Assignment 1: Grapheme to phoneme conversion
Due: Sept. 16

The goal of this assignment is to get familiar with some Perl syntax, while creating a program that does something like phonology. Your task is to write a script that converts Italian orthography into phonemic transcription. It should take an input file, read it, and perform whatever substitutions are necessary to produce a first approximation of a (broad) phonemic transcription. When you have completed your script, you can send it to me as an email attachment.

- If you have a particular interest in a language other than Italian, it is possible to take on a different language! (Consult with me first, though, so we can determine whether it looks an appropriate task.)

In addition, you should ponder the following questions, for discussion next week: (1) in some cases, two or more replacement rules must be ordered in a particular way. Is it possible to examine two rules and determine that their order might make a difference? How might you go about diagnosing that an ordering is necessary? Can you predict what sequences would be informative, so a learner could look out for them? (2) How would you write a program that could detect whether an input file was in orthography or transcription?

1 Background on Italian

Italian has (roughly) the following segments, which your transcription should make use of:

\[
P, B, T, D, K, G, m, n, N, \gamma, f, v, s, z, \tilde{s}, dz, \tilde{t}j, \tilde{d}z, \tilde{i}, \tilde{o}, l, r, \lambda, j, w
\]

- We'll ignore [r] and [ɔ] (they can't be predicted from spelling, and not all dialects have them anyway)
- We'll also ignore the phoneme [dz], and lump it together with [tʃ]

When creating a transcription for computational modeling, it is generally convenient to represent segments using symbols that can be typed and transmitted easily. I recommend the following phonemic “alphabet” for Italian (but feel free to modify as you see fit):

\[
P, b, t, d, k, g, m, n, n, N, G, \gamma, f, v, s, z, l, r, L, \lambda, j, w
\]

2 Pronunciation rules

Here is the core of what makes Italian orthography not “purely phonetic” (and what you need to undo to get back to phonetic representation). A summary can be found at:

- http://italian.about.com/library/nosearch/nblfare103a.htm

Grapheme-to-phoneme correspondence in Italian is largely, but not completely predictable. The following graphemes correspond straightforwardly to the equivalent phonemes: p, t, b, d, m, n, l, r. Here are the rules for more complicated cases: (simplified somewhat)
• *gli* is pronounced [gli] before another vowel (*meglio* = [mɛˈʎːo], *famiglia* = [famˈʎja]). It is pronounced [gli] at the end of a word, or before a consonant (*figli* = [fiʎˈʎi])
• *sci* is pronounced as [ʃ] before a vowel (*scolta* = [ʃkɔlt̚a]), [ʃi] before a consonant or word-finally (*pesci* = [peʃʃi], *sciroppo* = [ʃiroˈppo])
• *sc* is pronounced as [ʃ] before [e], [i] (*pesce* = [peʃʃe]) (See previous for rules about [i] in particular). It is pronounced as [sk] otherwise (*pesca* = [peska])
• *ci* and *gi* are pronounced as [tʃi] and [dʒi] before another vowel (*cielo* = [ʧeˈlo], *giusto* = [dʒustɔ], and as [ʃi] and [dʒi], respectively, before a consonant or word-finally (*cibo* = [ʧibo], *undici* = [ʊndɪˈʧi])
• *c* is pronounced [k] before back vowels ([a, o, u]), [tʃi] before front vowels ([e, i]) (e.g., *cera* = [ʧera])
• *ch* is pronounced as [k] before front vowels (it doesn’t occur before back vowels). Hence, *sch* is [sk] (*dischi* = [diski])
• *g* is pronounced [g] before back vowels, [dʒi] before front vowels (e.g., *gelo* = [dʒelo])
• *gh* is pronounced as [g] before front vowels (it doesn’t occur before back vowels)
• *gn* is pronounced as [ɲ]
• *n* is pronounced as [ɲ] before velar stops (*funghi* = [fʊŋɡi], *banca* = [baŋka])
• *ng* is pronounced as [ŋg] before back vowels (*mango* = [maŋgo]), [ndʒi] before front vowels (*mango* = [maŋdʒo]) (This is not a separate rule, but illustrates the interaction of two rules above)
• *qu* is pronounced as [kw]
• *h* is silent word-initially (pronounced as nothing)
• *s* is pronounced as [s] or [z]; for present purposes we’ll just assume the following rough distribution:
  − [z] intervocalically, and before voiced stops (*rosa* = [roza], * frase* = [fraze], *sbarco* = [zbaɾko], *sgarbato* = [zgaɾbartɔ])
  − [s] elsewhere (*sale* = [sale], *pasta* = [pasta], *pensa* = [pensa])
• *z* is pronounced somewhat unpredictable as [ts] or [dz]; we’ll cheat and call them all Z
• Italian also has geminates, which you can represent as a sequences ([pp], [ZZ], etc.)

**Watch out for sequences like cci, which should be CC, not KC (braccio = braʃʃo)**

• Vowels are sometimes written with accents to indicate stress (à, ò, etc.) There are also diphthongs (e.g., *ao* = [aʊ]). In addition, [ul], and [i] are sometimes realized as glides ([w, j]). Since the goal here is to learn how to implement a set of replacement rules, I recommend focusing on the consonants and ignoring these vowel issues.

A proposed alphabet of symbols to use for phonetic transcription can be found in ItalianPhone.txt on the course website. Here are some examples of transcriptions using this system:

<table>
<thead>
<tr>
<th>cara</th>
<th>kara</th>
<th>cents</th>
<th>Cera</th>
<th>cielo</th>
<th>CelO</th>
<th>mango</th>
<th>maʃʃo</th>
<th>banca</th>
<th>[baŋka]</th>
</tr>
</thead>
<tbody>
<tr>
<td>gusto</td>
<td>gusto</td>
<td>giusto</td>
<td>Justo</td>
<td>mango</td>
<td>maʃʃo</td>
<td>hannon</td>
<td>anno</td>
<td>pesce</td>
<td>peʃʃe</td>
</tr>
<tr>
<td>mangio</td>
<td>manJo</td>
<td>anche</td>
<td>aGke</td>
<td>braccio</td>
<td>braʃʃo</td>
<td>peSi</td>
<td>pazzO</td>
<td>pazzO</td>
<td>[pazzO]</td>
</tr>
<tr>
<td>sogno</td>
<td>soNo</td>
<td>questo</td>
<td>kHzesto</td>
<td>hannon</td>
<td>anno</td>
<td>pesche</td>
<td>peʃʃe</td>
<td>pazzO</td>
<td>paZzo</td>
</tr>
<tr>
<td>lascia</td>
<td>laSa</td>
<td>pesce</td>
<td>peʃʃe</td>
<td>pazzO</td>
<td>paZzo</td>
<td>pazzO</td>
<td>zbaLo</td>
<td>[zbaLo]</td>
<td></td>
</tr>
<tr>
<td>sciroppo</td>
<td>Siroppo</td>
<td>schivo</td>
<td>skivo</td>
<td>pazzO</td>
<td>paZzo</td>
<td>pazzO</td>
<td>[zbaLo]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>funghi</td>
<td>fuʃʃi</td>
<td>ghigno</td>
<td>giNo</td>
<td>sbaglio</td>
<td>zaBLo</td>
<td>esercizio</td>
<td>ezerCiZio</td>
<td>[zbaLo]</td>
<td></td>
</tr>
</tbody>
</table>

---

1In point of fact, both [i] and [j] are always long intervocally: [me.ɪ.o], [peʃʃi]. We will ignore this detail here.