The Unicode cookbook for linguists

Moran, Steven and Michael Cysouw** (*(University of Zurich, **University of Marburg)*

As a so-called standard, the International Phonetic Alphabet is often criticized because it is notoriously used inconsistently. However, like notational systems in other scientific disciplines, the IPA reflects facts and theories that have changed over time as more detailed information (about spoken languages) has been collected and analyzed. Consider the fact that the IPA was originally designed in the late 1800s to meet practical needs, e.g. for teaching literacy (especially in English and French) and for the development of practical orthographies of unwritten languages that were being increasingly described.

Obviously not all phoneticians agree, nor are they ever likely to agree, on all aspects of transcription practices and approaches. This subtly was built into IPA from the beginning. In fact, any IPA transcription is based on two premises: (i) that it is possible to describe the acoustic speech signal (sound waves) in terms of sequentially ordered discrete segments, and, (ii) that each segment can be characterized by an articulatory target. Both premises are theoretically problematic. Furthermore, over one hundred years after its inception and subsequent updates, the integration of the IPA into a digital encoding, into the Unicode Standard, brought with it a plethora of technological obstacles which most linguists still encounter in their daily working lives. Most users do not understand how to navigate these often intransparent pitfalls.

In this talk we present a brief history of the IPA, a description of the integration of the IPA into the Unicode Standard, and the common and not-so-common pitfalls that their marriage has created for linguists. We will present simple software packages that we have developed, in both the Python and R programming languages, that allow users to:

- characterize the distribution of graphemes and extended graphemes in text input
- to segment Unicode IPA
- to check for transcription errors in text input
- and to create and use "orthography profiles"

Orthography profiles are simple tab-separated tables where columns define the graphemes and orthographic rules in alphabetic-based writing systems. They allow users to easily segment their data and to transform it between different writing systems and encodings. As such they are a practical application for language scientists and they are being used in large-scale lexical databases to normalize across disparate transcription systems and orthographies.