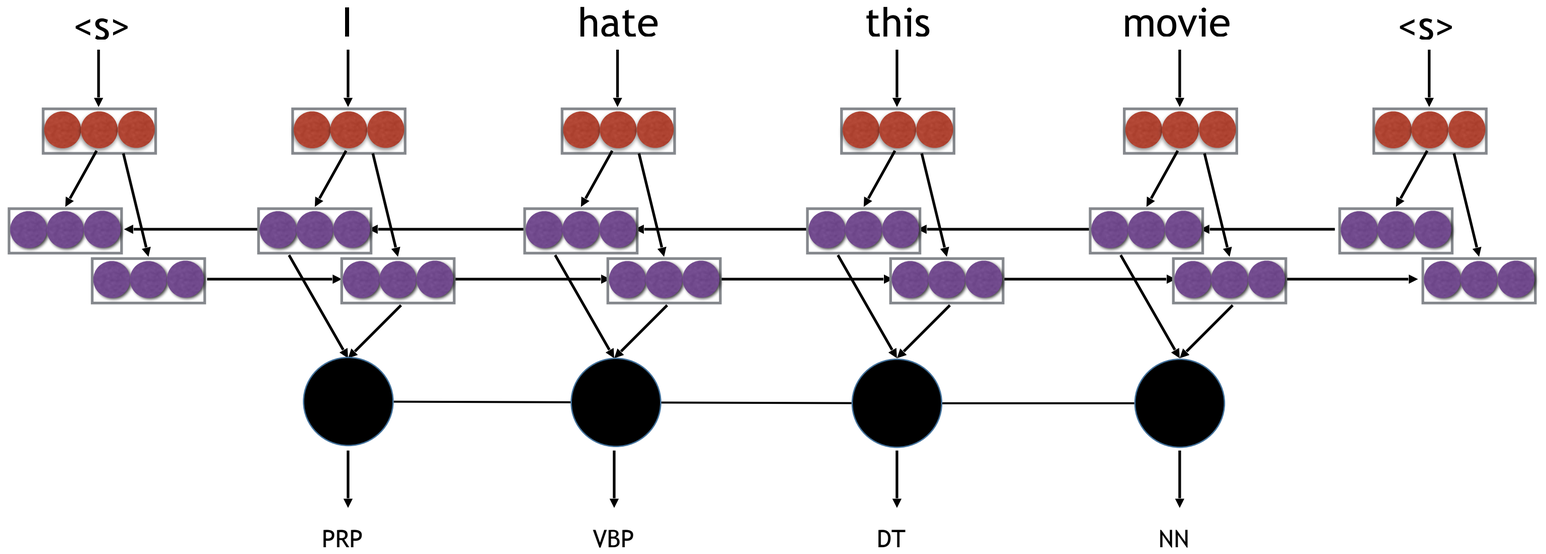


https://www.mentalfloss.com/article/68281/evolution-two-indo-european-language-family

Multilingual Applications

Sequence Labeling/Classification

- **Tasks:** language ID, POS tagging, named entity recognition, entity linking
- **Models:** sequence encoders, subword encoding



- **Data:** universal dependencies POS tags, wikipedia-based NER/linking

Morphology, Syntactic Analysis

- Morphological analysis

Much’ananayakapushasqakupuniñataqsunamá

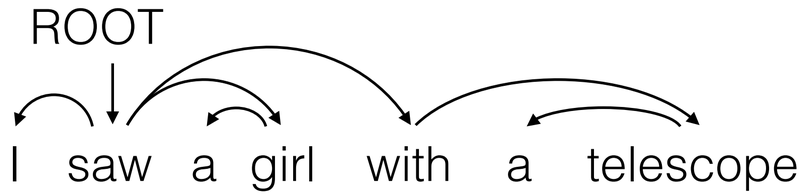
Much’a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

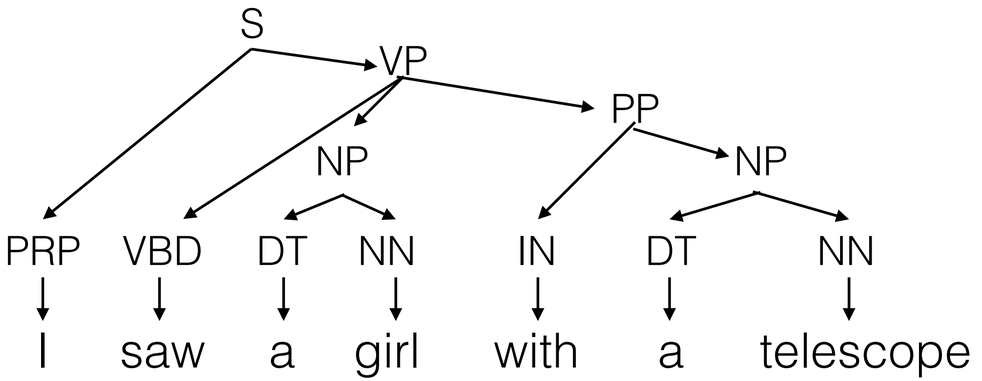
“So they really always have been kissing each other then”

Much’a	to kiss
-na	expresses obligation, lost in translation
-naya	expresses desire
-ka	diminutive
-pu	reflexive (kiss *eachother*)
-sha	progressive (kiss*ing*)
-sqa	declaring something the speaker has not personally witnessed
-ku	3rd person plural (they kiss)
-puni	definitive (really*)
-ña	always
-taq	statement of contrast (...then)
-suna	expressing uncertainty (So...)
-má	expressing that the speaker is surprised

(example from Quechua)

- Syntactic analysis

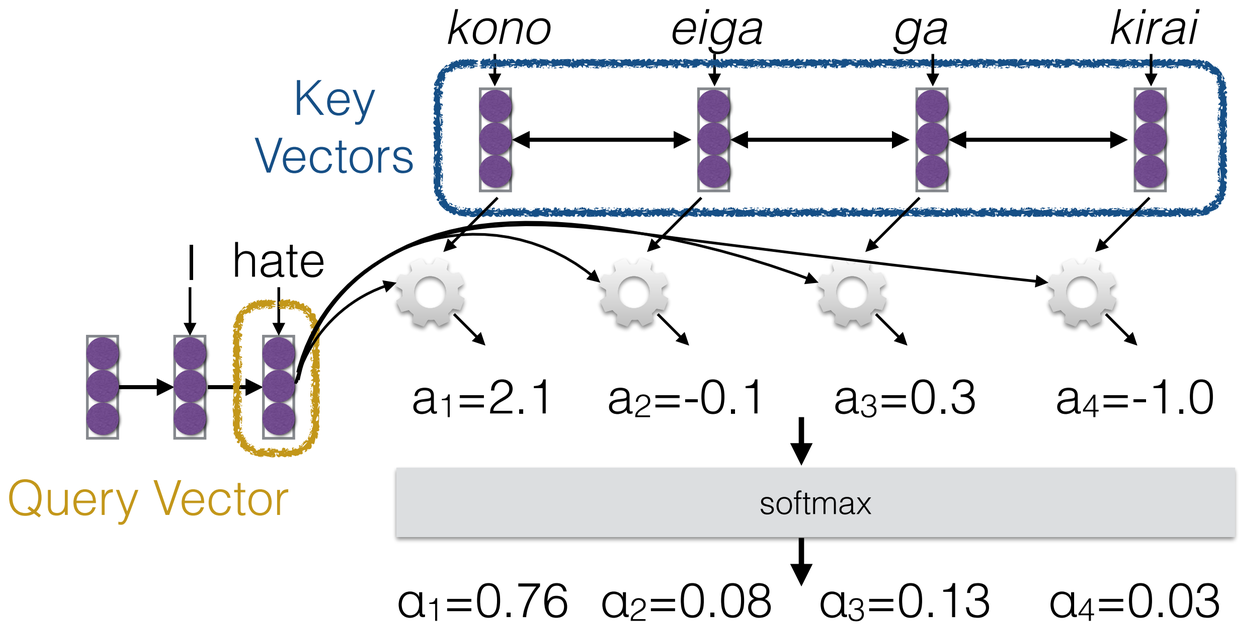




Machine Translation and Sequence-to-sequence Models

- Sequence-to-sequence problems
- Seq2seq models with attention
- Transformers
- Low-resource considerations

Input *X*	Output *Y*	Task
Text	Text in Other Language	Translation
Text	Response	Dialog
Speech	Transcript	Speech Recognitio
Text	Linguistic Structure	Language Analysis



Modeling Challenges

- Multilingual sharing of structure/vocabulary
- Balancing training over many languages
- Incorporating limited supervision for low-resource languages
- etc. etc.

Shinji: Speech

Logistics

Instructors/TAs

Instructors:

- Alan Black (code-switching, dialogue, speech synthesis)
- Graham Neubig (multilingual NLP, machine translation)
- Shinji Watanabe (speech recognition/synthesis, speech translation)

TAs:

- Xuankai Chang (speech recognition)
- Ting-Rui Chiang (machine translation, dialogue)
- Athiya Deviyani (number processing for speech synthesis)
- Patrick Fernandes (machine translation)
- Vijay Viswanathan (information extraction)

Class Format:

- ~30 minute **lecture**, with optional reading. There will be discussion questions.
- ~10 minute **language in 10**: introduce a language, in groups of 2-3.
- ~30 minute, **breakout room discussion** or **code/data/assignment walk-through**
- ~10 minute **summary**

Grading Policy

- Class/Discussion Participation: 15%
- Language in 10 Presentation: 5%
- Assignment 1 (Multilingual Sequence Labeling, individual): 15%
- Assignment 2 (Multilingual Translation, group): 20%
- Assignment 3 (Multilingual Speech Recognition, group): 20%
- Project: 25%

Discussion Period for 1/20

- No discussion for Thursday, but we will look at assignment 1/code walk

Alan: Language in 10!