ECE365: Introduction to NLP

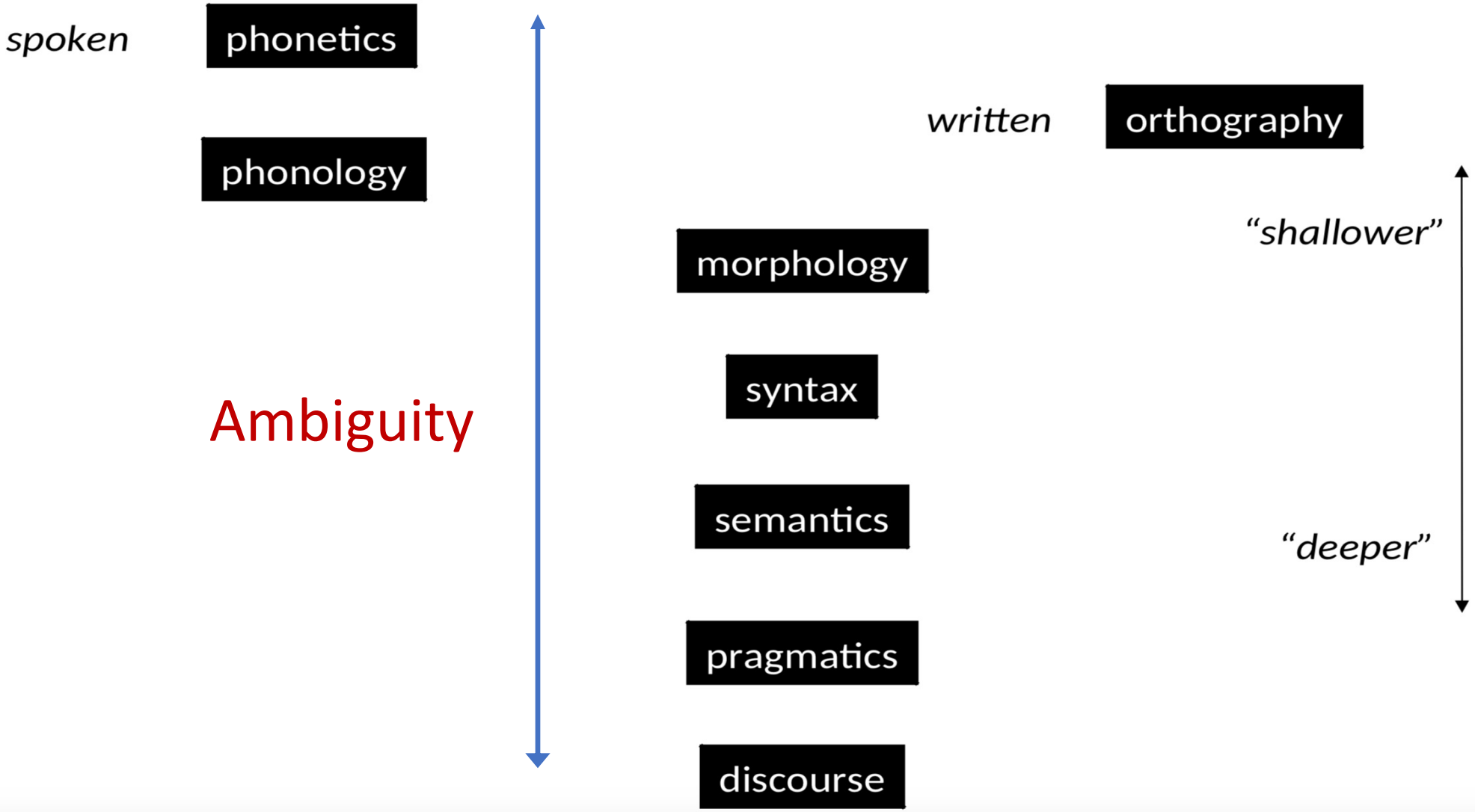
Spring 2021

Lecture 8: Summary, Non-English NLP, Ethics

Logistics

- No final exam for this course
- Quiz 2 tonight and conflict tomorrow

Levels of Linguistic Knowledge

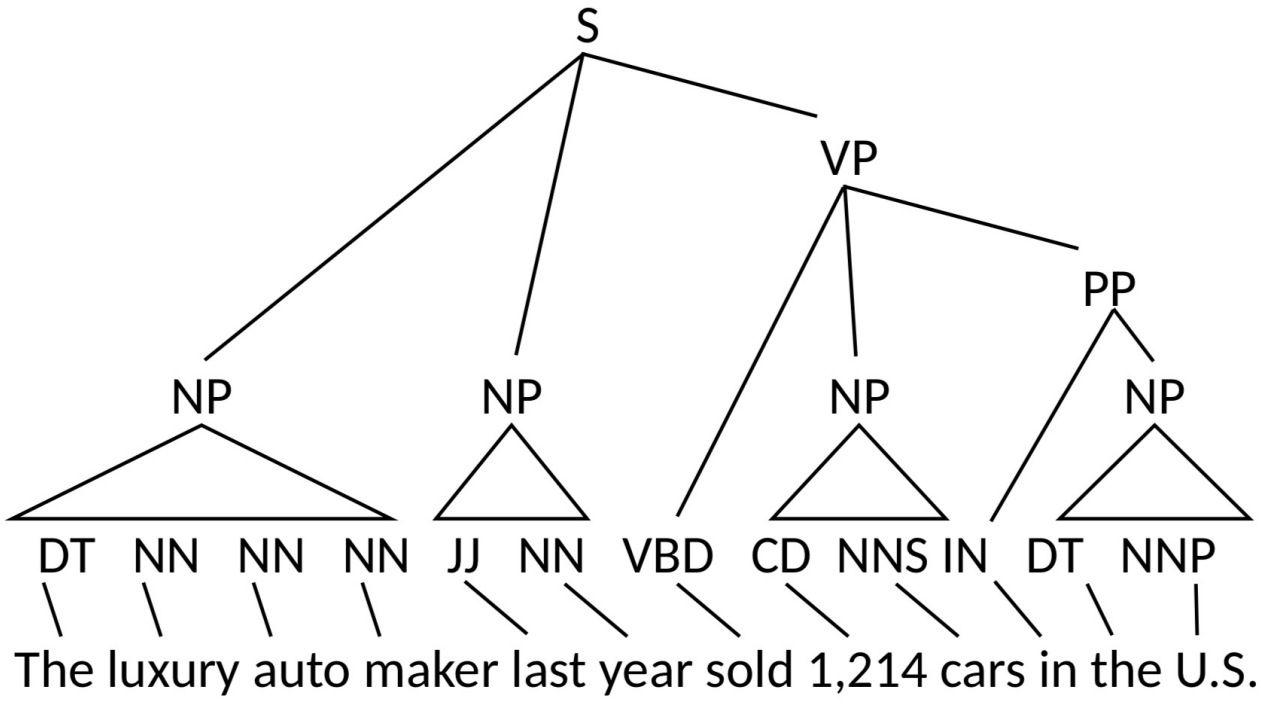


What we have seen

- Words as units of language, classification of text using words
- Word sequences as units of grammar, classification using word as sequence
- Word meaning using distributional hypothesis

What we have not seen

- Models of clause structure (syntax),
- sentence meaning (semantics), discourse (coherent structure of language above the level of clauses and sentences)



What we have not seen

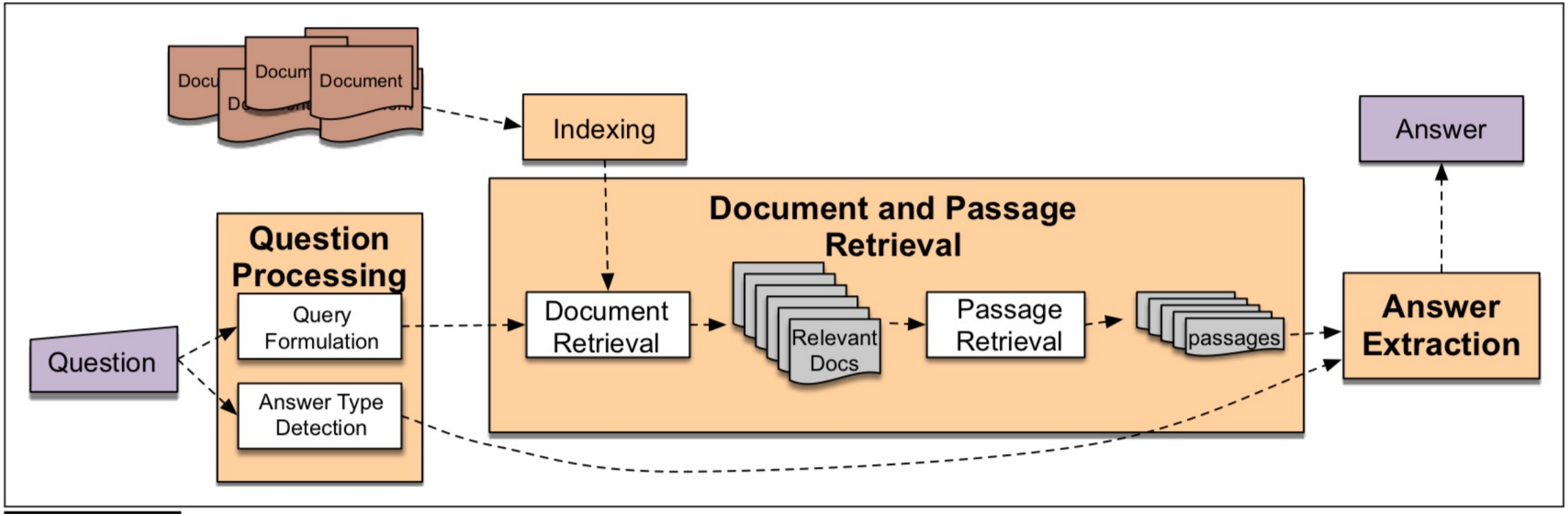
- Neural network models for NLP
- Other applications:
  - Information extraction
    - **Input**: text, empty relational database
    - **Output**: populated relational database

Senator John Edwards is to drop out of the race to become the Democratic party's presidential candidate after consistently trailing in third place.
In the latest primary, held in Florida yesterday, Edwards gained only 14% of the vote, with Hillary Clinton polling 50% and Barack Obama on 33%. A reported 1.5m voters turned out to vote.

State	Party	Candidate	Fraction
FL	D	Edwards	0.14
FL	D	Clinton	0.50
FL	D	Obama	0.33

What we have not seen

- Other applications:
  - Questions answering



More NLP-related courses

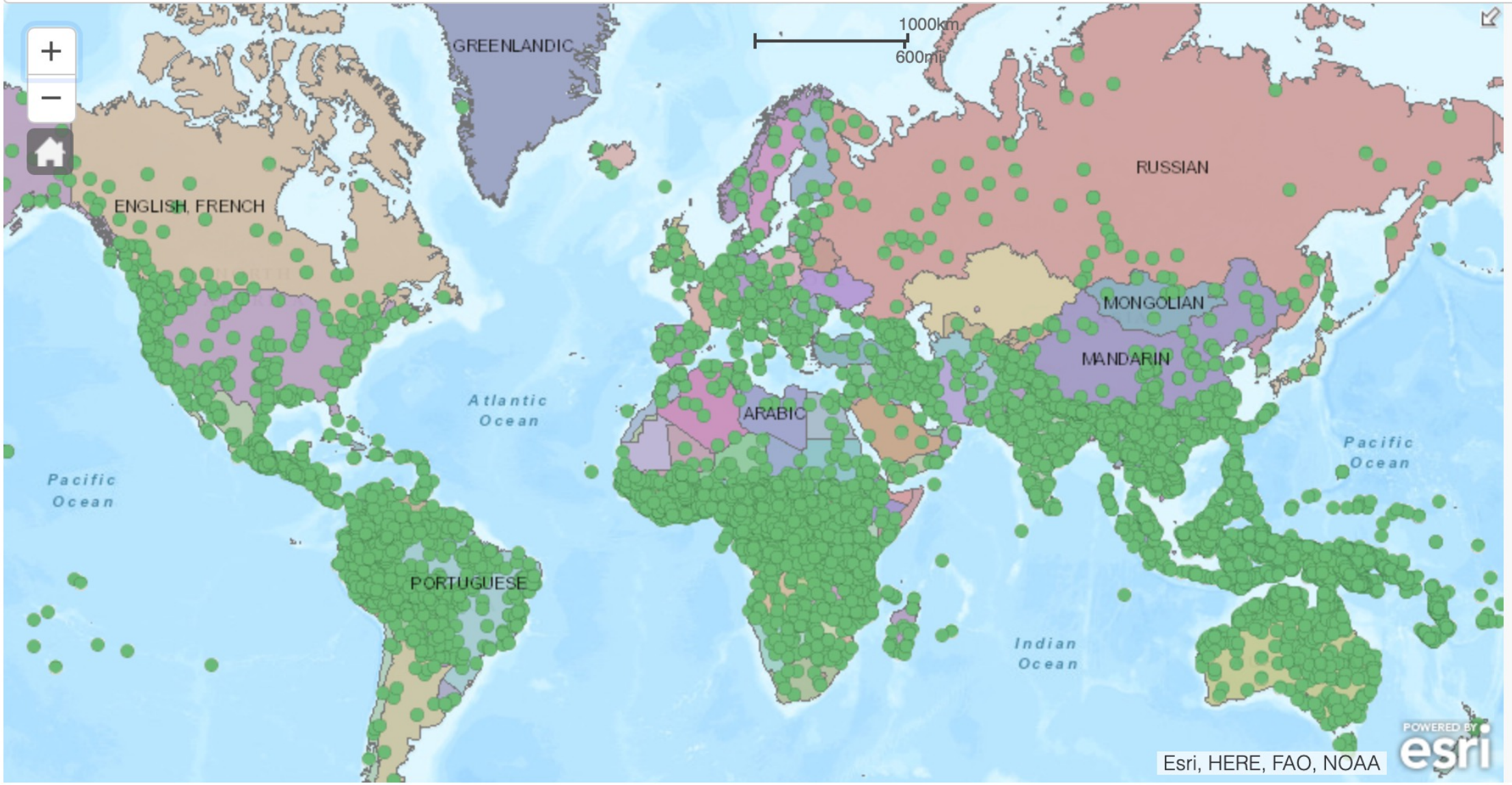
- CS 440 Artificial Intelligence
- CS447 Natural Language Processing
- CS410 Text Information Systems

Multilingual world

- How many languages?
  - Ethnologue: over 7400

LANGSCAPE DISCOVER ABOUT PARTNERS CONTRIBUTE BLOG

http://www.langscape.umd.edu/map.php



Why Language Technologies?

- Keyboard input
- Auto-complete
- Spellcheck
- Speech recognition
- Speech synthesis
- Information retrieval, search engines (and morphology)
- Grammar check
- Translation
- Question Answering

Upstream Language Technology

- Language detection
- Part of Speech tagging
- Parsing
- Semantic Role Labeling
- Named Entity Recognition
- Summarization
- Translation
- Question answering/ Information extraction

Enabling Technologies

- Character Encoding
- Fonts and rendering technologies
- Input methods
- Standard orthography/spelling
- Enough text/speech to train models

Languages with Significant Technologies

- Mandarin
- Spanish
- English
- Hindi
- Arabic
- Portuguese
- Bengali
- Russian
- Japanese
- Punjabi
- German
- Javanese
- Wu
- Malay/Indonesian
- Telugu
- Vietnamese
- Korean
- French
- Marathi
- Tamil

A Not-so-Uncommon Situation

Language

- Village language: **Kachai**
- Local language: **Tangkhul**
- Regional language: **Meithei**
- National language: **Hindi**
- Global language: **English**

Domain

- Family and village life
- Primary school, etc.
- Secondary school, etc.
- Military, etc.
- Higher education, etc.

Working with other languages

- Other languages present some problems not seen in English at all!

- Some of our algorithms have been specified to English
  - May not make sense for working with other languages
  - Neural methods typically tuned to English resources may not work for low-resource languages
- Questions
  - What other language-specific phenomena/challenges do we need to solve?
  - How can we leverage existing resources to do better in other languages without just annotating massive data?

Exclusion

- Primary source of annotated data is English
  - Dialects
  - Other languages (Non-European/Non-CJK)
  - Codeswitching

Take Away

In the near term, it is unlikely that current language technologies, no matter how clever the machine learning behind them, will be able to deal with languages that are not English in a perfectly language independent way.

ML and NLP

- Aggregate textual information to make predictions
- Hard to know why some predictions are made
- More widely used in various applications/sectors
  - Machine translation
  - Dialog systems

Ethics in NLP

What can go wrong?

- Bias amplification: systems exacerbate real-world bias rather than correct for it
  - AI system learning from training data 67% images involving cooking have women, predicts 80% women cooking at test time
  - Word2vec trained on news text:
    - ‘man’ - ‘computer programmer’ + ‘woman’ =
    - Father: doctor :: mother : nurse
  - Debiasing methods of importance

Unethical use

- **Dangers of automation**: automating things in ways we don't understand is dangerous
- **Exclusion**: underprivileged users are left behind by systems
- **Bias amplification**: systems exacerbate real-world bias rather than correct for it
- **Unethical use**: powerful systems can be used for bad ends
  - Generating fake news/misinformation

Dangers of relying on automation

What can go wrong?

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

BUSINESS NEWS OCTOBER 9, 2018 / 10:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Final Words

- You will face choices: what you choose to work on, what company you choose to work for, etc.
- Tech does not exist in a vacuum: you can work on problems that will fundamentally make the world a better place or a worse place (not always easy to tell)
- As AI becomes more powerful, think about what we *should* be doing with it to improve society, not just what we *can* do with it

Thank you!

- Good luck with the finals and projects
- Lab grading should be done by early next week