How we transcribed 150K Dutch podcasts

Sahra Mohamed

June 25th 2024

The next 20 minutes

01 Research idea

02 Research setup

03 Used tools

04 Executing with Snellius from SURF

05 Tips for further exploration

06 Questions

01

Research idea

Research idea

Datajournalist and poetry critic
sahra.site

Creative coder
haykranen.nl

Long-form content versus short-form

Utrecht University, Data School, Applied Data Science

2022/2023

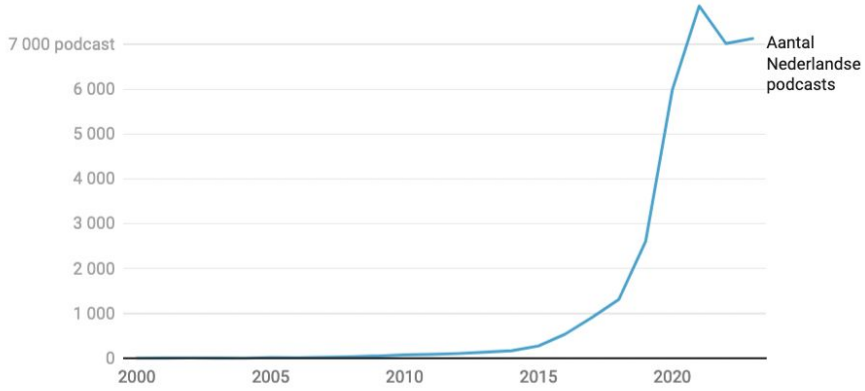
Whisper from OpenAI

Misinformatie en privacy

In 2023 voor 9 jaar aan Nederlandstalige pod-casts gepubliceerd

Vorig jaar werden 7.130 Nederlandstalige podcasts gepubliceerd

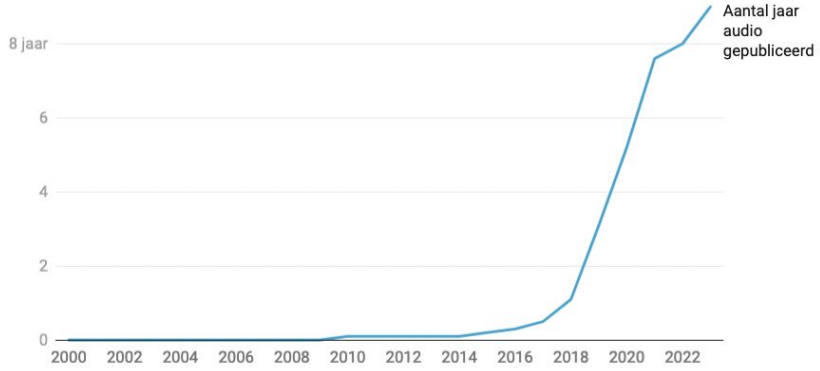
Dit zijn podcasts die in dat jaar minstens één aflevering hebben gepubliceerd.



Gegevens ophalen • Gecreëerd met Datawrapper

Sinds 2018 worden meer Nederlandstalige podcasts gepubliceerd dan je in een jaar kunt luisteren

Je hebt 9,3 jaar nodig om alle Nederlandstalige podcasts uit 2023 te beluisteren



Gegevens ophalen • Gecreëerd met Datawrapper

02

Research setup


XML, json
online link
mp3
Speech to text
RSS feeds
Podcast Index
https://podcastindex.org
PODCAST INDEX
Search for podcasts or music
All

Amberscript versus Whisper (at the time)

- **Punctuation**
  - Actual sentence:
    - “Hallo lieve luisteraar, Eline hier. Wat een rit hebben we samen gemaakt, hè? Het voelt vreemd om … ”
  - Amberscript:
    - “Hallo, lieve luisteraar, Eline hier, wat een rit hebben we samen gemaakt, hè, het voelt best vreemd om …”
  - Whisper:
    - “Hallo lieve luisteraar, Eline hier. Wat een rit hebben we samen gemaakt, hè? Het voelt best vreemd om …”

Amberscript versus Whisper (at the time)

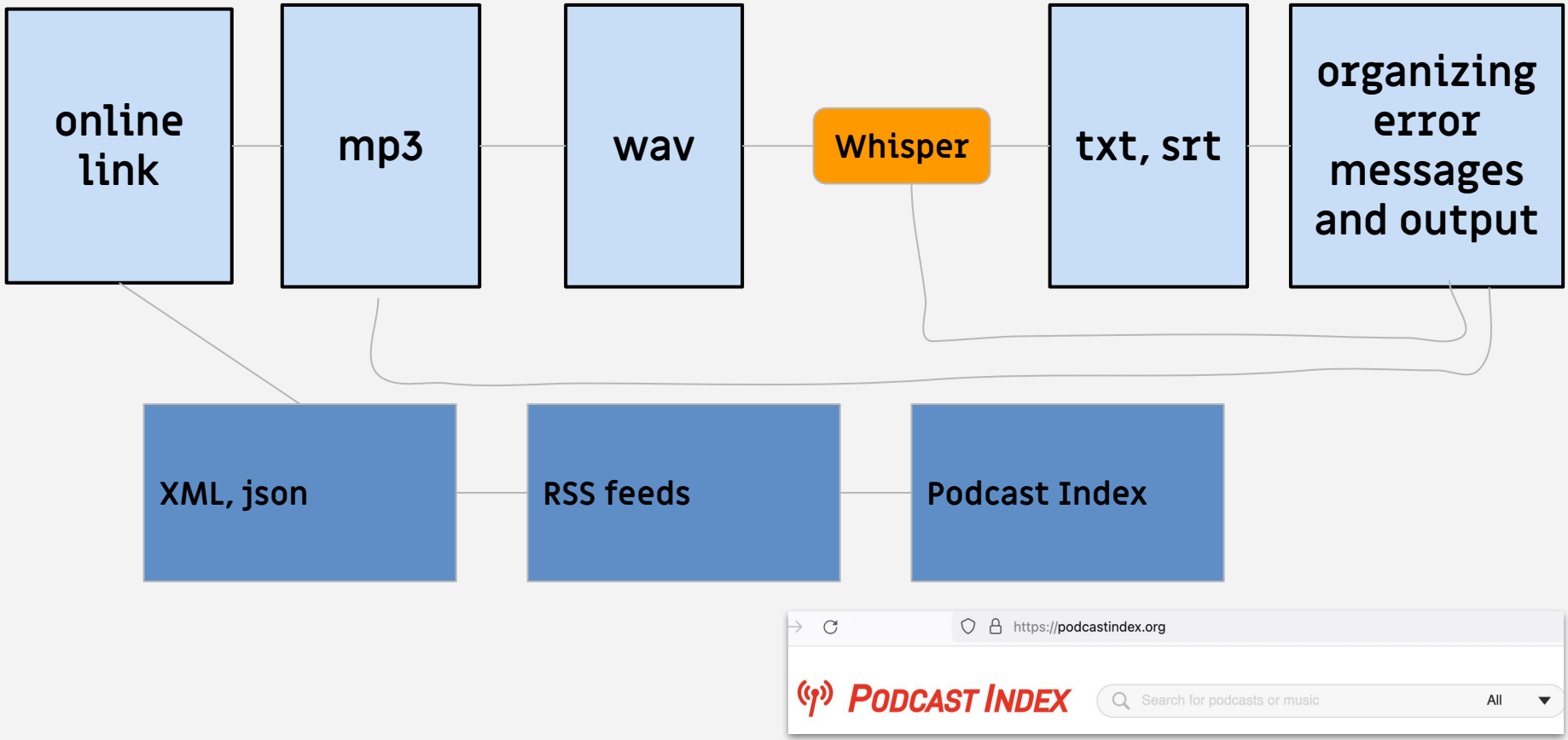
- **Random english words are difficult to handle**
  - Actual sentence:
    - “positief dat Tamar en ik in seizoen 2 samen als host aan de slag gingen.”
  - Amberscript:
    - “positief dat tamer en ik in seizoen twee samen als hoogste aan de slag gingen”
  - Whisper:
    - “positief dat Tamar en ik in seizoen 2 samen als hoogstaande slag gingen”

Amberscript versus Whisper (at the time)

- **New nouns**
  - Actual sentence:
    - “verslag uitbracht vanuit de Lappenmand.”
  - Amberscript:
    - “verslag uitbracht vanuit lopen”
  - Whisper:
    - “verslag uitbracht vanuit de Lapperman”

Amberscript versus Whisper (at the time)

- **Spelling**
  - Actual sentence:
    - “the time of COVID with Britney Wilson”“... my interview with Britney.””my interview with Britney.”
  - Amberscript:
    - “the time of COVID with Brittney Wilson”“My name is Brittany Wilson”“ ... my interview with Britney.” ”
  - Whisper:
    - “the time of COVID with Brittany Wilson”“My name is Brittany Wilson.”“.... my interview with Brittany.”


online link
mp3
wav
Whisper
txt, srt
organizing error messages and output
XML, json
RSS feeds
Podcast Index
https://podcastindex.org
PODCAST INDEX
Search for podcasts or music
All

03

Used tools

Used tools

- Snellius supercomputer at SURF
- https://github.com/hay/audio2text
- https://github.com/ggerganov/whisper.cpp
- https://github.com/m-bain/whisperX

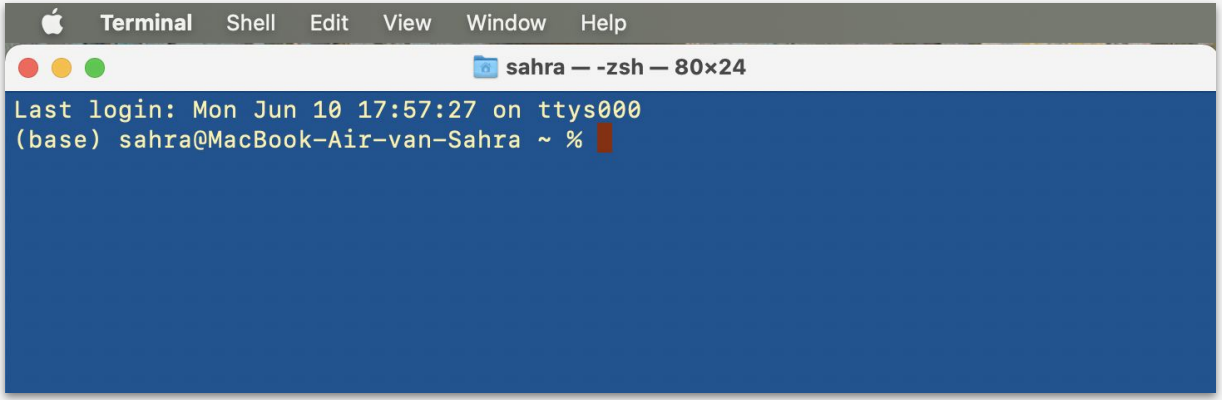
04

Executing with Snellius from SURF

Snellius

SBUs per 1 hour, full node	SBUs per 1 hour, smallest possible allocation
128 SBUs	16 SBUs

- Small grants from SURF
- Response within two weeks
- Maximum 1,000,000 CPU/GPU SBUs
- Few other specifications related to memory
- Connecting to Snellius is done via terminal



- Background in artificial intelligence and linguistics
- Jelle Treep, Research Engineer at Utrecht University

WhisperCPP

On CPU

```
sbatch --array 71-74:2 jobscript.sh
```

How we repeated the task

- Execute jobscript.sh with index 71
- Process current index + 1
- Execute jobscript.sh with index 73
- Process current index + 1

75 podcasts took a max wall time of 48 hours per node, because of the variety in lengths. Even if you hussle episodes to average out the lengths in the batches: it takes a large sample size to truly average things. Similar to flipping a coin.

WhisperX

On GPU

No more troubles with distributing podcast episodes of different lengths.

Pay attention to Python dependencies.

Example calculation:

You can transcribe 1 podcast in a minute on average. So 60 podcasts per hour.

16 hours on a node * 60 podcasts = 960 in a batch

102 nodes * 16 hours * 128 credits per hour = 208.896 credits. That is the cost for executing 102 nodes processing 960 podcasts each. That is almost 100K podcasts in total. Remember you can get 1 million credits in a small grant.

```
(base) [smohammed@int5 ~]$ squeue -u smohammed
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
         4202594_0       gpu x_job.sh smohamme  R      12:19      1 gcn20
       4202594_720       gpu x_job.sh smohamme  R      12:19      1 gcn49
      4202594_1440       gpu x_job.sh smohamme  R      12:19      1 gcn63
      4202594_2160       gpu x_job.sh smohamme  R      12:19      1 gcn43
      4202594_2880       gpu x_job.sh smohamme  R      12:19      1 gcn45
      4202594_3600       gpu x_job.sh smohamme  R      12:19      1 gcn10
      4202594_4320       gpu x_job.sh smohamme  R      12:19      1 gcn11
      4202594_5040       gpu x_job.sh smohamme  R      12:19      1 gcn12
      4202594_5760       gpu x_job.sh smohamme  R      12:19      1 gcn14
      4202594_6480       gpu x_job.sh smohamme  R      12:19      1 gcn31
      4202594_7200       gpu x_job.sh smohamme  R      12:19      1 gcn31
      4202594_7920       gpu x_job.sh smohamme  R      12:19      1 gcn33
      4202594_8640       gpu x_job.sh smohamme  R      12:19      1 gcn34
      4202594_9360       gpu x_job.sh smohamme  R      12:19      1 gcn35
     4202594_10080       gpu x_job.sh smohamme  R      12:19      1 gcn35
     4202594_10800       gpu x_job.sh smohamme  R      12:19      1 gcn36
     4202594_11520       gpu x_job.sh smohamme  R      12:19      1 gcn19
     4202594_12240       gpu x_job.sh smohamme  R      12:19      1 gcn20
     4202594_12960       gpu x_job.sh smohamme  R      12:19      1 gcn20
     4202594_13680       gpu x_job.sh smohamme  R      12:19      1 gcn22
     4202594_14400       gpu x_job.sh smohamme  R      12:19      1 gcn23
     4202594_15120       gpu x_job.sh smohamme  R      12:19      1 gcn23
     4202594_15840       gpu x_job.sh smohamme  R      12:19      1 gcn24
     4202594_16560       gpu x_job.sh smohamme  R      12:19      1 gcn26
     4202594_17280       gpu x_job.sh smohamme  R      12:19      1 gcn41
     4202594_18000       gpu x_job.sh smohamme  R      12:19      1 gcn51
     4202594_18720       gpu x_job.sh smohamme  R      12:19      1 gcn52
     4202594_19440       gpu x_job.sh smohamme  R      12:19      1 gcn62
(base) [smohammed@int5 ~]$
```

05

Tips for further exploration

For small amounts of audio:

- Local laptop with a GPU
- The Research Cloud (SURF)
- The official implementation of Whisper from OpenAI
- WhisperX (GPU)

For very large amounts of audio:

- Snellius (SURF)
- WhisperX (GPU)

In most cases WhisperX and the Research Cloud will be sufficient with the current technology.

Tips:

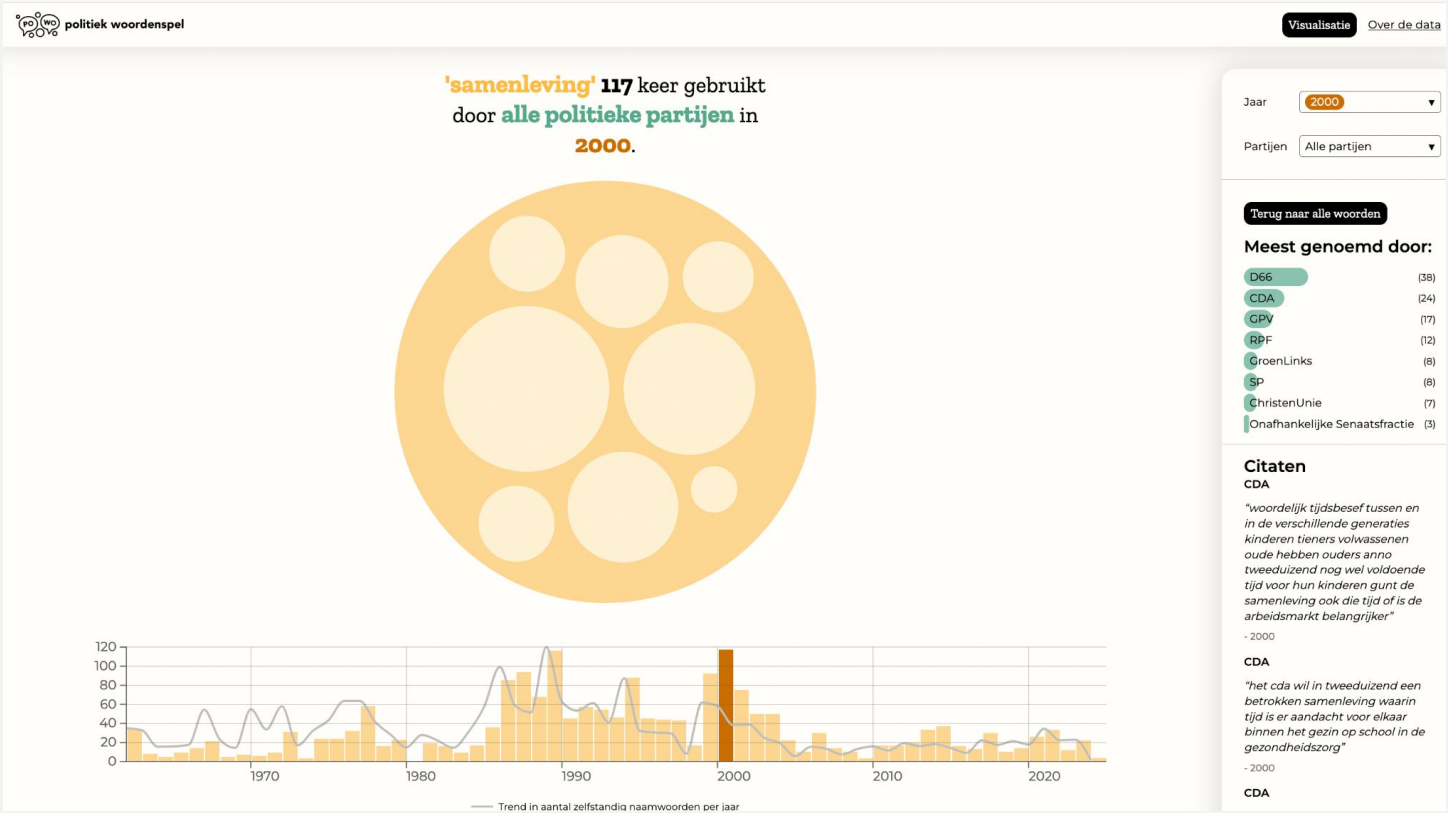
- Think about a margin for reserving time on a compute node.
  - For example: if you transcribe 10 episodes that each are between 15 minutes or 2 hours, the node needs to run for at least the maximum amount of time it takes to finish the longest episode.
  - In other words: adding up the time for each individual episode in the entire dataset will not give you the total time it takes to finish transcribing the entire dataset.
  - It helps to think about how you will distribute the episodes.
    - For example, by trying to figure out the length of the episode beforehand.
    - By processing podcasts in parallel within a node.
- Consider applying for consultancy hours at SURF when filling out the grant application form.
  - For example: if your project is planned to last X months, you can think about a sufficient number of Y milestones where you can check in with technical advisers and ask for advice. Then you can add that estimate to your grant application.

06

Questions?

The next article about Dutch podcasts will be published later this autumn at Pointer (KRO-NCRV)

For those interested in visual form and front-end: we used speech transcripts in this project as well to create a tool – politiekwoordenspel.nl



Thank you

If you get any questions later on, feel free to contact me!
sahra.site