Data Report

Conor Stuart Roe

September 13, 2019

The two main types of data I'll need are complete, annotated sets of inflected word forms in my chosen set of paradigms, and unstructured text corpora or wordlists.

For now, the inflectional paradigms I'm planning on collecting and analyzing are verbal tense/aspect/mood and subject marking in Spanish and Turkish and nominal case and number marking in Russian and Hungarian. All of these satisfy the criteria of transparent orthography and relative abundance of data, and can be scraped from Wiktionary. See examples with Spanish, Turkish, Russian, and Hungarian. Some amount of manual setup will be required to build scrapers for each language, but once that is done individual words should be automatically ingestible. Wiktionary tables for a particular language and paradigm generally all have identical layout, so I'll be able to get very consistent structured data from it. I'm not sure yet whether I'll need to manually identify lemmas for analysis, or in some way automate that process.

I've had trouble finding large, plain text corpora that are available wholesale instead of via a search interface. Corpora that are accessible via search include the Historical/Genre Corpus del Español, a collection of about 14,000 Spanish texts spanning several genres and time periods, and the Russian National Corpus. I can scrape Wikipedia or news sites for text data as well. It remains to be seen how important text corpora are to my process. The word inflection data is more important to begin with, and the purpose of text corpora is mainly to test my model's ability to identify words of relevant morphological class. I could probably still conduct the essential parts of my work without plain text corpus data.

I'd like to turn in some preliminary data scraped from Wiktionary for one of my four identified paradigms next week, as a proof of concept. This should be a pretty minor task in Python. I think that building scrapers for the other paradigms can wait until the implementation stage of my project. I might also want to identify a list of candidate sites to scrape textual data from, though I'm not interested in going too deep with that. Online text in all four languages certainly exists, and I don't yet have a detailed specification for what type of text would be most suited to the project.