The IFADV corpus: A free dialog video corpus

R.J.J.H. van Son[1], Wieneke Wesseling[1], Eric Sanders[2], Henk van den Heuvel[2]

(1) ACLC/IFA, University of Amsterdam (2) SPEX/CLST, Radboud University Nijmegen
The Netherlands

Abstract

Research into spoken language has become more visual over the years. Both fundamental and applied research have progressively included gestures, gaze, and facial expression. Corpora of multi-modal conversational speech are rare and frequently difficult to use due to privacy and copyright restrictions. A freely available annotated corpus is presented, gratis and libre, of high quality video recordings of face-to-face conversational speech. Within the bounds of the law, everything has been done to remove copyright and use restrictions. Annotations have been processed to RDBMS tables that allow SQL queries and direct connections to statistical software. From our experiences we would like to advocate the formulation of "best practises" for both legal handling and database storage of recordings and annotations.

1. Introduction

Fundamental and applied research have progressively included visual aspects of speech. Gestures, gaze, and facial expression have become important for understanding human communication. Such research requires corpora of multi-modal conversational speech. But such corpora are rare and frequently difficult to use due to privacy and copyright restrictions.

In the context of a research project into spoken language understanding in conversations, a corpus of visible speech was needed. Reaction time experiments were planned where experimental subjects watch and listen to manipulated recordings and react with minimal responses. For these experiments video recordings of informal conversations were needed. Neither ELRA (2004 2007) nor the LDC (1992 2007) had any conversational video material available. The corresponding entity in the Netherlands, the Dutch TST centrale (HLT-Agency, 2007), also had no conversational video corpus available. Nor were we able to obtain another video corpus.

In the world, several corpora exist that contain annotated video recordings of conversational speech. For instance, the HCRC Map Task Corpus (MAPtask, 1992 2007) does contain video recordings, but, according to their web-site, these have not been made generally available due to privacy concerns. Also, the French Corpus of Interactional Data, CID (Blache et al., 2007; Bertrand, 2007), is an annotated audio-video recording of conversational speech which seems to be available to other researchers, although their web-site does not give details about the conditions under which it is distributed.

Within our project, we have created a visual version of the friendly Face-to-Face dialogs of the Spoken Dutch Corpus (CGN, 2006). Within the bounds of our budget, the procedures and design of the corpus were adapted to make this corpus useful for other researchers of Dutch speech. For this corpus we recorded and annotated 20 dialog conversations of 15 minutes, in total 5 hours of speech. To stay close to the very useful Face-to-Face dialogs in the CGN, we selected pairs of well acquainted participants, either good friends, relatives, or long-time colleagues. The participants were allowed to talk about any topic they wanted.

In total, 20 out of 24 initial recordings were annotated to the same, or updated, standards as the original CGN. However, only the initial orthographic transcription was done by hand. Other CGN-format annotations were only done automatically (see below). As an extension, we added two other manual annotations, a functional annotation of dialog utterances and annotated gaze direction.

2. Recordings

For the recordings, the speakers sat face-to-face opposite of each other in a sound-treated room with a table in between (see Figure 1). The distance between the speakers was about 1m. Recordings were made with two gen-locked JVC TK-C1480B analog color video cameras (see table 1). Each camera was positioned to the left of one speaker and focused on the face of the other (see Figure 3). Participants first spoke some scripted sentences. Then they were instructed to speak freely while preferably avoiding sensitive material or identifying people by name.

Gen-lock ensures synchronization of all frames of the two cameras to within a half (interleaved) frame, i.e., 20 ms. Recordings were stored unprocessed on disk, i.e., in DV format with 48 kHz 16 bit PCM sound.

Recording the videos of the dialogs introduced some limitations to our participants. For technical reasons, all recordings had to be done in our studio, instead of in the participant's home, as was done for the CGN Face-to-Face recordings. The position of the cameras, as much as possible directly in front of the participants, did induce a static set-up with both participants sitting face-to-face at a table.

Figure 3 gives an example frame of each of the two cameras. Notice the position of the camera focussed on the other subject. The position of the head-mounted microphone was such that it would not obstruct the view of the lips. The posters on the back-ground were intended to suggest conversation topics when needed. In practise, subjects hardly ever needed any help in finding topics for conversation. They generally started before we were ready to record, and even tended to continue after we informed them that the session was over.

Figure 1: Recording room set-up. The distance between the speakers was around 1 m. Photograph courtesy of Jeannette M. van der Stelt.

The result of these procedures was that the conversations are probably as free-form as can be obtained in a studio setting. The quality of the sound and video is high and even the gaze direction can easily be identified. This makes this corpus useful for many types of research, from classical conversation analysis to automatically detecting gaze direction and emotion in facial expressions.

3. Materials

Annotated recordings are limited to 900 seconds (15 min). Each recorded DV file is around 4 GB in size. The diaphragm of the B camera overcompensated the lighting and most of the B recordings are, therefore, rather dark. However, there is enough range in the brightness left to compensate for this. Dropped frames during recording offset the synchrony of the two recordings, and all occurrences of frame drops have therefore been identified. For each recording, a SMIL (2008) file is available that specifies how the original frame timing can be restored by repeating frames to replace dropped frames.
For demonstration purposes, a set of MPEG 4 compressed and cropped movies with correct frame timing has been constructed from these SMIL files. These demonstration files are smaller, around 283 M byte for MP3 audio compression, and have also been equalized on brightness. That

Table 1: Recording equipment, two gen-locked JVC TK-C1480B analog color video cameras with following specifications and peripherals

Image pickup :	1/2 type IT CCD 752 (H) x 582 (V)
Synchronization :	Internal Line Lock, Full Genlock
Scanning freq. :	(H) 15.625kHz x (V) 50Hz
Resolution :	480 TV lines (H)
Screen size :	720x576 BGR 24-bit, 25 frames/s
Camera A :	Ernitec GA4V10NA-1/2 lens (4-10mm)
Camera B :	Panasonic WV-LZ80/2 lens (6-12mm)
AD conversion :	2 Canopus ADVC110 digital video conv.
Microphones :	Samson QV head-set microphones

```
F59H: heel melancholieke sfeer.
M65I: hoe was 't uh met de muziek op Kreta?
F59H: nou uh we zaten dit keer in 'n uh we
      hebben een huis gehuurd 'n
      traditioneel uh boerenhuis een stenen huis.
      en dat was een uh
M65I: wat je kende of niet zomaar uh?
F59H: nou we hebben 't van het internet
      geplukt en toen 'n beetje
      gecorrespondeerd met de eigenaar en
      dat leek ons wel wat.
      ja 't blijft natuurlijk altijd een gok.
      maar dat bleek dus heel erg leuk te zijn.
      in 'n heel klein boerendorpje*n
      helemaal noordwest uh Kreta.
```

Figure 2: Example transcription of recordings, formatted for readability (originals are in Praat textgrid format). Every utterance ends in a punctuation mark. M65I: Male subject, F59H: Female subject

is, the video frames and audio files of both recordings are synchronized and the brightness of both recordings is dynamically standardized.

4. Participants

The corpus consists of 20 annotated dialogs (selected from 24 recordings). All participants signed an informed consent and transferred all copyrights to the Dutch Language Union (Nederlandse Taalunie). For two minors, the parents too signed the forms. In total 34 speakers participated in the annotated recordings: 10 male and 24 female. Age ranged from 21 to 72 for males and 12 to 62 for females. All were native speakers of Dutch. Participants originated in different parts of the Netherlands. Each speaker completed a form with personal characteristics. Notably, age, place of birth, and the places of primary and secondary education were all recorded. In addition, the education of the parents and data on height and weight, were recorded, as well as some data on training or experiences in relevant speech related fields, like speech therapy, acting, and call-center work.
The recordings were made in-face with only a small offset (see Figure 3). Video recordings were synchronized to make uniform timing measurements possible. All conversations were "informal" since participants were friends or colleagues. There were no constraints on subject matter, style, or other aspects. However, participants were reminded before the recordings started that their speech would be published.

5. Annotations

20 conversations have been annotated according to the formalism of the Spoken Dutch Corpus (CGN, 2006) by SPEX in Nijmegen. A full list of the annotations can be found in table 2. The computer applications used for the automatic annotations were different from those used by the CGN, but the file format and labels were kept compatible with those in the CGN. The orthographic transliteration and rough time alignment of 5 hours of dialogs took approximately 150 hours (30 times real time).
The annotations are either in the same formats used by the CGN (2006) or in newly defined formats (*non-CGN*) for

Figure 3: Example frame of recordings (output camera A, left; output camera B right)

annotations not present in the CGN (table 2). As gaze direction, the timing of looking towards and away from the other participant has been segmented in ELAN (2002 2007). Other annotation files use Praat TextGrid format (Boersma and Weenink, 1992 2008).

The functional annotation was restricted to keep the costs within budget. A HRC style hierarchical speech or conversational acts annotation (Carletta et al., 1997; Core and Allen, 1997) was not intended. The idea behind the annotation was to stay close to the information content of the conversation. How does the content fit into the current topic and how does it function? The label set is described in table 3. The hand annotation of the chunk functions in context took around 140 hours (∼30 times real time).

Each utterance was labeled with respect to the previous utterance, irrespective of the speaker. Some labels can be combined with other labels, e.g., almost every type of utterance can end in a question or hesitation, i.e., *u* or *a*. Note that a speaker can answer (*r*) her own question (*u*). Labeling was done by naive subjects who were instructed about the labeling procedure. We are well aware that this annotation is impressionistic.

Gaze direction was annotated with ELAN (2002 2007). The categories were basically *g* for gazing at the partner and *x* for looking away. For some subjects, special labels were used in addition to specify consistent idiosyncratic behavior, ie, *d* for looking down and *k* for blinking. The start and end of all occurrences where one subject gazed towards their partner were indicated. This hand labelling took around 85 hours for 5 hours of recordings (two speakers, 17 times real time).

An identification code (ID) has been added to all linguistic entities in the corpus according to (Mengel and Heid, 1999; Cassidy, 1999; Van Son et al., 2001; Van Son and Pols, 2001). All entities referring to the same stretch of speech receive an identical and unique ID. See table 4 for an example[1]. Although the ID codes only have to be unique, they have been built by extending the ID of the parent item. That is, an individual phoneme ID can be traced back to the exact position in the recording session it has been uttered in. The gaze direction annotations run "parallel" to the speech and have been given ID's that start with *GD* (Gaze Direction) instead of *DV* (Dialog Video). In all other respects they are treated identical to speech annotations.

Table 2: Annotations in the IFA DV corpus. Annotations have been made by *Hand* and *Automatic*. Where possible, the annotations were made in a *CGN* format. Annotations *not* in the CGN used new formats

Orthographic transliteration:	Hand *CGN* chunk aligned
POS tagging:	Automatic, *CGN*
Word alignment:	Automatic, *CGN*
Word-to-Phoneme:	Automatic, *CGN*
Phoneme alignment:	Automatic, *CGN*
Conversational function:	Hand, *non-CGN*
Gaze direction:	Hand, *ELAN, non-CGN*

Table 3: Conversational function annotation labels. Both *u* and *a* can follow other labels

Label	Description
b:	Start of a new subject
c:	Continuing subject (e.g., follows b, or c)
h:	Repetition of content
r:	Reaction (to u)
f:	Grounding acts or formulaic expressions
k:	Minimal response
i:	Interjections
m:	Meta remarks
o:	Interruptions
x:	Cannot be labeled
a:	Hesitations at the end of the utterance
u:	Questions and other attempts to get a reaction

[1] Syllables are counted *S*, *T*, *U*, ... and divided into *O*nset, *K*ernel, and *C*oda using a maximum onset rule. So the ID of the first (and only) phoneme of the kernel of the first syllable in a word ends in *SK1*

```
SELECT
    avg(delay) AS Mean,
    stddev(delay) AS SD,
    sqrt(variance(delay)
       /count(properturnswitch.id)) AS SE,
    count(properturnswitch.id) AS Count
FROM
    properturnswitch
    JOIN
    fct
    USING (ID)
WHERE
    fct.value ~ 'u' AND fct.value ~ 'a';
```

Figure 4: Example SQL query. This query generates the results displayed in the *ua* row of table 7. *properturnswitch*: table with the chunk ID's and the turn switch delays; *fct*: table with the functional labeling

These codes are necessary to build RDBMS tables for database access (Mengel and Heid, 1999; Cassidy, 1999; Van Son et al., 2001; Van Son and Pols, 2001). Such tables are available for all annotations as tab-delimited lists. The RDBMS tables are optimized for PostgreSQL, but should be easy to use in other databases. Through the unique ID, it is possible to join different tables and perform statistics directly on the database (see Figure 4). For example, statistical scripts from *R* can connect directly to the database (R Core Team, 1998 2008). All numerical data in this paper have been calculated with simple SQL database queries and demonstrate their usefulness.

Transcripts are available in standard text form for easier reading (see Figure 2). Summaries were compiled from these transcripts (see Figure 5).

Meta data for all recordings are available. What is currently lacking are standard meta data records, ie, IMDI, and accessible documentation of the recordings. We propose to produce the IMDI (Isle Meta data Initiative) records and the documentation with the help of student assistants. We have applied for funding to convert the meta-data into IMDI (1999 2007) format.

Table 4: Example encoding scheme for item ID. The /e/ from the first word /neː/ (*no*) of the utterance "nee dat was in Leiden." (*no, that was in Leiden*) uttered by the left subject in the sixth session as her third chunk is encoded as:

Item	ID code	Description
phoneme	*DVA6F59H2C1SK1*	First vowel
syllable part	*DVA6F59H2C1SK*	Kernel
syllable	*DVA6F59H2C1S*	First syllable [1]
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

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 doesnt 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.

Figure 5: Example extract from a summary of a recording session. Female and Male subject

6. Copyright and privacy concerns

One of the aims of our corpus effort was to create a resource that could be *used*, *adapted*, and *distributed* freely by all. This aim looks deceptively simple. It is, however, fraught with legal obstacles. The law gives those who perform, create, or alter what is now often called *intellectual content* broad control over precisely *use*, *adaptation*, and *distribution* of the products of their works. In legal terms, "intellectual content" is described in the Berne Convention as (WIPO, 1979):

> ...every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression, ...

With the added requirement that it is "fixed in some material form" (WIPO, 1979). In practise, this can often be interpreted as anything that can be reproduced and is not automatically generated. It does not help that the relevant laws differ between countries. In addition, there are also *performance* and *editorial* rights for those who act out or process the production (WIPO, 2004) as well as *database* rights (Maurer et al., 2001; Kienle et al., 2004; EC, 2005). When creating corpora, these additional rights can be treated like copyrights. Most countries also allow individuals additional control over materials related to their privacy.

On the surface, the above problems could be solved easily. It only requires that all the subjects and everyone else involved in the creation and handling of the corpus, agree to the fact that the corpus should be free to be used and distributed by anyone. The copyright and privacy laws allow such an arrangement, provided that these agreements are put in writing and signed by everyone involved. And it must be clear that everybody, especially naive subjects, actually understood what they agreed to. Therefore, the problem shifts to what the written and signed agreements must contain to legally allow free *use*, *adaptation*, and *distribution* by all, and who must sign them.

In recent years, the interpretations of copyright and privacy laws have become very restrictive. The result is that the required written agreements, ie, copyright transfers and informed consents, have become longer and more complex

and have involved more people. There are countless examples of (unexpected) restrictions attached onto corpora and recordings due to inappropriate, restrictive, or even missing copyright transfer agreements or informed consent signatures. Experience has shown that trying to amend missing signatures is fraught with problems.

The solution to these problems has been to make clear, upfront, to subjects how the recordings and the personal data might be used. In practise, this has meant that the different options, eg, publishing recordings and meta data on the internet, have to be written explicitly into the copyright transfer forms. A good guide seems to be that corpus creators are specific about the intended uses whenever possible. At the same time, an effort should be made to be inclusive and prepare for potential, future, uses by yourself and others. All the "legal" information has to be made available also in laymans terms in an informed consent declaration. Obviously, subjects should have ample opportunity to ask questions about the procedures and use of the recordings.

For logistic reasons, signatures are generally needed before the recordings start. However, the courts might very well find that subjects cannot judge the consequences of their consent before they know what will actually be distributed afterwards. For that reason, subjects should have an opportunity to retract their consent after they know what is actually recorded and published.

As to who must all sign a copyright transfer agreement, it is instructive to look at movie credits listings. Although not authorative, the categories of contributers in these credits listings can be used as a first draft of who to include in any copyright transfer agreement. It might often be a good idea to include more people, but it is better to consult a legal expert before excluding possible contributers.

The requirements of privacy laws are different from those of copyrights. It is both polite and good practise to try to protect the anonymity of the subjects. However, this is obviously not possible for video recordings, as the subjects can easily be recognized. In general, this fact will be made clear to the subjects before the recordings start. In our practise we pointed out to the subjects that it might be possible that someone uses the recording in a television or radio broadcast. A more modern example would be posting of the recordings on YouTube. If the subjects can agree with that, it can be assumed that they have no strongly felt privacy concerns.

All our participants were asked to sign copyright transfer forms that allow the use of the recordings in a very broad range of activities, including unlimited distribution over the Internet. This also included the use of relevant personal information (however, excluding any use of participant's name or contact information). Participants read and accorded informed consent forms that explained these possible uses to them. To ensure that participants were able to judge the recordings on their appropriateness, they were given a DVD with the recordings afterwards and allowed ample time to retract their consent.

7. License

To be able to use or distribute copyrighted materials in any way or form, users must have a license from the copyright holder. Our aim of giving *free* (as in *libre*) access to the corpus is best served by using a Free or Open Source license (Ken Coar, 2006). We chose the GNU General Public License, GPLv 2 (FSF, 1991), as it has shown to protect the continuity and integrity of the licensed works. It has also shown to be an efficient means to promote use by a wide audience with the least administrative overhead. This license ensures the least restrictions and simplifies the continued build up of annotations and corrections.

In almost all respects, the GPLv2 is equivalent to, and compatible with, the European Union Public Licence, EUPL v.1.0 (IDABC , 2008). However, the GPLv2 is only available in English, while the EUPLv1 is available in all official EU languages where versions have the (exact) same legal meaning. So, future corpus building efforts in Europe might consider the EUPL for their license.

According to an agreement with the funding agency, the Netherlands Organization for Scientific Research (NWO), all copyrights were directly transfered to the Dutch Language Union (NTU). The Dutch Language Union distributes the corpus and all related materials under the GNU General Public License (FSF, 1991).

The GPLv2 allows unlimited use and distribution of the licensed materials. There is however a condition to (re-) distributing adapted or changed versions of the "works". Whenever such changes fall under copyright laws, ie, when they create a *derivative work* in the sense of of the law, they *must* be distributed under the same license, ie, the GPLv2. And that license requires the release of the "source" behind the works.

This condition raises the question of what the source of a corpus recording or annotation is. The short answer is, everything needed to reproduce the changes in whatever format is customary for making changes. Examples would be Praat TextGrid or ELAN EAF files. A long answer would include audio, video, and document formats and associated codecs. Basically, if the receiver has more problems making changes than the originator, there is reason to add additional sources.

Table 5: Distribution of utterances over conversational function. Labels *u* and *a* can be added to other labels and are counted separately (n =13,669). 52 Chunks did not receive a label when they should have.

Label	count	*description*
b	735	*begin*
c	8739	*continuation*
h	240	*repetition*
r	853	*reaction*
f	213	*functional*
k	2425	*minimal response*
i	27	*interjection*
m	61	*meta*
o	138	*interruption*
x	27	*unknown*
-	52	*unlabeled*
a	1374	*hesitation*
u	1028	*question etc*

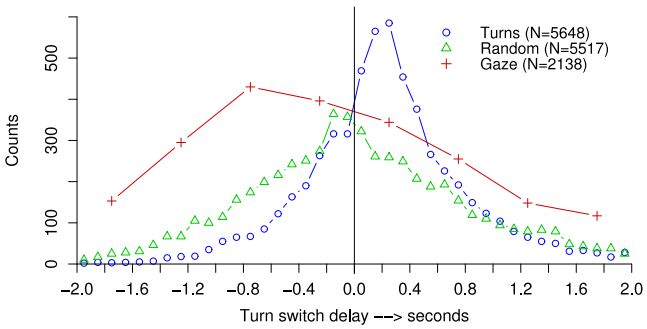

Figure 6: Distribution of turn switch delays (PSTS), circles, randomized turn switches, triangles, and gaze delays from the last speaker, plusses (see text).
Bin sizes: turn switch delays, 100ms; gaze delays, 500ms

8. Distribution

The corpus is currently freely available from the TST-centrale (HLT-Agency, 2007). This includes raw and processed video recordings, audio, and all annotations. In addition, there are derived annotation files available that combine different annotations. Summaries have been made for all annotated dialogs. IMDI metadata records are in preparation.

Relational database tables have been extracted from the annotations and stored in tab-delimited lists. These and all the scripts needed to process the annotations and tables are also available at the TST-centrale. All materials are copyrighted by the Dutch Language Union (Nederlandse Taalunie) and licensed under the GNU GPL (FSF, 1991). All materials are available free of charge. Pre-release development versions of all materials are available from the University of Amsterdam at URL http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/.

9. Results

In total, 13,373 verbal utterances with 69,187 words were recorded (excluding non-verbal noises). 589 words were transcribed as incomplete ('*a' in CGN). The original orthographic transliteration chunks were combined with the automatic word alignments to create word aligned chunks. Simplified *Proper Speaker Turn Switches* (PSTS) were defined as chunks where the next speaker started a verbal chunk *after* the start of the last verbal chunk of the previous speaker that continued beyond the end of that last chunk. Non-verbal noises were ignored.

Such PSTS events can be determined easily by sorting verbal chunks on their end time while requiring that the starting time of the new chunk is later than that of the current chunk. An important aspect of such PSTS events is the time delay between the two speakers. The distribution of the PSTS delay is given in figure 6 (circles). The modal turn switch delay time is visible around 300 ms. The distribution is broad and falls to half its height at delays of 0 and 500 ms.

The durations of utterances varies in intricate ways, as do pause durations. As a result, the statistics of the PSTS time delays are not straightforeward. For comparison, pseudo PSTS delays are calculated by cyclical shifting the annotations for one speaker by 100 seconds. This time shift should "randomize" turn-switch delays while keeping the duration and pause statistics intact. The resulting distribution shows a clear maximum close to a delay of 0s (triangles in figure 6). The differences between real and random PSTS delays are obvious, but the statistics might not be straightforeward.

The gaze direction annotation is combined with the speech annotation by linking every gaze event, starting to look towards or away from the dialog partner, to word annotations. For each start and end of a gaze label, the corresponding automatically aligned words or pauses are located that were annotated for the *same* (looking) and the *other* subject. The average delay between the speaker looking towards the partner and the end of the current turn of the speaker is presented in figure 6 (plusses). There were 5168 occurrences in total where one subject looked directly at the other.

Most of the annotations used in this corpus were taken from the CGN, and are well understood. Gaze direction is straightforeward and we do not expect problems with its interpretation. However, the functional annotation of the

Table 6: Distribution of utterance duration in seconds over the most important conversational function. Labels *u* and *a* can be added to other labels and are counted separately. Mean: mean delay; SD: Standard Deviation; SE: Standard Error; #: Number of occurrences; all: all functional labels

Label	Mean	SD	SE	#
b	1.535	0.648	0.024	735
c	1.367	0.667	0.007	8739
h	0.773	0.531	0.034	240
k	0.312	0.288	0.006	2425
r	0.937	0.687	0.024	853
f	0.539	0.318	0.022	213
a	1.194	0.667	0.018	1374
u	1.189	0.668	0.021	1002
ua	1.747	0.679	0.133	26
All	1.119	0.739	0.006	13669

Table 7: Distribution of Proper Speaker Turn Switch (PSTS) delays in seconds over the most frequent conversational functions. Labels *u* and *a* can be added to other labels and are counted separately. Mean: mean delay; SD: Standard Deviation; SE: Standard Error; #: Number of occurrences; all: all function labels

Label	Mean	SD	SE	#
b	0.425	0.633	0.039	262
c	0.233	0.670	0.011	3682
h	0.122	0.564	0.051	121
k	0.307	0.507	0.016	1009
r	0.251	0.644	0.032	409
f	0.271	0.713	0.075	90
a	0.167	0.754	0.038	388
u	0.278	0.613	0.023	733
ua	0.053	0.574	0.117	24
all	0.256	0.643	0.008	5752

dialog chunks was newly developed for this corpus. Therefore, the categories used have not yet been validated. The aim of this annotation was to add a simple judgment on the discourse function of individual chunks (utterances). We will try to find internal support in other annotations for the relevance of this functional labeling for the behavior of conversational participants.

The distribution of conversational function over utterances is given in table 5. Around 18% of all utterances are classified as minimal responses. A lot of non-verbal sounds (transcription: *ggg*) were labeled as minimal responses.

As expected, utterance duration depends on the functional label, as is visible in table 6. The most marked effect is expected between utterances adding content to the discourse, ie, *b*, *c*, and *h* (*begin*, *continuation*, and *repetition*).

These type labels are intended to describe those utterances that contribute directly to the subject matter of the discourse. Their difference lies in their relative positions with respect to content matter. *b* Indicates the introduction of a new topic at any level of the discourse. *c* Signifies utterances that contribute to an existing topic. *h* Labels utterances that mainly, word-by-word, repeat a message that has already been uttered before.

Obviously, it is expected that the predictability, or information content, of the utterances decreases from *b* to *c* to *h*. This should affect the duration, turn switches, and other behavior. The differences between the averages utterance durations are indeed significant for these categories (table 6, $p \leq 0.001$, Student's t-test: $t > 6.5$, $\nu > 8000$).

A distribution of the PSTS time delays over functional categories is given in table 7. Those for gaze timing in table 8. The PSTS delays in table 7 too show the marked effects of functional categories on dialog behavior. Less predictable chunks, like *b*, induce longer delays in the next speaker than more predictable chunks, like *c*. This difference goes beyond the mere effect of utterance duration as can be seen by comparing tables 6 and 7.

The gaze delays in table 8 show the opposite behavior to the turn delays. Where the next speaker tends to wait longer before starting to speak after a *b* utterance, the speaker that actually utters it starts to look towards her partner earlier. Again, it seems differences in utterance duration cannot completely explain this behavior.

Table 8: Distribution over the most important dialog functions of the time between the speaker looking towards the addressed dialog partner and the end of her turn (PSTS). Delay statistics calculated over the interval $[-2, 2]$ only. Labels *u* and *a* can be added to other labels and are counted separately. Mean: mean delay; SD: Standard Deviation; SE: Standard Error; #: Number of occurrences; all: all function labels

Label	Mean	SD	SE	#
b	-0.534	0.854	0.079	117
c	-0.328	0.916	0.024	1506
h	0.199	0.930	0.164	32
k	0.646	0.627	0.040	242
r	-0.116	0.850	0.071	142
f	0.254	0.730	0.141	27
a	-0.296	0.908	0.0718	160
u	-0.318	0.957	0.065	220
ua	-0.316	1.137	0.343	11
all	-0.181	0.935	0.020	2139

10. Discussion

A simple, low cost, functional annotation of dialogs into very simple content types was introduced for this corpus. A first look shows that these chosen categories seem to be relevant for interpersonal dialog behavior. But real validation will only come from successful use in explaining the behavior of the participants or experimental observers. The current results show the interaction between the functional annotation categories and the behavior of the speakers. These first results support the relevance of the functional label categories. These categories are at least predictive for some aspects of dialog behavior.

With the advent of large corpora, eg, the CGN (2006), speech communication science is becoming *big science*. With big science come new challenges and responsibilities, as distribution and access policies are required to unlock the collected data. For instance, see the discusion and references in Van Son et al. (2001; Van Son and Pols (2001).

At the moment, procedures for statistical analysis are urgently needed. For this project we have chosen to prepare the annotations for relational database access, RDBMS (Mengel and Heid, 1999; Cassidy, 1999; Van Son et al., 2001; Van Son and Pols, 2001). For many questions related to statistical tests and distributions such access is both required and sufficient. However, there are cases where the hierarchical nature of linguistic annotations (eg, syntax) would demand searching tree-like structures. We suggest that the use of XML databases would be studied for such use cases.

The above results show, again, that it is possible to integrate standard linguistic annotations and low cost dialog annotations into a searchable database. This opens an easy access to a host of statistical and analysis tools, from standard SQL to spreadsheets and *R*.

The method used to create a RDMS for the IFADV corpus is arguably ad-hoc, cf, (Mengel and Heid, 1999; Cassidy, 1999; Van Son et al., 2001; Van Son and Pols, 2001). We would prefer that *best practises* were formulated for preparing annotations for relational database access. With increasing corpus size, database storage will only increase in importance.

The bare fact that this paper spends more space on legal and license matters than on the annotations shows that, here too, there is a need for *best practises* for the handling of copyrights, informed consent, and privacy sensitive information in the context of corpus construction. Anecdotal reports emphasize the restrictions of the current laws where proper preparations might very well have prevented problems.

In the end it is the courts that decide on the boundaries of copyright and privacy laws. For a researcher of speech or language, little more can be done than listen to legal experts. During the construction of this corpus, we have tried to incorporate previous experiences with legal questions. This included attempts to inform our subjects about the

full possible extent of the distribution and use cases of the recordings, as well as about the legal consequences of their signatures. Moreover, we allowed our subjects ample time to review the recordings and retract their consent. None of the subjects did retract their consent. We used (adapted) copyright transfer forms that were prepared by legal staff of the Dutch Language Union for the CGN.

Copyright protects many aspects of recordings and annotations. It must be emphasized that almost everyone who has in any way contributed to, adapted, or changed the collected recordings or annotations has to sign copyright transfer forms.

11. Conclusions

A *free/libre* annotated corpus of conversational dialog video recordings is presented and described. For this corpus, it has been tried to overcome several known legal hurdles to freely sharing and distributing video recordings and annotations. With close to 70k words, there was a need for database storage and access for efficient analysis. This was tackled by using identification markers for every single item in the annotations that link the annotations together and to specific time points in the recordings.

12. Acknowledgements

The IFADV corpus is supported by grant 276-75-002 of the Netherlands Organization for Scientific Research. We want to thank Anita van Boxtel for transliterating the dialogs and labeling gaze direction, and Stephanie Wagenaar for compiling the summaries of the dialog transcripts.

13. References

R. Bertrand. 2007. Corpus d'interactions dilogales (CID). http://crdo.up.univ-aix.fr/corpus.php?langue=fr.

P. Blache, S. Rauzy, and G. Ferré. 2007. An XML Coding Scheme for Multimodal Corpus Annotation. In *Proceedings of Corpus Linguistics*.

P. Boersma and D. Weenink. 1992–2008. Praat: doing phonetics by computer. http://www.praat.org/.

J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23:13 – 31.

S. Cassidy. 1999. Compiling multi-tiered speech databases into the relational model: Experiments with the EMU system. In *Proceedings of EUROSPEECH99, Budapest*, pages 2239–2242.

CGN. 2006. The Spoken Dutch Corpus project. http://www.tst.inl.nl/cgndocs/doc_English/topics/index.htm.

M. Core and J. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28 – 35.

EC. 2005. First evaluation of Directive 96/9/EC on the legal protection of databases, DG INTERNAL MARKET AND SERVICES WORKING PAPER. http://europa.eu.int/comm/internal_market/copyright/docs/databases/evaluation_report_en.pdf.

ELAN. 2002–2007. ELAN is a professional tool for the creation of complex annotations on video and audio resources. http://www.lat-mpi.eu/tools/elan/.

ELRA. 2004–2007. European Language Resources Association: Catalogue of Language Resources. http://catalog.elra.info/.

FSF. 1991. GNU General Public License, version 2. http://www.gnu.org/licenses/old-licenses/gpl-2.0.html.

HLT-Agency. 2007. Centrale voor Taal- en Spraaktechnologie (TST-centrale). http://www.tst.inl.nl/producten/.

IDABC . 2008. European Union Public Licence (EUPL v.1.0). http://ec.europa.eu/idabc/eupl.

IMDI. 1999–2007. ISLE Meta Data Initiative. http://www.mpi.nl/IMDI/.

Ken Coar. 2006. The Open Source Definition (Annotated). http://www.opensource.org/docs/definition.php.

Holger M. Kienle, Daniel German, Scott Tilley, and Hausi A. Müller. 2004. Intellectual property aspects of web publishing. In *SIGDOC '04: Proceedings of the 22nd annual international conference on Design of communication*, pages 136–144, New York, NY, USA. ACM.

LDC. 1992–2007. The Language Data Consortium Corpus Catalog. http://www.ldc.upenn.edu/Catalog/.

MAPtask. 1992–2007. HCRC Map Task Corpus. http://www.hcrc.ed.ac.uk/maptask/.

Stephen M. Maurer, P. Bernt Hugenholtz, and Harlan J. Onsrud. 2001. Europe's database experiment. *Science*, 294:789–790.

A. Mengel and U. Heid. 1999. Enhancing reusability of speech corpora by hyperlinked query output. In *Proceedings of EUROSPEECH99, Budapest*, pages 2703–2706.

R Core Team. 1998–2008. The R Project for Statistical Computing. http://www.r-project.org/.

SMIL. 2008. W3C Synchronized Multimedia Integration Language. http://www.w3.org/AudioVideo/.

R.J.J.H. Van Son and L.C.W. Pols. 2001. Structure and access of the open source IFA Corpus. In *Proceedings of the IRCS workshop on Linguistic Databases, Philadelphia*, pages 245–253.

R.J.J.H. Van Son, D. Binnenpoorte, H. van den Heuvel, and L.C.W. Pols. 2001. The IFA corpus: a phonemically segmented Dutch Open Source speech database. In *Proceedings of EUROSPEECH 2001 Aalborg*, pages 2051–2054.

WIPO. 1979. Berne Convention for the Protection of Literary and Artistic Works. http://www.wipo.int/treaties/en/ip/berne/index.html.

WIPO, 2004. *WIPO Handbook on Intellectual Property: Policy, Law and Use*, chapter 5: International Treaties and Conventions on Intellectual Property, pages 237–364. WIPO, 2 edition. Date of access: March 2008.