UNIVERSITY OF GOTHENBURG
DEPARTMENT OF SWEDISH

SweCLARIN – Infrastructure for Processing Transcribed Speech

Dimitrios Kokkinakis, Kristina Lundholm Fors and Charalambos Themistokleous

Department of Swedish, University of Gothenburg, Sweden

Introduction

- We present spoken language resources, including transcriptions, under development within the project *Linguistic and extra-linguistic parameters for early detection of cognitive impairment*.

- The focus is on the resources that are being produced and the way in which these could be used to pursue research in dementia prediction, an area in which more scientific investigations are required in order to raise the predictive value and improve early diagnosis and therapy.

Procedure
Recording, Transcription, Segmentation and Population

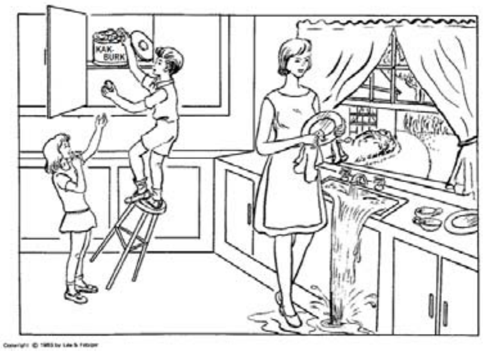


Figure 1. The Cookie Theft picture from the *Boston Diagnostic Aphasia Exam*.

	HC (n=36)	SCI (n=23)	MCI (n=31)
Age (years)	67.9 (7.2)	66.3 (6.9)	70.1 (5.6)
Education (years)	13.2 (3.4)	16.1 (2.1)	14.1 (3.6)
Sex (F/M)	23/13	14/9	16/15
MMSE (/30)	29.6 (0.61)	29.5 (0.90)	28.2 (1.43)

Table 1. Demographic data for the project's cohort. **HC**: Healthy Controls; **MCI**: Mild Cognitive Impairment; **SCI**: Subjective Cognitive Impairment; **MMSE** (MiniMental State Exam): a test of general cognitive ability; max is 30, a score of ≤24 is been proposed for cognitive impairment; a score between 25-27 indicates possible cogn. impairment which should be further evaluated.

Recording Characteristics, Tools and Linguistic Annotation

Recordings were made in two rounds: round 1 and round 2.

The first in 2016 and the second, same as the first plus 3 new tasks, in 2018, and the same population.

- a *picture description task,* the 'Cookie theft' (round 1&2; Fig. 1),
- a *read aloud task* (round 1&2)
- a *read silent task* (round 1&2)
- a *complex planning task* (round 2)
- a *map task* (round 2)
- a *semantic verbal fluency task*, category 'animals' (round 2)

The transcriptions during round-1 were made manually by professional transcribers; while for the round-2, automatically, with a speech-to-text system, "THEMIS-SV" (Fig. 2).

For the linguistic annotation we use *Sparv*, a major infrastructure for Swedish processing, part of the SweCLARIN. Sparv consists of several NLP tools e.g., tools for lexical and compound analysis, POS-tagging and comes with a web interface (Fig. 3).

Case Study

The transcriptions require some pre-processing before using automatic annotation tools such as Sparv. In written text, sentences are typically delimited by punctuation marks, whereas in spoken language, boundaries are indicated by for e.g. pauses and prosodic patterns.

The transcribers (phase-1) were therefore asked to identify appropriate segmentation points and to manually add a full-stop according to their own judgement (most NLP tools require sentence boundaries).

During transcription a number of other phenomena were also annotated, e.g. filler words (such as *uhm*), corrections and false starts (where speakers begin a sentence but change their plan of what they want to say and continue different).

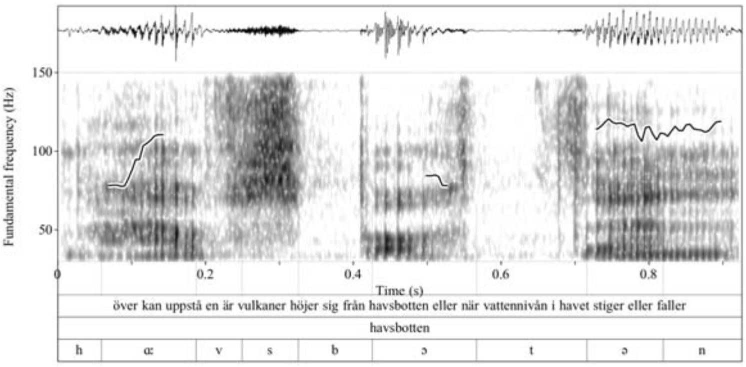
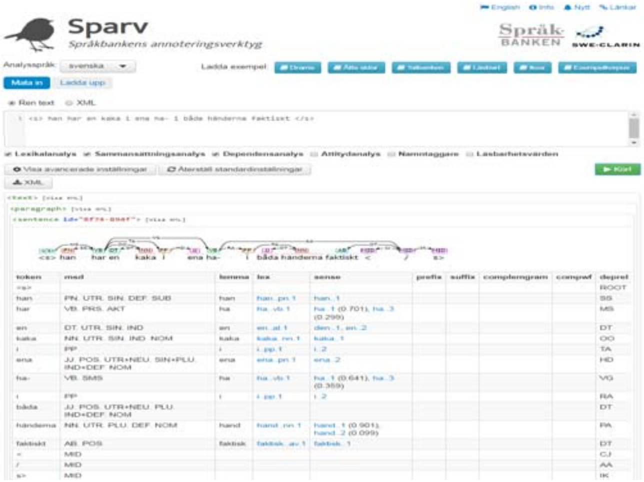


Figure 2. Automatically transcribed speech samples from Cookie Theft picture descriptions.



```
1 <s> han har en kaka i ena ha- i båda händerna faktiskt </s>
```

```
<text> [visa en]
<paragraph> [visa en]
<sentence id="8f74-894f"> [visa en]
```

<s> han har en kaka i ena ha- i båda händerna faktiskt < / s>

token	msd	lemma	lex	sense	prefix	suffix	complemgram	compwf	deprel
<s>									ROOT
han	PN.UTR.SIN.DEF.SUB	han	han..pn.1	han..1					SS
har	VB.PRS.AKT	ha	ha..vb.1	ha..1 (0.701), ha..3 (0.299)					MS
en	DT.UTR.SIN.IND	en	en..al.1	den..1, en..2					DT
kaka	NN.UTR.SIN.IND.NOM	kaka	kaka..nn.1	kaka..1					OO
i	PP	i	i..pp.1	i..2					TA
ena	JJ.POS.UTR+NEU.SIN+PLU.IND+DEF.NOM	ena	ena..pn.1	ena..2					HD
ha-	VB.SMS	ha	ha..vb.1	ha..1 (0.641), ha..3 (0.359)					VG
i	PP	i	i..pp.1	i..2					RA
båda	JJ.POS.UTR+NEU.PLU.IND+DEF.NOM								DT
händerna	NN.UTR.PLU.DEF.NOM	hand	hand..nn.1	hand..1 (0.901), hand..2 (0.099)					PA
faktiskt	AB.POS	faktisk	faktisk..av.1	faktisk..1					DT
<	MID								CJ
/	MID								AA
s>	MID								IK

Figure 3. Processing results using Sparv and the analysis in a table format. Here the truncated token *ha-* is erroneously pos-annotated as VB ([compound] verb) instead of noun which also produces an erroneous dependency annotation.

Conclusions

- The corpus we have presented represents a valuable resource for early detection of a neurodegenerative diseases currently exploited for our research agenda.
- Spoken data become more and more accessible as automatic speech recognition mature, so needs for time-intensive manual transcription decrease.
- With such data, we see a need for language tools to be adapted to accommodate transcriptions of spoken language.

GÖTEBORGS UNIVERSITET

The CLARIN Annual Conference 2018
8-10 October 2018, Pisa, Italy