HoMed project *(Homo Medicinalis)* https://homed.ruhosting.nl

An Innovative Methodology Utilizing AI-based Automatic Speech Recognition for Transcribing Dutch Patient-Provider Consultation Recordings

DH Benelux Conference 2024 – 11th Edition – June 5-7, 2024 | Leuven (Belgium)

- **Dr. Cristian Tejedor-García** - *Assistant professor, Radboud University*

  cristian . tejedorgarcia @ ru . nl

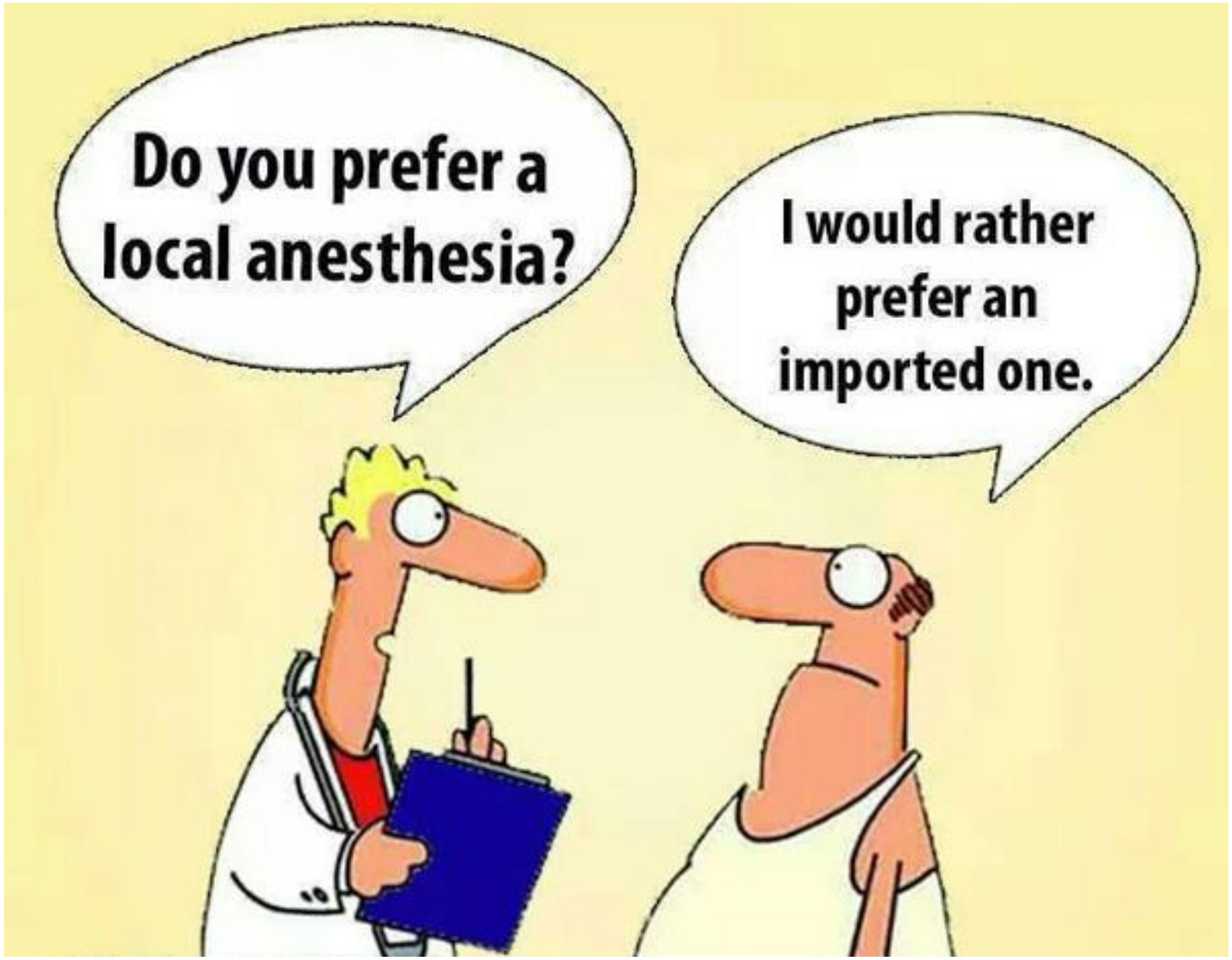
  https://www.ru.nl/en/people/tejedor-garcia-c

Platform Digitale Infrastructuur Social Sciences & Humanities

NIVEL Research for better care

CLST | Centre for Language and Speech Technology Radboud University

Motivation



15k cases / year – [Report, 2017]

Vervolgonderzoek medicatieveiligheid: eindrapport. *Rotterdam/Utrecht/Nijmegen: Erasmus MC, NIVEL, Radboud UMC, PHARMO, 2017. 129 p.*

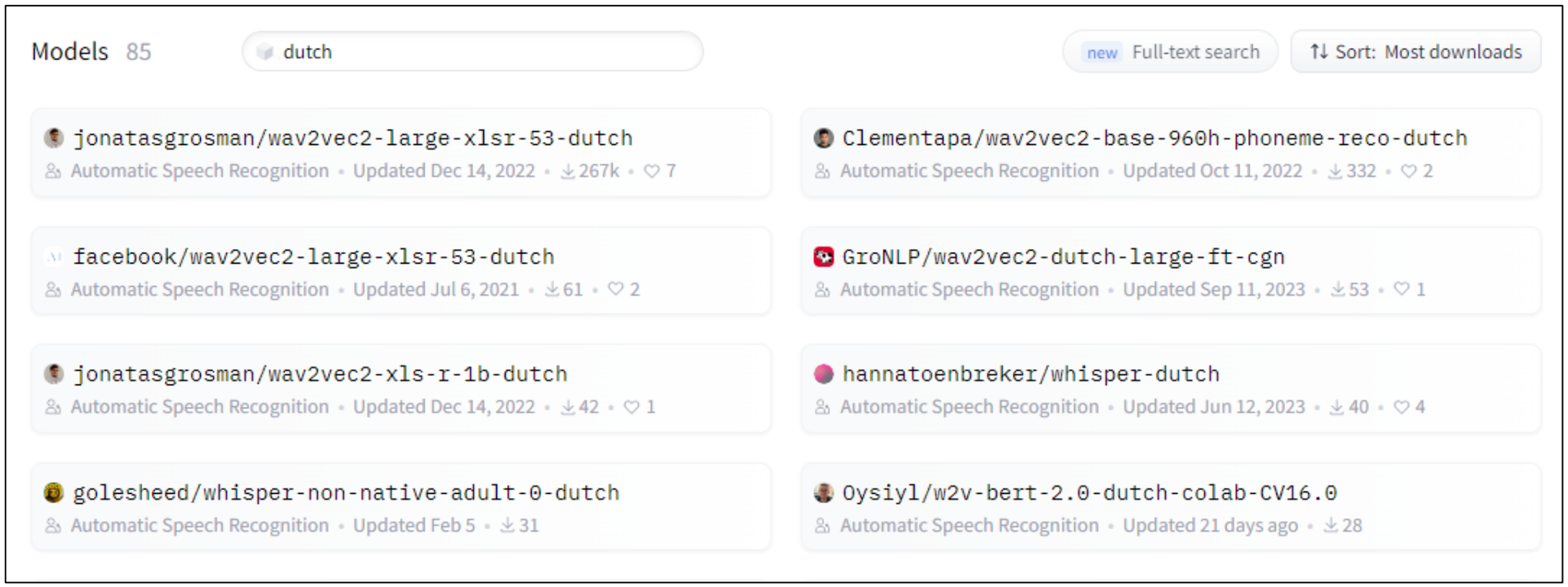
Index of contents

1. Speech Recognition in Healthcare

2. Speech Recognition & Dutch language (I)

- Kaldi_NL: Classical ASR system
  - ~900 speech hours

- **New Hugging-face era: Transformer-based systems**
  - 500k – 15M of speech hours



2. Speech Recognition & Dutch language (II)

SURF

Diensten ▾ Thema's ▾ Nieuw

NIEUWS

02 november 2023

Nederland start bouw GPT-NL, als eigen AI-taalmodel

Nederland gaat een eigen open taalmodel ontwikkelen: GPT-NL. Dit model is nodig voor het ontwikkelen, versterken en bestendigen van de digitale soevereiniteit. Non-profitpartijen TNO, NFI en SURF gaan samen het model ontwikkelen om zo een belangrijke stap te zetten richting transparant, eerlijk en toetsbaar gebruik van AI naar Nederlandse en Europese waarden en richtlijnen en met respect voor het eigenaarschap van data. Financiering van het model is afkomstig van het ministerie van EZK.

Nederland trekt 13,5 miljoen euro uit voor ontwikkeling van eigen AI-taalmodel

Onderzoeksinstituut TNO gaat samen met SURF en het Nederlands Forensisch Instituut een eigen AI-taalmodel ontwikkelen, GPT-NL. Hiermee willen de partijen naar eigen zeggen een 'veilig alternatief' ontwikkelen voor buitenlandse taalmodellen als ChatGPT.

2. Speech Recognition & Dutch language (III)

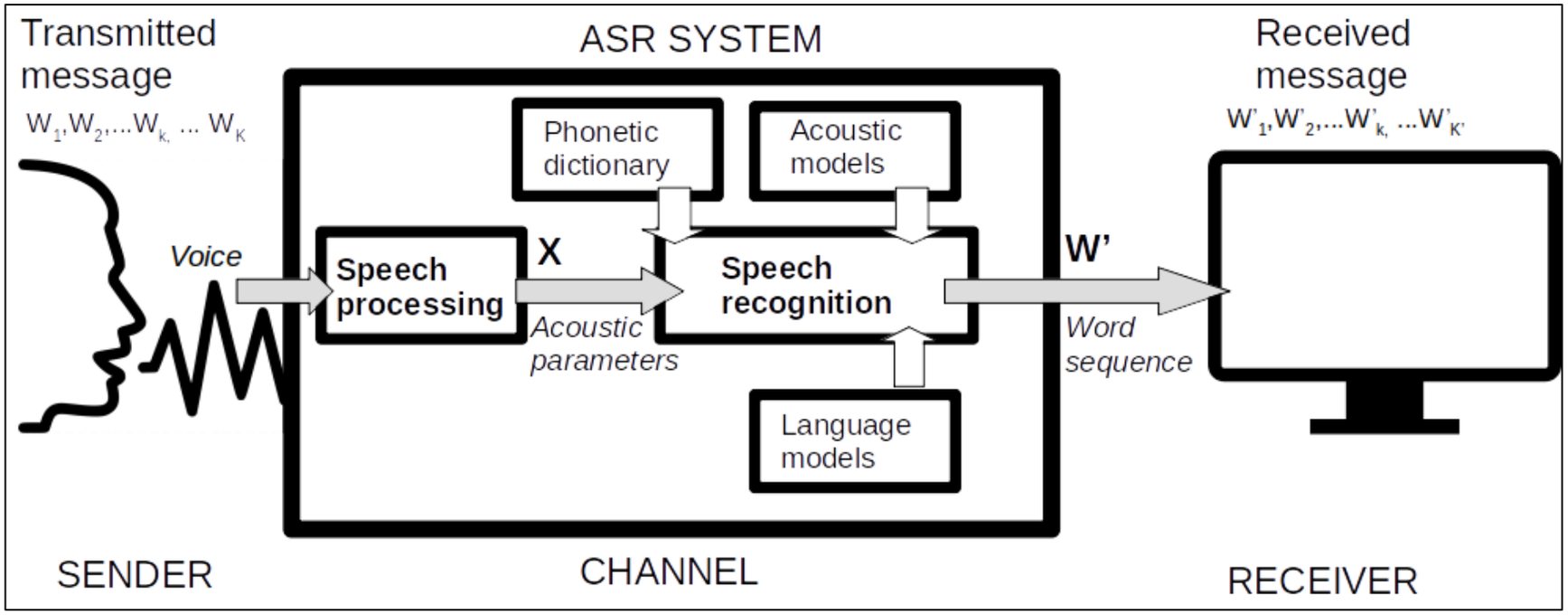
- ASR_NL_Benchmark:
  - https://opensource-spraakherkenning-nl.github.io/ASR_NL_results/
    - University of Twente
    - Radboud University

2nd Dutch Speech Day

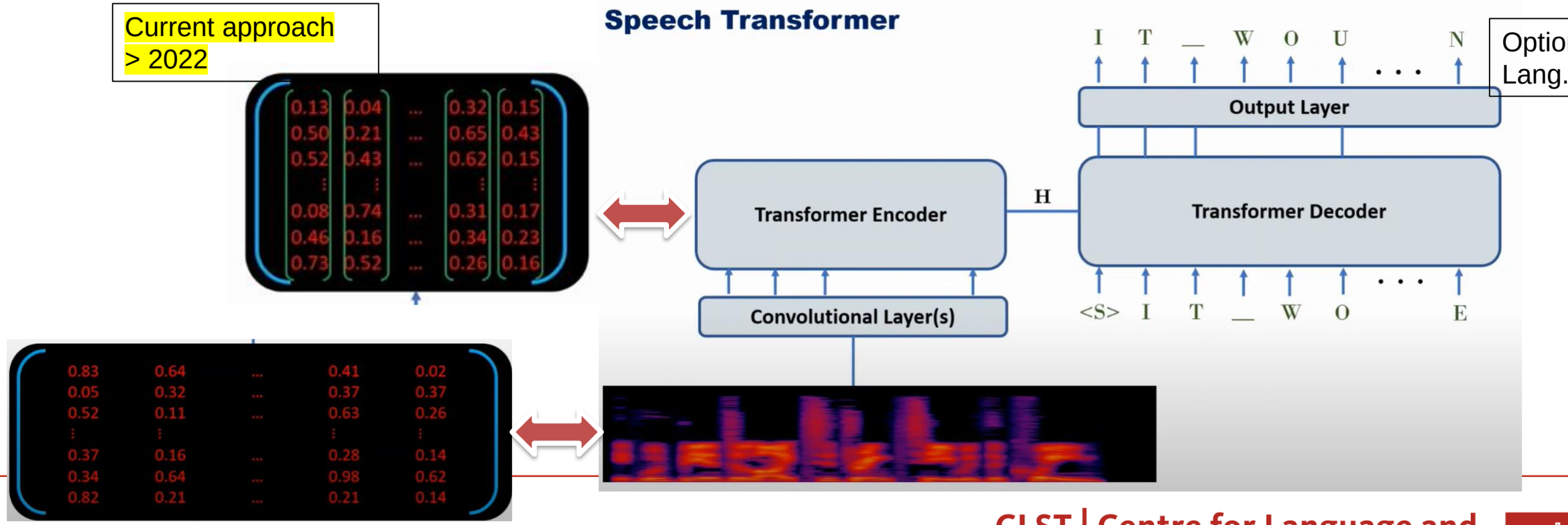
Model\Dataset	Jasmin_q_1	Jasmin_q_2	Jasmin_q_3	Jasmin_q_4	Jasmin_q_5
Kaldi_NL	28.1%	16.2%	43.6%	45.3%	20.9%
Whisper v2	22.6%	18.0%	36.5%	37.3%	22.2%
Whisper v3	34.2%	29.4%	50.4%	58.5%	34.4%
Whisper v2 w/ VAD	20.1%	12.4%	30.2%	33.4%	14.9%
Whisper v3 w/ VAD	34.7%	27.5%	46.7%	53.0%	30.2%
faster-whisper v2	20.3%	11.3%	29.9%	30.6%	13.7%
faster-whisper v3	28.1%	25.2%	50.9%	62.6%	27.6%
faster-whisper v2 w/ VAD	**19.1%**	**11.1%**	**29.5%**	**30.0%**	**12.8%**
faster-whisper v3 w/ VAD	27.5%	22.4%	42.6%	49.4%	25.2%

3. Methodology (I)

Classical approach

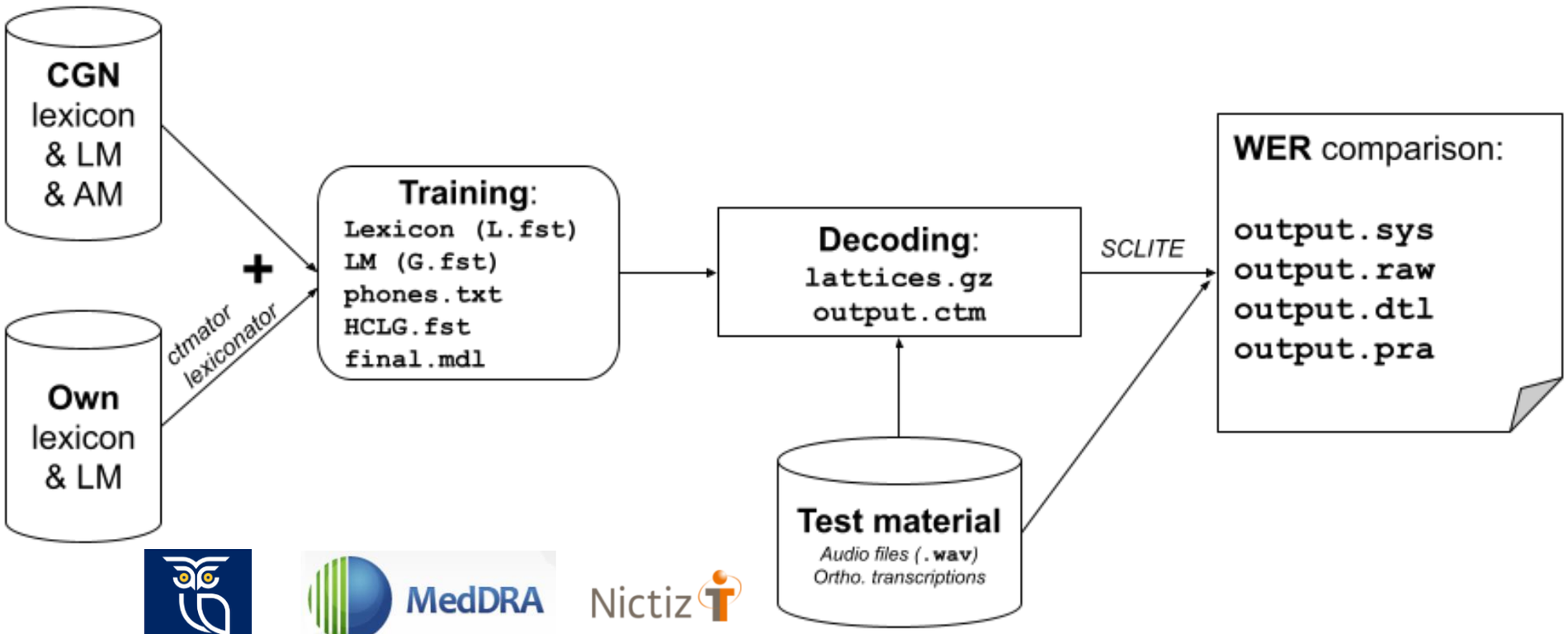


Current approach
> 2022



3. Methodology (II)

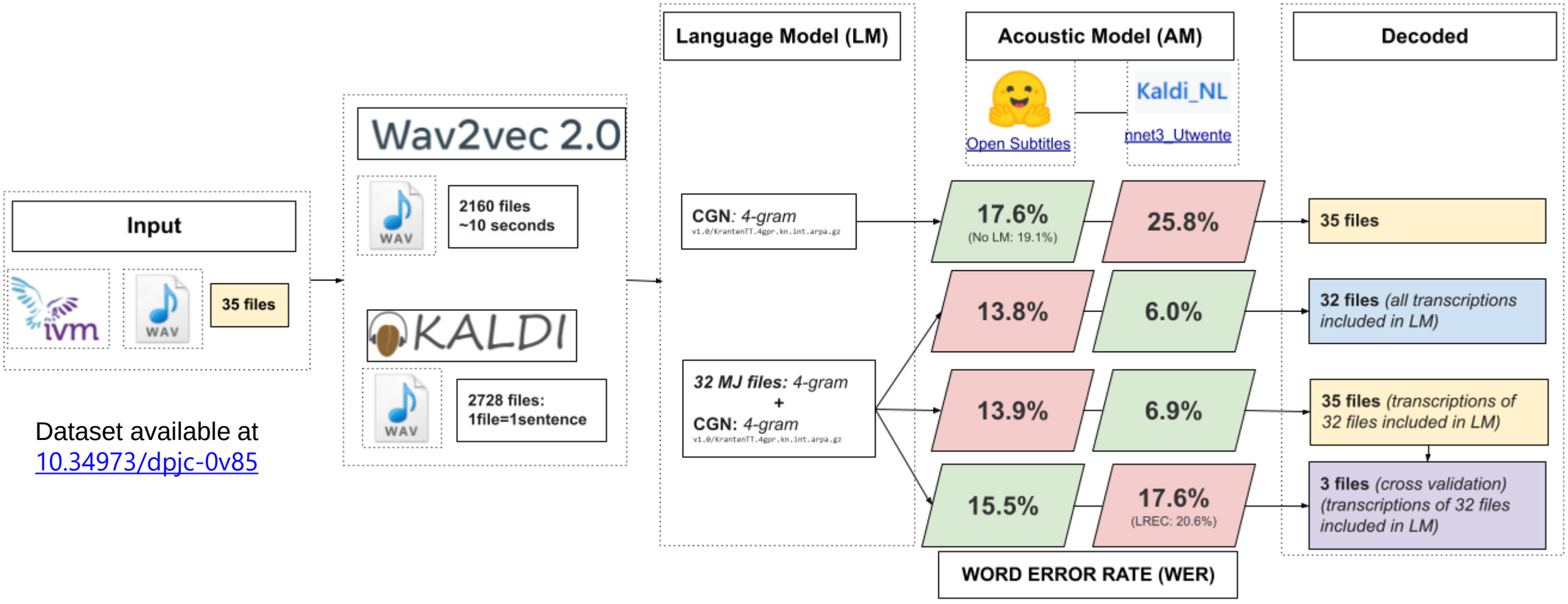
- HoMed: Language model fine-tuning (2021-2022) – Kaldi-NL



LREC2022 - Tejedor-Garcia et al. 2022a
DHBENELUX2022 - van der Molen et al. 2022

3. Methodology (III)

HoMed: 2022 Wav2vec2.0 exploration

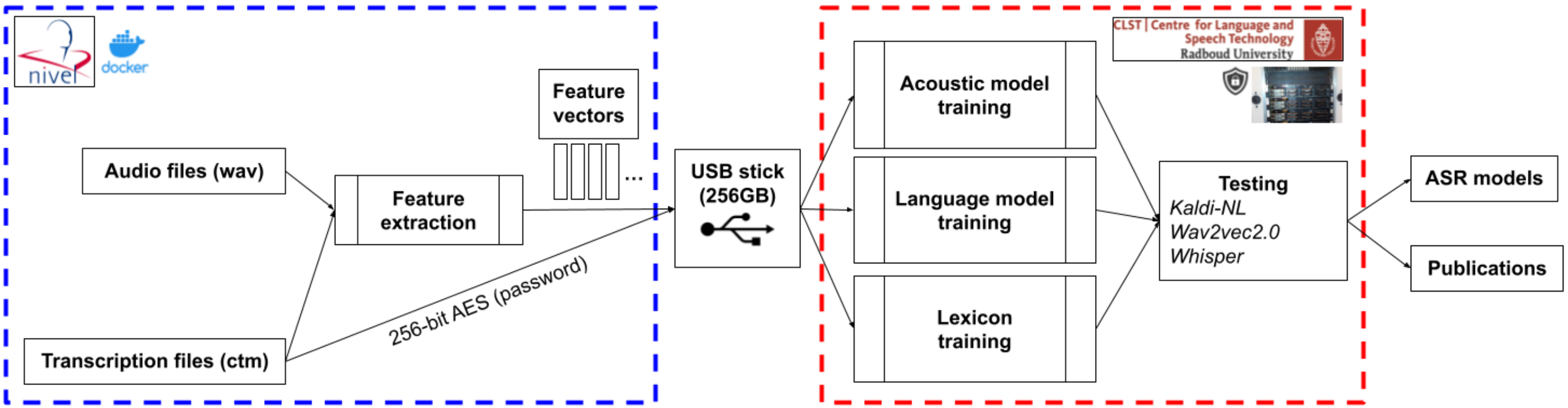


Dataset available at
10.34973/dpjc-0v85

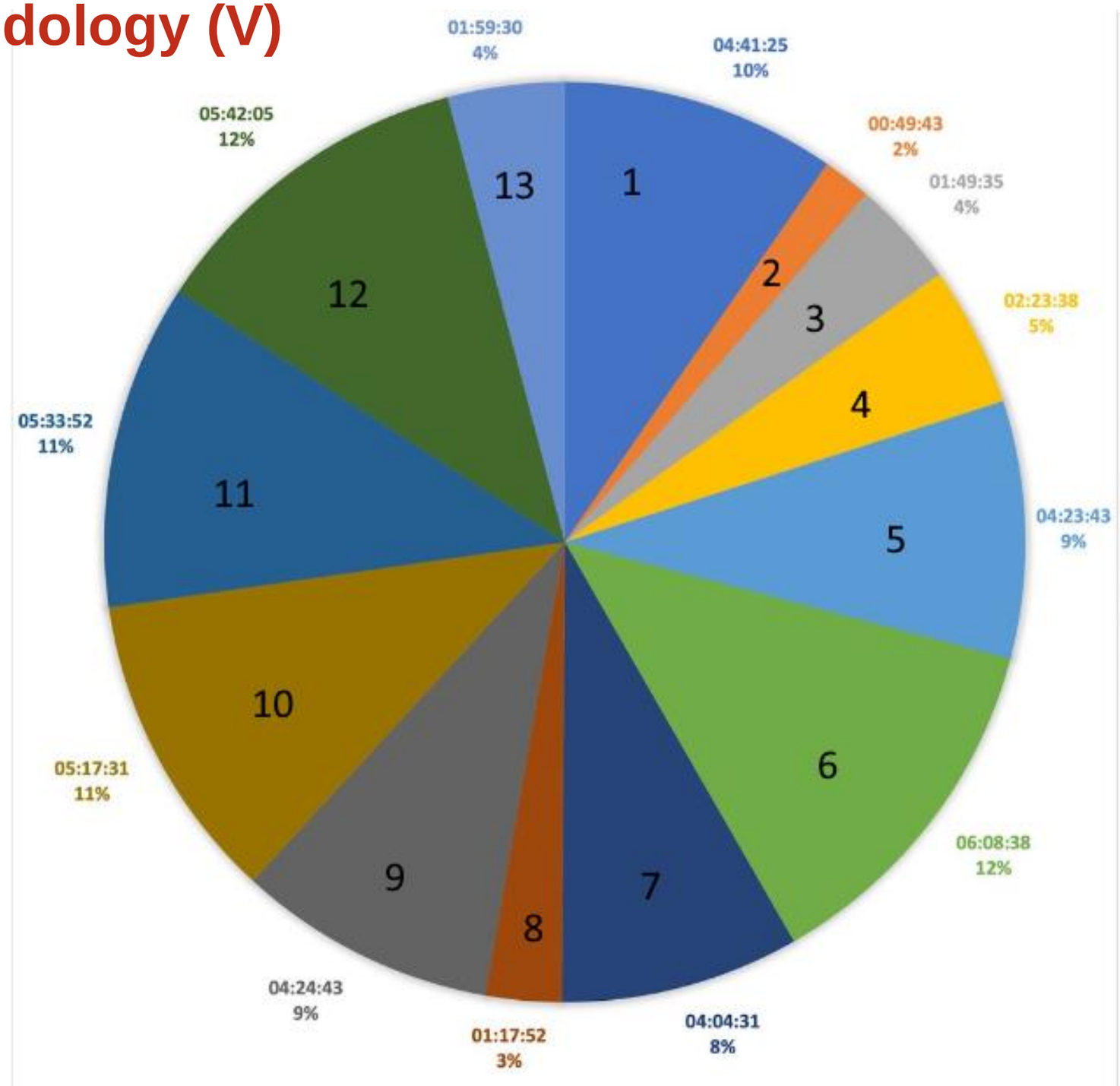
LREC2022 - Tejedor-Garcia et al. 2022a

3. Methodology (IV)

HoMed: 2022-2023: An innovative method for building special purpose speech recognisers for sensitive data domains built on evidence-based pilot projects.



3. Methodology (V)



Total: 50 hours

-- Audio quality

1 Pulmonology
2 Oncology
3 Palliative care
4 Radiology
5 Pharmacy
6 Nephrology
7 Obstetrics
8 Home care
9 Oncology
10 Paediatrics
11 Internal medicine
12 Gynecology
13 Geriatry

4. Results (I)

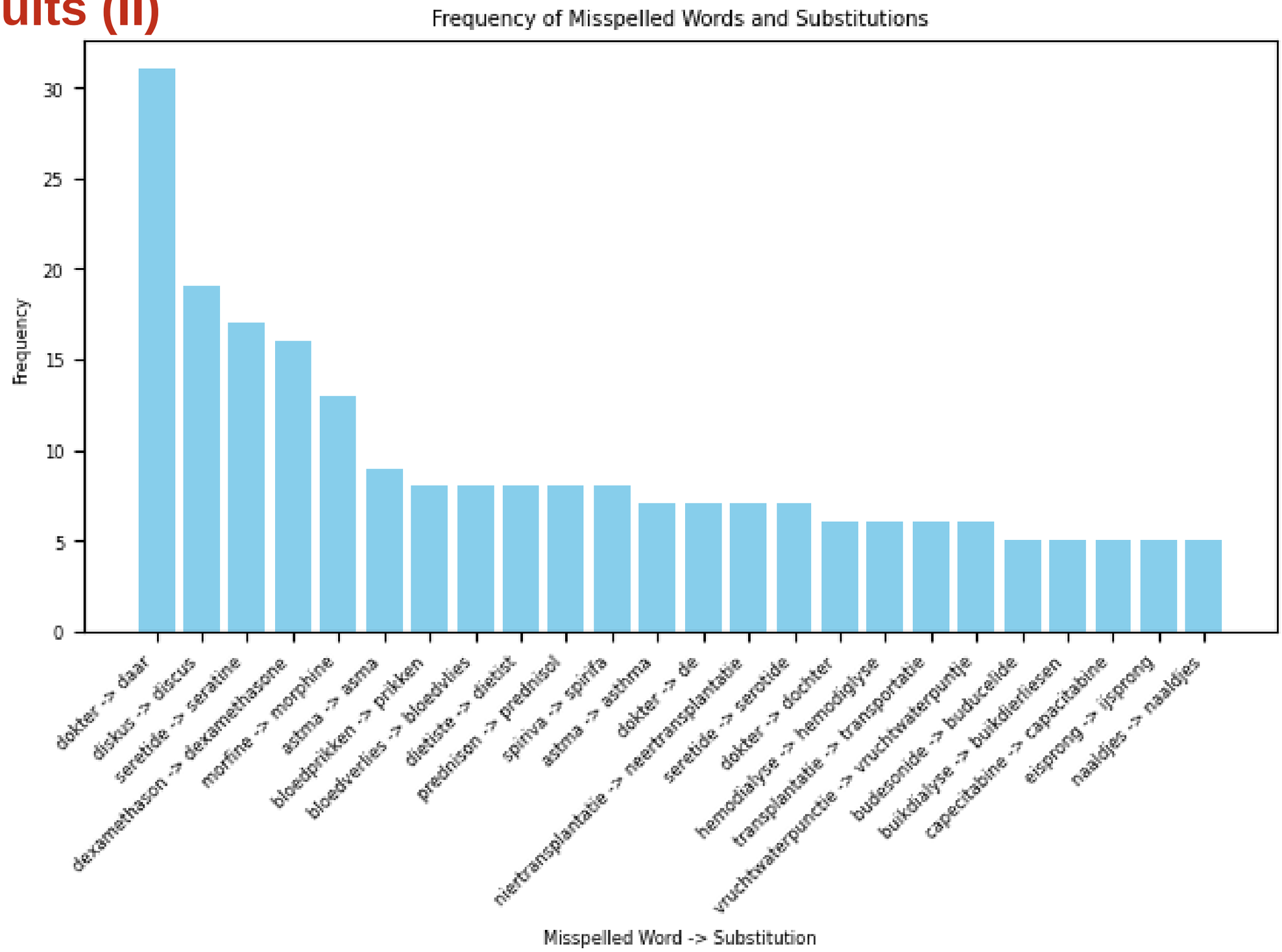
TEST-40H	# Fil	# Wrd	Corr	Sub	Del	Ins	Err
Whisper-large-v2	110	421847	71.3	10.9	17.8	5.4	34.1
Kaldi_NL	110	421847	33.9	22.8	43.3	2.4	68.5

Subs:
1: 942 -> daar ==> er
2: 454 -> da's ==> is
3: 397 -> dat ==> het

Ins:
1: 1251 -> dat
2: 1208 -> het
3: 1152 -> ja

Del:
1: 8568 -> ja
2: 2776 -> dat
3: 2429 -> ik

4. Results (II)


Frequency of Misspelled Words and Substitutions
Frequency
0
5
10
15
20
25
30
dokter -> daar
diskus -> discus
seretide -> seratine
dexamethason -> dekamethasone
morfine -> morphine
astma -> asma
bloedprikken -> prikken
bloedverlies -> bloedvlies
dietiste -> dietist
prednison -> prednisol
spiriva -> spirifa
astma -> asthma
dokter -> de
niertransplantatie -> neertransplantatie
seretide -> seroxide
dokter -> dochter
hemodialyse -> hemodiglyse
transplantatie -> transportatie
vruchtwaterpunctie -> vruchtwaterpuntje
budesonide -> buducelide
buikdialyse -> buikdiarliesen
capecitabine -> capacitabine
eisprong -> ijsprong
naaldjes -> naaldjes
Misspelled Word -> Substitution

5. Conclusions

1. AI has drastically changed the speech recognition field > 2022
   - Companies now have a predominant role
   - Huge open-source models are available: Quality, on-domain adaptation

2. A tailor-made methodology is key:
   - Standardized input and output is necessary following a well-stablished protocol
   - Words out of vocabulary (jargon) not a problem for most new models
   - A post-processing step (LLM) improves the performance results

3. In the upcoming years the speech recognition results will significantly improve:
   - New AI techniques, + data + open-source collaboration + multilingual approach
   - In-domain adaptation (fine-tuning) will still be the key
     - Low quality audio signal will still be the challenge
   - Richer information will be provided by ASR: emotions, personal information
     - Responsible AI & GDPR

6. Publications

1. Dutch ASR benchmark (**HoMed, 2024**): https://opensource-spraakherkenning-nl.github.io/ASR_NL_results/RU/wer.html

2. Tejedor-García, et al. (**2024a**). *An Innovative Methodology Utilizing AI-based Automatic Speech Recognition for Transcribing Dutch Patient-Provider Consultation Recordings.* **DH Benelux 2024** (June 5-7, Leuven) Conference. In press.

3. Tejedor-García, et al. (**2024b**). *Comparative analysis of state-of-the-art automatic speech recognition systems for transcribing medical consultation audio recordings* . **Health by Tech** Conference 2024 (May 30-31, Groningen). In press.

4. Van der Molen et al. (**2022**, May 4). *Challenges on the Promising Road to Automatic Speech Recognition of Privacy-Sensitive Dutch Doctor-Patient Consultation Recordings.* **DH Benelux 2022** - ReMIX: Creation and alteration in DH (hybrid), Belval Campus, Esch-sur-Alzette, Luxembourg and online. https://doi.org/10.5281/zenodo.6517157

5. Tejedor-García, et al. (**2022a**). *Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain. Proceedings of the 13th International Conference on Language Resources and Evaluation (**LREC2022**), pp. 1032-1039.* http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.110.pdf

6. Tejedor García, C., Molen, B. & Heuvel, H. van den (**2022b**). *Homed Transcriptions Medicijnjournaal. Radboud Data Repository* [**Dataset**]. doi: 10.34973/dpjc-0v85.

7. Davelaar et al. *Spraakherkenning mogelijk zinvol bij medicatiegesprek.* Pharmaceutisch Weekblad, 12-09-2022

HoMed project *(Homo Medicinalis)* https://homed.ruhosting.nl

An Innovative Methodology Utilizing AI-based Automatic Speech Recognition for Transcribing Dutch Patient-Provider Consultation Recordings

DH Benelux Conference 2024 – 11th Edition – June 5-7, 2024 | Leuven (Belgium)

- **Dr. Cristian Tejedor-García** - *Assistant professor, Radboud University*

  cristian . tejedorgarcia @ ru . nl

  https://www.ru.nl/en/people/tejedor-garcia-c

Platform Digitale Infrastructuur Social Sciences & Humanities

NIVEL Research for better care

CLST | Centre for Language and Speech Technology Radboud University