Analyzing (-r) with R

Lívia OUSHIRO (USP)

Quantitative analyses of sociolinguistic variation (Guy, 1993; Bayley, 2002) often involve handling hundreds or thousands of tokens of a variable, especially in studies of phonetic variation. Analyses in softwares such as GoldVarb X and RBrul should be preceded by the identification, isolation, coding, and extraction of variants within a variable context. These tasks are mechanical, time-consuming, tiresome, and subject to errors due to lapses of attention on the part of the researcher.

This paper presents a method employed in the analysis of variable coda (-r) in Paulistano Portuguese, in words such as "porta" ('door') and "mulher" ('woman'), in which the tap, retroflex and zero (deletion) variants alternate. The corpus consists of 89 one-hour-long sociolinguistic interviews (about one million words), which yielded 63,994 tokens of variable coda (-r). The aforementioned tasks for handling such a number of tokens were greatly minimized by the use of the software R (Gries, 2009; Hornik, 2011), which was employed to automatically: (a) clean and prepare the transcripts; (b) identify tokens of variants in the speech of informants; and (c) extract tokens with preceding and subsequent context into a precoded spreadsheet file. The script is largely based on Gries 2009 and the internet discussion list "CorpLing with R" (https://groups.google.com/group/corpling-with-r).

After setting a working directory (setwd), identifying the transcript files (dir), and loading them into a list (scan) with a for-loop, the next step was to clean the files of double spaces, tab stops, and parentheses (since GoldVarb X reads "(" as the beginning of a token). This was performed with the gsub function.

```
for (i in 1:length(files)) {
clean.files<-gsub("\\\t", "", all.corpus[[i]])
}
```

Identifying tokens of coda (-r) was probably the trickiest part of the script. In this analysis, we wanted to identify and mark the instances of the variable with the symbols "<" and ">" right after the word, so that a token word such as "porta" would be marked "porta <>". This was again accomplished with the function gsub as shown below:

```
words.with.R<-gsub("(\\b.*?[aáâeéêiíoóôuú]r[bcçdfgjklmnpqstvwxz/\\.,:;!\\? ].*?\\b)", "\\1
> ", corpus.R[[i]], ignore.case=T)
```

This command line instructs R (i) to look for instances of the grapheme "r" preceded by a vowel [aáâeéêiíoóôuú] and followed by a consonant or end of a word [bcçdfgjklmnpqstvwxz/\\.,:;!\\?], and (ii) to substitute the word by the word itself and "<>" (\\1 <>). However, since we are only interested in tokens of coda (-r) in the speech of informants, the list had to be further cleaned from "<>" in the speech of the researcher (D1):

```
only.S1<-gsub("(^[D].*)<>", "\\1", words.with.R, ignore.case=T)
```

In order to code the instances of coda (-r) as a tap <T>, a retroflex <R>, or zero variant <A> we listened to the recordings and individually marked each token identified by R. After tokens were coded, R was again utilized for extracting each instance into a spreadsheet file, using the functions unlist, strsplit, gsub, and grep. The output provided each token in a different line, its preceding context, subsequent context, the variant employed by the speaker, the interview from which the token was extracted, and precodings for each speaker’s social characteristics (sex/gender, age group, level of education, area of residence, zone of residence, geographic mobility).

Therefore, with as few as seven different functions in R, we were able to greatly minimize the strenuous tasks of identifying, isolating, and extracting thousands of tokens of a phonetic variable. The present script, applied to the study of variable coda (-r) in Paulistano Portuguese, can be easily adapted to studies of other variables, such as the pronunciation of nasal /e/ (e.g. "faze(i)nda") and nominal agreement (e.g. "as casas" vs. "as casa"), currently under study by our research group. The script allows the researcher to handle data in a more consistent manner and, by reducing the time spent in preparing the token file, it allows more time to perform statistical analyses.

References

Bayley, R. (2002) The quantitative paradigm. In: Chambers, J.K., P.Trudgill, N. Schilling-Estes (eds.) The Handbook of Language Variation and Change. Malden, MA: Blackwell, p.117−141.

Gries, S.Th. (2009) Quantitative Corpus Linguistics with R. NewYork: Routledge.
Guy, G.R. (1993) The quantitative analysis of linguistic variation. In: Preston, D. (ed.), American Dialect Research. Amsterdam: Benjamins, p. 223−249.

Hornik, K. (2011) R FAQ. Available at <http://cran.r-project.org/doc/FAQ/R-FAQ.html>.