A BUSZI XML formátuma

Az interjúk szerkezeti elemei és a lejegyzési útmutatóban meghatározott egyes kódoknak megfeleltetett XML elemek:

- Teljes interjú: `<div type="interview">`
- Egyes modulok: `<div type="modul" id="MODULID.MODNUM">`
- Megnyilatkozás: `<u who="beszelo.ID" id="interview.ID">`
  folytatólagos megnyilatkozás (pl. `a>`) esetén további attribútum:
  `prev="interview.ID-2"`
- Szünet: `<pause>`
- Szünet szótesten belül: `<seg type="pause">`
- Megjegyzés (`[xxx]`): `<event desc="xxx">`
- Nem értett kifejezés (`Z Z Z`):
  `<gap reason="incomp" extent="number_of_Zs"/>`
- Hezitáció
  - `ö; öööö`: `<vocal desc="o_hesitation" iterated="y|n"/>`
  - `nnn; mmm`: `<vocal desc="[n|m]_hesitation" iterated="y"/>`
    Ha utána '=' jel áll, a `desc` attríbútum értéke
    `desc="[n|m]_hesitation_slip"`.

  Szótesten belül is szerepelhet.
- Hiányzó elem:
  - bizonytalan (`<0xxx?>`):
    `<add cert="low" resp="enc.id">`xxx`</add>`
  - biztos (`<0xxx>`):
    `<add cert="high" resp="enc.id">`xxx`</add>`
- Nem sztenderd névelő (`<=0az>`):
  `<orig resp="id" reg="az">`a`</orig>`
- egyidejű beszéd: `<anchor id="ID"/>` illetve `<anchor synch="ID">`
  ID jelek típusai lehetnek:

- *Szünet nélkül megszakított megnyilatkozás*:
  `<anchor type="split_utterance" id="ID"/>`
- *Szünettel megszakított megnyilatkozás*:
  `<anchor type="split_paused_utterance" id="ID"/>`

- szóalakok: `<w>`. Attribútumai:
  - `lemma,msd,ctag` mint MNSZ-ben
  - `skel`: *regularizált szótő* CV váza, magánhangzók BNF[1] alakban.
  - `phon`: *elhangzott* szóalak fonetikai reprezentációja
- központozás: `<c>`
- megszakított alak: `<seg type="slip">`
- Egyéb kódok. Általánosan új elem bevezetésével:
  `<annot type="tipus1 tipus2 ..." resp="enc.id" reg="regularized_form">`
  original_form
  `</annot>`

  A jelenség tokenen belüli helyét az egyes típushoz fűzött további specifikáció jelöli, ahol ez releváns. Lehetséges értékei:
  - `final`: szóvégi jelenség (pl. `type="l_drop_final"`)
  - `iv`: intervokális jelenség (pl. `type="comp_length_iv"`)
  - `ic`: interkonszonentális jelenség (pl. `type="comp_length_ic"`)
  - `prewov`: magánhanzó előtti
  - `precons`: mássalhangzó előtti
  - `short`: nem teljes kiesés, csak rövidülés (pl. `type="l_drop_final_short`, `<annot type="l_drop_prevow_short" reg="kellene">kelene</annot>`)

  Az egyes kódok típusai:

  1. `< >` kódok eredetileg az aktuális szóalakon:
     - magánhangzóharmónia-sértés: `vh_violation`
     - hiperkorrekt IK: `hyper_ik`
     - hiperkorrekt BAN: `hyper_ban`
     - BAN helyett BA: `ba_ban`
     - -NÁK: `substandard_e1`

[1] *back, neutral, front.*

- pótlónyúlás: `comp_length`
- hangkiesések:
  * T kiesés: `t_drop`
  * D kiesés: `d_drop`
  * L kiesés: `l_drop`
  * LY kiesés: `ly_drop`
    Azonos hang kettős kiesése esetén kettőzött a prefixum is, pl. `dd_drop`.
    Kettős kiesések nem azonos hanggal:
  * LT kiesés: `l_drop_t_drop`
  * LD kiesés: `l_drop_d_drop`
- betűejtés (<*x*-*x*>): `spelling_pron_x-x`
- nem állítmányi 'e': `nonpred_clitic`
- hosszan ejtett *s*: `long_s`

2. < > kódok eredetileg a szóalak után
   - *-suk/-sük*: `suk`
   - hiperkorrekt kijelentő móddal: `hyper_suk`
   - *-szuk/-szük*: `szuk`
   - hiperkorrekt kijelentő móddal: `hyper_szuk`
   - egyéb, nem meghatározott kód: `unspec`

   Mindent típusnak létezik `_uncertain` végződésű változata, amennyiben a kérdéses kód "()" között szerepel (`(<xxx>)`), illetve a magyarázó kód kérdőjelre végződik (`<=xxx?>`). A nyelvi bizonytalanság (`<x?>`) jelölése: `unres`.

3. egyéb jelenségek
   - hezitációs hangzónyújtás (`aaalma`): `hesit_length_x`, ahol `x`=nyújtott hang Prószéky kódban
   - félbehagyott token (`xxx=`): `slip`