# Turinys

**1 Apie** **1**

**2 Statistika** **1**

**3 Struktūra** **2**

**4 Failų pavadinimai** **2**
4.1 Informaciniai laukai . . . . . 3
4.2 Vardinimo pavyzdžiai . . . . . 4

**5 Duomenų aprašas** **5**
5.1 Įrašų žymės . . . . . 6
5.2 Triukšmų žymėjimas . . . . . 7

# 1 Apie

Čia yra Liepa-2 projekto metu surinktas ir anotuotas ∼1000 val. trukmės įvairios lietuviškos šnekos garsynas. Licencija: Creative Commons CC BY 4.0.

# 2 Statistika

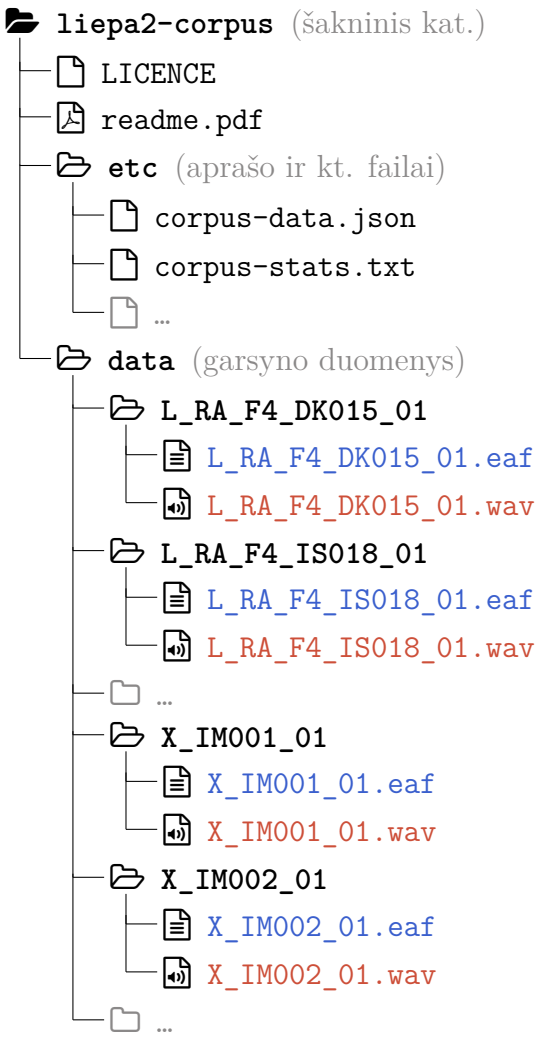
Viso garsyne yra:

| | |
|---|---|
| EAF (Reg., X-failų) | **3656** (3535, 121) |
| Kalbėtojų | **2621** |
| Unikalių žodžių (formų) | **420 710** |
| Takelių (šnekos/triukšmų) | **5882** (4086/1796) |
| Segmentų (šnekos/triukšmų) | **1 874 648** (1381180/493468) |
| Anot. trukmė (šnekos/triukšmų) | **1000:46:21** (939:58:11/60:48:10) |

Duomenų pasiskirstymas pagal kalbėtojų lytį ir amžiaus grupes:

| Amž. grupė | Kalb. | Žodžių | Takelių | | Segmentų | | Anotuota trukmė | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Šn. | Tr. | Šn. | Tr. | **Viso** | Šnekos | Triukšmų |
| F1 (0-12) | 100 | 39822 | 147 | 111 | 47471 | 20439 | **27:11:36** | 24:56:54 | 02:14:42 |
| F2 (13-17) | 39 | 29081 | 57 | 30 | 25187 | 8447 | **15:16:33** | 14:35:46 | 00:40:47 |
| F3 (18-25) | 335 | 145672 | 582 | 222 | 226696 | 66607 | **165:09:16** | 157:06:12 | 08:03:03 |
| F4 (26-60) | 853 | 229156 | 1336 | 642 | 439657 | 182790 | **343:05:35** | 321:05:43 | 21:59:51 |
| F5 (60+) | 145 | 78819 | 246 | 173 | 81972 | 51056 | **63:12:52** | 57:00:12 | 06:12:40 |
| M1 (0-12) | 101 | 36914 | 153 | 104 | 42103 | 17195 | **24:15:16** | 22:23:36 | 01:51:39 |
| M2 (13-17) | 36 | 29506 | 53 | 38 | 25770 | 10011 | **15:48:31** | 14:42:53 | 01:05:38 |
| M3 (18-25) | 144 | 86967 | 260 | 77 | 107708 | 26372 | **69:15:32** | 65:57:13 | 03:18:19 |
| M4 (26-60) | 751 | 187081 | 1089 | 344 | 331006 | 93725 | **240:29:24** | 227:18:00 | 13:11:23 |
| M5 (60+) | 117 | 53174 | 163 | 52 | 53610 | 16715 | **37:01:20** | 34:51:38 | 02:09:41 |

# 3 Struktūra




`./data` kataloge yra garsyno duomenys. Katalogų medis yra „plokščios“ struktūros, t.y. garsyno failai nėra suskirstyti į atskirus katalogus pagal įvairias įrašų/šnekos kategorijas ar atributus. Katalogais atskirti tik susiję failai: atskiras katalogas – kiekvienai įrašo ir anotacijos failų porai (susieti, to pačio pavadinimo failai) ir kt. susijusiams failams. Garsyno įrašų failai – **WAV** formato (*16 kHz, 16 bitų, mono*), o jų anotacijos failai – **EAF** (*XML*), kuriuos galima atverti ELAN programoje.

# 4 Failų pavadinimai

Įvairi informacija apie įrašus ir kalbėtojus yra užkoduota pačių garsyno failų/katalogų pavadinimuose (informaciniuose laukuose), kurie – apibrėžto formato, dviejų tipų:

1. kai anotuota vieno kalbėtojo šneka, tai įrašų/anotacijų failų pavadinimai yra tokio, 7 informacinių laukų formato:

   {N}_{KT}{ĮT}_{L}{AG}_{ID}_{SN}[.wav|.eaf]

2. kai anotuota kelių skirtingų kalbėtojų šneka, tai įrašų/anotacijų failų pavadinimai yra tokio formato:

   X_{ID}_{SN}[.wav|.eaf]

   o pačių anotacijos takelių pavadinimai – pirmojo tipo (be failų plėtinio).

## 4.1 Informaciniai laukai

1. {**N**} – Įrašo formato/glaudinimo (ne)nuostolingumas (L | R):
   - **L** (Lossy) – nuostolingas (pvz.: mp3 ar kt. nuostolingi audio kodekai)
   - **R** (Raw) – nenuostolingas (pvz.: studijiniai ir kt. analoginiai įrašai)
2. {**KT**} – Kalbos tipas (R | S):
   - **R** (Read) – labiau/daugiausiai skaitytinė kalba
   - **S** (Spontaneous) – labiau/daugiausiai spontaninė kalba
3. {**ĮT**} – Įrašo/šaltinio tipas (A | D | P | R | S | T ):
   - **A** (Audiobook) – audio knygos
   - **D** (Dictaphone) – diktofoniniai įrašai
   - **P** (Phone) – telefoninės šnekos įrašai/intarpai
   - **R** (Radio) – radijo laidų įrašai
   - **S** (Studio) – studijiniai įrašai
   - **T** (TV) – televizijos laidų įrašai
4. {**L**} – Kalbėtojo lytis (F | M):
   - **F** (Female) – moteris
   - **M** (Male) – vyras.
5. {**AG**} – Kalbėtojo amžiaus grupė (1 | 2 | 3 | 4 | 5):
   - **1**: 0-12 m.
   - **2**: 13-17 m.
   - **3**: 18-25 m.
   - **4**: 26-60 m.
   - **5**: 60+ m.
6. {**ID**} – Kalbėtojo/įrašo ID (formatas: *XX000*), kur:

   **XX** – dviejų raidinių simbolių anotuotojo kodas

   **000** – trijų skaitmenų kalbėtojo ar įrašo ID kodas/numeris
7. {**SN**} – Įrašo/sesijos eil. nr. (formatas: *00*): dviejų skaitmenų eil. nr, kai yra kelios to pačio įrašo dalys ar to pačio kalbėtojo įrašų sesijos.

**Pastabos**:

- pirmojo failų tipo atveju (kai anotuota vieno kalbėtojo šneka), {**ID**} laukas unikaliai koduoja kalbėtoją vieno/paskirojo anotuotojo failų kontekste. Daroma prielaida, kad tie patys kalbėtojai retai (atsitiktinai) pasikartoja skirtingų anotuotojų failuose. Tokie pasikartojimai garsyne neišskirti, t.y. tas pats kalbėtojas kartais gali būti pavadintas skirtingais ID, nes jo įrašus anotavo skirtingi anotuotojai, neatsižvelgdami į kitų anotuotojų duomenis.
- antrojo failų tipo atveju (kai anotuota kelių skirtingų kalbėtojų šneka), {**ID**} laukas nekoduoja kalbėtojo, o yra įrašo identifikatorius (skaitinė dalis – eilės ar kt. numeris)
- skirstymas į skirtingas kategorijas yra objektyviai ir subjektyviai sąlyginis, tad gali būti netikslumų (pvz. kai nežinomas įrašo tipas, šaltinis, kalbėtojo amžius, nevienareikšmiškas kalbos tipas ir kt.)

## 4.2 Vardinimo pavyzdžiai

- `R_RS_F3_VP101_01[.wav|.eaf]`:
  - *pirmasis failo pav. tipas (anotuota vieno kalbėtojo šneka)*;
  - **R** (Raw) – nenuostolingo formato/kodeko įrašas;
  - **RS** (Read, Studio): skaitytinės kalbos tipas, studijinis įrašas;
  - **F3** (Female, 3): moteris, 18-25 m. (trečia amžiaus grupė);
  - **VP101**: unikalus šio kalbėtojo ID, anotuotojas – VP;
  - **01**: pirma įrašo dalis/sesija.
- `L_RA_M4_AS027_02[.wav|.eaf]`:
  - *pirmasis failo pav. tipas (anotuota vieno kalbėtojo šneka)*;
  - **L** (Lossy) – nuostolingo formato/kodeko įrašas;
  - **RS** (Read, Audiobook): skaitytinės kalbos tipas, audio knyga;
  - **M4** (Male, 4): vyras, 26-60 m. (ketvirta amžiaus grupė);
  - **AS027**: unikalus šio kalbėtojo ID, anotuotojas – AS;
  - **02**: antra įrašo dalis (garsyne turėtų būti ir kitos AS027 kalbėtojo įrašo dalys).
- `X_IS002_01[.wav|.eaf]`:
  - *antrasis failo pav. tipas (anotuota kelių skirtingų kalbėtojų šneka)*;
  - **IS002**: unikalus failo identifikatorius, anotuotojas – IS;
  - **01**: pirma įrašo dalis.

  Šiame faile anotuota, tarkime, trijų skirtingų kalbėtojų šneka, tad anotacijos faile (.eaf) yra tokie anotaciniai takeliai (angl. *tiers*):

- `R_SR_M3_IS103_01`: {**R**aw}, {**S**pont., **R**adio}, {**M**ale, **3**}, {**IS103**}, {**01**}
  Nenuostolingo kodeko/formato įrašas, spontaninės kalbos tipas, radijo laidos įrašas, kalbėtojas – vyras, trečios amžiaus grupės (18-25 m.), kalbėtojo ID: IS103, pirma įrašo dalis.
- `R_SR_F4_IS323_01`: {**R**aw}, {**S**pont., **R**adio}, {**F**emale, **4**}, {**IS323**}, {**01**}
  Nenuostolingo kodeko/formato įrašas, spontaninės kalbos tipas, radijo laidos įrašas, kalbėtojas – moteris, ketvirtos amžiaus grupės (26-60 m.), kalbėtojo ID: IS323, pirma įrašo dalis.
- `L_SP_M5_IS104_01`: {**L**ossy}, {**S**pont., **P**hone}, {**M**ale, **5**}, {**IS104**}, {**01**}
  Nuostolingo kodeko/formato įrašas, spontaninės kalbos tipas, telefoninės šnekos įrašas/intarpas, kalbėtojas – vyras, penktos amžiaus grupės (60+ m.), kalbėtojo ID: IS104, pirma įrašo dalis.

# 5 Duomenų aprašas

`./etc/corpus-data.json` – visų garsyno duomenų aprašas (JSON formatas). Šį aprašą paranku naudoti atsirenkant reikiamus duomenis, formuojant tikslinius garsynus. Aprašo struktūra:

**"[Pavadinimas].eaf"** (object): Anotacijos failo (EAF) pavadinimas – aprašo žodyno[1] raktas.

- **"name"** (string): EAF failo pavadinimas
- **"path"** (string): EAF failo kelias (reliatyvus šakninio katalogo atžvilgiu)
- **"sha1"** (string): EAF failo SHA1 maišos kodas
- **"media"** (object): Informacija apie įrašo failą (WAV)
  - **"name"** (string): WAV failo pavadinimas
  - **"path"** (string): WAV failo kelias (reliatyvus šakninio katalogo atžvilgiu)
  - **"sha1"** (string): WAV failo SHA1 maišos kodas
  - **"len"** (integer): WAV failo trukmė (ms – milisekundės)
- **["tags"]** (array): Susietos žymės (įrašo tipai/požymiai). Nebūtinasis laukas (tik jei yra žymių)
- **["tiers"]** (array): Takelių sąrašas. Nebūtinasis laukas (yra tik antrojo failų tipo atveju)
- **"speech"** (array/object): Anotuoti šnekos segmentai; pirmojo failų tipo atveju šis objektas yra šnekos segmentų masyvas, o antrojo – žodynas[1], kurio raktai yra anotacijos takelių pavadinimai, o reikšmės – šnekos segmentų masyvai. Šnekos segmentų masyvo elementai[1]:
  - **"sid"** (string): Anotacijos/EAF segmento ID
  - **"beg"** (integer): Anotacijos segmento pradžia (ms)
  - **"end"** (integer): Anotacijos segmento pabaiga (ms)
  - **"len"** (integer): Anotacijos segmento trukmė (ms)
  - **"val"** (string): Anotacijos segmento reikšmė (tekstas)

[1] angl.: Dictionary, Mapping (key: value) duomenų tipas

**["noise"]** (array): Anotuoti triukšmų segmentai. Nebūtinasis laukas (yra jei anotuoti triukšmai). Triukšmų masyvo elementai[1]:

**"sid"** (string): Anotacijos/EAF segmento ID

**"beg"** (integer): Triukšmo segmento pradžia (ms)

**"end"** (integer): Triukšmo segmento pabaiga (ms)

**"len"** (integer): Triukšmo segmento trukmė (ms)

**"val"** (string): Triukšmo segmento reikšmė

Žemiau pateiktas šio aprašo fragmentas, kuriame vaizdžiau matyti aprašyta duomenų struktūra:


```
{
  "L_RA_F4_DK015_01.eaf": {
    "name": "L_RA_F4_DK015_01.eaf",
    "path": "data/L_RA_F4_DK015_01/L_RA_F4_DK015_01.eaf",
    "sha1": "602A239B923DBE8F69724AE23F27E10A5E7AE1ED",
    "media": {
      "name": "L_RA_F4_DK015_01.wav",
      "path": "data/L_RA_F4_DK015_01/L_RA_F4_DK015_01.wav",
      "sha1": "50AB8E9BBED516E2C2B8F4E23C7FC378C3A5C125",
      "len": 1607000
    },
    "speech": [
      {"sid": "a1", "beg": 404, "end": 1058, "len": 654,
       "val": "pirmas"},
      {"sid": "a2", "beg": 1863, "end": 4820, "len": 2957,
       "val": "kaip ekstravertas tampa visų mėgstamu vyruku "},
      <...>
    ],
    "noise": [
      {"sid": "a6", "beg": 15368, "end": 15613, "len":245, "val":"+BREATH+"},
      {"sid": "a12", "beg": 20825, "end": 21084, "len":259, "val":"+BREATH+"},
      <...>
    ]
  },
  <...>
}
```


## 5.1 Įrašų žymės

Dalis garsyno įrašų yra specialios paskirties ir pažymėti duomenų apraše šiomis žymėmis:

- **abc**: lietuviškos abėcėlės raidės
- **123**: įvairūs skaičiai ir skaitmenys
- **addr**: įvairūs adresai
- **TK**: projekto paslaugos reikmėms.

## 5.2 Triukšmų žymėjimas

Anotuojant triukšmus buvo apsiribota šiomis triukšmo garsų klasėmis:

- **+BREATH+**: (iš/į)kvėpimai, atodūsiai
- **+COUGH+**: kosėjimas, krenkštimas, čiaudėjimas
- **+LAUGH+**: juokas, kikenimas
- **+SMACK+**: čepsėjimas ir kt. lūpiniai, liežuvio garsai
- **+AH+**: beprasmiai [A:] tipo garsai
- **+EH+**: beprasmiai [E:], [Ė:] tipo garsai
- **+MM+**: mykimas
- **+GARBAGE+**: visi kiti lingvistiniai, kalbos trakto triukšmai
- **+NOISE+**: ne lingvistiniai triukšmai