Building a database of phonetic transcriptions from a speech corpus

Maarten Janssen, IULA; Fabíola Santos, ILTEC
maarten@iltec.pt; fabiola.santos@iltec.pt

This paper presents a process for the semi-automatic transcription of a spoken corpus that makes use of a database of phonetic transcriptions to help along the process of providing transcription for Praat [1] text data.

This process was developed in the course of the Oral-Phon project, developed at ILTEC, which provides the normative phonetic (wide and narrow) transcription of a portion of a spoken corpus of standard European Portuguese [2], as well as lemmatization, syllable division, stress marking and POS tagging for all words in the text.

To successfully complete the time-consuming task of annotating the corpus, several automatization methods were created. This rendered simple time-saving procedures that simplify the task and that may be taken a step further to create other resources making the time spent on the task even more profitable in the future. This process aims only at obtaining the normative transcription for the words in the corpus, given that the narrow transcription is being manually performed.

Processes for automatically annotating corpora with a phonetic transcription are not new, and rely on the prior existence of either of two things: a rule system for automatically generating transcription, or a complete pronunciation lexicon. The rules can be used for automatic transcription, but, in most languages, many rules allow for exceptions. In the particular case of this corpus, a dialogues corpus, with most speakers between the 20- 35 years old strip, the informal style of speech was full of colloquial words and loan words and would require a lot of manual work during the revision.

The second option is to use those rules to create a pronunciation lexicon, which in turn can be used to transcribe the words automatically. This lexicon has to be revised and corrected but with a full pronunciation lexicon, you can perform the task more easily. In our case, we used a combination of the two methods. From a previous project, we had both a set of rules and a small pronunciation lexicon of around 55 thousand words. The words that were already in our database were the most frequent in a written journalistic corpus called CETEMPúblico.

The process consists of the following steps:

1. Starting from a Praat transcription file, or any other compatible transcription file that can be imported into Praat, the desired tier is written into a TextGrid file.
2. The orthographic tier in this Textgrid file is first tagged morphossyntatically using a POS tagger, in our case a Brill tagger [3] trained previously for (written) Portuguese. This tier has to be revised manually, and the corrected output is read back into the Praat file.
3. After this, a first script splits the transcription into words assigning each word an interval, simply by dividing the lines (in our case speaker turns) on the token boundaries created by the tagger, and assigning each token an equal part of the interval for the line. These intervals are added as a new tier to the TextGrid file.
4. The automatically created word boundaries are adjusted manually by the transcriber that performs the narrow transcription, who aligns these boundaries with the segments she

identifies. Since the current number of boundaries is already provided, the boundaries only have to be moved to their correct place in the spectrum.
5. Using a second script, the words in the transcription are then looked up in the previously existing database and imported into a tier. In this lookup, the POS tag is used to look-up the correct transcription in those cases where there are non-homophonous homographs.
6. The transcription of the words that are not found in the database is automatically generated using the rules [4]. Since the generated transcriptions are less reliable than the stored transcriptions, the generated transcriptions are adorned with a prefix to speed up the correction process.
7. In our database, the phonetic transcription contains syllable boundaries. These are used, in a further step, to provide automatically generated boundaries for phonetic syllables analogous to the separation into words.
8. Finally the information on the word lemma also contained in the database is used to generate a lemmatization tier.
9. The end result is a Praat file with an orthographic tier, and word-delimited tiers for orthography, part-of-speech and lemma, broad phonetic transcription, as well as syllable-delimited broad phonetic transcription.
This full Praat file implicitly contains a lot of corrected phonetic transcriptions. These corrected transcriptions are exploited after the transcription to add new transcriptions to the phonetic database, which will increase the rate of known transcriptions higher over time.
The current rate of words that are found in the database is higher then what we previously expected, and is around 77%. This means that in a 50 hours corpus like ours, the time of work necessary to accomplish the task, can be lowered from around 500 hours of work to around 150, which represents a very significant reduction of 70% of cost in time. The percentage of words found in the database will be successively higher which means that the time needed for the correction will decrease along the course of the project.
There are, of course, other ways of creating this pronunciation database, however, if you need to make a phonetic transcription and you don't have such a database, you can make your work more productive by thus progressively creating the database for later use.

References

[1] Boersma, Paul and Weenink, David (2010): *Praat: Doing phonetics by computer* [programa de computador]. http://www.fon.hum.uva.nl/praat/
[2] Freitas, Tiago and Santos, Fabíola (2008): Corp-*Oral: Spontaneous speech corpus for European Portuguese*. *In* Proceedings of LREC VI. ELRA: Marrakesh, Morroco.
[3] Brill, Eric (1992): *A simple rule- based Part-of-Speech Tagger*. HLT'91: Proceedings of the Workshop on Speech and Natural Language, Morristown, NJ, USA: Association for Computational Linguistics, pp 112 – 116.
[4] Janssen, Maarten and Barbosa, Sílvia (2009): *Two-Step Grapheme-to-Phoneme Conversion for Portuguese*. *In*: Proceedings of PaPI 2009