

# The JASMIN-CGN Corpus: Recording Speech of Children, Non- Natives and Elderly People for HLT Applications

Catia Cucchiarini, Joris Driesen  
Hugo Van hamme and Eric Sanders

# Outline

- aim of the project
- partners
- motivation
- speakers
- speech material
- annotations
- lessons learned

# Aim of the project

Extending the Spoken Dutch Corpus (CGN)  
along three dimensions:

1. age → speech of children and elderly people
2. mother tongue → speech of non-natives
3. modality → speech in human-machine interaction

# JASMIN-CGN: partners

- Radboud Universiteit Nijmegen (CLST)
- Katholieke Universiteit Leuven (ESAT)
- TalkingHome

## **Radboud Universiteit Nijmegen**

1. Catia Cucchiarini
2. Andrea Diersen
3. Olga van Herwijnen
4. Leontine Aul
5. Eric Sanders

## **Katholieke Universiteit Leuven**

6. Hugo Van hamme
7. Maarten Van Segbroeck
8. Alain Sips
9. August Oostens
10. Joris Driesen

## **TalkingHome**

11. Felix Smits
12. Barry van der Veen
13. Erik Stegeman
14. Chantal Mülders
15. Koen Snijders

# Motivation

- ASR applications require dedicated corpora.
- Speech in human-machine interaction is characterized by phenomena that still pose serious challenges to state-of-the-art speech recognizers (disfluencies, hesitations, hyper-articulation etc.)

# Speakers

- native children aged 7 - 11
- native children aged 12 – 16
- non-native children aged 7 - 16
- non-native adults
- native adults older than 65

# Selecting variables

- region of origin (NL or FL)
- nativeness
- dialect region
- gender
- age
- proficiency in Dutch

# Speech material

- 12 minutes of speech per speaker
- about 50% of the material is read speech
- about 50% extemporaneous speech recorded in the human-machine interaction modality

# Read speech

- phonetically rich sentences
- texts of different reading levels

# Human-machine interaction

TalkingHome, a company involved in the project, developed:

- a WoZ-based platform for recording speech in the human-machine interaction mode.
- a dialogue development tool (XML) for designing the dialogues.

# Human-machine dialogues

different dialogues were developed for children, non-natives and elderly people.

same dialogue steps, but on a different (more suitable) topic.

# Human-machine dialogues

We specifically aimed at non-perfect dialogues, since we wanted to elicit data containing typical HMI phenomena:

- hesitations
- syllable lengthening
- loud speech
- accent shift
- restart
- filled pause
- self talk
- repetition
- paraphrasing
- hyper-articulation

# Human-machine dialogues

Various states of mind in the speaker will cause different HMI phenomena:

**Confusion:**

speakers start talking to themselves/machine

**Uncertainty:**

longer pauses, filled pauses and repetitions in order to gain more time

**Frustration:**

yelling, hyper-articulation, paraphrasing, accent shift, cursing the machine

# Human-machine dialogues

We induced these states of mind by:

- asking unexpected or ambiguous questions
- providing insufficient information on what is coming
- asking questions with higher cognitive load
- refusing to understand the speaker

# IPR issues

- permission from publishers
- permission from speakers (or parents)

# Collecting speech material

Speaker recruiting:

- more time-consuming than envisaged
- problematic in schools
- especially for non-native speakers

# Speech material collected

- native children (7– 1): NL: 15h 10m; FL: 7h 50m
- native children (12–16): NL: 10h 59m; FL: 8h 01m
- non-native children (7–14): 12h 34m; FL: 9h 15m
- non-native adults: NL:15h 01m; FL: 8h 02m
- native adults (> 65): NL: 16h 22m; FL: 8h 26m

# Annotations

- orthographic transcription
- annotation of HMI phenomena
- automatic part-of-speech tagging
- automatic phonemic transcription

# Validation and dissemination

The JASMIN-CGN corpus

- is now being validated by BAS
- will be made available through the HLT-Agency for research and commercial purposes.

Further information:

<http://homes.esat.kuleuven.be/~spch/projects/JASMIN/>

# Lessons learned and recommendations for the future

- how to implement budget cut
- mortality of personnel
- speaker recruiting

# Acknowledgments

The JASMIN-CGN project was carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (<http://taalunieversum.org/taal/technologie/stevin/>).