Speech Acquisition
next steps in a Meertens tradition

Maarten van der Peet
Menzo Windhouwer

Structured Data
Digital Infrastructure
KNAW- Humanities Cluster

Surveys

CENTRAAL BUREAU VOOR NEDERLANDSCHE E

ONDER LEIDING VAN DE

DIALECTENCOMMISSIE DER KONINKLIJKE AKADEMIE

VRAGENLIJST No 1. (1931).

Dialect van Roderwolde, provincie Drente

gem. Roden

GESPROKEN DOOR:

Naam:

Leeftijd:

Beroep: landbouwer

Geboorteplaats: Gem. Roden

Geboorteplaats van den vader: Gem. Roden

Geboorteplaats van de moeder: Gem. Roden

OPGETEEKEND DOOR:

Naam:

Beroep:

Geboorteplaats:

Woonplaats: Roderwolde

238

N.B. Van alle woorden wordt ook de meervouds- en de verkleiningsvorm (-en) gevraagd. Als vormen niet gebruikelijk is, gelieve men dit te vermelden. Als meer dan een verklei gebruikelijk is, met verschil in beteekenis, gelieve men dit eveneens te vermelden. Ook woorden voor de lichaamsdeelen worden gevraagd. Bij woorden, die in het algemeen-besc klinken, teekene men aan, of deze in het dialect als plat bedoeld zijn.
Overigens zie men de „Aanwijzingen voor de medewerkers".

NAMEN VAN LICHAAMSDEELEN VAN DEN MENSCH

1. hoofd.

Ook platte woorden!
Ik kan er m'n hoofd (het hoofd) niet meer bijhouden!

kòp - {kòppen / kòpp

de kòp lòpt mij o

Meertens Instituut Meertens Panel

90 JAAR

KNAW Humanities Cluster /DI

LimeSurvey

Taalkunde

De volgende vragen zijn een voorbeeld van het huidige taalkundige onderzoek aan het Meertens Instituut. Door deze vragen proberen we in kaart te brengen hoe de taal van Nederlanders en Vlamingen er uitziet: hoe zinnen en woorden worden opgebouwd, hoe de taal klinkt, welke woorden mensen gebruiken. We doen dat door u bijvoorbeeld te vragen woorden of zinnen te vertalen in uw eigen variant van het Nederlands of om aan te geven of u een voorkeur heeft voor een bepaalde variant. Hiermee worden wij heel wat wijzer over de vele varianten die het Nederlands rijk is.

*Geeft u alstublieft per woord aan, hoe deze "nd" klinkt in het dialect van uw woon-plaats, of in uw eigen variant van het Nederlands.

hond

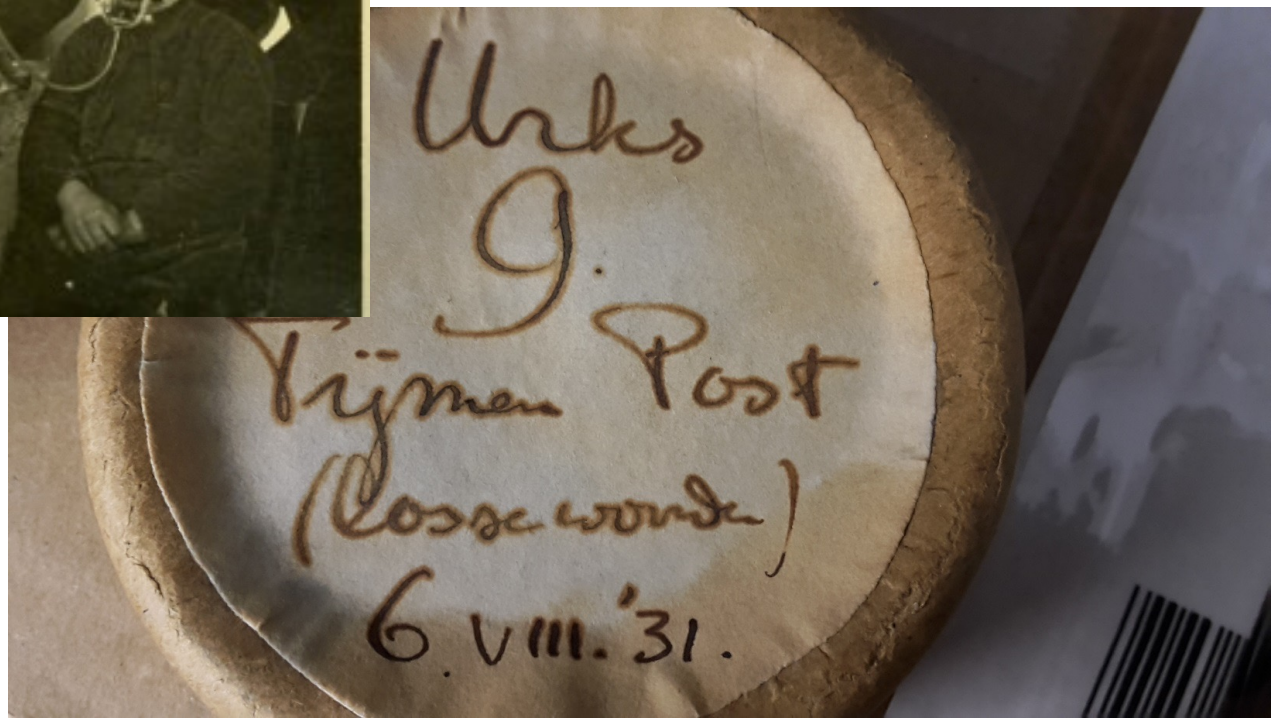
- ○ nt (als in Standaard-Nederlands 'aantikken')
- ○ nd (als in 'aanduwen')
- ○ nj (als in 'oranje')
- ○ ng (als in 'bang')
- ○ nk (als in 'bank')
- ○ ntj (als in 'wint Jan')
- ○ ntj (als in 'plantje')
- ○ n (als in 'plan')

hondje

- ○ nt (als in Standaard-Nederlands 'aantikken')

Speech Acquisition campaigns

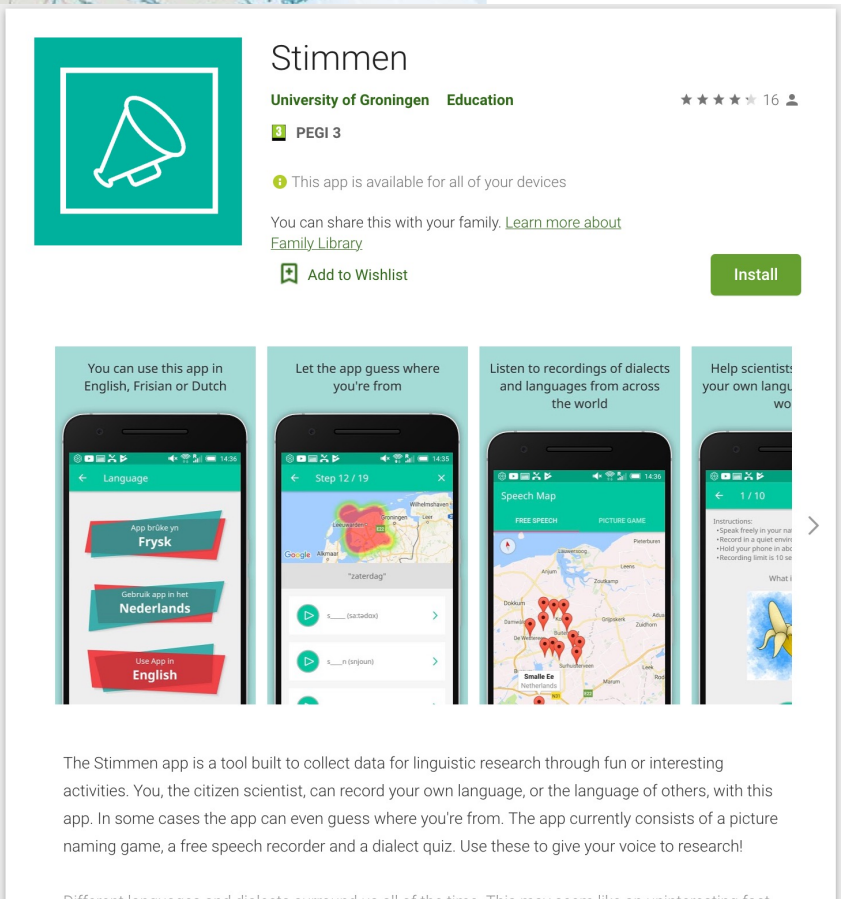
- 1930 P.J. Meertens on Urk



KNAW Humanities Cluster /DI

1960's: Jo Daan in America

- Nicoline van der Sijs & Vertrokken Nederlands



CLARIAH+ WP3 SPAQ:
a combination of these two traditions

- Adding audio recording to LimeSurvey

1. Research the support for the Web Audio API in the major browsers
2. Develop & test a demo app
3. *Extend LimeSurvey for SPAQ type questions*

whoami

- mvdpeet
- musicologist => developer
- KNAW Meertens => HuC DI Structured Data
- mostly LAMP stack, PHP, MariaDB & JavaScript/ JQuery
- lots of structured databases & surrounding code for Meertens Institute
- last couple of years Surveys for linguistics & ethnology
- 70%: LimeSurvey 30%: custom build

Web Audio

HTML5 2011 **Web Audio** support

"The Web Audio API is a high-level JavaScript API for processing and synthesizing audio in web applications. "

A few years ago:

- unreliable (FireFox) or not supported at all (Safari, Internet Explorer)
- documentation out-of date, changing api
- unreliable polyfills

Browsers

Chrome 

Safari 

FireFox 

Edge 

- Stats from different sources. https://en.wikipedia.org/wiki/Usage_share_of_web_browsers
- Analyze the access-data of the latest Meertens Panel Survey held December 2020.

These are the results, roughly (Statcounter)

Platforms World / Netherlands:
Desktop (50%), Mobile (44%), Tablet (6%)

Platforms Meertens Panel (dec 2020)
Desktop (68%), Mobile (28%), Tablet (4%)

Desktops:

Chrome (70%), Safari (9%), FireFox(9%), Edge (7%), Rest (5%)

Mobile:

Chrome (65)%, Safari (22%), Samsung (7%), Rest (5%)

Tablet:

Safari (47%), Chrome (40%) AOSP(11%), Rest (2%)

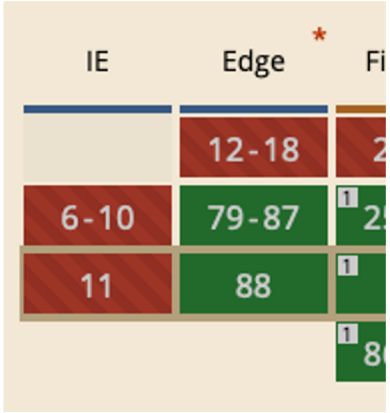
iOS VS Android
world: 15% vs 85%
Netherlands: 40% vs 60%
USA: 55% vs 45%

Web Audio & Browsers

- support in different browsers is growing and becoming reliable for this project caniuse.com
- I mostly use a part of the web audio-specification that's prety well supported now: MediaRecorder
- documentation is pretty good now: https://developer.mozilla.org/
- it's still complex sometimes now with the many apis that seem to do the same

The **MediaStream Recording API**, sometimes referred to as the *Media Recording API* or the *MediaRecorder API*, is closely affiliated with the Media Capture and Streams API and the WebRTC API.

The most problematic browsers : Safari & Internet Explorer



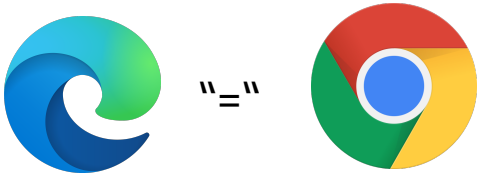
Problematic browsers?

Apple - Safari

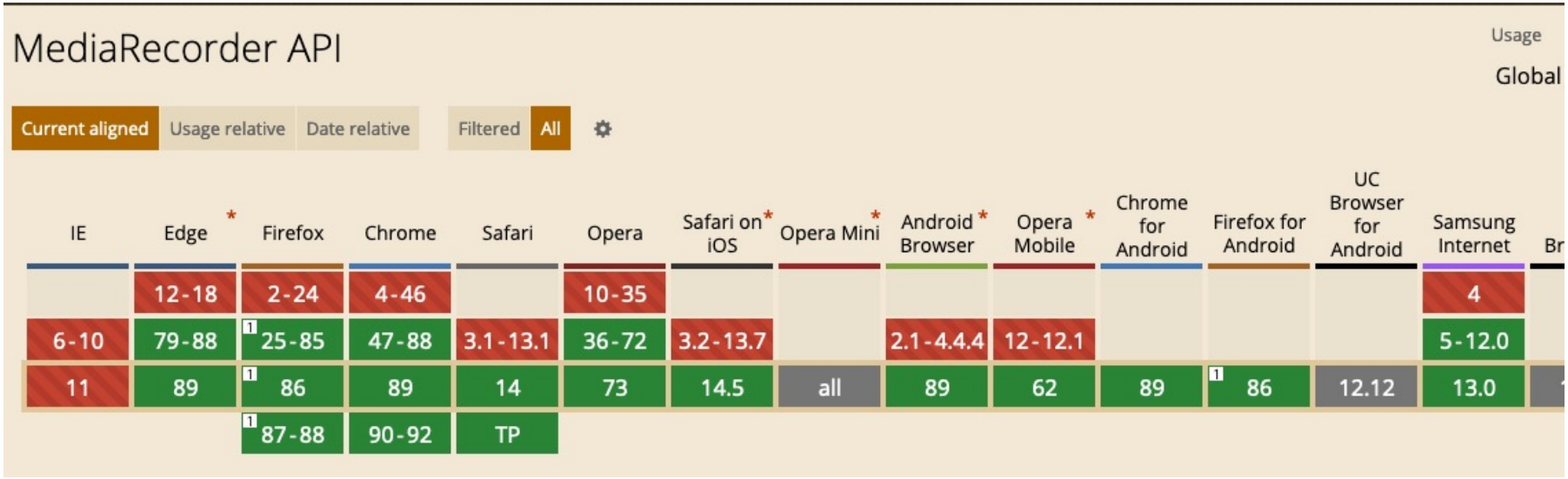
- Since sept 2020 it does support MediaRecorder out-of-the-box
- Apple is pushing iOS-updates aggresively
- also working on 6 year old iPads and IPhones (informal tests)

MicroSoft - Edge

- no more Internet Explorer
- support for Edge Legacy stopped March 13, 2021
- MicroSoft is pushing Edge updates aggresively
- the latest Edge is 'identical' to Chrome



Current Support (MDN browser-compat-data)



Quality of the recording

The best recording

- speaker is well prepared
- controlled environment
- a recording engineer
- quality equipment
- one microphone for all subjects
- etc.etc.

Quality of the recording

The real recording

depends on:

- the quality of the recording device
- the ability of the user
- the environment of the user
- the quality of the network
- the quality of the recording
- extremities in dynamics
- extremities in duration
- audio feedback loops
- audio-vandalism
- trust in permission questions

Mitigate with technical means:

At least:

- give the user, visual feedback of the audio signal
- give the user a possibility to listen to his/her recording
- give the user a possibility to re-record
- take care of time-limited recording
- safety net in the backend against abuse (not yet implemented)

Maybe:

- noise cancelling for reducing background noise
- compression or expansion
- audio bufferering in storage in the browser

Demo time: front-end

What do we need:

- obtain access to the hardware of the audience (**mediaStream api**)
- give visual feedback to the user of the audio signal (**Canvas api**)
- record an audio-stream in the browser (**mediaRecorder api**)
- make it an object after recording (**Blob api**)
- give an option to delete the file/blob (**js**)
- give an option to store the file somewhere (**Fetch api**)

Demo time: back-end

What do we need:

- store the recording in it's rawest form, storage space
- convert the file (**ffmpeg**) to a common format, all browsers have ideal 'native' container/codec formats (FF: .ogg (opus), Safari: .mp4 (aac), Chrome / Edge: .webm (opus)) => .mp4 (aac)
- simple viewer on the uploaded files, for demo purposes

Written in PHP, LimeSurvey is also written in PHP

Demonstration Time

Future work

- clean integration in/extension of LimeSurvey
  - Still be able to upgrade LS
- How to help users with wide variety of the permission dialogues?
- Desktops still won't have microphones most of the time
  - Temporarily switch to mobile with a QR code?
    - Store recording at the right place, return to desktop …

- Of course the results of this work will be available to the whole CLARIAH++ community not only The Meertens Institute ☺

More info

- Later:
  - https://github.com/knaw-huc/SPAQ
  - https://www.neerlandistiek.nl/2021/03/de-collecties-van-het-meertens-instituut/
  - Menzo Windhouwer
    - Slack
    - Menzo.Windhouwer@di.huc.knaw.nl
  - Maarten van der Peet
    - Maarten.van.der.Peet@di.huc.knaw.nl
- Now:
  - Questions/suggestions? ☺