LING83800: Methods in Computational Linguistics 2 **Selin Alkan, 8/7/2022**

Term Paper

In this final project, I wrote a grapheme-to-phoneme (g2p) conversion program for Swedish and evaluated it against the Swedish pronunciation dictionary from WikiPron.[1]

Swedish has 9 vowels and 20 consonants in its alphabet; 26 of these letters are the same as those in English while the three others, <å>, <ä>, and <ö>, are additional, commonly used letters in the language. In my opinion, despite being relatively more phonetic than English, Swedish presents a difficult case for such tasks like g2p especially due to its large vowel phoneme inventory and specific segmental rules in addition to the tonal word accent system (i.e., pitch accent) that plays a very important role in pronunciation helping with situations such as differentiating *anden* 'the duck' (pronounced with accent 1 where the pitch on the stressed syllable (i.e., *án*) goes up and then down again in the unstressed syllable (i.e., *den*) from *anden* 'the spirit' (pronounced with accent 2 where the pitch rises in the stressed syllable (i.e., *án*) as in accent 1, and it also rises again on the following, (un)stressed syllable (i.e., *dén*) as well) (Riad 2014). Stress placement is also a major factor in determining the pronunciation of any given word – in particular, the vowel length. However, according to Riad (2014), it is mainly morphology that controls where stress falls within words while there is not a strong phonological generalization for stress placement.

There is a wide range of variation when it comes to dialects in Swedish. To the best of my knowledge, all the rules I implemented are based on the information that is said to be valid for Central Sweden Swedish. In my g2p rules, I did not take the pitch accent into consideration. Another assumption I made is closely related to vowel pronunciation: Vowels in unstressed syllables are always short; however, since vowel length is such an important aspect of Swedish, I wanted to write rules to reflect short and long vowels. Since a vowel in a stressed syllable can be either short or long, the natural assumption I ended up making is that all the syllables are stressed, and it is only the phonetic environment that differentiates short vowels from their long versions and vice versa. I also made major simplifications and some minor adjustments in some rules to make them work with each other some of which I detail further below. Lastly, I left out some rules that were too complicated to handle and tried to keep the rules I used as general as possible without losing the overall "Swedishness".

The first major issue I encountered was the phoneme inventory of Swedish. Different sources used slightly different phonetic symbols especially for some vowel sounds (e.g., /e/ vs. /ɛ/ for <e>), and there is a vowel lowering rule that I decided to use, which added an initial 4 phonemes to the set of 18 phonemes that most sources used. I had to place the vowel lowering rule before the short- and long-vowel rules so that nothing is overridden by those two rules.

Another major issue was with some consonant rules, especially regarding the graphemes <c>, <k>, <ck>, <g>, <gn>, and <j>. Most rules ended up being overridden by some others, and I fixed the failures by reordering the rules and making small adjustments to some phonetic environments making them more specific (e.g., in *Rule #11 – <gn> as /ŋn/*, the preceding environment for the rule to apply is "elsewhere" according to the Swedish Alphabet Wikipedia

[1] The file is found here: https://github.com/kylebgorman/wikipron/blob/master/data/scrape/tsv/swe_latn_broad.tsv

page; however, I restricted it to only vowels since it would otherwise override *Rule #10* giving us an incorrect pronunciation of [ŋnɪst] for the word *gnist* 'telegraph'.)
Lastly, there are lots of exceptions to all these rules, some sounds, such as /ɧ/, are very difficult to handle due to being very common and being represented by many different spellings, and there is also optionality for some letter-pronunciation correspondences. For example, the consonant combination <ch> can be pronounced both as /ɧ/ and /ɕ/. Initially, I thought I could create a union of possible outputs, but I got a "Multiple top writes" error, which led me to pick the most plausible phoneme based on my personal experiences with the language, namely /ɧ/.

Overall, I ended up creating 24 rules, which can be found here. Once I finished my rules, I wrote my tests in g2p_test.py and all the word-pronunciation pairs I picked for the tests were successful.

To my surprise, it was the second half of the project – evaluating the grammar against the Swedish pronunciation dictionary from WikiPron – that I struggled the most with. I worked with the broad transcription data, which needed the following changes:

- Since I ignored accent in my grammar, I needed to remove the ˈ and ˌ used to mark accent in the file. For example:
  - abbe ˌa b ɛ
- There were too many repeated entries that I had to remove.
  - After cleaning the data, I deleted the TSV file that I had cloned from here previously, which had many repeated words and WikiPron links as pronunciations for some words, but now the data in the hyperlinked page seems to be devoid of all those problems that I had to clean by hand before making me unable to provide examples here.
- There were IPA symbols such as /ɾ, ◌͡◌, ◌̯, œ̞/ and some more in the pronunciations from the WikiPron data that I did not include in my grammar. Therefore, I had to leave out words that had such symbols in their pronunciations. For example:
  - adelssläkt ɒ: d ə l s ◌͡◌ s l ɛ k: t

The dictionary that I ended up with has 3328 word-pronunciation pairs, which is 542 pairs fewer than the original WikiPron file. There is one other change I had to make in the transcription file that I am going to mention below.

Since I was unable to figure out how to do the evaluation and then compute the Word-Error Rate (WER), it took me a long while to get back to the project, but I have finally finished the evaluation script, accessible as g2p_eval.py, which gave a composition failure error on the first run. Since I did not want to end this whole project resorting to using `try:/except:` (which is still included in the program) and being unable to compute the WER, I tried the following:

- I created a TSV file out of the word-pronunciation pairs I used in my `g2p_test.py` script, ran the `g2p_eval.py` on it to receive a WER of 0.0, which gave me the idea that there may be something wrong with the WikiPron data that I cleaned up before.
- I realized the white spaces between the segments in the WikiPron data were at least one of the problems, so I removed them using regex.

- Since I still got the same Composition Failure error, I decided to try to evaluate portions of the WikiPron data; I got a WER of 92.00 from the first 50 words in the file, which was the first successful step!
    - I first tried to divide the file into halves, but since I kept getting the same error with both halves and halves of halves and so on, which soon became too difficult to handle, I move on with portions of 20 pairs after the first 50, and found the problem:
        - The 76th word in the list has the letter <é>, which was not included in my SIGMA_STAR, since it is not a part of the alphabet.[2]

andré	andre:

        - I added the letter to my grammar and ran the program without any failures this time.

While the accuracy, which I am aware was not asked to be reported in the project, is 0.22, the WER is 78.46 as seen below:

```
(base) Selins-Air-2:LING83800_Term_Project selinalkan$ ./g2p_eval.py --input sv_desegmented.tsv
Accuracy:       0.22
WER:            78.46
```

I have been studying Swedish for some years now, and not having any native speakers to interact with around me make it hard to check and/or confirm the pronunciations of words. Among the different resources which sometimes provide the chance to listen to pronunciations, Wiktionary has always been my first choice since it also mostly provides the IPA transcriptions. However, there are entries with no audio files to listen to and the IPA transcriptions are not always available and/or accurate, which causes problems for me and other learners to practice the language. Therefore, even though this is a very specific and practical issue, it still constitutes a good enough reason to automate and improve the g2p technology in Swedish as it is the case for most other languages, too. My work on Swedish has also proved how difficult it can be to capture all the pronunciation rules and numerous exceptional pronunciation patterns in languages.

Lastly, the graphemes I used are based on the Swedish Alphabet Wikipedia page. For the phonemes, I consulted three sources: the book titled The Phonology of Swedish by Tomas Riad, the Swedish Phonology Wikipedia page, and the Swedish IPA Wikipedia page. The phonological rules are based on the following: the Swedish Alphabet Wikipedia page, the Swedish Phonology Wikipedia page, The Phonology of Swedish by Tomas Riad, Swedish Phonology on Glottopedia, this blogpost on the Language Learners website.

[2] In the *Usage Notes* for Swedish regarding <é>, this Wiktionary page also indicates the letter in question is not an independent letter in the Swedish alphabet.

References

Help:IPA/Swedish. (2022, May 18). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Help:IPA/Swedish&oldid=1088478735

Nario, J. (2021, August 24). Tricks to master the Swedish accent: Common Swedish pronunciation problems students have & how to solve them. *Language Trainers.* https://www.languagetrainers.ca/blog/swedish-accent-pronunciation/#:~:text=In%20Swedish%2C%20speakers%20use%20single,to%20distinguish%20between%20two%20words.&text=Double%20consonants%20also%20affect%20pronunciation,(vägen%2C%20the%20road).

Riad, T. (2013). *The phonology of Swedish*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199543571.001.0001

Swedish alphabet. (2022, May 5). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Swedish_alphabet&oldid=1086274115

Swedish Phonology. (2018, March 2). In *Glottopedia*. http://www.glottopedia.org/index.php/Swedish_Phonology#:~:text=Swedish%20makes%20use%20of%20nine%20short%20vowels.&text=In%20many%20cases%20and,%5Bɛ%5D%20are%20only%20allophones.

Swedish phonology. (2022, May 20). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Swedish_phonology&oldid=1088877117

é. (2022, August 6). In *Wiktionary*. https://en.wiktionary.org/wiki/é#Usage_notes_2