Appendix C

PHOIBLE SEGMENT CONVENTIONS

In this appendix I describe the general conventions that were used to encode segments in inventories that were added to the PHOIBLE data set. I begin by explaining the segment and diacritic ordering that was used. I then address general consonant- and vowel-specific decisions, including which symbols were used to indicate sounds not officially in the International Phonetic Alphabet (IPA; International Phonetic Association 2005).[1] Lastly, I briefly discuss marginal sounds and how they are marked in PHOIBLE.

C.1 Diacritic ordering

Each segment type that is composed of more than one character is first normalized into a canonical decomposition form that adheres to the Unicode Normalization Form D (NFD; The Unicode Consortium 2007).[2] The diacritic ordering conventions I describe below deal with Unicode characters that are not in the "Combining Diacritical Marks" block. The logical ordering of Combining Diacritical Marks is handled by normalization into NFD. Characters sequences that are not handled by NFD must be explicitly ordered, including characters from the "Spacing Modifier Letters" block, which may appear as diacritics to the user. The ordering is influenced by the linguistic literature and to my knowledge the IPA does not explicitly state in which order diacritics should appear in segments.

If a segment type contains more than one rightward diacritic, I use this order:

- unreleased/lateral release/nasal release → palatalized → labialized → velarized → pharyngealized → aspirated/ejective → long

[1] See also Appendices E and F for SPA and UPSID_{451} specific notes. Appendix D provides a list of the Unicode IPA characters used in segments in inventories in PHOIBLE.

[2] See discussion in Section 2.1.4.

For example, a labialized aspirated long alveolar plosive: < t^{wh}: >. If a segment type contains more than one diacritic below the base segment:

- the place feature is applied first (dental, laminal, apical, fronted, backed, lowered, raised), then the laryngeal setting (voiced, voiceless, creaky voice, breathy voice), and finally the syllabic or non-syllabic marker (for vowels, ATR gets put on between place and laryngeal setting)

For example, a creaky voiced syllabic dental nasal: < $\underset{\tilde{r}}{n}$ >.

C.2 Consonants

There are some common encoding errors that occur when linguists use the (Latin-based) keyboard to input certain IPA symbols that Unicode has assigned to different code points. These include:

- the IPA symbol <g> LATIN SMALL LETTER SCRIPT G (U+0261) is not the same code point as keyboarded <g> LATIN SMALL LETTER G (U+0067)

- the IPA symbol <!> LATIN LETTER RETROFLEX CLICK[3] (U+01C3) is not the same code point as keyboarded <!> EXCLAMATION MARK (U+0021)

- the IPA symbol <|> LATIN LETTER DENTAL CLICK (U+01C0) is not the same code point as keyboarded <|> VERTICAL LINE (U+007C)

- the IPA symbol <’> MODIFIER LETTER APOSTROPHE (U+02BC) is not the same code point as keyboarded <’> APOSTROPHE (U+0027)

- the IPA symbol <ː> MODIFIER LETTER TRIANGULAR COLON (U+02D0) is not the same code point as keyboarded <:> COLON (U+003A)

Other segment conventions relevant to consonants are given below by subsection.

---

[3]In the IPA, the <!> is an alveolar or postalveolar click, not a retroflex click as stated in the Unicode Standard.

C.2.1 Aspiration

For aspiration, the conventions include:

- Aspirated: pʰ
- Preaspirated: ʰt
- Breathy release: tɦ

C.2.2 Double articulations

I do not currently use a "tie bar", i.e. COMBINING DOUBLE INVERTED BREVE (U+0361) or COMBINING DOUBLE BREVE BELOW (U+035C), to signal double articulations (e.g. affricates, clicks and diphthongs). So for example, <k͡p> and <t͟s> appear as <kp> and <ts> in inventories in PHOIBLE.

Affricates are marked for homorganic place of articulation. For example, in SPA the "t/s-hacek-prenasalized" is indicated by the symbol <n̪tʃ> and the "voiceless retroflex sibilant affricate" in UPSID_{451} is signaled by <ʈʂ>.

C.2.3 Fricatives

I use a lowered diacritic, the <o̞> COMBINING DOWN TACK BELOW (U+031E), with a fricative to make an approximant, e.g. SPA's "beta-approximant" looks like <β̞>. The raised diacritic is also used with the pharyngeal fricative to indicate a voiced pharyngeal plosive <ʕ̝>.

All "affricated" trills and clicks are marked with the non-IPA diacritic <o̝> COMBINING X BELOW (U+0353), which I use to indicate "frictionalized". For example "r-flap-fricative" in SPA and "voiced alveolar fricative flap" in UPSID_{451} are both indicated as <ɾ̝>.

UPSID_{451} forces the distinction between sibilant and non-sibilant fricatives, so another non-IPA diacritic was selected. To mark "non-sibilant" fricatives, I use the < o̼ > COMBINING EQUALS SIGN BELOW (U+0347), e.g. "r-fricative" is <z̼>.

C.2.4 Glottalization

Glottalization conventions include:

- Preglottalized: ˀd
- Glottalized / postglottalized: dˀ
- Creaky voiced / laryngealized: d̰

C.2.5 Nasalization

For prenasalized consonants, i.e. homorganic nasals, I use <NC> where <N> is a nasal that agrees in place of articulation with the following consonant, e.g. <mb>, <nd>, <ŋg>, etc. The character <ⁿ> SUPERSCRIPT LATIN SMALL LETTER N (U+8319) is used to indicate nasal release, e.g. the "d-nasal-release" in UPSID_{451} is given as <dⁿ>.

C.2.6 Clicks

Clicks are ordered with the voice setting first:

- <k> indicates voiceless
- <g> indicates voiced
- <ŋ> indicates nasal

Following the voice setting, the place/manner of the click is indicated, e.g. a voiceless alveolar click is encoded as <k!>. Laryngeal modifiers are placed on the voice setting and diacritics for place are placed on the symbol for the click. For example, a "voiceless nasal palatoalveolar click": <ŋ̥!̠>.

C.2.7 Labialized

Labialized segments are represented with the <ʷ> MODIFIER LETTER SMALL W (U+02B7), e.g. the "labialized voiceless labio-velar plosive" in UPSID_{451} is <kpʷ>. For velarized segments I use the <ˠ> MODIFIER LETTER SMALL GAMMA (U+02E0), e.g. SPA's "d-velarized" is <dˠ>. Labiovelarized segments use the combination of both space modifying characters in this order: <ʷˠ>.

C.3 Vowels

When a low back unrounded vowel appears in a phonological description, I use the character <ɑ> LATIN SMALL LETTER ALPHA (U+0251), even if the author used the keyboard <a> in his or her phoneme inventory chart (which seems to be the case more often than not).

For diphthongs I use <i> or <u> and not <j> or <w> to indicate the glide component of the diphthong. In cases in which this leads to a sequence of two identical vowels, I use the non-syllabic diacritic marker <◌̯> COMBINING INVERTED BREVE BELOW (U+032F), e.g. SPA's "i/yod" is marked with <i̯i>. Long vowels are marked with the length diacritic <ː>, e.g. SPA's "iota-creaky voice-long" is <i̯ː>.

C.4 Marginal phonemes

Marginal phonemes are those that behave notably different phonologically than the majority of segments found in a particular language. Language contact factors contribute to marginal phonemes. For example, loanwords containing non-native sounds can introduce maringal phonemes into the borrowing language. There are varying degrees of marginalism; see discussion in Jelaska and Machata (2005). For PHOIBLE it would be ideal to create a ranking or vocabulary for varying degrees of marginal status.[4] To do so, I have collected any remarks about the marginality of segments as described in the resources from which I extracted inventories. However, since different authors use different descriptions of marginality, these have to be fit into some type of ranking. I propose adding this information in a future release of PHOIBLE. Currently I simply mark any type of phoneme

[4]Perhaps along the line of "anomalous" segments in UPSID (Maddieson, 1984, 170).

described as marginal or loan by an author of a language description by enclosing those segments in less-than and greater-than symbols < >.