Research team: Ghadeer, Ira
Possible project title: Elision of vowels in written representation of an impaired English speech variant

I Hypothesis

Elision of vowels is more likely to happen in unstressed syllables, comparing with stressed ones. Possible universal justification is that there's often a neutral sound in unstressed syllables of an English word and therefore such syllables are often shorter (= pronounced faster) than stressed syllables with a non-neutral vowel. That's why we expect these shorter neutral sounds to experience elision more often and consequently be replaced by apostrophe in written speech.

II Research design

What data we need to test the hypothesis. We are going to use a corpus of ~25000 English fanfiction works of different length, a lot of which feature a fictional character with a slight speech impairment (we don't know exactly how many of the works feature this character's peculiar speech, but as this character is very popular we assume that some works definitely include his speech). We are going to investigate this impairment represented in written speech by fanfiction authors (authors mainly use apostrophe to imitate this impairment).
First we'll need to match and extract from the corpus all tokens which can possibly represent the impairment that we're going to study (i.e. all tokens with unusual apostrophe position). Then we'll need to manually filter the tokens to make sure that we use only those that are part of impaired speech. Second, we'll need to manually reproduce the original spellings of the tokens. Then we'll be working with the original word forms (and their corresponding distorted spellings) which we are planning to annotate on several levels. The final dataset will luckily include all possible syllables of collected impaired tokens (it's important that impairment can happen both inside the word and at the word boundary, so we'll be interested not only in unigrams but also in bigrams). Each syllable will have the following information associated with it:

- original token that the syllable is part of,
- transcription in IPA notation,
- whether this syllable is distorted in the written form of the word (= whether it experiences elision),
- whether the syllable is stressed,
- whether the syllable is open or closed (possibly),
- position of the syllable in the original token (1st, 2nd, 3d etc).

For the syllables that were distorted in written speech we assume that they experienced elision and thus introduce additional information about them that we're going to collect manually and using some python tools described further (hopefully, we'll manage to collect all this information, but it will depend on the amount of interesting tokens we'll be able to detect in our corpus):

- how many letters were replaced by apostrophe in the syllable,

- how many sounds were represented by the omitted letters of original spelling,
- how many sounds were represented by the omitted part of the word,
- if it was one sound, was it a consonant or a vowel,
- if it was a cluster of sounds, how many sounds are in this ommited cluster.

This is just some possible choices of potential variables and it's highly likely that we'll have to choose only some of them. We will decide what information will be necessary for our research when we have a closer look at our data.

Final dataset will definitely include syllables annotated as stressed or unstressed – this is the critical piece of information that we'll need to test the alternative hypothesis. But it will also include other variables (we are just not sure which ones exactly) to make our model more interesting and hopefully explore other effects that are less obvious than that of stress mentioned in our alternative hypothesis.

Formal hypotheses. H0: Elision is equally likely to happen (=be marked by apostrophe) in unstressed and stressed syllables of English words which represent the character's speech impairment. H1: Elision is more likely to happen (=be marked by apostrophe) in unstressed syllables of English words which represent the character's speech impairment than in stressed syllables.

Model. Decision tree and possibly a regression model.

Statistical tools. R package *party* and its *ctree* function.

III Description of data collection method

The fanfiction corpus of ~25000 archiveofourown.org/ works is collected with the help of python libraries *requests* and *bs4* . Possibly we'll have to collect additional information with the help of unofficial API python library (github.com/alexwlchan/ao3), if we decide to introduce one more categorical variable (author, i.e. English speaker that produced the distorted token) in our research. Collected tokens that are in focus of our research will be transcripted in IPA notation with github.com/mphilli/English-to-IPA python library. We'll also need to investigate pyhon libraries that perform syllable division.