Corpus-linguistic methods for improving the lexicographic treatment of pronunciation variants

Peter Uhrig, Universität Leipzig (at the time of the conference) / FAU Erlangen-Nürnberg (now)

The major pronunciation dictionaries on the market - the *Cambridge English Pronouncing Dictionary* (EDP18) and the *Longman Pronunciation Dictionary* (LPD3) - attempt to offer all relevant pronunciation variants for Standard British and American English, as illustrated in the following excerpt from the entry *advertisement* in EPD18:

ədˈvɜː.tɪs.mənt, -tɪz-, -təs-, -təz-, US ˌæd.vəˈtaɪz.mənt, ədˈvɝː.ţəs-, -ţəz-

The British pronunciation comes first and is printed in red, then, after the label US, the American variants are printed in blue. If there is no difference between the varieties, the variants are printed in black without a label, as for *aggregate*:

ˈæg.rɪ|.geɪt, -rə-

LPD3 employs a similar approach, but marks non-RP variants with §, separates British and American usage with **‖** and prints the most common pronunciation in a blue bold face font:

əd ˈvɜːt ɪs mənt -ɪz-, -əs-, -əz-; §ˈæd və taɪz mənt, §ˌ••ˈ•• **‖ ˌæd v ə r ˈtaɪz mənt** əd ˈvɝː ţ əs-, -əz- *(*)*

If the main pronunciation variant for American English is the same as for British English, it is not repeated after the ‖, which is also not printed in a blue/bold faced font, as in the example of *Muslim*:

ˈmʊz lɪm ˈmʌz-, ˈmʊs-, -ləm ‖ ˈmuːz-, ˈmuːs-, ˈmʌs

For a small number of entries, LPD3 offers the results of preference polls, often visualized in pie charts. Thus for *Muslim*, the dictionary contains the following information, visualized in the charts in Figure 1 below.

— Preference poll, British English: ˈmʊ- 70%, ˈmʌ- 30%, -z- 89%, -s- 11%, -lɪm 91%, -ləm 9%.

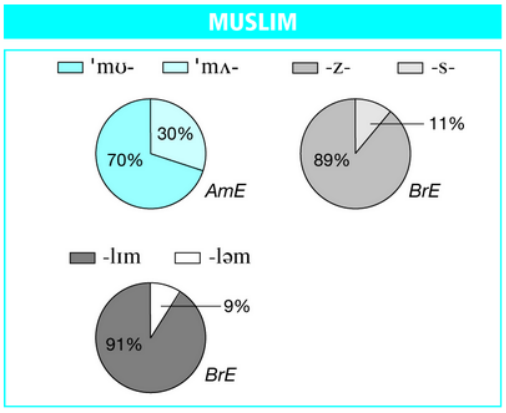

Figure 1: Pie charts visualizing the pronunciation preference poll for *Muslim* in British English; LPD3 (CD-ROM edition)[1]

[1] Note that the charts in the CD-ROM version often contain errors. Thus even though it says AmE for the first chart in Figure 1, the written version is more trustworthy here.

For users of the dictionaries, it would of course be interesting to obtain similar statistics for all entries with pronunciation variation, not just the ones selected for the preference polls in LPD3.[2] While it should be possible to determine the most frequent pronunciation from the order in which they are presented in the dictionaries, EPD18 and LPD3 do not seem to agree as to what the most important American pronunciation of, for instance, *Muslim* is (see the excerpt from EPD18 below), indicating the need for further research.

ˈmʊz.lɪm, ˈmʊs-, -ləm, US ˈmʌz.ləm, ˈmʌs-, ˈmʊz-, ˈmʊs-, ˈmuːz-, ˈmuːz-, -lɪm

In this presentation, a range of corpus-linguistic methods for the distinction of pronunciation variants in a corpus of American TV (see Uhrig 2018 for details) news will be presented. This will include an optimized manual workflow for the fast classification of audio snippets as well as the extraction of pronunciation variants identified by speech recognition or forced alignment software. Furthermore, statistical methods such as the automatic clustering of formant frequency vectors and simple threshold-based metrics will be evaluated.

Works Cited:

Cambridge English Pronouncing Dictionary. 18th edition. 2011. Edited by Peter Roach, Jane Setter and John Esling. Cambridge etc.: Cambridge University Press. [=EPD18]

Longman Pronunciation Dictionary. 3rd edition. 2008. Edited by John C. Wells. Harlow: Longman. [=LPD3]

Uhrig, Peter. 2018. "NewsScape and the Distributed Little Red Hen Lab – A digital infrastructure for the large-scale analysis of TV broadcasts." In: Anne-Julia Zwierlein, Jochen Petzold, Katharina Böhm and Martin Decker (eds.), *Anglistentag 2017 in Regensburg: Proceedings. Proceedings of the Conference of the German Association of University Teachers of English.* Trier: Wissenschaftlicher Verlag Trier.

[2] Also note the methodological problem that preference polls may not at all accurately reflect the frequency distribution.