



SIGNSPEAK

Scientific understanding and vision-based technological development for continuous sign language recognition and translation

Grant Agreement Number 231424

**Small or medium-scale focused research project (STREP) FP7-ICT-2007-3.
Cognitive Systems, Interaction, Robotics**

Project start date: 1 April 2009

Project duration: 36 months

Deliverable D.1.3: Enriched annotation documents for the existing CNGT corpus containing sentence boundaries units.

Dissemination Level: **Public**

Date: 16 September 2010

Version: 3

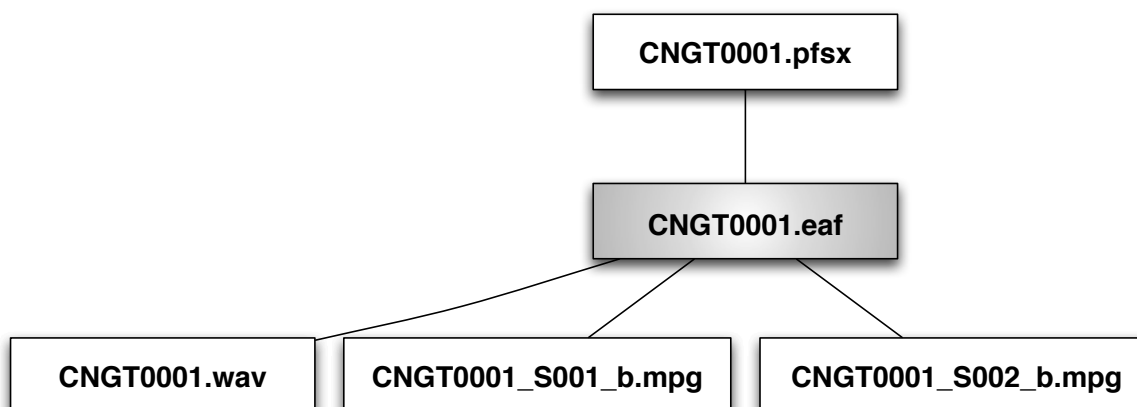
Author: Onno Crasborn
Centre for Language Studies
Radboud University Nijmegen

1. Introduction

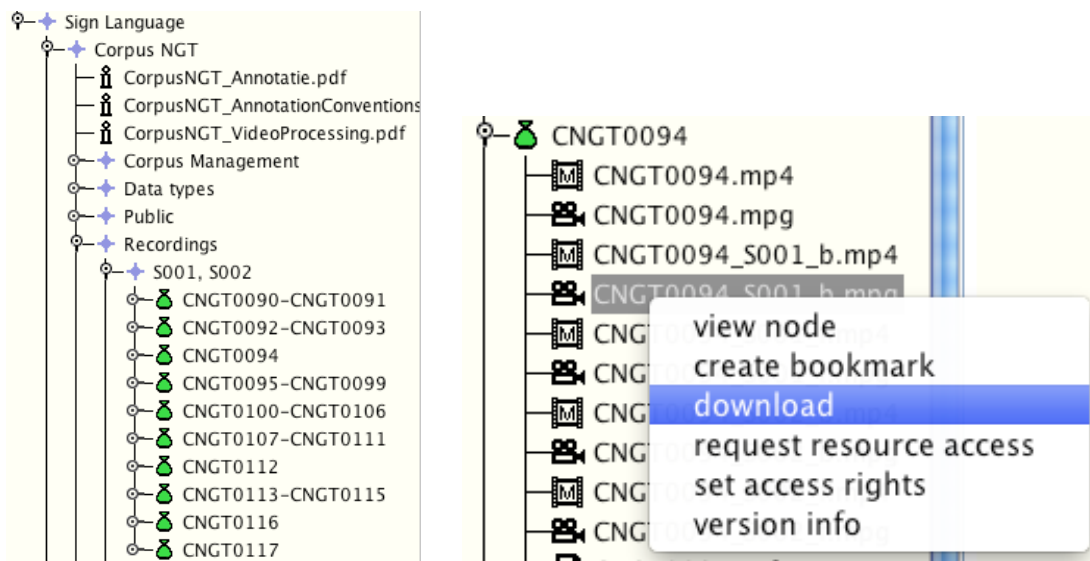
The compressed file **CNGT_EAF_31mar2010.zip** contains the full set of 2375 annotation files of the Corpus NGT, organized in folders per 100 files. A subset of the whole corpus has been annotated; appendix 1 lists all files that contain gloss annotations, indicating whether they also have sentence boundary annotations or not.

The annotations have been created with the open source software ELAN, developed at the Max Planck Institute for Psycholinguistics (<http://www.lat-mpi.eu/tools/elan/>). The annotation format EAF is rarely changed with new releases of the software, but where necessary, all corpus files have been updated to conform to the latest version of the format.

Each ELAN file is connected to two MPEG-1 video files, one for each signer in the dialogue. Additionally, some files have a voiceover by an interpreter in a separate WAV file. Preferences per document are stored in a separate file. This is represented in the following schema, including the file name templates used for the Corpus NGT.



The video and audio files are altogether more than 100 GB, and are not centrally distributed through SignSpeak. They can be located for public access at the archive of the Max Planck Institute for Psycholinguistics, at <http://corpus1.mpi.nl/>, under Sign Language > Corpus NGT. The subnode Recordings contains all 2375 sessions sorted by signer. Nearly all sessions are publicly accessible. Movies can be downloaded by scrolling down to the appropriate session, and right-clicking on the appropriate video file (see screen shots). Higher resolution versions of the movies are not publicly accessible, and have been distributed project-internally.



The final version of the annotation files are currently not yet located in the online corpus (as of Sept. 2010), so the .eaf documents in the online archive are the initial release versions of the Corpus NGT project that ended in 2008. In April 2011, all the updated files and updated annotation conventions will be ingested in the online archive, and thus be accessible for anyone to download and use.

Aside from downloading and viewing in ELAN, the MPI archive also allows for browsing annotations within a web browser using the web application ANNEX. This application looks much like ELAN in its interface, but in the present version does not allow for editing. It can be tested with the archive versions of the annotation files by selecting “view” from the contextual menu linked to the .eaf files. The MPEG-4 versions of the movies are used to stream the content. Once the new (April 2011), this will be the easiest way to inspect the annotation files that have been created within SignSpeak.

2. Content of the annotation files

The documents contain the following tiers; the right column lists the number of files that have been annotated up to March 2010.

GlosL S1	Gloss annotations (present in 200 files)
GlosR S1	
GlosL S2	
GlosR S2	
SignSpeakZin S1	Sentence boundary annotations (present in 43 files)
SignSpeakZin S2	
SS_zinsopmerking	
SS_Head S1	Head movement annotations (present in 2 files)
SS_Head S2	
SS_Mouth S1	

SS_Mouth S2	Mouth action annotations (see below)
-------------	--------------------------------------

With respect to the initial release of the corpus, the documents contain extra tiers (all of the above except the Glos tiers), with better and more glosses, sentence translations, and sentence boundary units.

A mouth tier per signer was added in March 2010 as this was a recurrent wish expressed by the annotators during the glossing process. Especially in cases where the mouth articulation contributes specifying information with respect to the semantics of the manual sign, this makes the glosses more consistent as annotators do not try to change the semantics of the Dutch gloss anymore. The mouth actions will not be systematically annotated, however, as that would be highly time-consuming process and because it is not required by the other Work Packages.

An overview of the present state of annotation is listed in the following table.

	Start of project	M12
Gloss annotations	64.288	91.194
Signs		
tokens	49.463	70.061
types	5.191	7.127
singletons	2.919	4.184
Sentence boundary annotations with translations	0	1276
Sentence boundary annotations	0	758

Appendix 1 lists the files that had been annotated until April 1st, 2010, and indicates whether they have been glossed during SignSpeak or were already part of the initial release of the Corpus NGT in December 2008. These latter files have seen some modifications within SignSpeak, reflecting changes in the annotation conventions. Ongoing changes during year 2 of SignSpeak will be reflected in the annotation manual that will be published in English with the final release of the SignSpeak-enhanced version of the Corpus NGT (Deliverables 1.5 and 1.7).

The present version of the annotation guidelines is only available as a wiki site for the RU, in Dutch so as to facilitate the work of the deaf annotators. The English version will be created

3. How to view and search the annotation documents

The full manual for ELAN can be accessed online at <http://www.lat-mpi.eu/tools/elan/manual/> or downloaded from <http://www.mpi.nl/corpus/manuals/manual-elan.pdf>.

Here are some basic hints that should make it easy to view annotated documents.

- When opening an ELAN file (extension .eaf), the associated media should automatically open. For the Corpus NGT, these are by default the two MPEG-1 movies of the two signers in the dialogue (resp. S1 on the left and S2 on the right). If ELAN cannot locate the media files, the user is asked to locate them himself. In the preferences of the tool, the user can set a default directory where ELAN should look for media files if they are not in the same folder as the .eaf file. This makes it easy to put all media files to the same folder (or on an external drive, for example).
- The media can be played back by using the standard controls below the video windows. In addition to the standard options (jump by frame or second), there are buttons for moving one screen forward or backward, and one pixel forward or backward. The 'screen' in this case refers to the time segment that is visible in the so-called *timeline viewer* at the bottom. It's time-domain can be changed by zooming in and out (options in the contextual menu, clicking anywhere in the timeline pane). In the standard 100% view, few glosses are fully displayed, given the short duration of most signs. Zooming out to 200% is often most comfortable.
- To the right of the standard video controls, there are three buttons that manipulate the time selection: by dragging in the timeline viewer or selecting an annotation, a time segment is selected that can then be played back.
- Aside from in the timeline viewer, annotations can be viewed in several different ways in the tabs at the top right of the screen: the *grid* (list), *text* (running text), and *subtitle* viewers. For each, a tier has to be selected the contents of which is then presented in the window.
- A final useful presentation of the annotations is in the *annotation density viewer*, the bar above the timeline viewer. Here, the total duration of the media files is presented from left to right in the window, and every annotation is represented by a small vertical black bar. In this way, pauses in the conversation can be easily spotted. Clicking on the timeline viewer also changes the location of the cursor.
- Finally, dragging up or down the double arrow on the middle right of the window allows one to change the display size of the media files.

For further information, please see the manual or online manual that were already referred to above.

Appendix 1: overview of annotated files

Session Number	Glosses	Sentence Segmentations	Head Annotations
0001	Corpus NGT Dec. 2008		SignSpeak D1.3

0004	Corpus NGT Dec. 2008		SignSpeak D1.3
0005	Corpus NGT Dec. 2008		
0006	Corpus NGT Dec. 2008		
0007	Corpus NGT Dec. 2008		
0008	Corpus NGT Dec. 2008		
0009	Corpus NGT Dec. 2008		
0010	Corpus NGT Dec. 2008		
0011	Corpus NGT Dec. 2008		
0012	Corpus NGT Dec. 2008		
0013	Corpus NGT Dec. 2008		
0014	Corpus NGT Dec. 2008		
0015	Corpus NGT Dec. 2008		
0016	Corpus NGT Dec. 2008		
0017	Corpus NGT Dec. 2008		
0018	Corpus NGT Dec. 2008		
0044	Corpus NGT Dec. 2008		
0046	Corpus NGT Dec. 2008		
0047	Corpus NGT Dec. 2008		
0048	Corpus NGT Dec. 2008		
0049	Corpus NGT Dec. 2008		
0050	Corpus NGT Dec. 2008		
0055	Corpus NGT Dec. 2008	SignSpeak D1.3	
0056	Corpus NGT Dec. 2008	SignSpeak D1.3	
0057	Corpus NGT Dec. 2008	SignSpeak D1.3	
0058	Corpus NGT Dec. 2008	SignSpeak D1.3	
0059	Corpus NGT Dec. 2008	SignSpeak D1.3	
0060	Corpus NGT Dec. 2008	SignSpeak D1.3	
0061	Corpus NGT Dec. 2008	SignSpeak D1.3	
0062	Corpus NGT Dec. 2008	SignSpeak D1.3	
0063	Corpus NGT Dec. 2008	SignSpeak D1.3	
0064	Corpus NGT Dec. 2008	SignSpeak D1.3	
0065	Corpus NGT Dec. 2008	SignSpeak D1.3	
0066	Corpus NGT Dec. 2008	SignSpeak D1.3	
0067	Corpus NGT Dec. 2008	SignSpeak D1.3	
0068	Corpus NGT Dec. 2008	SignSpeak D1.3	
0069	Corpus NGT Dec. 2008	SignSpeak D1.3	
0090	Corpus NGT Dec. 2008		
0091	Corpus NGT Dec. 2008		
0092	Corpus NGT Dec. 2008		
0093	Corpus NGT Dec. 2008		
0094	Corpus NGT Dec. 2008		
0095	Corpus NGT Dec. 2008		
0096	Corpus NGT Dec. 2008		
0097	Corpus NGT Dec. 2008		
0098	Corpus NGT Dec. 2008		
0099	Corpus NGT Dec. 2008		
0117	Corpus NGT Dec. 2008		
0118	Corpus NGT Dec. 2008		
0119	Corpus NGT Dec. 2008		
0120	Corpus NGT Dec. 2008		
0121	Corpus NGT Dec. 2008		
0124	Corpus NGT Dec. 2008		
0128	Corpus NGT Dec. 2008		
0129	Corpus NGT Dec. 2008		
0130	Corpus NGT Dec. 2008		
0131	Corpus NGT Dec. 2008		
0132	Corpus NGT Dec. 2008		

0133	Corpus NGT Dec. 2008		
0134	Corpus NGT Dec. 2008		
0135	Corpus NGT Dec. 2008		
0136	Corpus NGT Dec. 2008		
0137	Corpus NGT Dec. 2008		
0138	Corpus NGT Dec. 2008		
0139	Corpus NGT Dec. 2008		
0154	Corpus NGT Dec. 2008		
0159	Corpus NGT Dec. 2008		
0170	Corpus NGT Dec. 2008		
0205	Corpus NGT Dec. 2008		
0206	Corpus NGT Dec. 2008		
0207	Corpus NGT Dec. 2008		
0208	Corpus NGT Dec. 2008		
0215	Corpus NGT Dec. 2008		
0245	SignSpeak D1.3	SignSpeak D1.3	
0250	Corpus NGT Dec. 2008		
0251	Corpus NGT Dec. 2008		
0252	Corpus NGT Dec. 2008		
0254	SignSpeak D1.3		
0255	SignSpeak D1.3		
0256	SignSpeak D1.3		
0258	Corpus NGT Dec. 2008	SignSpeak D1.3	
0259	Corpus NGT Dec. 2008	SignSpeak D1.3	
0260	Corpus NGT Dec. 2008		
0274	Corpus NGT Dec. 2008		
0279	Corpus NGT Dec. 2008		
0280	Corpus NGT Dec. 2008		
0283	Corpus NGT Dec. 2008		
0284	Corpus NGT Dec. 2008		
0295	Corpus NGT Dec. 2008		
0296	Corpus NGT Dec. 2008		
0297	Corpus NGT Dec. 2008		
0298	Corpus NGT Dec. 2008		
0299	Corpus NGT Dec. 2008		
0313	Corpus NGT Dec. 2008		
0316	Corpus NGT Dec. 2008		
0318	SignSpeak D1.3		
0319	Corpus NGT Dec. 2008		
0320	Corpus NGT Dec. 2008		
0328	Corpus NGT Dec. 2008		
0329	Corpus NGT Dec. 2008		
0330	Corpus NGT Dec. 2008		
0331	Corpus NGT Dec. 2008		
0332	Corpus NGT Dec. 2008		
0333	Corpus NGT Dec. 2008		
0334	Corpus NGT Dec. 2008		
0335	Corpus NGT Dec. 2008		
0336	Corpus NGT Dec. 2008		
0337	Corpus NGT Dec. 2008		
0338	Corpus NGT Dec. 2008	SignSpeak D1.3	
0339	Corpus NGT Dec. 2008	SignSpeak D1.3	
0340	Corpus NGT Dec. 2008		
0341	Corpus NGT Dec. 2008		
0362	Corpus NGT Dec. 2008		
0363	Corpus NGT Dec. 2008		
0364	Corpus NGT Dec. 2008		
0369	Corpus NGT Dec. 2008		
0370	Corpus NGT Dec. 2008		
0371	Corpus NGT Dec. 2008		

0386	SignSpeak D1.3	SignSpeak D1.3	
0387	SignSpeak D1.3	SignSpeak D1.3	
0388	SignSpeak D1.3	SignSpeak D1.3	
0389	SignSpeak D1.3	SignSpeak D1.3	
0411	Corpus NGT Dec. 2008		
0412	Corpus NGT Dec. 2008		
0413	Corpus NGT Dec. 2008		
0414	SignSpeak D1.3		
0415	Corpus NGT Dec. 2008		
0416	Corpus NGT Dec. 2008		
0417	Corpus NGT Dec. 2008		
0418	Corpus NGT Dec. 2008		
0419	Corpus NGT Dec. 2008		
0427	Corpus NGT Dec. 2008	SignSpeak D1.3	
0428	Corpus NGT Dec. 2008	SignSpeak D1.3	
0429	Corpus NGT Dec. 2008	SignSpeak D1.3	
0430	Corpus NGT Dec. 2008	SignSpeak D1.3	
0431	Corpus NGT Dec. 2008	SignSpeak D1.3	
0432	SignSpeak D1.3	SignSpeak D1.3	
0434	SignSpeak D1.3	SignSpeak D1.3	
0435	SignSpeak D1.3	SignSpeak D1.3	
0436	SignSpeak D1.3	SignSpeak D1.3	
0460	Corpus NGT Dec. 2008		
0466	Corpus NGT Dec. 2008		
0467	Corpus NGT Dec. 2008		
0476	SignSpeak D1.3	SignSpeak D1.3	
0510	Corpus NGT Dec. 2008		
0511	Corpus NGT Dec. 2008		
0512	SignSpeak D1.3		
0513	SignSpeak D1.3		
0514	SignSpeak D1.3		
0515	Corpus NGT Dec. 2008		
0516	Corpus NGT Dec. 2008		
0517	Corpus NGT Dec. 2008		
0518	Corpus NGT Dec. 2008		
0519	Corpus NGT Dec. 2008		
0529	SignSpeak D1.3		
0531	SignSpeak D1.3	SignSpeak D1.3	
0532	SignSpeak D1.3	SignSpeak D1.3	
0534	SignSpeak D1.3		
0541	Corpus NGT Dec. 2008		
0546	Corpus NGT Dec. 2008		
0557	SignSpeak D1.3		
0564	SignSpeak D1.3		
0570	SignSpeak D1.3		
0571	SignSpeak D1.3		
0592	Corpus NGT Dec. 2008		
0597	Corpus NGT Dec. 2008		
0616	SignSpeak D1.3		
0641	Corpus NGT Dec. 2008		
0642	Corpus NGT Dec. 2008		
0694	SignSpeak D1.3		
0695	SignSpeak D1.3	SignSpeak D1.3	
0697	Corpus NGT Dec. 2008		
0704	Corpus NGT Dec. 2008		
0747	Corpus NGT Dec. 2008		
0748	SignSpeak D1.3	SignSpeak D1.3	
0749	SignSpeak D1.3		
0752	SignSpeak D1.3		
0798	SignSpeak D1.3	SignSpeak D1.3	

0805	Corpus NGT Dec. 2008		
0806	Corpus NGT Dec. 2008		
0814	SignSpeak D1.3	SignSpeak D1.3	
0831	Corpus NGT Dec. 2008		
0832	Corpus NGT Dec. 2008		
0838	Corpus NGT Dec. 2008		
0847	Corpus NGT Dec. 2008		
0848	Corpus NGT Dec. 2008		
0862	Corpus NGT Dec. 2008		
0877	SignSpeak D1.3		
0904	SignSpeak D1.3	SignSpeak D1.3	
0905	SignSpeak D1.3		
0947	Corpus NGT Dec. 2008		
0954	Corpus NGT Dec. 2008		
0955	Corpus NGT Dec. 2008		
0958	Corpus NGT Dec. 2008		
0961	Corpus NGT Dec. 2008		
0962	Corpus NGT Dec. 2008		
0981	Corpus NGT Dec. 2008		
1004	Corpus NGT Dec. 2008		
1006	SignSpeak D1.3		
1008	Corpus NGT Dec. 2008		
1028	Corpus NGT Dec. 2008		
1046	Corpus NGT Dec. 2008		
1047	Corpus NGT Dec. 2008		
1048	Corpus NGT Dec. 2008		
1055	Corpus NGT Dec. 2008		
1056	Corpus NGT Dec. 2008		
1057	Corpus NGT Dec. 2008		
1058	Corpus NGT Dec. 2008		
1059	Corpus NGT Dec. 2008		
1060	Corpus NGT Dec. 2008		
1071	Corpus NGT Dec. 2008		
1072	Corpus NGT Dec. 2008		
1073	Corpus NGT Dec. 2008		
1074	Corpus NGT Dec. 2008		
1075	Corpus NGT Dec. 2008		
1076	Corpus NGT Dec. 2008		
1086	Corpus NGT Dec. 2008		
1105	SignSpeak D1.3	SignSpeak D1.3	
1157	SignSpeak D1.3		
1261	SignSpeak D1.3		
1415	SignSpeak D1.3		
1474	SignSpeak D1.3		
1771	SignSpeak D1.3	SignSpeak D1.3	
1789	SignSpeak D1.3		
1831	SignSpeak D1.3	SignSpeak D1.3	
1840	SignSpeak D1.3		
1853	SignSpeak D1.3		
1854	SignSpeak D1.3		
1894	SignSpeak D1.3	SignSpeak D1.3	
2072	SignSpeak D1.3	SignSpeak D1.3	

