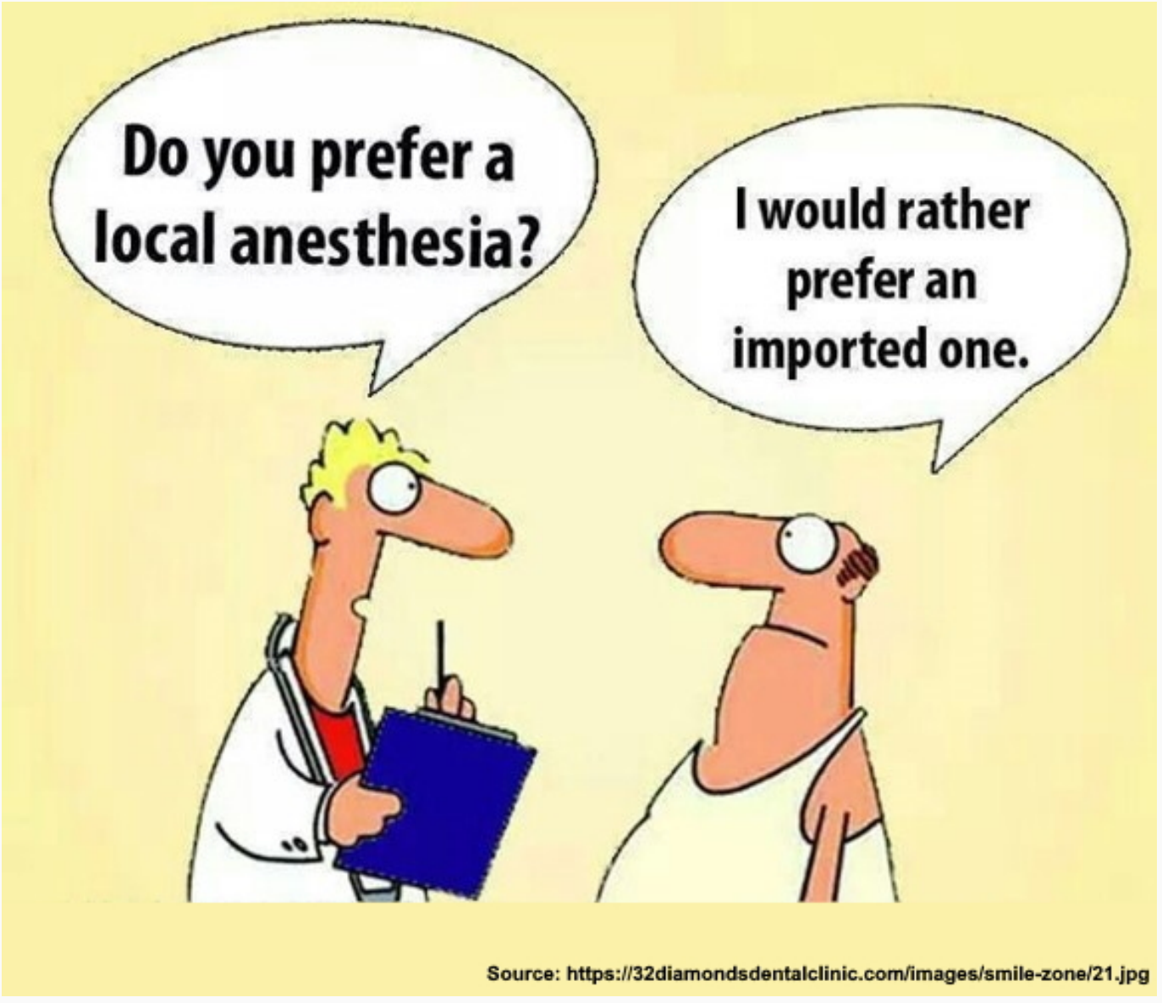
Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain

C. Tejedor-García (**cristian.tejedorgarcia@ru.nl**)*, B. van der Molen†, H. van den Heuvel*, A. van Hessen‡, T. Pieters†

*CLST (RU, Nijmegen),† Freudenthal Institute (UU, Utrecht), ‡ HMI (UT, Twente), the Netherlands

Context

- Annually +**15,000 hospital admissions** in the Netherlands
  - Avoidable **misuse** of **medicines**
- Main **reasons**:
  - Functional illiteracy
  - Forgetfulness
  - Misuse of prescribed usage
- **Consequences**:
  - Diverse + inappropriate forms of use
  - Low levels of adherence
  - Waste of scarce financial resources


Do you prefer a local anesthesia?
I would rather prefer an imported one.

Source: https://32diamondsdentalclinic.com/images/smile-zone/21.jpg

- **How** to **solve it**?
  - Better understanding of the explicit and implicit attribution of meaning to medicines as part of the information processing
  - Effective + efficient **transcriptions** of doctor-patient interviews: **ASR** technology
    - Context with considerable **privacy-sensitive constraints**

Our proposal: HoMed Project

- Proposes a **new** research **infrastructure** and **method**:
  - Automatic transcription of sensitive audio-visual (AV) recordings
  - General Data Protection Regulations (**GDPR**)
- Current largest open-source generic ASR for Dutch (Kaldi_NL):
  - Vocabulary: not healthcare jargon: needs **adaptation**:
    1. **Semantic adaptation (LM)**: *this paper*
       Material: Medicijnjournaal + lists of medical terms
    2. **Acoustic adaptation (AM)**:
       Material: Nivel AV recordings + previous material
- **INPUT**: healthcare-related material
  - Transcription files + healthcare word lists
  - AV-recording files (highly sensitive)
- **OUTPUT**: ASR models + methodology to other domains:
  1. CLARIAH's Infrastructure (Media Suite)
  2. Stichting Open Spraaktechnologie
  3. Nivel: standalone version

Funding & Useful Links

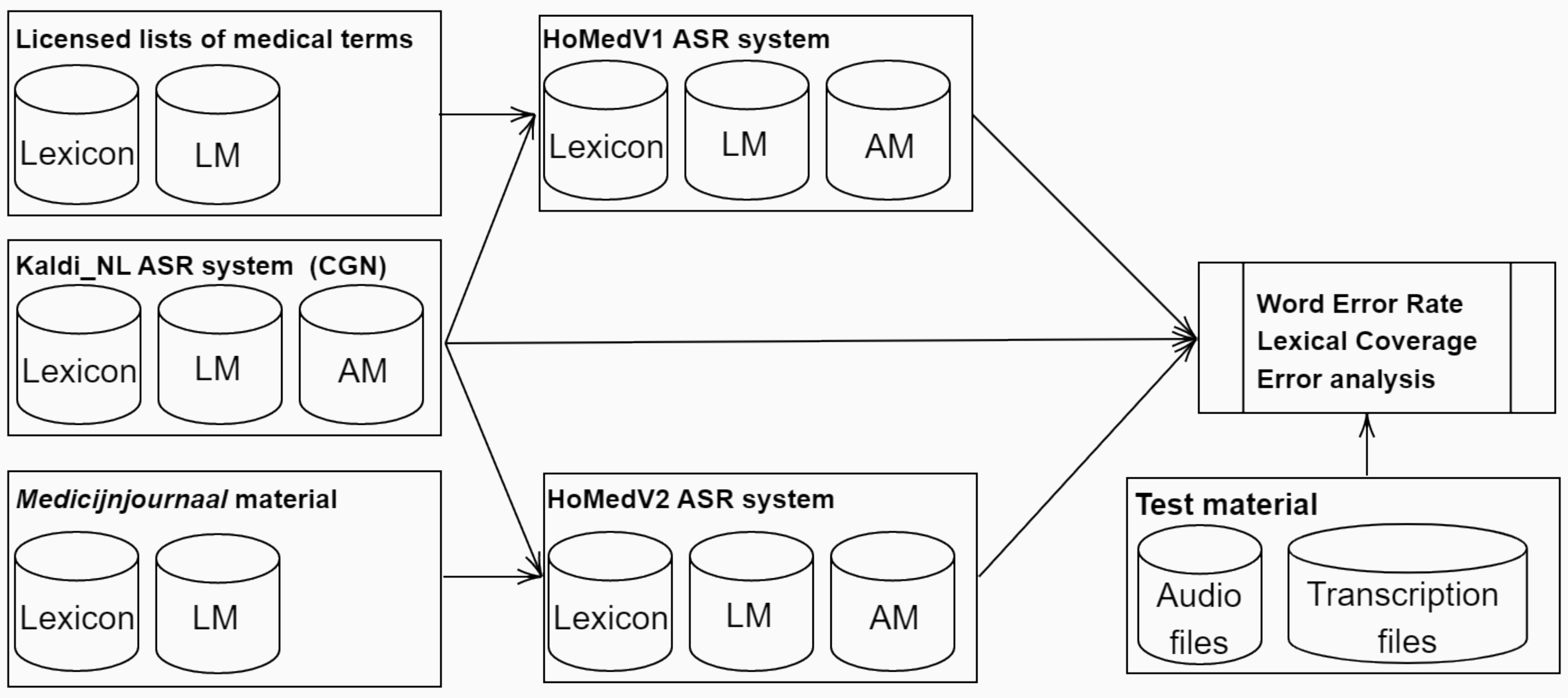
- Platform Digitale Infrastructuur Social Science and Humanities **PDI-SSH** 2020: `https://pdi-ssh.nl`
- **Project** webpage: `https://homed.ruhosting.nl`

ASR in (Dutch) Healthcare

- ASR Advantages:
  - Increases medical staff's productivity
  - Facilitates the completeness of medical documentation
  - Inspires patient engagement
- **CGN** (*Corpus Gesproken Nederlands*):
  - Generic/daily conversations
  - **WER**: ∼7-8%
- **Commercial** ASR systems:
  - Jargon + Data privacy + Costs
- Related current research projects (*Google ASR*) in the NL
  - Care2Report, CAIRE-lab

ASR Systems Development & Evaluation

1. **CGN**: **LM**: General conversations, **AM**: Adult speech, **Lexicon**: 255000 tokens
2. **HoMedV1**: **LM**: CGN+Lists of medical terms, **AM**: CGN, **Lexicon**: CGN+13934 tokens
3. **HoMedV2**: **LM**: CGN+Medicijnjournaal, **AM**: CGN, **Lexicon**: CGN+5342 tokens


Licensed lists of medical terms

Lexicon, LM

Kaldi_NL ASR system (CGN)

Lexicon, LM, AM

Medicijnjournaal material

Lexicon, LM

HoMedV1 ASR system

Lexicon, LM, AM

HoMedV2 ASR system

Lexicon, LM, AM

Word Error Rate
Lexical Coverage
Error analysis

Test material

Audio files, Transcription files

Test material: 360 minutes of speech recordings, 47912 transcribed words, 35 transcribed bulletins.

WER & Error Analysis (Categories)

ASR system	WER	Lexical coverage
Kaldi_NL	25.8	94.9
HoMedV1	24.7	96.1
HoMedV2	20.6	97.2

Table: Comparison of the ASR systems performance

Type of error	Kaldi_NL	HoMedV2
1. Spelling variant	457	798
2. Compound word	158	19
3. Morphological variant	21	31
4. Error within lexicon	598	872
5. OOV	286	78

Table: Categorization of main ASR output confusions

Future Work

Netherlands Institute for Health Services Research (**Nivel**)

Source: https://shorturl.at/kqvyF