The IFADV corpus: A free dialog video corpus

Rob van Son, Wieneke Wesseling
Eric Sanders, Henk van den Heuvel

ACLC/IFA, University of Amsterdam
SPEX/CLST, Radboud University Nijmegen
The Netherlands

LREC 2008, Marrakech

Introduction

Looking for a Video corpus of informal conversational speech for *Perception* and *Reaction Time* Experiments on dialogs

Requirements

- To see the gaze and lips
- High quality for perception experiments
- Aligned transliteration, broad phonetic transcription, POS tags
- Both speakers synchronized
- Aligned transliteration
- Free access and use
- Dutch

What was available

Corpora containing (conversational) video recordings

In order of decreasing suitability

- Corpus d'interactions dilogales, fully annotated (CID, R. Bertrand)
- HCRC Map Task Corpus, fully annotated (task related conversations)
- Nothing on ELRA
- Nothing on LDC
- Audio Visual Speech Technologies, IBM Research (not conversational)
- UTDrive: The Smart Vehicle Project (Human Machine)

not available to the public

No Dutch accessible to us, and no corpus was freely available

Solution

Do-It-Yourself corpus building and sharing

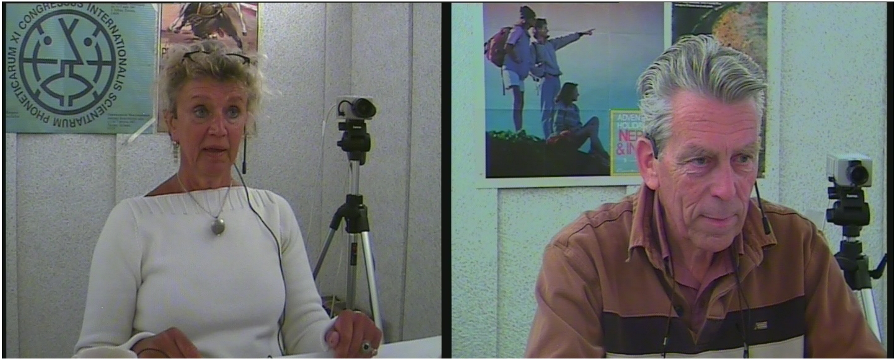
- Informal conversational speech
- Well acquainted speaker pairs → lively speech
- High quality recordings
- High visibility of eyes and lips
- Synchronized, frontal, face recordings → two cameras
- No copyright or privacy restrictions → *Free/Libre* license

Video recordings

Two color cameras facing the speakers

- Seated at a table facing each other
- Separation ~ 1 m, camera to their left
- Sound treated room
- Full face recordings down to the shoulders
- Gen-locked recordings → *synchronized* frames
- Head mounted microphones
- Free view of lips

Examples



Example frame of recordings

output camera A, left; output camera B right

▸ Recording room

Materials and Participants

In total 20 out of 24 dialogs of 900 s each have been annotated (5h)

- 24 female and 10 male speakers
- 12-72 years of age
- Relevant personal (meta-) information recorded
- Long time colleagues or friends
- Informal and unrestricted dialogs (lively speech)
- → warnings about privacy given beforehand
- 69 kWords, 13669 utterances, 5752 Turn Switches (simplified)

▸ Example summary

Annotations in the IFA DV corpus

Annotations (by SPEX) have been made by *Hand* and *Automatic*

Where possible, the annotations were made in a *Spoken Dutch Corpus* (*CGN*) format. Other annotations used new formats (*non-CGN*)

Annotation type	Performed by	Format	$\times RT$	Sec/Word
Orthographic translit.:	Hand, aligned	*CGN*	*30*	*8*
POS tagging:	Automatic	*CGN*		
Word alignment:	Automatic	*CGN*		
Word-to-Phoneme:	Automatic	*CGN*		
Phoneme alignment:	Automatic	*CGN*		
Conversational function:	Hand	*non-CGN*	*30*	*7*
Gaze direction:	Hand	*non-CGN*	*17*	*4.5*

Legalities

A corpus that can be freely *used*, *adapted*, and *distributed*

Copyrights:

- Transfered to central legal entity → *Dutch Language Union*
- Includes: Speakers, technicians, assistants, annotaters, and ourself
- Legal documents adapted from *Spoken Dutch Corpus*

Subjects:

- Consent for use of “personal” (meta-)data and portrait rights
- Speakers made aware that the recordings could be “broadcasted”
- Speakers read and signed an *Informed Consent* document
- Speakers could retract their consent *after* assessing the recordings

Free/Libre License: GNU GPL v2

▸ Skip

Database queries

Prime use would be statistical analysis of data (∼ 69 kW)

- All “text” data stored in database tables
- Each item in the annotations must be *uniquely* labeled ▸ Unique id’s
- Link items with their identification labels
- Join tables in database to determine relative position of items
- All data for this talk were obtained by way of SQL queries

(Mean, SD, SE, and N of the turn delays)

```
SELECT
    avg(delay) AS Mean,
    stddev(delay) AS StdDev,
    sqrt(variance(delay)/count(properturnswitch.id)) AS StdError,
    count(properturnswitch.id) AS Count
FROM
    properturnswitch JOIN fct USING (ID)
WHERE
    fct.value ∼ ’u’ AND fct.value ∼ ’a’;
```

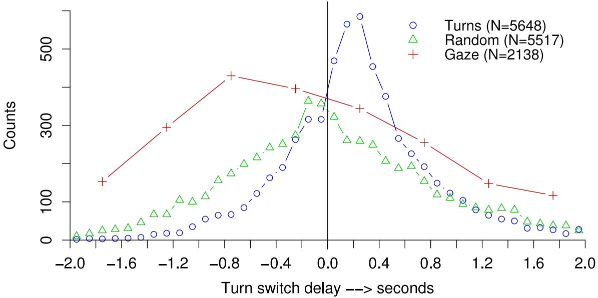
▸ Skip

Envisioned uses for the IFA Dialog Video corpus

Research in Human-Human/Machine dialogs

- Analysis of informal multimodal conversations
  - Speech in the presence of visual feed-back
  - Use and distribution of visual cues, eg, frowning, gaze, head posture
  - Expressive speech in "normal" conversations
- Conversation *shadowing* by experimental subjects
  - Cognitive processing → RT experiments
  - Use of available visual and audible cues
  - Eye tracking
- Automatic recognition of Audio-Visual speech
  - Eye tracking, lip and face "reading"
  - Emotion and attention recognition
  - Repair, hesitations, and other "problems"

Analysis: Turn switches and gazes



Turn switch delays and gaze direction

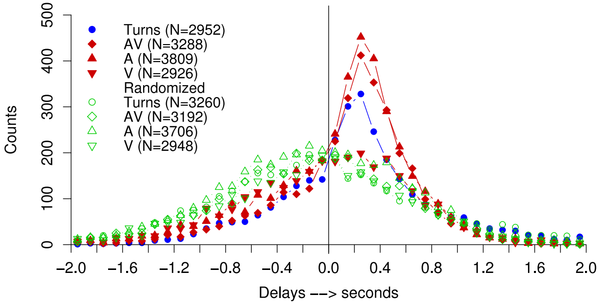
- Turn switch delays
- Randomized turn delays for statistics
- Gaze movement timings

Conversation shadowing: RT experiments

Present subjects with running Audio-Visual dialogs

- Record minimal responses
- Audio/Visual, Audio-only, Video-only
- Look at effects of gaze direction and dialog function
- (all figures created with *R* linked to the database)
- 14344 RT responses of 30 listeners analyzed

RT delays



Response time distribution: Effect of visual mode, eg, gaze

- Turn switches
- Reaction times relative to utterance end
- Randomized delays for statistics

▸ more

Discussion and $\overset{\text{our}}{\curlyvee}$ conclusions

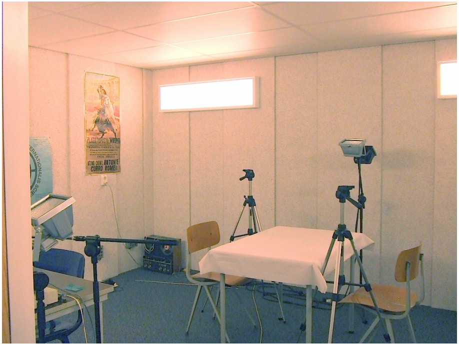
Useful, shareable, Free/Libre Audio Visual corpus

- More AV corpora needed
- Good quality video and audio is cheap
- Value and costs are in the annotations ($\sim 75 \times RT$)
- Free/Libre licenses can handle copyright and privacy laws
- Academic funding agencies like corpus sharing
- Database & internet access really unlock the data

▸ more advice

Thank You

Recording room



Recording room set-up

The distance between the speakers was around 1 m

Photograph courtesy of Jeannette M. van der Stelt

◂ Return

Example summary extract

Summary of a recording session. Female and Male subject

Summary *DVA6H+I*
Relation Speakers: *Colleagues*
List of Topics: *Leiden, Russian, Storage of documentation, Edison Klassiek, Crete, Greek, Restoration, Noord/Zuidlijn, Sailing*

Summary: *2 Speakers (F59H and M65I)*

. . .

Then they discuss the chaos on Amsterdam Central. A tunnel for a new metro line, the 'Noord/Zuidlijn', is built there. F59H says to M65I that he doesn't have to take a train anymore. He says that he will take the train to Amsterdam every now and then. M65I is going sailing soon. He describes the route that they are going to take.

◂ Return

Functional annotation

Simplified Conversational Function

Description and distribution of utterances over conversational function. All labels are interpreted wrt the previous utterance(s) *(N=13,669)*

Label	n	Description
b:	735	Start of a new topic
c:	8739	Continuing topic (e.g., follows b, or c)
h:	240	Repetition of content
r:	853	Reaction (to u)
f:	213	Grounding acts or formulaic expressions
k:	2425	Minimal response
i:	27	Interjections
m:	61	Meta remarks
o:	138	Interruptions
x:	27	Cannot be labeled
a*:	1374	Hesitations at the end of the utterance
u*:	1028	Questions and other attempts to get a reaction

* Labels *u* and *a* can be added to other labels.

◂ Return

License

Widest possible use: GNU GPL v2

- Free and Open Source license
- Written statements from grant agency, employers, and *Dutch Language Union*
- All materials can be obtained from the *Dutch HLT Agency* or http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/
- Raw and processed video recordings, audio, annotations, tables
- Distributed changes must themselves be licensed under the GNU GPL
- Alternative license would have been the *European Union Public Licence*, EUPL v.1.0

◂ Return

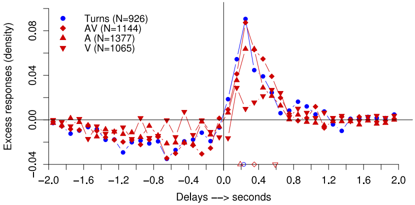
Example ID coding

Item	ID code	Description
phoneme	*DVA6F59H2C1SK1*	First vowel
syllable part	*DVA6F59H2C1SK*	Kernel
syllable	*DVA6F59H2C1S*	First syllable
word	*DVA6F59H2C1*	First word
chunk	*DVA6F59H2C*	Third chunk
Tier name	*DVA6F59H2*	-
Recording	*DVA6F59H2*	(this subject's)
Speaker	*DVA6F59H*	Female H
Session	*DVA6*	Recording session 6
Camera	*DVA*	Left subject
Annotation	*DV*	Dialog Video Audio

◂ Return

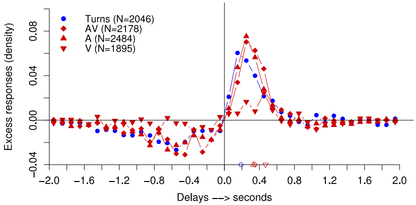
RT delays with *Random responses* subtracted

Gaze



Speaker starts gazing at listener

Nogaze



Speaker does *not* start gazing

Preliminary results: Visual cues suppress “noise” in responses, gaze speeds up responses a little

- Turn switches
- Reaction times relative to utterance end

◂ Return

: Summary of advice

Don't create a new corpus unless it is absolutely unavoidable

- Minimize overhead with appropriate license
- Organize *Copyright Transfers* before you start
- Use *Informed Consent* forms
- Get *everyone* to sign the papers
- Outsource annotations ($\sim 75 \times RT$)
- Go for best quality video and *never* compress
- Ensure visibility of eyes and lips (microphone placement!)
- Hands and eyes difficult to combine
- Plan for database access
- Plan for Internet distribution

◂ Return