

CLARIN

Common Language Resources and Technology Infrastructure Network



ieskats runas korpusu vēsturē

Runas korpusa izveide

Ilze Auziņa

ilzea@latnet.lv

LU MII Mākslīgā intelekta laboratorija

Semināra plāns

1. Ieskats runas korpusu izveides vēsturē; to daudzveidība
2. Runas korpusa izveide
3. Runas datu atšifrēšana un marķēšana

KORPUSA JĒDZIENS

Korpuss – reprezentatīvs rakstīta teksta un / vai transkribētas runas kopums elektroniskā formā, ko izmanto datorizētā valodas analīzē un aprakstīšanā

Tekstu korpuss – apjomā liels daudzveidīgs tekstu kopums, kas parasti uzkrāts elektroniski un saistīts ar programmatūru, kura atvieglo tā lingvistisko analīzi

Runas korpuss – runas ierakstu kopums, kas pieejams mašīnlasāmā formā, ir marķēts un dokumentēts

Runas korpusu nepieciešamība un izmantošanas iespējas

- Korpusi ļauj pētīt reālo valodu un atklāt līdz šim nepamanītas lietas, pamanīt tipisko
- Korpusu var izmantot lingvistiskiem pētījumiem
- Leksikogrāfijā – bez korpusa un korpusa rīkiem nevar mūsdienās uzrakstīt labu vārdnīcu (statistikas dati, vārdu savienojumu analīze)
- Terminoloģijas izstrādē
- Valodas mācīšanā
- Tulkošanas studijās un tulku/tulkotāju apmācībā
- Valodas tehnoloģiju izstrādē
- Psiholingvistikā, sociolingvistikā
- Humanitārajās zinātnēs vispār

Runas korpusu tipi

- **Vispārīgie runas korpusi**
 - Runas korpusi
 - Zviedru runātās valodas korpuss u.c.
 - Runas korpusi – vispārīgo/nacionālo korpusu daļa
 - Britu nacionālais korpuss
 - Krievu nacionālais korpuss
 - Čehu nacionālais korpuss
 - u.c.
- **Speciālie runas korpusi**
 - Bergenas Londonas pusaudžu runātās valodas korpuss
 - Igaņu valodas dialektu korpuss
 - 22 valodu telefonu sarunu korpuss
 - bērnu valodas korpusi
 - u. c

Runas korpusu vēsture (1)

- Pirmie runātas valodas korpusi

Londonas Lundas runātās angļu valodas korpuss

(The London-Lund Corpus of Spoken English)

- **Pamatā - *The Survey of English Usage corpus (SEU)***
- Izstrāde sākta 20.gs. 50.gados, datorizēts 20.gs.70. gadu otrajā pusē
- ~500 tūkst. vārdlietojumu

Runas korpusu vēsture (2)

Piemērs. *The Survey of English Usage corpus*

		S.1.12.8
	* D	+ə f/ Ī ə / Ī f/ # +.
	* B	you / applied to go didn't *you#m#*
	D	*Ī I* / found ;N òut# - - - p that / ðL: Ī# .
?non-nuc		!Royal N?Àir Force# . m/ did Ī re!search on
m:nar		plàstics#m# in [?ðne ?place]Ī# and òne
?xNy		place ' only# . and / that was ma!terials
		de'partment 'R A ;È#p# - - - f/non
		me,tàllic#f# . in / Ībràckets# . and I / wrote
		'up to the !chief gáffer# - Ī/ and I °says
?non-nuc		I !want to ?Ncòmē#Ī# - aa m to do m#aa#
m:lax		re/séarch on 'plastics# - / ānd# . / this is
q:lau	*	the plàce# . q/ and he said W éh# q' / HW ís
q':fal		it#q'#q# . * - * ^ and m/he 'didn't knòw#m#
m:scan,	*	
wide		

Runas korpusu vēsture (3)

Piemērs. Londonas Lundas runātās angļu valodas korpus

D

B

70 you ||APPLIED to go DIDN'T ☆ you ■ ☆

D

71 ☆ I ☆ ||found OUT ■ - - - 72 that ||[ði:] · ΔRoyal AIR Force ■ · 73 ||did reΔsearch on
PLASTICS ■ 74 in ||{ONE place} and ONE place ▷only ■ · 75 and ||that was maΔterials
de'partment 'RAE ■ - - - 76 ||NON-METALLIC ■ · 77 in ||BRACKETS ■ · 78 and I ||wrote
'up to the Δchief GAFFER ■ - 79 ||and I ▷says I Δwant to COME ■ - 80 to do
RE||SEARCH on 'plastics ■ - 81 ||AND ■ · 82 ||this is the PLACE ■ ·
83 ||and he said EH ■ 84 ||Is it ■ · 85 ☆ - ☆ and ||he 'didn't KNOW ■ - -

Runas korpusu vēsture (4)

Spoken Corpus of the Survey of English Dialects

- Tips: speciāls korpus
- Apjoms: 60 ieraksta stundas, 800 tūkst. vārdu
- Izstrādes laiks: datu uzkrāšana no 1948. – 1961.g.
- Nav brīvi pieejams, tīmeklī iespējams noklausīties tikai dažus fragmentus (tīmekļa vietne:
<http://sounds.bl.uk/Browse.aspx?collection=Survey-of-English-dialects>)

Pasaules pieredze (1)

- **COBUILD korpuss** (*The COBUILD (Collins Birmingham University International Language Database) corpus*)
 - Tīmekļa vietne:
<http://www.collins.co.uk/Corpus/CorpusSearch.aspx>
 - Tips:
 - vispārīgs korpuss
 - runas (25%) un tekstu (75%) korpuss
 - pirmais lielais, mašīnlasāmais korpuss
 - Apjoms: Internetā izdevniecība *Harpers Collins* piedāvā izmantot 56 milj. vārdlietojumu lieli paraugu
 - COBUILD korpuss ir daļa no Angļu valodas bankas (*the Bank of English*)
 - izstrāde sāka 20.gs. 70. gados, runas dati pievienoti 1994. gadā; **COBUILD runas korpuss** (*The COBUILD Spoken Corpus*) ietverti 10 milj. vārdlietojumu

Pasaules pieredze (2)

- **BASE, *British Academic Spoken English corpus***
 - Tīmekļa vietne:
<http://www2.warwick.ac.uk/fac/soc/al/research/collect/base>
 - Tips:
 - sinhronisks vienvalodas korpuss
 - speciālais (“akadēmiskās” runas un arī videoierakstu) korpuss
 - Apjoms: 160 lekciju un 40 semināru ieraksti, 1644942 vārdlietojumi
 - Izstrādes laiks : 2000. – 2005. gads

Pasaules pieredze (3)

- *The Longman Spoken American corpus*
 - Tīmekļa vietne:
<http://www.pearsonlongman.com/dictionaries/corpus/spoken-american.html>
 - Tips:
 - sinhronisks vienvalodas korpuss
 - speciālais korpuss (ikdienas sarunvaloda)
 - Apjoms: 5 milj. vārlietojumu, vairāk nekā 1000 runātāju no 30 dažādiem ASV štatiem
 - Izstrādātājs : izdevniecība *Longman*, Kalifornijas universitāte

Pasaules pieredze (4)

- **Čehu nacionālais korpuss**
 - Tīmekļa vietne: <http://uncnk.ff.cuni.cz/english>
 - Tips:
 - sinhroniskais korpuss
 - Sabalansēts mūsdienu tekstu korpuss
 - Runas korpuss
 - Dialektu korpuss
 - Diahroniskais korpuss
 - Papildu
 - Mašīnlasāmas vārdnīcas un citas datu bāzes
 - Plašs neapstrādātu, elektronisku tekstu masīvs
 - Apjoms: teksti – 100 milj. vārdlietojumu, runa – 700 tūkst. vārdlietojumu

Pasaules pieredze (5)

- **Britu nacionālais korpuss**

- Tīmekļa vietne: <http://www.natcorp.ox.ac.uk>
- Tips: sinhronisks vispārīgs vienvalodas runas un tekstu korpuss
- Apjoms: 100 milj. vārdlietojumu – teksti 90%, runa – 10%. Korpuss ir pabeigts, netiek papildināts.
- Marķējums: metadati; morfológija; sintakse; fonētiskā transkripcija

Pasaules pieredze (6)

- **Krievu nacionālais korpuss**

- Tīmekļa vietne: <http://www.ruscorpora.ru>

- Tips:

- sinhroniskais un diahroniskais korpuss

- Runātā un rakstītā valoda

- mūsdienu literārās valodas korpusam (“galvenajam korpusam”) ir 3 apakškorpusi:

- Agrīno tekstu korpuss (18.gs. vidus – 20.gs. vidus)

- Mūsdienu tekstu korpuss (20.gs vidus – 21. gs. sāk.)

- **Mūsdienu runātās krievu valodas korpuss** (20.gs. 50-tie gadi – 21.gs. sāk.)

- Apjoms: paredzēts sasniegt 200 milj. vārdlietojumu

- Pētniecības nolūkiem brīvi pieejams tīmeklī

Pasaules pieredze (7)

- **Bergenā Londonas pusaudžu valodas korpuss – COLT** (Britu nacionālā korpusa daļa)
 - Tīmekļa vietne: <http://torvald.aksis.uib.no/colt>
 - Tips: speciāls korpuss (13-17 gadus vecu pusaudžu runātās valodas korpuss)
 - Izstrādes laiks: 1993. – 1994. gads
 - Apjoms: ~ 500 tūkst. vārdlietojumu, 50 stundu ieraksta
 - Nav brīvi pieejams
 - Marķējums: morfoloģija, fonētiskā transkripcija

Pasaules pieredze (8)

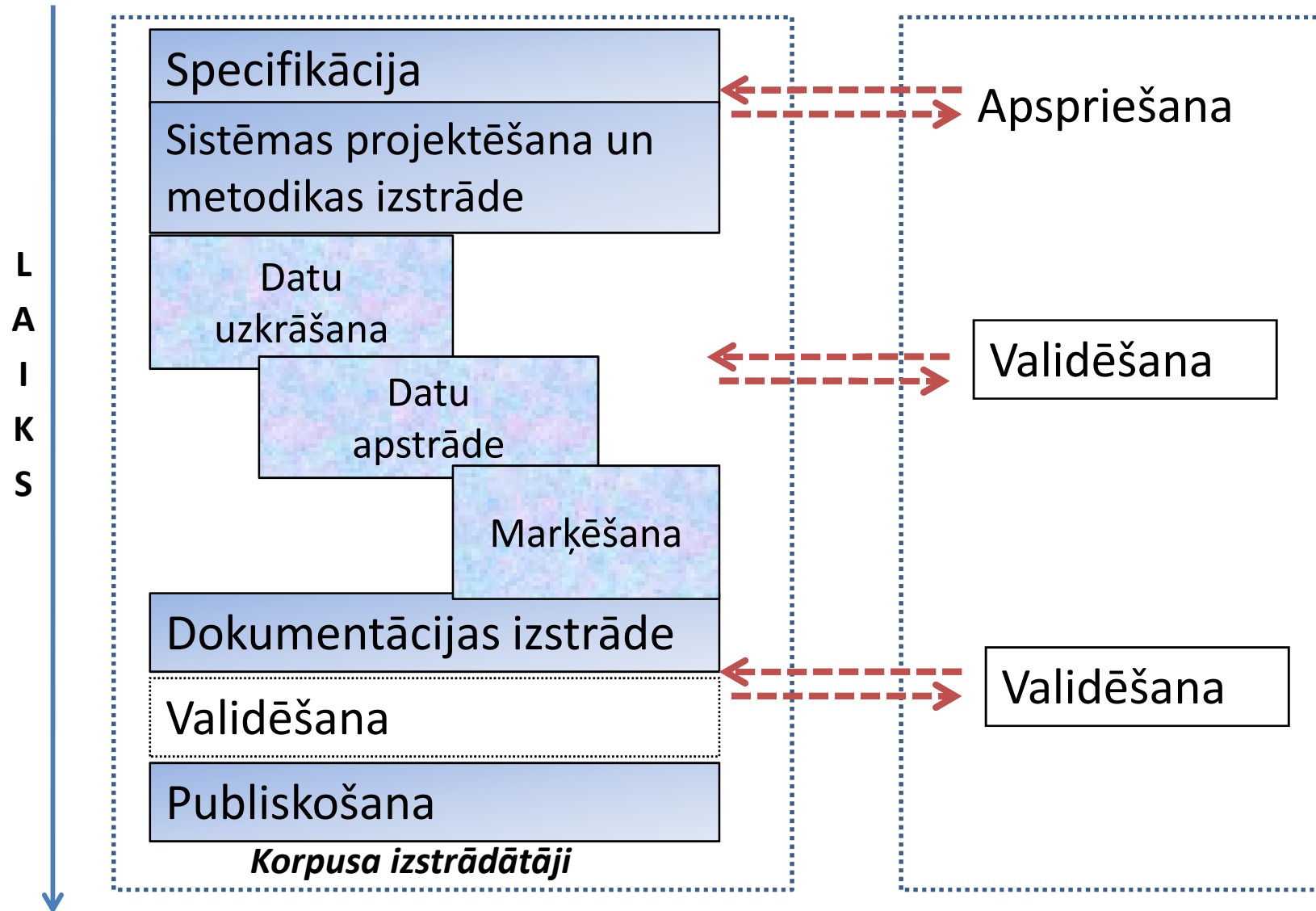
- ***The CHILDES corpus***

- Tīmekļa vietne: <http://childes.psy.cmu.edu>

- Tips:

- speciāls korpuss
 - runas un audiovizuālais korpuss
 - daudzvalodu: dati angļu, vācu, krievu, igauņu u.c. val.

Runas korpusa izveides grafiks



Runas korpusa izveide

I. Korpusa specifikācija

II. Runas datu uzkrāšana

III. Transkribēšana

IV. Marķēšana

Korpusa specifikācija (1)

- Kāds būs korpusa veids: vienvalodas, divvalodu, daudzvalodu?**
- Kāds būs lietojuma mērķis:** tulkošanas pētījumi, studentu valodas prasmju izvērtēšana, gramatikas rakstīšana, sinhroni vai diahroni valodas pētījumi, vārdnīcu veidošana, kāda noteikta valodas stila izpēte u.c.

Korpusa specifikācija (2)

□ Runātāju izvēle

- dzimumu proporcijas: parasti 50:50
- dalījums vecuma grupās, piem.,
 - līdz 16 gadu vecumam, virs 50 gadu vecuma
 - vienlīdzīgs dažādu vecumgrupu runātāju skaits: 12-22, 23-30, 31-40, 41-55
- dzimtā valoda
- runātāja pārstāvētais dialekts, izloksne
- izglītība/prasmes/nodarbošanās

Korpusa specifikācija (3)

□ Runas veids

- pazīme, kas ir svarīga, definējot runas korpusa izmantošanas iespējas
- **Iespējamais dalījums (1):**
 - **Monologi**
 - lasīšanai iepriekš sagatavoti teksti, piemēram, ziņas, referāti, grāmatas lasīšana u. tml.
 - sagatavota runa – lekcijas, sprediķi, parlamentāriešu runas u.tml.
 - spontāni monologi – sporta sacensību komentēšana, stāstījums (piem., cilvēku atmiņu stāstījums)
 - **Dialogi**
 - telefona sarunas
 - privātas sarunas, cilvēkiem atrodoties vienā telpā
 - publiskās diskusijas
 - intervijas

Korpusa specifikācija (4)

– iespējamais dalījums (2)

- lasīta runa (*read speech*)
- jautājumi un atbildes (*answering speech*)
- komandas (*command/control speech*)
- aprakstoša runa (*descriptive speech*)
- nesagatavota (neuzrakstīta) runa (*non-prompted speech*)
- spontāna runa
- neitrāla/emocionāla runa

Korpusa specifikācija (5)

□ Metadati

- Metadati – informācija par ieskaņoto runu, ziņas par ierakstu, t. i., dati par datiem
- Runas korpusa metadatu grupas:
 - ieraksta protokols
 - runātāja raksturojums
 - dažādi komentāri

Korpusa specifikācija (6)

❖ Ieraksta protokols

- Ieraksta identifikators
- Runātāja identifikators
- Ieraksta datums
- Apkārtnes raksturojums
- Tehniskie ieraksta apstākļi
 - mikrofons
 - ieraksta iekārta
 - ierakstītā signāla tehniskā specifikācija
- Citi svarīgi parametri

Korpusa specifikācija (7)

❖ Runātāja raksturojums:

- Obligātie parametri
 - runātāja identifikators
 - dzimums
 - dzimšanas datums
- Citi svarīgi parametri
 - runātāja dzimtā valoda
 - svešvalodu zināšanas
 - vecāku dzimtā valoda
 - patoloģijas
 - dialekts
 - u.c.

❖ Dažādi komentāri

Runas datu uzkrāšana

□ Runas datu ierakstīšana

- *atklāti vai slepeni* ierakstīta runa
- akustiskā vide:
 - studija, telefonsarunas, dzīvojamā istaba, pilsēta u.tml.
- kontrolēti fona trokšņi
- mikrofonu veids, skaits, novietojums u. tml.
- runātāja darbība ieraksta laikā, piem.,
 - ieraksta laikā nemaina pozīciju,
 - runātājs vada auto,
 - runātājs rāda uz objektu, par kuru stāsta u. tml.
- video

Transkripcija (1)

- Transkripcija – ortogrāfisks runas datu pieraksts.
- Runas transkripcijas 3 galvenie principi:
 - Jāievēro konkrētas transkripcijas principi (ortogrāfiskā, fonētiskā, prosodiskā)
 - Transkripcijai jābūt saprotamai visiem interesentiem (pētniekiem); jāizmanto noteikti standarti
 - Transkripcijai jābūt mašīnlasāmai

Transkripcija (2)

Latviešu valodas runas transkripcijā izmantotie apzīmējumi

- i] — ieelpa (an. *inspiration*)
- [r] — elpošana (an. *respiration*)
- [e] — izelpa (an. *expiration*)
- [khm] — „iztīra kaklu”, nokremšļojas
- [...] — pauze (ja pauze ir garāka par 0,1 sekundi, aiz daudzpunktes iekavās var dot pauzes ilgumu, piem., [...(0,3)])
- [sm] — smieklis
- (()) — neskaidrs teksts
- [apl] — aplausi
- [tr] — troksnis
- = — pagarināta zilbe, kāda skaņa zilbē tiek izrunāta garāka nekā nepieciešams
- { } — vārds vai vārda daļa tiek atkārtots vai pateikts pilnībā, pārteikšanās
- [muzika]
- [reklama]

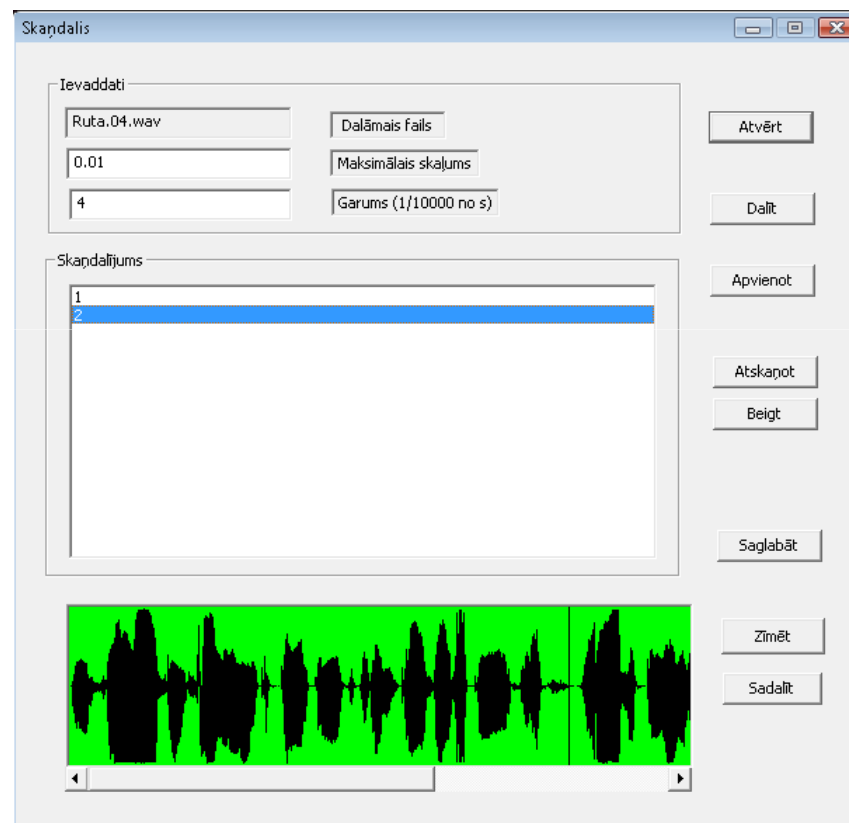
Situācija Latvijā

- **Tiek uzkrāti un digitalizēti plaši audio (un video) materiāli:**
 - Daugavpils Universitāte Mutvārdu vēstures centrs: projekts “Mutvārdu vēsture 20. gs. Latvijas vēstures pētīšanā. Mūsdienu Latvijas vēstures avotpētnieciskās bāzes pilnveidošana”, digitalizēti ~200 audioieraksti
 - Liepājas Universitātes Folkloras un valodas centrs: tiek atšifrēti izlokšņu materiāli
 - LU Filozofijas un socioloģijas institūts: projekts “Nacionālā mutvārdu vēsture”, krājumā ~3000 audio ierakstu (www.dzivesstasts.lv)
 - u.c.
- **Ir sāкта runas korpusa veidošana**

Programmriki skaņu failu apstrādei un marķēšanai (1)

- **Skandalis**

- Izstrādāts LU MII Mākslīgā intelekta laboratorijā
- Audiofaili tiek segmentēti īsākos fragmentos, tiek norādīta runātāju maiņa

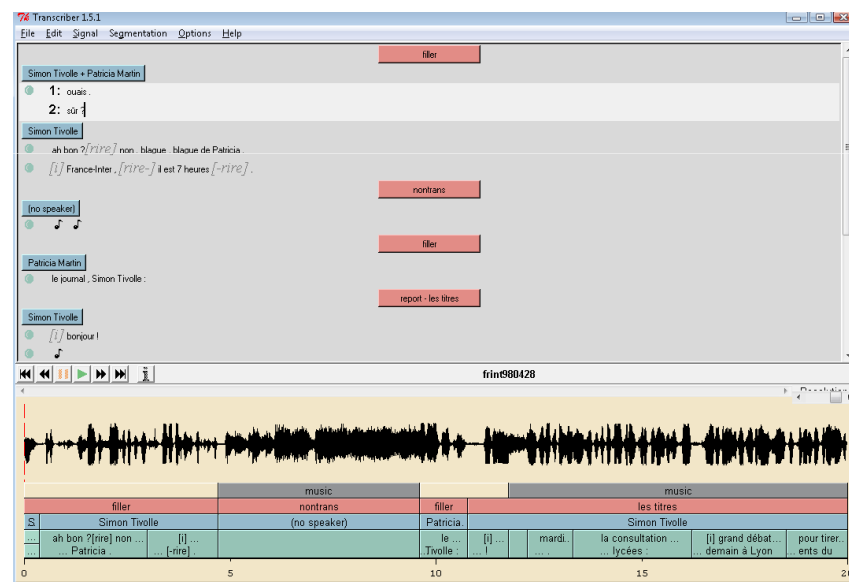


Programmāriki skaņu failu apstrādei un marķēšanai (2)

- **Transcriber**

<http://trans.sourceforge.net/en/presentation.php>

- brīvi pieejams rīks manuālai runas signāla transkribēšanai un marķēšanai
- iespējams apstrādāt un transkribēt arī garus skaņu failus, norādot runātāju un sarunas temata maiņu, akustiskos apstākļus
- pirmā versija - 1998. gadā



Programmrīki skaņu failu apstrādei un marķēšanai (3)

- **WaveSurfer**

<http://www.speech.kth.se/wavesurfer/download.html>

- programma runas analīzei
- brīvi pieejams rīks manuālai runas signāla transkribēšanai un marķēšanai

