The Importance of NLP in other Languages

- **Challenges assumptions**
  - Features and statistics thought to be language independent may not actually be so
    - For example, word n-grams can be less robust for languages with very free word order
  - Problems that seem really hard in English might not be so hard in general
    - Strings to sounds are very hard in English

- **Requires the study of other tasks**
  - For languages like Mandarin, Japanese, Thai, and others, words are not as reliably segmented as they are in other languages
    - Word segmentation becomes an important task
    - What we learn might help in other tasks, like multi-word expression segmentation

- **Systems might need to be built very quickly**
  - English to Haitian-Creole translation systems in the aftermath of the 2010 earthquake
Resources

- **OLAC: Open Language Archives Community**
  - Examples: Nahuatl, Powhatan, Slovenian
  - Unified search for many different sources of language data, including
    - WALS
    - ELRA
    - Ethnologue

- **Ethnologue**
  - Provides basic data about almost any language
    - Language Family
    - Number of Speakers
    - Status
    - Dialects

- **World Atlas of Language Structures (WALS)**
  - Currently consists of 192 features, grouped into 151 chapters
  - Great reference to learn about different variations in languages, and browse languages by feature
  - Useful for “borrowing” existing NLP solutions
Three Example Situations

- **Traditional Natural Languages**
  - Example: You need to build a constituency parser for a language that doesn't currently have one, or at least one easily available, such as Bengali or Javanese.

- **Code-Switching**
  - Example: You need to build a spellchecker for Spanglish or Tamil-French with a little English mixed in.

- **Dialectal/Register Variation**
  - You need to build a part of speech tagger for text messages (SMS).
  - You want to build a translator from academic English to everyday English.
Building NLP for a “traditional” language

- May be a language with a large number of speakers and an active web community
  - Amharic
    - About 25 million Speakers
    - 13,372 articles on wikipedia
    - Relatively easy to find newspapers and text in it.

- May be a critically endangered language, with very little digital data available
  - Wampanoag/Massachusett
    - About 4,000 Speakers (<10 as first language)
    - Textbooks and formal grammars exist

- Or somewhere in between
  - Shanghainese (Major Dialect of Wu Chinese)
    - About 80 million speakers
    - 5,695 articles on wikipedia
    - At least one corpus of spoken text
What can we do with Amharic?

- Building a language model shouldn’t be too difficult
- What about a part of speech tagger?
  - UniversalDependencies.org says Amharic is “coming soon”
  - There is a bilingual corpus available, but alignment becomes an issue.
  - Some prior work out there, so the corpora exist
    - But can we get it?
- What about a constituency parser
  - Crazy Idea: We know Amharic is a Semitic language like Arabic and Hebrew.
    - Assuming we have a part of speech tagger, use that for the first iteration of CKY and then Hebrew and Arabic parsers for the rest
    - Probably won’t work very well, but it is a start.
What can we do with Wampanoag?

- Not much digital data
- Could use data collected in theses like *An Introduction to Wampanoag Grammar*
- For a serious research project, the best bet would be to contact [Wôpanâak Language Reclamation Project](https://www.wpanaak.org/) for any digital teaching materials they had
- Alternatively, a language identification system could be trained to recognize Wampanoag
  - Once this was working well enough, just start scraping the web and hope you come across some text
What can we do with Wampanoag?

- Not much digital data
- Could use data collected in theses like *An Introduction to Wampanoag Grammar*
- For a serious research project, the best bet would be to contact [Wôpanâak Language Reclamation Project](https://www.wopanak.org) for any digital teaching materials they had
- Alternatively, a language identification system could be trained to recognize Wampanoag
  - Once this was working well enough, just start scraping the web and hope you come across some text
What can we do with Wu (Shanghai Dialect)

- Uses Chinese characters, so cross our fingers and hope word segmentation for Mandarin work
  - Could be a really bad assumption
  - We could also use a model trained on Mandarin as a base, and then augment with one trained on limited Wu corpus.
- IF the segmentation goes ok, building a language model shouldn't be too difficult, but will not have very wide coverage.
- A large number of people speak the knowledge, so with the right setup, annotating an initial corpus through crowdsourcing should be possible.
Notes about Crowdsourcing

● Using non-experts to label your data over the web has become very popular in NLP over the last few years.

● Some really great info on crowdsourcing is on the course website for the [crowdsourcing class](http://example.com) at UPenn, taught by Chris Callison-Burch (ccb).

● The quality and ethics of crowdsourcing are still up for debate
  ○ My rule of thumb: get some friends and family to do your task and figure out how long it will take them. Pay so that if someone did your task for a whole hour (unlikely) they would be making at least minimum wage.
  ○ There are some really cool methods to check the quality of your labels - as long as you consider how you will evaluate the quality of your labels ahead of time you should be ok

● Crowdsourcing requires IRB approval (at least currently!)
  ○ Can take a few weeks to go through, so probably not appropriate for this class
The register of a language refers to the social use of a dialect, i.e. legal, scientific, extremely informal
- This is mostly solved with more data, but on the extreme end of informalness, alternative or additional processing may be necessary
  - E.g. expanding abbreviations in SMS/Twitter

Dialect is a fuzzy concept that is somewhere between being a completely separate language and a slight variation among speakers of a language
- A lot of the time, dialect can be handled with more data, as it might be reflected in new words or words with different meanings
- It is possible for dialect to change the syntax and semantic structure of a language
  - For example, in African-American Vernacular English (AAVE) is grammatical to say “I been bought it” and it conveys additional information than the phrase “I bought it”
Building NLP for codeswitching

- Code switching is when a conversation takes place in more than one language or registers
  - Spanglish
    - Sometimes I'll start a sentence in Spanish y termino en español\(^1\)
  - Hinglish (with a splash of Marathi)
    - Sometimes main hindi may chalu karke and finish it in English\(^2\)
  - Finnish-Sami
  - Very common, but not on the web
    - In text messages and personal conversations they are more common
  - How do we borrow from the two languages involved.

- Code switching introduces some new tasks not seen in other situations
  - Language identification at the word or phrase level
  - Language disambiguation when a word is possible in both languages

2. A patient labmate
In addition to targeting specific languages, there is research being done to abstract away from specific languages

- We will cover some of this in semantics
- Work in syntax is being done by Emily Bender at UW and others
  - Goal is to great language independent description of grammar that can be used to bootstrap NLP resources for any language in combination with typological information (WALS)