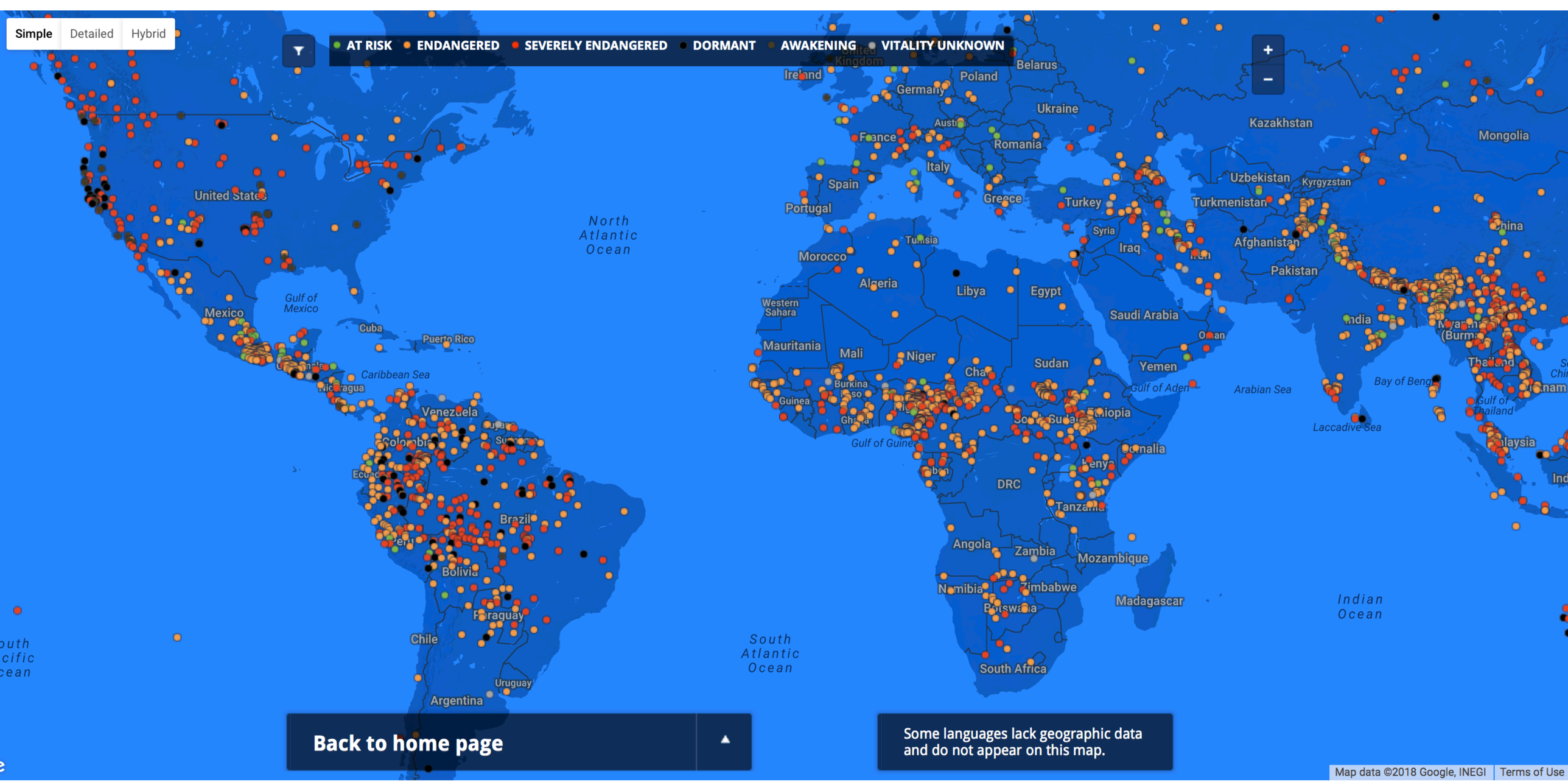
CS11-737
Multilingual Natural Language Processing

Lei Li

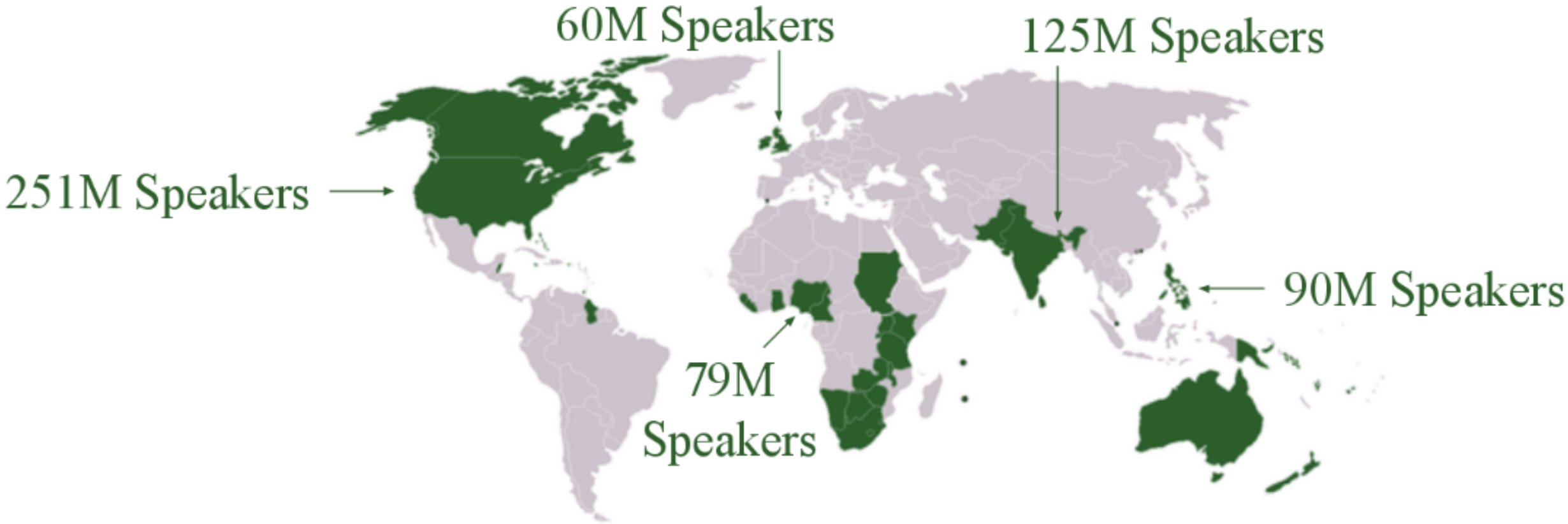
https://lileicc.github.io/course/11737mnlp23fa/

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

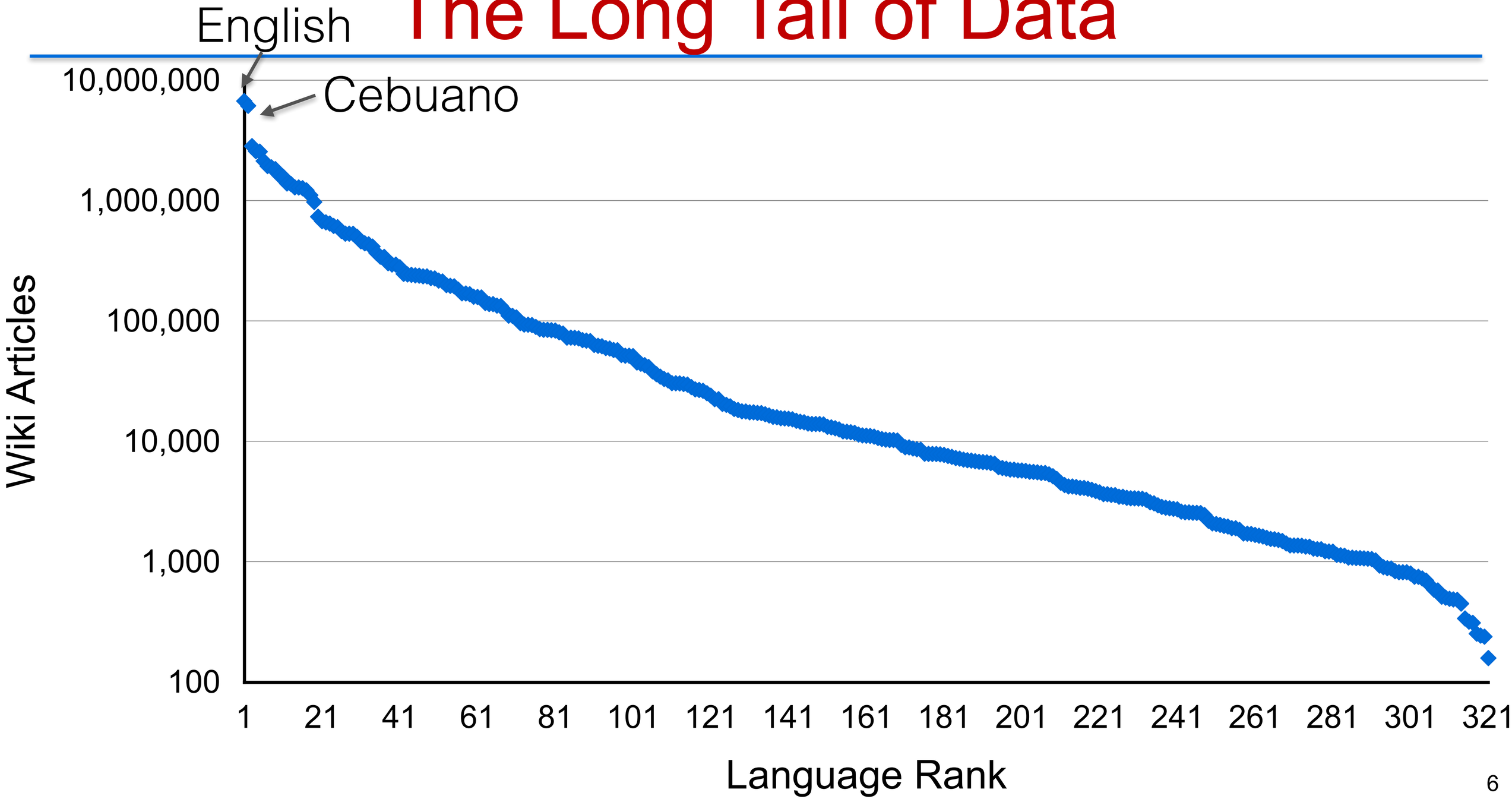
Machine Learning Models

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

Input X	Output Y	Task
Text in src lang	Text in Other Language	Translation
Text	Response	Dialog
Speech	Transcript	Speech Recognition
Speech in src lang	Text in other lang	Speech to text translation
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 NLP system in a low-resource language that you do not know

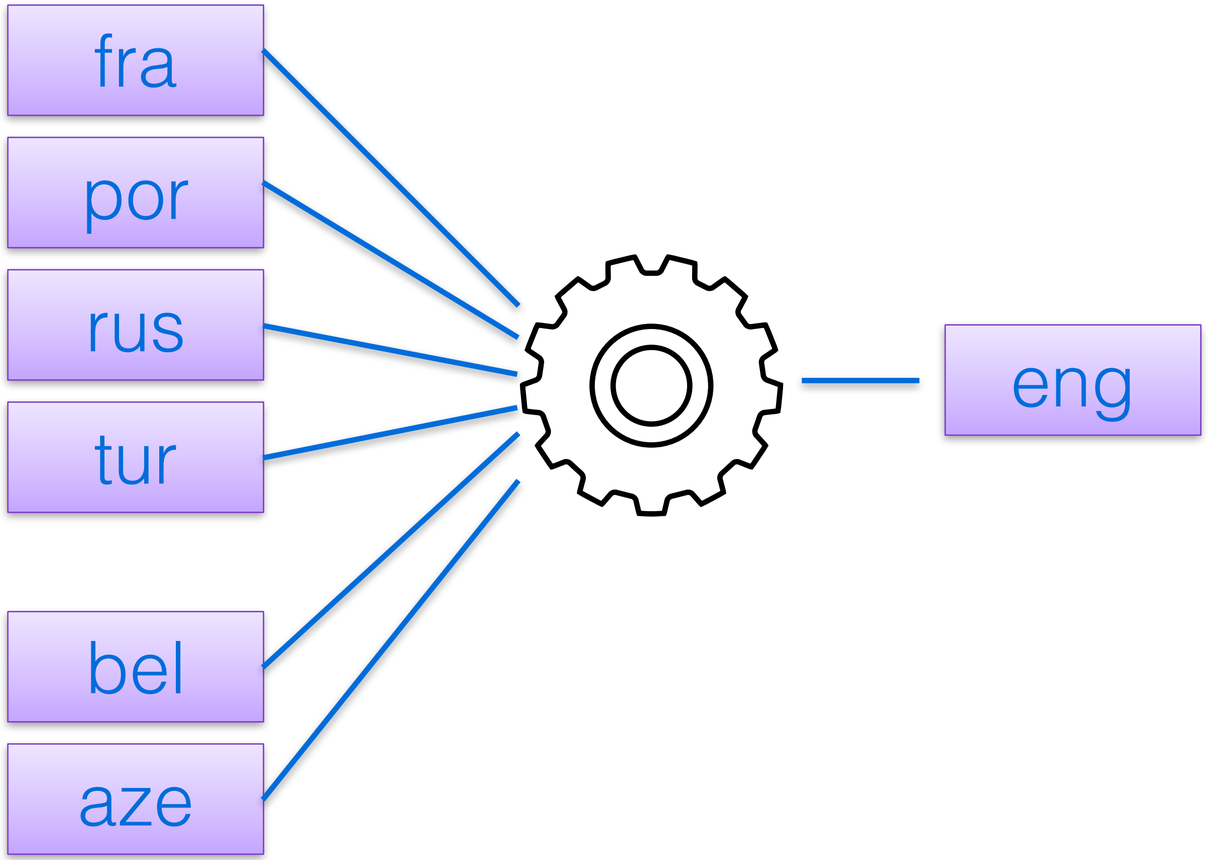
Training Multilingual NLP Systems

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

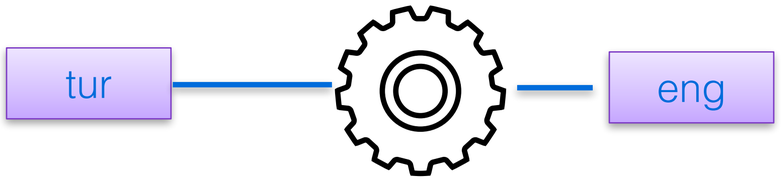
- Train a large multi-lingual NLP system

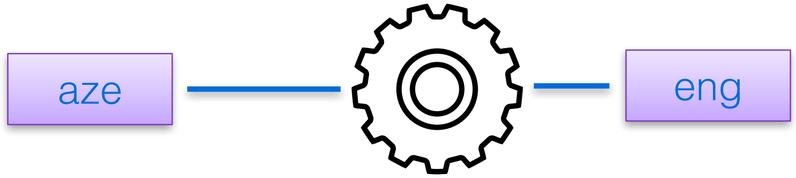


- Challenges: how to train effectively, how to ensure representation of low-resource languages

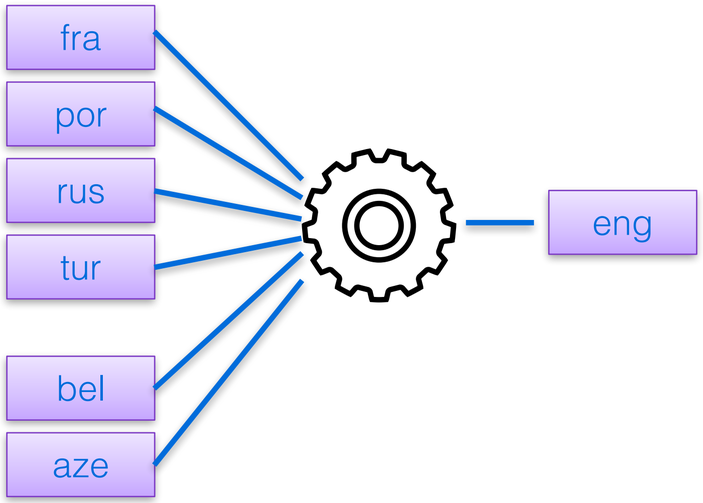
Transfer Learning

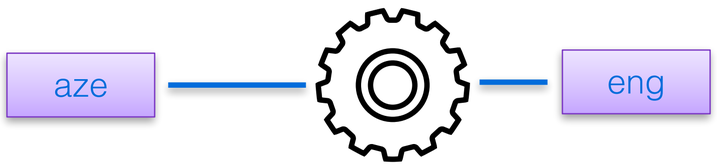
- Training on one (pair) language, transfer to another





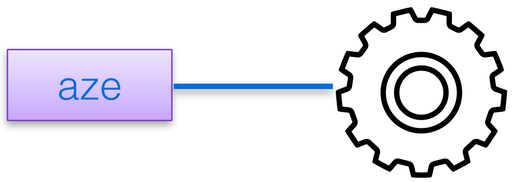
- Train on many languages, transfer to another

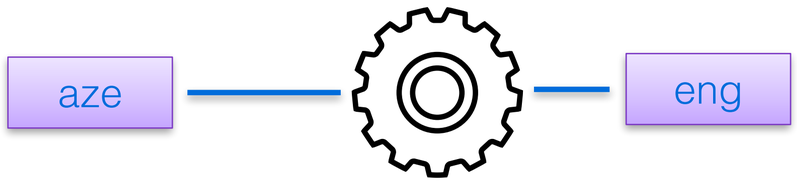




Pre-training

- Unsupervised or Self-supervised training on unannotated data, then fine-tuning on annotated data



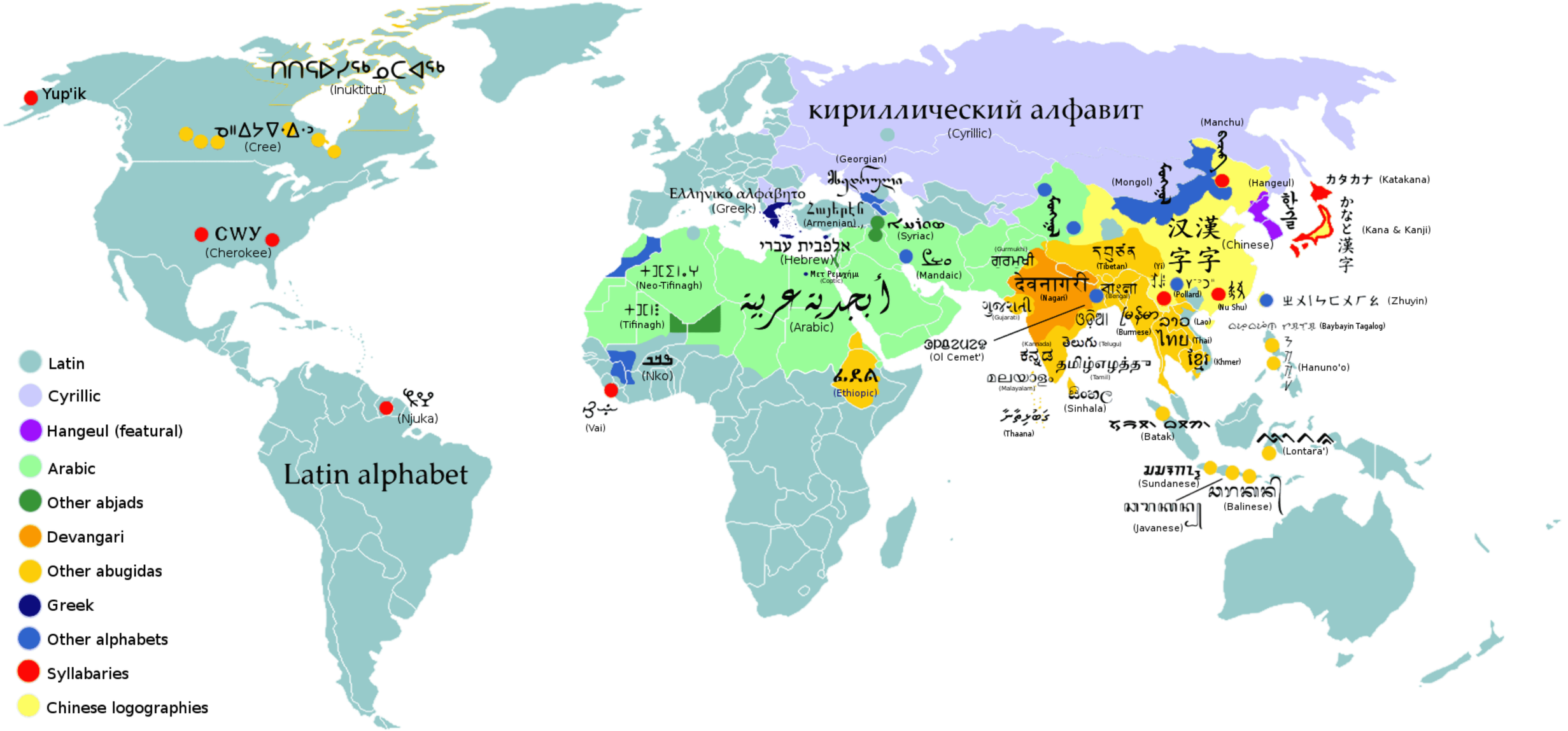


Multilingual Linguistics

Typology: The Space of Languages

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

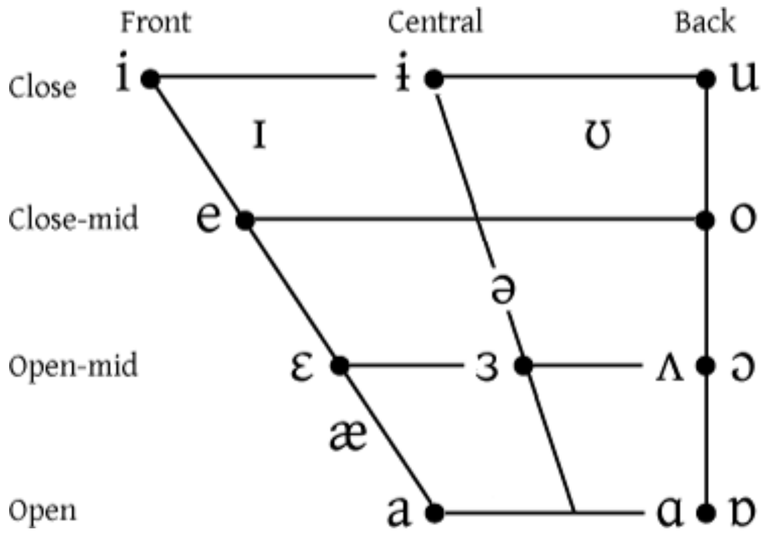
Scripts / Writing System



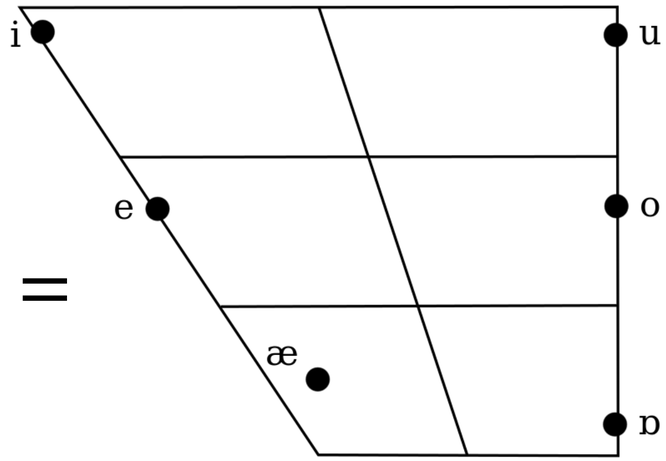
Phonology

- 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

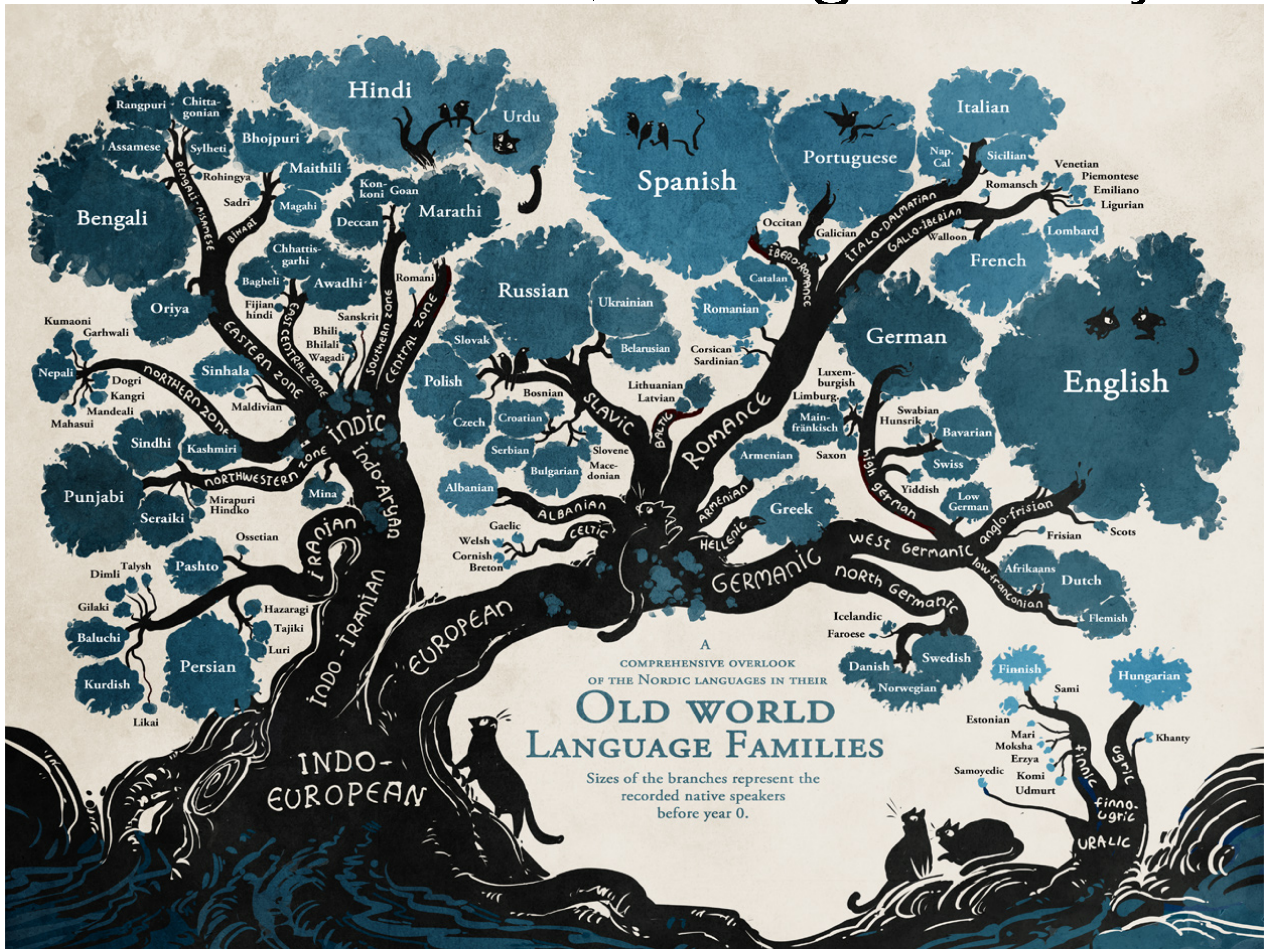
Japanese = SOV: kare wa kuruma wo katta

Irish = VSO: cheannaigh sé carr

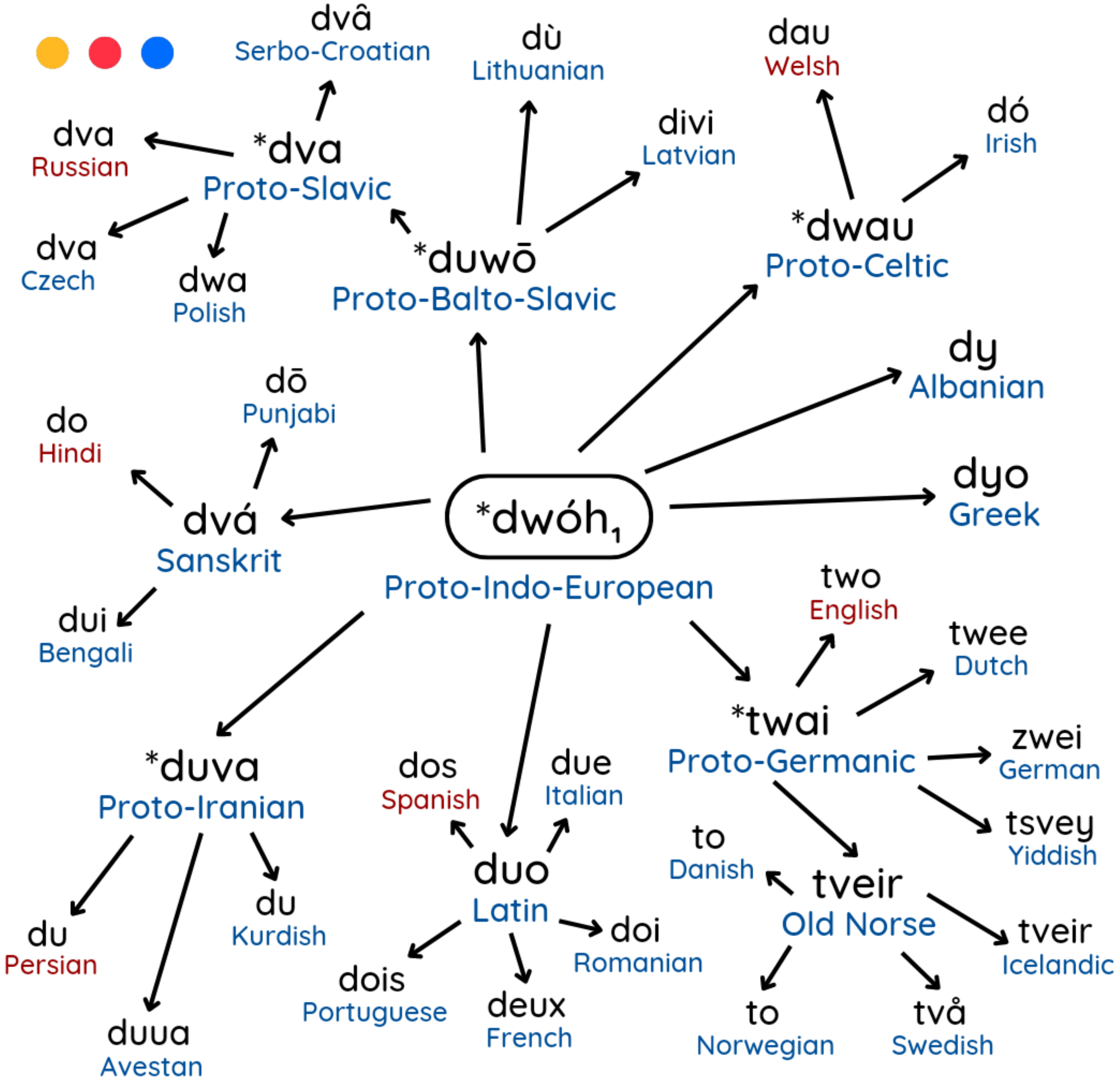
Malagasy = VOS: nividy fiara izy

Language Varieties, Contact, and Change

- Languages contact from one-another, and gradually



- Similarity in structure, but also words

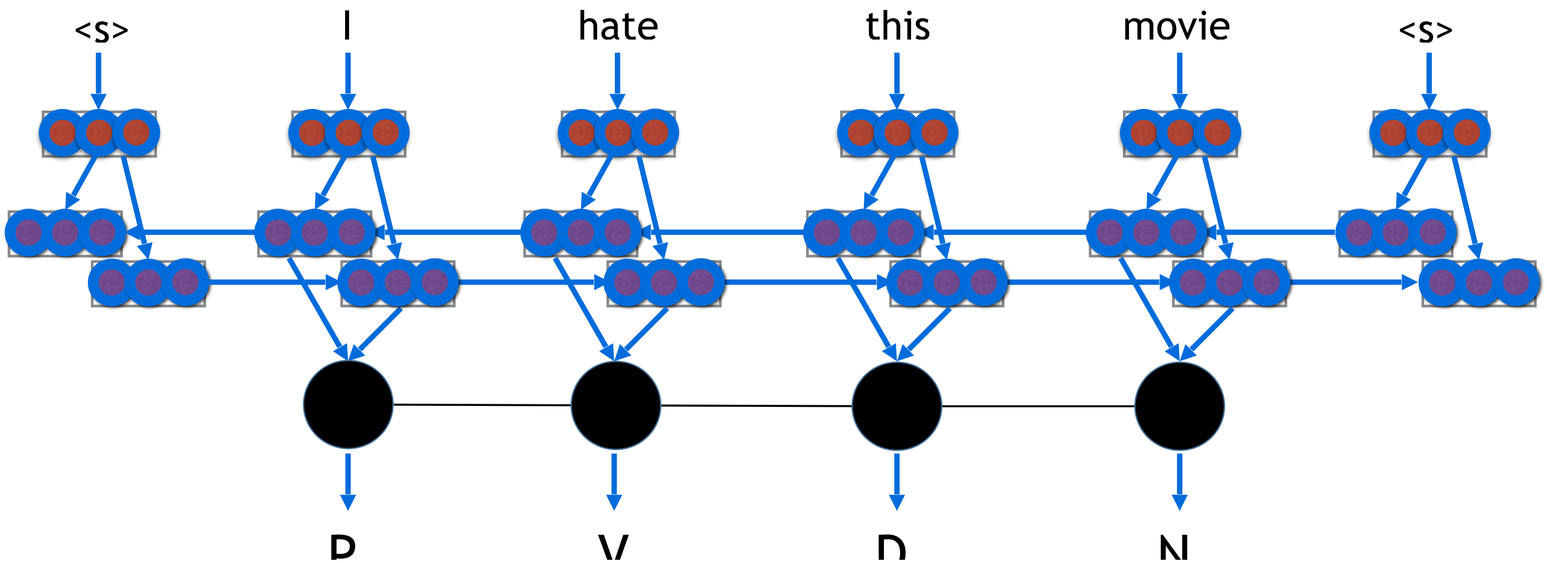


w.sssscomic.com/ … s.com/article/68281/evo

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

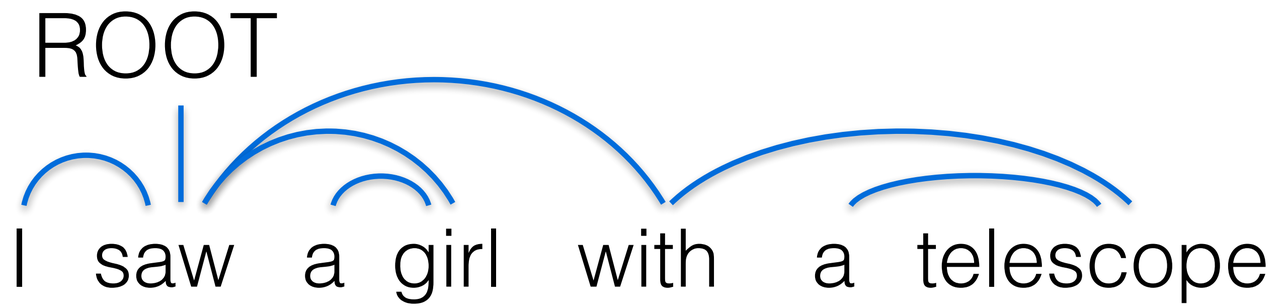
Much'a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

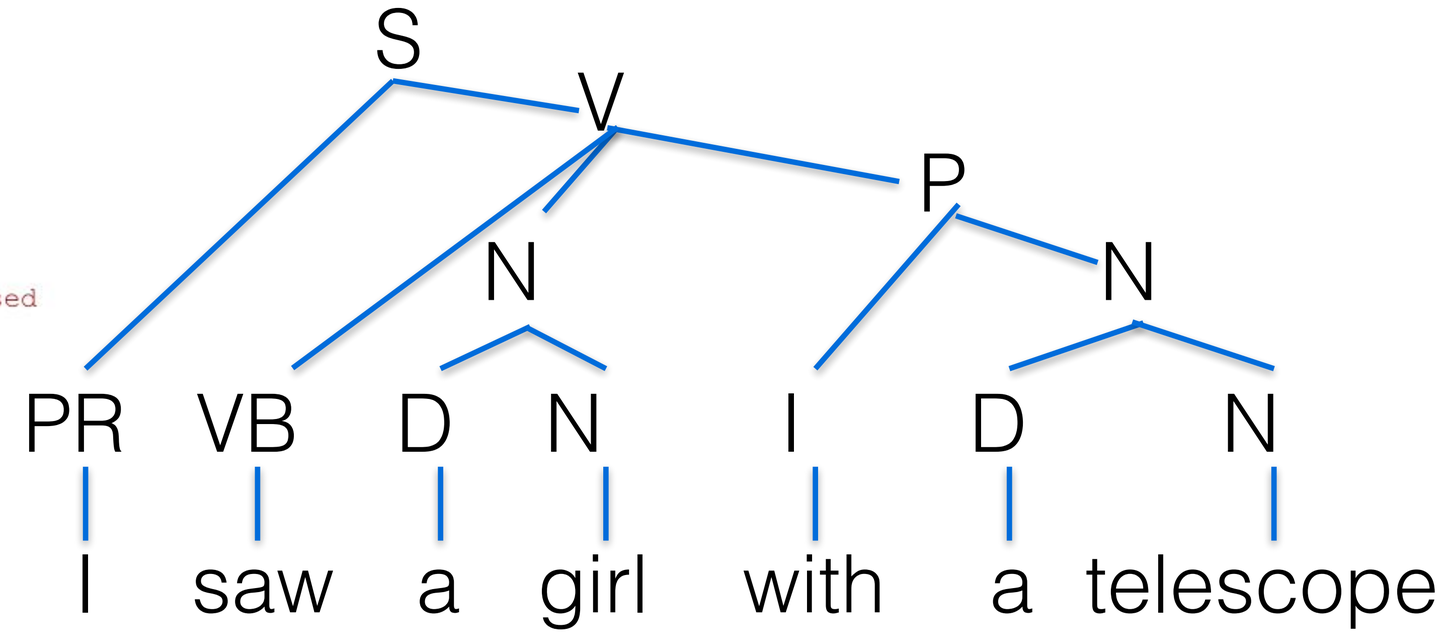
"So they really always have been kissing each other then"

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)

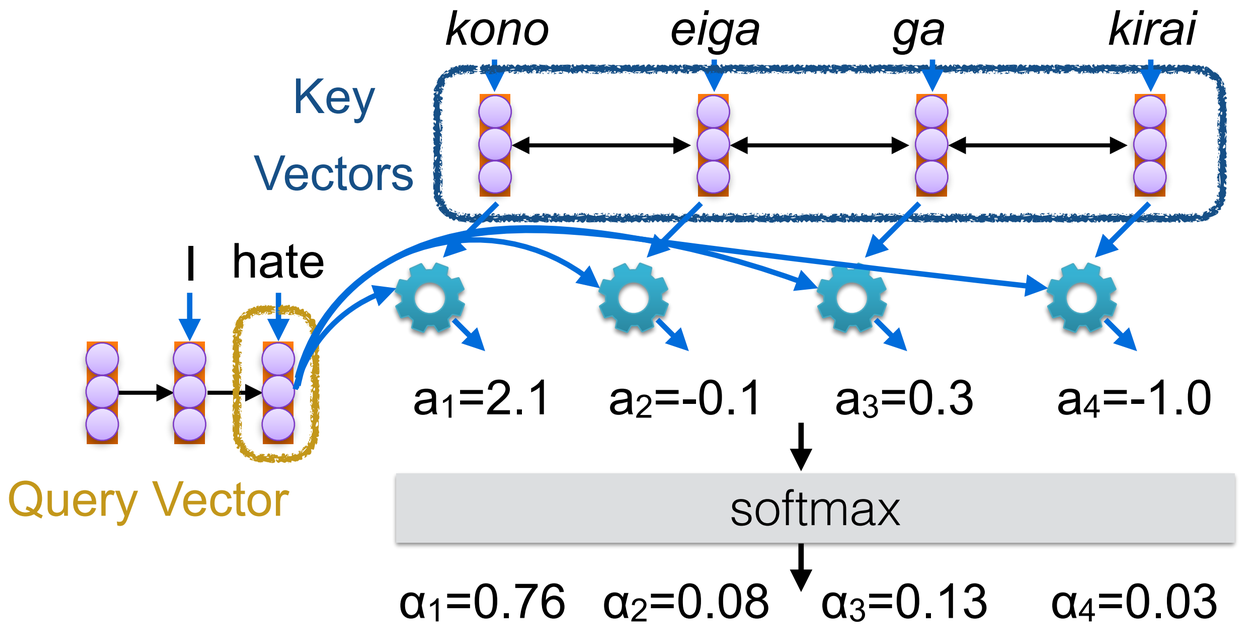
- Syntactic analysis





Machine Translation

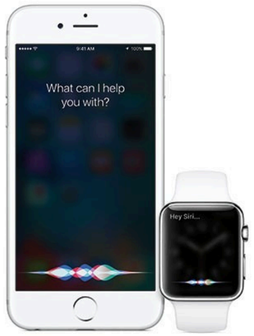
- Sequence-to-sequence problems
- Seq2seq models with attention
- Transformers
- Low-resource domains



Modeling Challenges

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

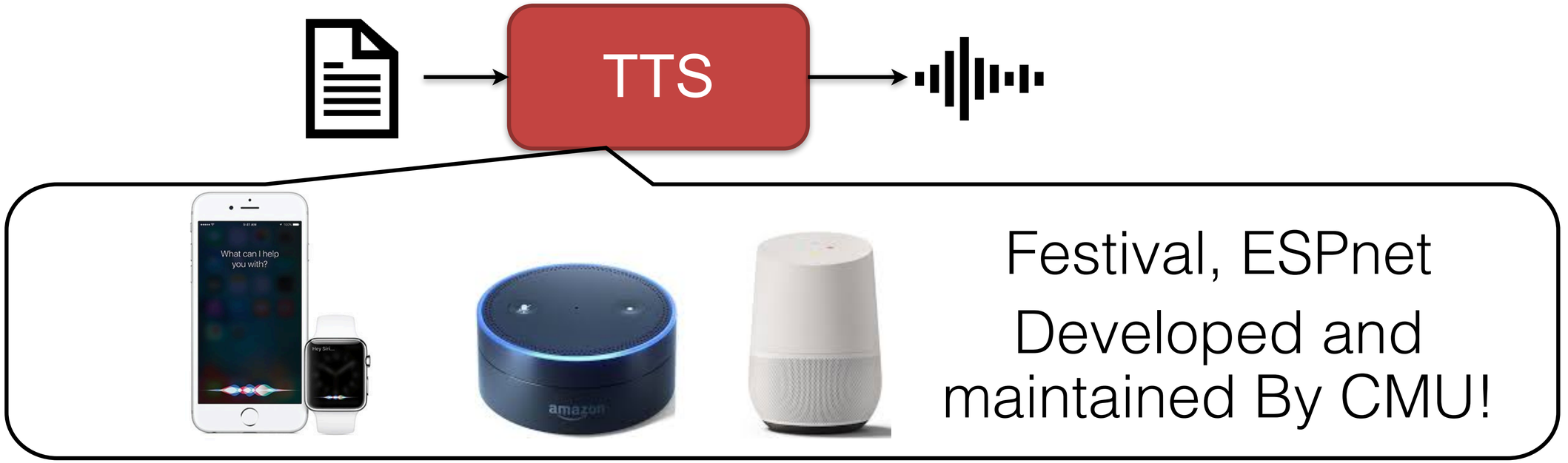
Automatic Speech Recognition (ASR)



Sphinx, Janus, ESPnet, etc. Developed and maintained By CMU!

Widely used in many applications!

Speech Synthesis (Text to Speech, TTS)

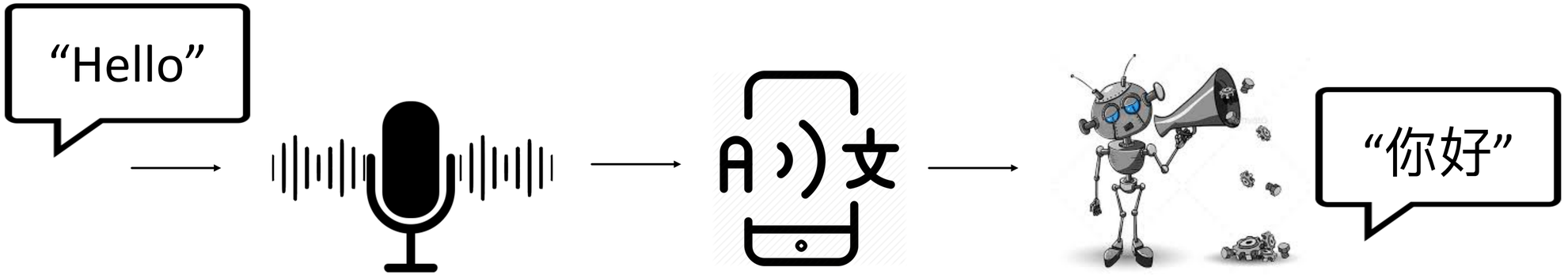


Inverse problem of ASR

Speech Translation (ST)

- source language ***speech(audio)*** → target lang ***text***

- source language ***speech*** → target lang ***speech***



Ultimate goal is crossing all language barriers in human communication

Relation to other courses

- The course mainly covers high-level explanations and system descriptions of ASR, TTS, and related technologies
  - If you want to know more about them, please consider to take “11-751 Speech Recognition and Understanding” and “11-492 Speech Processing” ☺
- Most of MT, ASR and TTS technologies are studied with major languages (English, Chinese, German, French, Japanese, etc.)
  - Rich resources, accumulated knowhow, marked priority
- What can you learn? The lectures will focus toward how to build NLU/MT/ASR/TTS/ST systems in any language

Logistics

Instructors/TAs

- Instructors:
  - Lei Li (Machine Translation, Multilingual NLP, LLM)
  - CMU -> UC Berkeley -> Baidu Research -> ByteDance —> UC Santa Barbara —> CMU
  - You may use my translation system on Tiktok/Lark, or WeiChat app (火山翻译, VolcTrans） or translate.volcengine.com
- TAs:
  - Simran Khanuja (multilingual LM, multimodal translation)
  - Sayali Kandarkar (multimodal generation)
  - Possibly another TA

Class Format

- 45 minute lecture, with optional reading. There will be discussion questions.
- ~10 minute language in 10: introduce a language, in groups of 2.
- ~25 minute (once every week), breakout discussion or code/data/assignment/project walk-through

Grading Policy

- Class/Discussion Participation: 5%
- Language in 10 Presentation: 5%
- Assignment 1 (Multilingual Translation, individual): 20%
- Assignment 2 (Multilingual Speech Recognition, group): 20%
- Assignment 3 (A blog post on recent papers related to multilingual NLP, group of 2): 15%
- Project: 30% (5% for mid-term report, 25% final presentation + report)

HW3 Blog

- Group of 2
- Read one paper and write a popular science or step-by-step cooking article about one paper in Multilingual NLP
  - choose from the suggested list (no overlap)
  - https://lileicc.github.io/course/11737mnlp23fa/multiling_reading.html
  - You may choose other paper but need to be confirmed with Instructor.
- Try to Reproduce results
  - no need to re-train
  - but need to use their published model to inference on same or extra data
  - Case study
- Indicate whether ok to put public

HW3 Blog

- Writing suggestion
  - In Markdown (with math support), or HTML (w/ javascript, no php)
  - more than 1/4 of content (the problem, challenge, intuition etc) should be understood by high school students (layman's term, non-expert)
  - about 1/2 of content understood by college students
  - no more than 1/4 of content understood by NLP researchers
  - Use visualization, figures, tables, and show-case examples
  - Interactive (e.g. via js) could be helpful as well
  - ChatGPT allowed if used in the same way as Grammarly (grammar correction). You should create your original content.

HW3 Blog

- Writing Template:
  - VuePress template: https://github.com/lileicc/blog
    - You may create and edit a new markdown file under blogs/ directory
- Example:
  - https://lileicc.github.io/blog/mt/mrasp/
  - https://towardsdatascience.com/byte-pair-encoding-subword-based-tokenization-algorithm-77828a70bee0
  - https://lena-voita.github.io/posts/nmt_training_through_smt_lens.html
  - https://lileicc.github.io/blog/mt/VOLT/

Project

- Group of 2~3 (should be different from other assignment group)
- A semi-research project on MT, ST, ASR, TTS, Multilingual transfer, etc.
- Proposal: no grade but we will provide feedback
  - Please include: project description, data, evaluation procedure/metric, estimated computation, other resources
- Mid-term report: 5%
  - Everything in proposal with adjustment, project description, data, evaluation procedure/metric, computation, a baseline model and baseline results.
- Final Project: 25%
  - Poster presentation in-class
  - Final report (content similar to a conference/workshop paper)

Project inspiring ideas

- Develop a working MT/ASR/TTS/ST system for some new (no high-quality available MT) and low-resource languages (e.g. Spanish-to-Tamil), explore and solve challenges along the way
- Improving methods to better utilize monolingual data
- Extending and improving Vocabulary and Tokenization for NMT
- Improving evaluation quality and efficiency, certain human-assisting tools for evaluation, conduct study.
- Computer-assisted and interactive translation methods
- MT for multimodal data, e.g. video translation, speech translation
- Integrating domain knowledge into MT/ASR/TTS system
- Novel hardware-based MT/ASR/TTS/ST system, e.g. Compress MT model to very small size and build a system (with inference but not training) on mobile phones, or extending existing CUDA library (e.g. LightSeq) to support more complex models.
- Extending a massive NMT (e.g. LegoMT) to a few more languages

Language in 10

- History, geography, social position
- Linguistic: morphology, grammar, phonology
- Examples of something (linguistically) interesting about the language
- Status with respect to resources (data, software)
- Influences, social use, issues that may affect collection/ access
- Example:
  - https://www.youtube.com/watch?v=JpOJiL9ZF7w (towards the end)

Computing Resource

- We will distribute complementary AWS credits for course project
- Please let me (TA) know your AWS account id if interested.
- Additionally, you may use colab/kaggle.

Discussion for Thursday August 31

- Reading Assignment:
  Ponti, E.M., O'horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E. and Korhonen, A., 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. Computational Linguistics, 45(3), pp.559-601.
- Discussion Question:
  What are some unique typological features of a language that you know, regarding phonology, morphology, syntax, semantics, pragmatics?