Promoting *free* dialog video corpora: the IFADV corpus example

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

[1] ACLC/IFA, University of Amsterdam, the Netherlands, R.J.J.H.vanSon@uva.nl
[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. In contrast, *Free-and-Libre* corpora would allow anyone to add incremental annotations and improvement, distributing the cost of construction and maintenance. A freely available annotated corpus is presented with high quality video recordings of face-to-face conversational speech. An effort has been made to remove copyright and use restrictions. Annotations have been processed to RDBMS tables that allow SQL queries and direct connections to statistical software. A few simple examples are presented to illustrate the use of a databases of annotated speech. 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. Creating such a corpus is an expensive and time-consuming effort. Free, as in *freedom*, corpora would allow anyone to add incremental annotations and improvement. This way the burden of construction and maintenance of the corpus can be distributed over a wider user community. Such a distribution of efforts is also seen in other communities that freely share expensive information resources, e.g., Genbank [1], web based community resources like Wikipedia [2], and Free and Open Source software like the Linux kernel [3,4,5].

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 [6] nor the LDC

[7] had any conversational video material available. The corresponding entity in the Netherlands, the Dutch TST centrale [8], also had no conversational video corpus available. Nor were we at the time 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 [9] 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 [10,11], is an annotated audio-video recording of conversational speech which is currently available to other researchers at no cost. It is distributed under a license that intends to "*guarantee the undividedness of data distributed by CRDO and the follow-up of its utilisation for the benefit of its producers*". As such, the license does not allow redistribution and sharing of the corpus and requires that the distribution of upgrades and changes should go through the CRDO [12].

Within our project, we have created a visual version of the friendly Face-to-Face dialogs of the Spoken Dutch Corpus, also known as CGN [13]. 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.

The remainder of this paper is organized as follows. Sections 2 to 5 will describe the construction and structure of the corpus. Sections 6 and 7 contain a discussion on the legal aspects of creating and distributing (spoken) language corpora. Section 8 presents some illustrative examples of corpus use. The last sections, 9 and 10, contain a general discussion and conclusions drawn from our experiences with creating this corpus.

2 Recordings

For the recordings, the speakers sat face-to-face opposite of each other in an audio studio with a table in between (see Figure 1) The recording studio had a sound-treated box-in-a-box design and noise levels were low. 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 2). 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 digitized, and then 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 2 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 practice, 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. After the interruption by the instructions and scripted sentences that started each recording, the subjects in all dialogs initiated a new conversation on a new topic. Recordings were cut-off 900 seconds after the start of the conversations following the scripted sentences. Consequently, no conversation open and closing parts were recorded.

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.

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

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

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

during recording offset the synchrony of the two recordings, and all occurrences of frame drops have therefore been identified. For each recording, a SMIL [14] file is available that specifies how the original frame timing can be restored by repeating frames to replace dropped frames.

3 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 a small offset (see Figure 2). Although participants never looked directly into the camera, it is immediately obvious when watching the videos whether a person looks at the other participant or not. 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.

Fig. 2. Example frame of recordings (output camera A, left; B, right). In the frames shown, the speakers are looking at each other, i.e., a Gaze annotation label g (see text).

4 File formats

Like archives, corpora are constructed with an aim of long term access. Obviously, the first and foremost concern is the preservation of the original recordings, metadata, and other materials. This principle extends to the annotations, which should be added cumulative. The combined requirements of source preservation and cumulative annotations leads to the principle of *stand off annotation*, the separation of annotations and source materials [15].

For long term access, all data must be available in well understood, preferably open, formats [16,17]. It is essential that access and processing of the files is not restricted to specific applications or computer platforms as this will compromise the long term access and integrity of the corpus. Data stored using proprietary codecs and file formats faces a considerable risk of losing support from the technology's "owner" at some moment. Data stored in such legacy formats might become inaccessible in only a few years [16]. Being proprietary, it is often impossible to find or build supporting software from other sources.

Exclusionary proprietary restrictions are obviously a problem with codecs or file formats that are only available from a single vendor for selected platforms, e.g., Microsoft's VC-1 codecs and ASF file format. In the distribution of on-line broadcast media content, such exclusionary choices of formats and codecs are quite common. For instance, the codec chosen for the on-line broadcast of the 2008 Olympics in Beijing was at the time only available for Microsoft Windows and Apple OSX and excluded users of other platforms. Therefore, the designers of multi-modal corpora should be wary to take an example from media distribution on the internet.

Where possible, international standards should be supported in corpus construction and distribution, e.g., SMIL, MPEG, PDF, or ODF. However, the use of some standards, like MPEG, is restricted by patents which might be an is-

```
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.
```

Fig. 3. Example transcription of recordings, formatted for readability (originals are in Praat textgrid format). The transcription of a chunk ends with a punctuation mark. M65I: Male subject, F59H: Female subject

sue for some users. It is therefore advisable to include an option for accessing copies that are unencumbered by Intellectual Property (IP, e.g., copyrights and patents) claims. For the current IFADV corpus this was an issue only for the processed, i.e., compressed, versions of the recordings. Therefore, these are offered in several different formats, one of which was always "open" [18,19], e.g., Ogg formats and codecs .

For the IFADV corpus, we chose to preserve the original DV format recordings with audio and video as they were obtained from the AD converters. For each recording, a SMIL markup file [14] was created that described the frames that were dropped. These SMIL files will recreate the original timing by redoubling frames to stand in for the lost ones. The original recording files are large ($>$ 3GB) and rather cumbersome to distribute and manage. Therefore, frame corrected files are made available in DV format. These are also available as compressed files in avi (DivX3) and Ogg (Theora) format, with normalized brightness and contrast levels. The audio is available seperately as 48 kHz WAV (RIFF) files and a selection of compressed versions (FLAC, Ogg, Speex, MP3). All these file formats, and codecs, are widely used and available for all platforms.

There is currently no accepted standard file format for speech and video annotations. Work on such an international (exchange) standard has only just been presented [20]. For practial reasons, all annotations were stored in the Praat TextGrid format and the ELAN EAF file format (original gaze annotations). Both applications are widely used and sources are available. These annotation file formats are well documented and easy to parse automatically.

Where possible, the annotation labels and procedures of the CGN were used (see table 2).

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.

Fig. 4. Example extract from a summary of a recording session. Female and Male subject

5 Annotations

20 conversations have been annotated according to the formalism of the Spoken Dutch Corpus, CGN [13], 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 manual orthographic transliteration and rough time alignment of 5 hours of dialogs took approximately 150 hours (30 times real time).

The basic unit of the transliteration was the utterance-like *chunk*. This is an *inter-pausal unit* (IPU) when short, up to 3 seconds. Longer IPUs were split on strong prosodic breaks based on the intuition of the annotators. For practical purposes, these chunks can be interpreted as *utterances*, c.f., figure 3. To improve readability, we will refer to *utterances* in this text when we, strictly speaking, are referring to *chunks*, as defined in the Spoken Dutch Corpus [13]. The annotations are either in the same formats used by the CGN [13] or in newly defined formats (*non-CGN*) for annotations not present in the CGN (table 2).

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*
End intonation:	Automatic, *non-CGN*

Table 3. Conversational function annotation labels and their distribution in the corpus. Both *u* and *a* can follow other labels. 52 Chunks did not receive a label when they should have. Labels *u* and *a* can be added to other labels and are counted separately (n =13,669).

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

The functional annotation was restricted to keep the costs within budget. A HRC style hierarchical speech or conversational acts annotation [21,22] 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 [23]. 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, i.e., *d* for closing the eyes and *k* for stereotypical 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 (17 times real time).

The intonation at the end of an utterance is an important signal for potential turn switches, or Transition Relevance Places (TRP) [24]. Therefore, an automatic annotation on utterance end pitch was added (*low*, *mid*, and *high*, coded as 1, 2, 3), determined on the final pitch (in *semitones*) relative to the utterance mean and global variance, i.e., $Z = \left(F_0^{end} - \overline{F_0}\right) / stdev(F_0)$ with boundaries for *mid* $-0.5 \leq Z \leq 0.2$ [25].

Table 4. Example encoding scheme for item IDs. 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 (*A*) 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

An identification code (ID) has been added to all linguistic entities in the corpus according to [26,27,28,29]. 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.

These codes are necessary to build RDBMS (Relational Database Management System) tables for database access
[26,27,28,29]. 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 5). For example, statistical scripts from *R* can connect directly to the database [30]. 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 3). Summaries were compiled from these transcripts (see Figure 4). Meta data for all recordings are available. These have been entered into IMDI format [31].

[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';
```

Fig. 5. Example SQL query. This query generates the results displayed in the right hand (PSTS) side of the *ua* row of table 5. *properturnswitch*: table with the chunk ID's and the turn switch delays; *fct*: table with the functional labeling

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 by the Berne Convention as [32]:

> ... 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" [32]. 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 [33] as well as *database* rights [34,35,36]. 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, i.e., 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, up-front, to subjects how the recordings and the personal data might be used. In practise, this has meant that the different options, e.g., 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 layman's 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 authoritative, the categories of contributors 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 *in*clude more people, but it is better to consult a legal expert before *ex*cluding possible contributors.

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 and distribution

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 [19]. We chose the GNU General Public License, GPLv 2 [37], 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 [38]. 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 transferred to the Dutch Language Union (NTU). The Dutch Language Union distributes the corpus and all related materials under the GNU General Public License [37].

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

The corpus is currently freely available from the TST-centrale [8]. This includes raw and processed video recordings, audio, and all annotations. In addition, there are derived annotation files available that combine different annotations. Summaries and IMDI metadata records have been made for all annotated dialogs. Relational database tables have been constructed 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 GPLv2 [37]. All materials are available free of charge. Pre-release

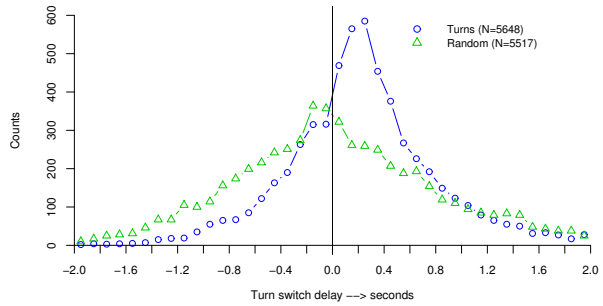

Fig. 6. Distribution of turn switch delays (PSTS), circles, and randomized turn switches, triangles. Bin sizes: 100ms

development versions of all materials are available from the University of Amsterdam at URL http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/.

8 Examples

An important aim that guided the construction of the IFADV corpus was to make the corpus easy accessible for research purposes. The approach chosen was to store all meta-data and annotations in database tables (RDBMS). As a result, many research questions can be directly answered using standard SQL queries using common tools. Such an approach to spoken language corpora is not yet standard practice. To illustrate the usefulness of annotation databases, we will present three practical examples of simple research questions that can be tackled using database searches.

To start with the contents of the database. In total, 13,373 (IPU based) verbal *chunks* 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.

8.1 Example: Simplified *Proper Speaker Turn Switches*

Many important aspects of conversations are localized around, potential, speaker turn switches. Determining such places is non-trivial, but it is possible to automatically mark all overt speaker switches. Simplified *Proper Speaker Turn Switches* (PSTS) were defined as succeeding pairs of utterances from different speakers where the next speaker started an utterance *after* the start of the last utterance of the previous speaker that continued beyond the end of that last

Table 5. Distribution of durations in seconds over the most important conversational functions. Chunks (Chk, left) and PSTS delays (right). Labels *u* and *a* can be added to other labels and are counted separately. Mean: mean duration; SD: Standard Deviation; SE: Standard Error; #: Number of occurrences; All: all functional labels

Label	Chk Mean	SD	SE	#	PSTS Mean	SD	SE	#
b	1.535	0.648	0.024	735	0.425	0.633	0.039	262
c	1.367	0.667	0.007	8739	0.233	0.670	0.011	3682
h	0.773	0.531	0.034	240	0.122	0.564	0.051	121
k	0.312	0.288	0.006	2425	0.307	0.507	0.016	1009
r	0.937	0.687	0.024	853	0.251	0.644	0.032	409
f	0.539	0.318	0.022	213	0.271	0.713	0.075	90
a	1.194	0.667	0.018	1374	0.167	0.754	0.038	388
u	1.189	0.668	0.021	1002	0.278	0.613	0.023	733
ua	1.747	0.679	0.133	26	0.053	0.574	0.117	24
All	1.119	0.739	0.006	13669	0.256	0.643	0.008	5752

utterance. Non-verbal noises were ignored. These PSTS events were automatically determined by sorting verbal chunks on their end times and selecting those turns that start after the start of the preceding turn.

Such PSTS events are cardinal places in dialogs and the delay between the end of the last turn and the start of the new turn contains important information about the dynamics of the conversation. We will use the distribution of PSTS (turn) delays to illustrate the use of the IFADV corpus. The basic distribution of the PSTS delays as found in the IFADV corpus 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.

To be able to evaluate the distribution of PSTS delays, the statistics of observed turn switch delays must be compared to some null hypothesis of random, or unrelated, turn switches. Features of interest, e.g., the variance, can be extracted from the PSTS delays and related to some other aspect of the conversation, e.g., the expected mental processing effort based on the transcription. This relation would than have to be compared to the null hypothesis, that there is no relation between the two features investigated.

To do this comparison, the statistics of the delays under random delay timings should be known. That is, the statistics of a real conversation will be compared to the same statistics for a sample of random, or unrelated, turn switches. The statistical difference between the feature measurements in the real, observed and the randomized turn delays can then be used to indicate whether the presence or size of the feature is convincingly shown. For instance, the question whether the PSTS delay distribution in figure 6 indicates that speakers wait for each other to end a turn can only be answered if we have a distribution of PSTS delays of

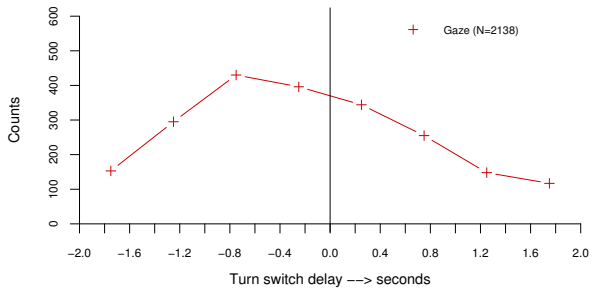

Fig. 7. Distribution of gaze delays from the last speaker (see text). Bin sizes: 500ms

people who do *not* wait for each other but start speaking randomly. That would be our null hypothesis.

Such random turn switch delays can be generated from the utterances of speakers from different conversations who are compared as if they were participating in the same conversation. As the distribution of utterances and pauses is specific for each conversation [39], a randomized PSTS delay distribution should actually be generated from the same conversation it models. This can easily be done by calculating PSTS delays after offsetting the transcribed utterances from one of the participants by a large time interval.

That is, if we have participants A and B, we add 100 seconds to all utterance starts and ends of B, modulo the length of the conversation (i.e., wrap around). After 100 seconds, the utterance starts and ends of A and B have become uncorrelated and the new "PSTS" delays are a model of the distribution of unrelated, random, turn switches. 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 seem obvious, and can be quantified with measures like the Kullback-Leibler divergence, but the statistics are not always straightforward.

8.2 Example: Gaze direction

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 distribution of gaze delays between the speaker looking towards the partner and the end of the nearest turn of the speaker is presented in figure 7. There were 5,168 discrete time spans in total where one subject looked directly at the other (and equally many where they

Table 6. 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

looked away). Dialog participants gazed at each other for almost 75% of the time and at the end of 70% of all of their utterances. However, speakers gaze at the listener preceeding 79% of the turn switches (PSTS), which is more than expected ($p \leq 0.001$, χ^2 test).

8.3 Example: Functional annotation

Most of the annotations used in this corpus were taken from the Spoken Dutch Corpus (CGN) [13], and are well understood. Gaze direction is straightforward and we do not expect problems with its interpretation. However, the functional annotation of the 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 judgement on the discourse function of individual 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 verbal chunks over conversational function is given in table 3. 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 5. The most marked effect is expected between utterances adding content to the discourse, i.e., *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 average utterance durations for these conversational function categories, *b*, *c*, and *h* are indeed statistically significant (table 5, $p \leq 0.001$, Student's t-test: $t > 6.5$, $\nu > 8000$). Indeed, the average duration of a type *b*, topic start, utterance is twice that of a simple repetition utterance, type *h*.

A distribution of the PSTS time delays over functional categories is given in table 5. Those for gaze timing in table 6. The PSTS delays in table 5 too show the marked effects of functional categories on dialog behavior. Less predictable utterances, like *b*, induce delays in the next speaker that are almost twice as long as more predictable utterances, like *c*. The difference in delay duration is much larger than the corresponding difference in utterance duration, as can be seen in table 5. However, interpreting the delays is complicated by the generally negative correlation between stimulus length and response times.

The gaze delays in table 6 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, the relation between gaze delay and utterance lengths might not be simple.

More work is obviously needed to disentangle the effects of utterance duration and conversational function, e.g., *b*, *c*, and *h*, on the gaze and next speaker timing.

9 Discussion

With the advent of large corpora, e.g., the Spoken Dutch Corpus [13], speech communication science is becoming a *big data* science [40]. With big science come new challenges and responsibilities, as distribution and access policies are required to unlock the collected data, e.g., [1]. For language corpora, see also the discussion and references in [28,29].

At the moment, comprehensive mechanisms for statistical analysis are urgently needed. For the IFADV corpus, we have chosen to prepare the annotations for relational database access, RDBMS [26,27,28,29]. 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, e.g., syntax, would demand searching tree-like structures. We suggest that the use of XML databases would be studied for such use cases. The above examples show, again, the usefulness of integrating 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 Query Language (SQL) to spreadsheets and *R* [30].

The method used to create a RDMS for the IFADV corpus is arguably ad-hoc, c.f., [26,27,28,29]. 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.

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 (see section 8.3). 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.

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.

10 Conclusions

The speech and language community can gain a lot from widely available corpora of language behavior. Experience in other fields have shown that gains in efficiency can be obtained by sharing information resources in a *free/libre* fashion. An example of 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. A few simple examples are presented to illustrate potential uses of such a database of annotated speech.

Corpus construction has only recently been finished, so there are currently no data about any effect of it's liberal license on use and maintenance.

11 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, Stephanie Wagenaar for compiling the summaries of the dialog transcripts, and Maaike van Naerssen for the IMDI records.

References

1. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L.: Genbank. Nucleic Acids Research **35**(Database-Issue) (2007) 21–25
2. Kolbitsch, J., Maurer, H.: The transformation of the web: How emerging communities shape the information we consume. Journal of Universal Computer Science **12**(2) (feb 2006) 187–213
3. Lerner, J., Tirole, J.: Some simple economics of open source. Journal of Industrial Economics. **50** (06/2002 2002) 197–234
4. Ciffolilli, A.: The economics of open source hijacking and declining quality of digital information resources: A case for copyleft. Development and Comp Systems 0404008, EconWPA (April 2004)
5. Rullani, F.: Dragging developers towards the core. how the free/libre/open source software community enhances developers' contribution. LEM Papers Series 2006/22, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy (September 2006)
6. ELRA: European Language Resources Association: Catalogue of Language Resources. http://catalog.elra.info/ (2004–2007)
7. LDC: The Language Data Consortium Corpus Catalog. http://www.ldc.upenn.edu/Catalog/ (1992–2007)
8. HLT-Agency: Centrale voor Taal- en Spraaktechnologie (TST-centrale). http://www.tst.inl.nl/producten/ (2007)
9. MAPtask: HCRC Map Task Corpus. http://www.hcrc.ed.ac.uk/maptask/ (1992–2007)
10. Blache, P., Rauzy, S., Ferré, G.: An XML Coding Scheme for Multimodal Corpus Annotation. In: Proceedings of Corpus Linguistics. (2007)
11. Bertrand, R.: Corpus d'interactions dilogales (CID). http://crdo.fr/voir_depot.php?langue=en&id=27 (2007)
12. CRDO: Licences. http://crdo.up.univ-aix.fr/phpwiki/index.php?pagename=Licences (2008)
13. CGN: The Spoken Dutch Corpus project. http://www.tst.inl.nl/cgndocs/doc_English/topics/index.htm (2006)
14. SMIL: W3C Synchronized Multimedia Integration Language. http://www.w3.org/AudioVideo/ (2008)
15. Ide, N., Romary, L.: Outline of the international standard linguistic annotation framework. In: Proceedings of the ACL 2003 workshop on Linguistic annotation, Morristown, NJ, USA, Association for Computational Linguistics (2003) 1–5

16. Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., Hinrichs, E.: Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources. In: Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art, Lansing, Michigan (2006)
17. Rehm, G., Witt, A., Hinrichs, E., Reis, M.: Sustainability of annotated resources in linguistics. In: Proceedings of Digital Humanities 2008, Oulu, Finland (2008) 27–29
18. IDABC : European Interoperability Framework for Pan-European eGovernmentservices. http://europa.eu.int/idabc/en/document/3761 (2004)
19. Ken Coar: The Open Source Definition (Annotated). http://www.opensource.org/docs/definition.php (2006)
20. Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., Magnusson, M., Rose, T., Sloetjes, H.: An exchange format for multimodal annotations. In (ELRA), E.L.R.A., ed.: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco (may 2008)
21. Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., Anderson, A.: The reliability of a dialogue structure coding scheme. Computational Linguistics **23** (1997) 13 – 31
22. Core, M., Allen, J.: Coding dialogs with the damsl annotation scheme. In: AAAI Fall Symposium on Communicative Action in Humans and Machines. (1997) 28 – 35
23. ELAN: ELAN is a professional tool for the creation of complex annotations on video and audio resources. http://www.lat-mpi.eu/tools/elan/ (2002–2007)
24. Caspers, J.: Local speech melody as a limiting factor in the turn-taking system in dutch. Journal of Phonetics **31**(2) (April 2003) 251–276
25. Wesseling, W., van Son, R.J.J.H.: Early Preparation of Experimentally Elicited Minimal Responses. In: Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue. (2005) 11–18
26. Mengel, A., Heid, U.: Enhancing reusability of speech corpora by hyperlinked query output. In: Proceedings of EUROSPEECH99, Budapest. (1999) 2703–2706
27. Cassidy, S.: Compiling multi-tiered speech databases into the relational model: Experiments with the EMU system. In: Proceedings of EUROSPEECH99, Budapest. (1999) 2239–2242
28. Van Son, R., Binnenpoorte, D., van den Heuvel, H., Pols, L.: The IFA corpus: a phonemically segmented Dutch Open Source speech database. In: Proceedings of EUROSPEECH 2001 Aalborg. (2001) 2051–2054
29. Van Son, R., Pols, L.: Structure and access of the open source IFA Corpus. In: Proceedings of the IRCS workshop on Linguistic Databases, Philadelphia. (2001) 245–253
30. R Core Team: The R Project for Statistical Computing. http://www.r-project.org/ (1998–2008)
31. IMDI: ISLE Meta Data Initiative. http://www.mpi.nl/IMDI/ (1999–2007)
32. WIPO: Berne Convention for the Protection of Literary and Artistic Works. http://www.wipo.int/treaties/en/ip/berne/index.html (1979)
33. WIPO: 5: International Treaties and Conventions on Intellectual Property. In: WIPO Handbook on Intellectual Property: Policy, Law and Use. 2 edn. WIPO (2004) 237–364 Date of access: March 2008 http://www.wipo.int/about-ip/en/iprm/.

34. Maurer, S.M., Hugenholtz, P.B., Onsrud, H.J.: Europe's database experiment. Science **294** (2001) 789–790
35. Kienle, H.M., German, D., Tilley, S., Müller, H.A.: Intellectual property aspects of web publishing. In: SIGDOC '04: Proceedings of the 22nd annual international conference on Design of communication, New York, NY, USA, ACM (2004) 136–144
36. EC: 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 (2005)
37. FSF: GNU General Public License, version 2. http://www.gnu.org/licenses/old-licenses/gpl-2.0.html (1991)
38. IDABC : European Union Public Licence (EUPL v.1.0). http://ec.europa.eu/idabc/eupl (2008)
39. ten Bosch, L., Oostdijk, N., Boves, L.: On temporal aspects of turn taking in conversational dialogues. Speech Communication **47**(1-2) (2005) 80–86
40. : Community cleverness required. Nature **455**(7209) (2008) 1–1