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 attach-ment,

- 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	ɲ	ŋ
f, v	s, z	ʃ	
	t͡s, d͡z	t͡ʃ, d͡ʒ	
	l, r	ʎ	j, w

i	u
e	o
ɛ	ɔ
a	

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

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	G
f, v	s, z	ʃ	
	Z	t͡ʃ, d͡ʒ	
	l, r	L	j, w

i	u
e	o
a	

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 [ʎ][1] before another vowel (*meglio* = [meʎo], *famiglia* = [famiʎa]). It is pronounced [ʎi] at the end of a word, or before a consonant (*figli* = [fiʎi])
- *sci* is pronounced as [ʃ] before a vowel (*sciolta* = [ʃolta]), [ʃi] before a consonant or word-finally (*pesci* = [peʃi], *sciroppo* = [ʃiroppo])
- *sc* is pronounced as [ʃ] before [e], [i] (*pesce* = [peʃe]) (See previous for rules about [i] in particular). It is pronounced as [sk] otherwise (*pesca* = [peska])
- *ci* and *gi* are pronounced as [t͡ʃ] and [d͡ʒ] before another vowel (*cielo* = [t͡ʃelo], *giusto* = [d͡ʒusto]), and as [t͡ʃi] and [d͡ʒi], respectively, before a consonant or word-finally (*cibo* = [t͡ʃibo], *undici* = [undit͡ʃi])
- *c* is pronounced [k] before back vowels ([a, o, u]), [t͡ʃ] before front vowels ([e, i]) (e.g., *cera* = [t͡ʃera])
- *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, [d͡ʒ] before front vowels (e.g., *gelo* = [d͡ʒelo])
- *gh* is pronounced as [g] before front vowels (it doesn't occur before back vowels)
- *gn* is pronounced as [ɲ]
- *n* is pronounced as [ŋ] before velar stops (*funghi* = [fuŋgi], *banca* = [baŋka])
- *ng* is pronounced as [ŋg] before back vowels (*mango* = [maŋgo]), [nd͡ʒ] before front vowels (*mangio* = [mand͡ʒo]) (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* = [zbarko], *sgarbato* = [zgarbato])
  - [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* = `braCCo`)
- Vowels are sometimes written with accents to indicate stress (*à*, *è*, etc.) There are also diphthongs (e.g., *ao* = [aʊ]). In addition, [u], 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 `ItalianPhones.txt` on the course website. Here are some examples of transcriptions using this system:

cara	`kara`	*cera*	`Cera`	*cielo*	`Celo`
gusto	`gusto`	*giusto*	`Justo`	*mango*	`maGgo`
mangio	`manJo`	*anche*	*aGke*	*braccio*	`braCCo`
sogno	`soNo`	*questo*	`kwesto`	*hanno*	`anno`
lascia	`laSa`	*pesce*	`peSe`	*pesci*	`peSi`
sciroppo	`Siroppo`	*schivo*	`skivo`	*pazzo*	`paZZo`
funghi	`fuGgi`	*ghigno*	`giNo`	*sbaglio*	`zbaLo`
chiuso	`kiuzo`	*festa*	`festa`	*esercizio*	`ezerCiZio`

[1] In point of fact, both [ʎ] and [ʃ] are always long intervocalically: [meʎːo], [peʃːi]. We will ignore this detail here.