- **Exclusion of words with less than two Unicode characters or less than two phone segments** [?] add an explanation

- **Separation by script** [?]: It is very straightforward why this is done. There is no obvious connection between the different scripts of a language and its pronunciation. It makes sense to treat different scripts as different languages.

- **Exclude foreign words with foreign pronunciations** [?]: Foreign words in a language with their original pronunciation can add phonemes that are not in that language's phoneme inventory. If they were to be included it would make sense to include a pronunciation adapted to the actual language.

- **Words with multiple pronunciations in word lists**: **?** excluded those words, however, it might also be possible to add **pos!** (**pos!**) tags or other linguistic information to distinguish these words.

- **Consistent broad transcriptions** [?]: With broad transcriptions it is important to be consistent and not use allophones. **?** did this specifically for Bulgarian.

- **Linguistic variation and processes** [?]: Some transcriptions include examples for monophthongization or deletion which are ongoing linguistic processes but should not be part of a dataset representing a standard variation. **?** dealt with monophthongization by choosing the longer to two transcriptions as this logically exclude the monophthonged version. This does of course only work if there are more than one pronunciations available.

- **Tie bars**: **?** notice that some languages (English and Bulgarian) have inconsistent use of tie bars. This can be correct by replacing all inconsistencies by the tie-bar-version.

- **Errors in the transcriptions**: **?** noticed many errors in the WikiPron English data. They identified errors by looking at the least frequent phones and then check the word-pronunciation pairs where those phones occurred in. As the number of phones in a language is often known this can be used to check the phones in the datasets and identify uncommon ones.

References