Validation report for the CGN Database
Release 3

Authors	Chr. Draxler F. Schiel
Affiliation	BAS Bayerisches Archiv für Sprachsignale Institut für Phonetik Universität München
Postal address	Schellingstr. 3 D 80799 München
e-mail	{draxler|schiel}@phonetik.uni-muenchen.de
Telephone	+49-89-2180-2758
Fax	+48-89-2800362
Version	2.0
Date	29.Oct.2001
Status	final
Comment	This version 2.0 of the original validation report defines more clearly the procedures for the orthographic and POS validation, corrects the figures for the error rates, and adds two more log files to the appendix list.

1 Preliminary notes

This document uses different type faces to distinguish literal quotes, references to files, directories, and named entities, and regular text. For literal quotes, a proportional italic font is used and the quotation is enclosed by quotation marks, e.g. *"da's"*. File names, directories, and named entities are given in a monospaced typwriter font, e.g. `/META/CODES/` or `&euml;`, and a proportional regular font is used for all other text.

The file system hiearchy is presented with the UNIX separator symbol, i.e. the slash "`/`".

2 Formal validation of data

The formal validation consists of checks that can be carried out automatically. It does not include checks that regard the contents of the data. Details on what was checked are given in each section.

The original validation log files are given in the appendix and they are named as follows:

`FormalValidation.log<L>total`	Formal validation of signal files
`AnnotValidation.log<L>`	Formal validation of annotation files

where: `<L>` = language NL/VL

The following remarks apply to the CGN database as a whole:

1) Depending on which platform the CDs are used the file names appear in upper or lower case. ISO9660 specifies upper case letters, and the file names appear in upper case letters under Windows and MacOS; however, under Linux most CDs mount as RockRidge-formatted CDs, i.e. with lower case names. This may cause problems on operating systems that distinguish upper and lower case names, e.g. Unix.

2) many signal files show 100% maximum level. This may have two reasons:
a) some signal file were normalized to 100% level (unlikely because there seems to be no pattern for this happening
b) these signals are clipped
It is interesting to note that some volumes do not contain any file that are on 100% level. Please refer to the log files in the appendix for a complete listing of the maximum levels for each signal file.

3) The naming of the volumes makes handling somewhat awkward. It would have been better to have a straight row of volume numbers, that is CGN_R2NL_01 should be better CGN_R2NL_11 and so on. It make sense to distinguish the languages; it does not make much sense to start the numbering in each release from 1 again.

4) The documentation files on the data CDs are not consistent across the corpus; in the NL releases 1 and 2 the files are named `readme` and `copyrigh.txt`; in the NL release 3 only a `readme.txt` is found; in the VL release 1 it is `info.txt` and `copyrght.txt`.

5) Several CD cases and labelings were damaged; whenever possible they were replaced by BAS.

2.1 Signal data

The data CDs were checked as follows

1) mountable as IS9660
2) correct language labels on all signal data
3) any other data that does not conform to file naming
4) correct WAV RIFF header structure
5) each data file contains a signal
6) maximum level of each data file
7) check for readme and copyright files

2.1.1 Results: NL

R1NL_03 : missing from the original delivery; it was shipped separately and arrived later.

R1NL_08 : no standard Rockridge format; in contrast to all other data CDs this volume does not mount with lower case file naming (not even when forced to)

R2NL_13 : files FN000425-429.WAV do not have a RIFF standard header

R3NL_09 : i/o error at the end of the volume; this causes the last signal data file FN000598.WAV not to be readable

R3NL_24 : i/o error in file FN001400.WAV; not readable

2.1.2 Results: VL

No errors found.

2.2 Signal vs. FRG-files

Checks have been performed whether

1) each signal file is listed correctly in the FRG file
2) each listed signal file in the FRG file can be found on the data CD

2.2.1 Results: NL

R2NL_16 : FN001139.WAV is listed as belonging to R2NL_20

R2NL_16 : FN001140.WAV is listed as belonging to R2NL_20

2.2.2 Results: VL

No errors found.

2.3 Annotation data

Both annotation CDs CGN_R3_ANN1 and CGN_R3_ANN2 have been checked for

3) each listed annotation file in FRG file does actually exist on CD
4) all annotation files found on CD are listed correctly in FRG file
5) any superfluous files found not matching the required nomenclature
6) any utterance that has an ORT annotation file but no PRI or SKP files

7) any utterance that has an POS annotation file but no TAG file

2.3.1 Results: CGN_R3_ANN1 (NL)

FN000245.ORT but no FN000245.PRI

FN000397.ORT but no FN000397.PRI

FN000397.ORT but no FN000397.SKP

FN000398.ORT but no FN000398.PRI

FN000398.ORT but no FN000398.SKP

FN000399.ORT but no FN000399.PRI

FN000399.ORT but no FN000399.SKP

A total of 97 annotation directories ANNOT/??? were found which contain no annotation files but a single, empty file named '_'.

2.3.2 Results: CGN_R3_ANN2 (VL)

No errors found.

2.4 Annotation data vs. FRG

Checks have been performed whether

1) each annotation file is listed correctly in the FRG file
2) each listed annotation file in the FRG file can be found on the annotation CD

2.4.1 Results: CGN_R3_ANN1 (NL)

FN000163.POS not listed in file FRG_NL.TXT

2.4.2 Results: CGN_R3_ANN2 (VL)

FV600014.SYN is listed in FRG file but does not exist

FV600029.SYN is listed in FRG file but does not exist

2.5 Structural validation of documentation files

The documentation files consist of plain text (.TXT), Excel-formatted (.XLS), PostScript (.PS), Portable Document Format (.PDF), and HTML (.HTM) files plus auxiliary image files. The documentation directories are INFO/ and META/CODES/ on the annotation disks CGN_R3_ANN{1|2}.

2.5.1 Root directory

README.TXT contains two small errors:

1) in "*(in .ps en .pdf format)*" the "*en*" must be replaced by "*and*".
2) the name of directory on CD is FRQ_LST, not FREQ_LST as stated.

In LEESMIJ.TXT the directory FRQ_LST is given as FREQ_LST.

2.5.2 INFO/ Directory

The HTML structure in the INFO/ directory contains many broken links. These broken links are either real broken links, e.g. the relative links to the background image in the files PAGINAS/COREX.HTM, PAGINAS/FREQ_LST.HTM, PAGINAS/LANDCODE.HTM, PAGINAS/OTS.HTM, and PAGINAS/PRAAT.HTM should be ../IMAGES/ACHTERGR.JPG instead of simply ACHTERGR.JPG, or they are broken because of the upper vs. lower case file name problem of the CD-ROMs, e.g. achtergr.jpg in the link in the HTML file vs. ACHTERGR.JPG on CD. Furthermore, there are many anchors in HTML files that are not being referenced, e.g. fon, pos, syn, wav, wrd, xml in FORMATS.HTM. For a full list of errors see the appendix.

The frame version of the HTML structure does not work on all browsers or platforms; apparently the script in BANNER.HTM is incorrect.

Browser	Platform	works with frames	works without frames
Internet Explorer 5	Windows	no	yes
	Mac	no	yes
Netscape Navigator 4 or newer	Windows	yes	yes
	Mac	no	yes
	Linux	yes	yes
Opera 5	Windows	no	yes
	Mac	no	yes
	Linux	no	yes

Table 1 Browser and Platform compatibility for HTML-structure

In file FORMATS.HTM an incorrect LaTeX code for &euml; is given: \"{e| instead of \"{e}.

2.5.3 META/CODES/ Directory

The files LANDEN, POSTC_NL, and POSTC_VL were found to be correct in both XLS and TXT format.

2.5.4 META/FRG_SPR/ Directory

We checked the consistency of the FRG_??.TXT files with the orthographic annotation files. A number of errors were found in the Dutch (NL) files:

3) there are ill-formatted speaker IDs with 6 digits instead of the required 5, and
4) speaker IDs were missing from FRG that are in the corresponding ORT file, and there were speaker IDs in the ORT file which do not occur in the ORT file.

The list of errors is in the appendix.

2.5.5 META/FRQ_LST/ Directory

For the validation of the frequency count files we implemented perl scripts. The results obtained with these scripts differ considerably from the frequency tables provided on the disks.

The deviations can be classified as follows:

1) Words marked as new, e.g. "*spijkerfabriek*n*" which occur in the annotations, but not in the frequency tables
2) Illegal escape sequences instead of ISO8859 characters or named entities, e.g. "*\e'\e'n*" instead of `&eacute;&eacute;n`
3) Typos, e.g. "*Background*" vs. "*background*" or "*Reykjav&iacute;k*" vs. "*Reykjavik*" in `TOTALPH.FRQ`
4) Non-letter characters, unmatched pairs of quotation marks or brackets, e.g. "*(de*", "*"bleeps*"
5) Other words, e.g. "*claxons*", "*da's*" which are not in the frequency tables.

The character sets for the lexicon, the frequency tables, and the orthographic annotation files differ considerably. The lexicon contains 85 different characters (7-bit ASCII characters plus 20 named entities), the frequency table `TOTALPH.FRQ` contains 99 characters, and the annotation files contain 110 characters. The character sets are given in the appendix.

Because of these differences, the frequency tables were not validated further.

3 Content validation

The content validation covers all checks that concern the contens of signal and annotation data.

A sub-sample of 181 minutes and 25 secs of speech together with corresponding anntotation files were selected using a random scheme as described below.

These data were checked acoustically as well as with regard to the existing annotation files.

3.1 Selection scheme

According to table 2 of the specs for this evaluation the following distribution of samples has been derived randomly from the 14 different components of the corpus:

	Component	Number of samples
NL	1	12
	2	6
	6	6
	13	12
	14	6
VL	1	12
	2	3
	5	3

	6	3
	7	3
	9	4
	11	4
	12	4
	14	6

Table 2 Selection of samples per component

Using a pseudo-random sequencer and selection probabilities based on the above distribution the following samples were selected for detailed evaluation.

NL		VL	
FN000055.ORT	FN000161.ORT	FV400027.ORT	FV600134.ORT
FN000056.ORT	FN000172.ORT	FV400036.ORT	FV600135.ORT
FN000057.ORT	FN000211.ORT	FV400050.ORT	FV600180.ORT
FN000060.ORT	FN000250.ORT	FV400061.ORT	FV600196.ORT
FN000061.ORT	FN000316.ORT	FV400062.ORT	FV600201.ORT
FN000062.ORT	FN000324.ORT	FV400063.ORT	FV600255.ORT
FN000063.ORT	FN000342.ORT	FV400066.ORT	FV600285.ORT
FN000064.ORT	FN000353.ORT	FV400067.ORT	FV600306.ORT
FN000065.ORT	FN000358.ORT	FV400069.ORT	FV600317.ORT
FN000068.ORT	FN000390.ORT	FV400070.ORT	FV600338.ORT
FN000072.ORT	FN000419.ORT	FV400072.ORT	FV600358.ORT
FN000074.ORT	FN000450.ORT	FV400073.ORT	FV600390.ORT
FN000086.ORT	FN000457.ORT	FV400074.ORT	FV600580.ORT
FN000094.ORT	FN000519.ORT	FV400076.ORT	FV600647.ORT
FN000105.ORT	FN000602.ORT	FV400079.ORT	FV600740.ORT
FN000110.ORT	FN001022.ORT	FV400134.ORT	FV800008.ORT
FN000114.ORT	FN001070.ORT	FV400136.ORT	FV800098.ORT
FN000127.ORT	FN001115.ORT	FV400142.ORT	FV800247.ORT
FN000145.ORT	FN001142.ORT	FV600039.ORT	FV800352.ORT
FN000153.ORT	FN001160.ORT	FV600059.ORT	FV800528.ORT
FN000159.ORT	FN001197.ORT	FV600111.ORT	FV800728.ORT

Table 3 Files selected for validation

This resulted in a total number of 42 NL and 42 VL samples of each maximum 140 sec length (if a selected sample had more than 140 sec length, only the first 140 sec were selected).

3.2 Structural validation of annotation files

As a first step, the ORT, POS and PRI annotation files were checked whether they conform to the structure given in the documentation files, namely the `FORMATS.HTM` pages.

3.2.1 ORT files

1) The documentation on the HTML pages is not sufficient: other identifiers than speaker name are allowed, e.g. "*COMMENT*", "*UNKNOWN*", "*BACKGROUND*"
2) Illegal value for begin time: "*0*" instead of "*0.000*"
3) Strange decimal values for many time marks: 14 decimal digits, the last 10 being "*9999999999*"
4) The following list contains files that were not saved in DOS but in UNIX format

```
FN000132.ORT
FN000175.ORT
FN000206.ORT
FN000076.ORT
FN001178.ORT
FN000494.ORT
FN000737.ORT
FN000738.ORT
FN001074.ORT
FN001086.ORT
FN001090.ORT
FN001115.ORT
FN001123.ORT
FN001131.ORT
FN001134.ORT
FN001135.ORT
FN001138.ORT
```

3.2.2 POS files

The POS files are all in UNIX format, not DOS.

6) `<mu .*>` lines not mentioned in `FORMATS.HTM`
7) The following list contains files that contain misformatted lines (less than three tab-delimited items)

```
FN000029.POS
FN000043.POS
FN000055.POS
FN000196.POS
FN000199.POS
FN000368.POS
```

3.2.3 PRI files

The PRI files were checked with an XML parser for syntactic correctness. Note that this parser was not a validating parser. No errors were found, but one file (`FV400138.PRI`) caused the parser to run out of memory. Note that this may not be an error in the file, but a limitation of the XML parser used.

3.2.4 COREX software

We installed the COREX software on a Windows NT 4.0 PC according to the installation guidelines given on the CD. The installation procedure worked well. We did not use COREX extensively, but during our use we found the following problems:

1) Software requires directory `C:\TMP`, even if installed somewhere else

2) "*Stop Search*" button does not stop search – at least the progress bar keeps progressing.

3) Using a regular expression does not work properly, e.g. search for "*Nij**” found "*(ni)euw...*” etc., but also "*(nar)coti...*”. The brackets indicate the part of the word that was highlighted as a result of the search.

3.3 Content validation of annotation files

The content validation consists of a verification of the orthographic transcription (as found in the `ORT` files), and a verification of the Part-Of-Speech annotations (as found in the `SYN` files).

For the content verification native speakers of Dutch and Flemish were recruited. They were either known to the Institute because they had participated in previous projects, or acquaintances of these persons, or recommended by Prof. Carel ter Haar of the "Institut Deutsch als Fremdsprache”, who teaches Dutch at Munich university.

3.3.1 Orthographic Transcription

For the verification of the orthographic transcription the `ORT` files were read into a DBMS. The corresponding audio files were split according to the chunk segment boundaries specified in the `ORT` files.

For the 84 selected files, a total of 6306 segments were extracted from the ORT files. 1820 of them contained empty segments because they correspond to speech pauses or to BACKGROUND and UNKNOWN tiers. 4486 non-empty segments were validated, 2685 for Dutch, and 1801 for Flemish.

In a web-based editor, the audio signal was presented both graphically and acoustically; the validator could listen to the audio file as often as wanted. The orthographic transcription was presented beneath the signal display in its original form, together with two groups of check boxes to enter the verification result.

The validator was asked to perform a segmental and a content check. If necessary, the validator could enter his or her modifications to the original transcript. This input is analyzed for formal consistency by a tokenizer built into the validation tool.

The segmental validation checked the following issues:

1) Are the segment boundaries at chunk boundaries and between words?

2) Is the signal quality sufficient for auditory verification?

3) Does the transcription match the speech presented or is it totally different?

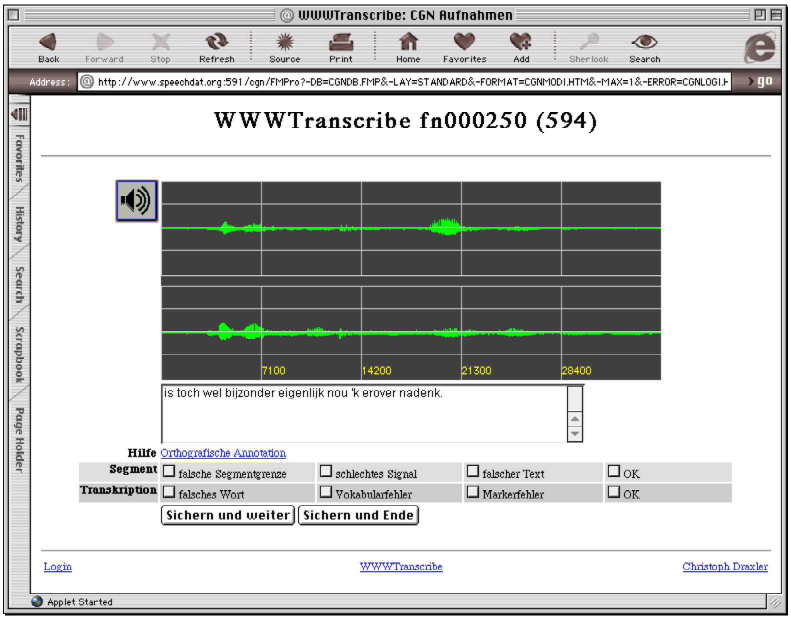
The content validation checked the following issues:

1) Are there missing or extra words in the transcription?

2) Are the words used in the transcription valid Dutch or Flemish words?

3) Have the markers (if any) been used correctly?

Once a verification was submitted to the DBMS the next utterance was presented. It was not possible for the validators to return to a previous validation and change it.



http://www.speechdat.org:591/cgn/FMPro?-DB=CGNDB.FMP&-LAY=STANDARD&-FORMAT=CGNMODI.HTM&-MAX=1&-ERROR=CGNLOGI.H

WWWTranscribe fn000250 (594)

is toch wel bijzonder eigenlijk nou 'k erover nadenk.

Hilfe Orthografische Annotation

Segment	□ falsche Segmentgrenze	□ schlechtes Signal	□ falscher Text	□ OK
Transkription	□ falsches Wort	□ Vokabularfehler	□ Markerfehler	□ OK

Sichern und weiter | Sichern und Ende

Login | WWWTranscribe | Christoph Draxler

A total of 4486 chunk segments were verified by at least two validators to result in 5370 verified orthographic transcriptions for Dutch, and 3695 for Flemish. If the validators clicked on a button other than "ok" the corresponding transcription was considered as containing an error. If the validators clicked on more than one button in a button group this was counted as one error only.

Language	Verifications	Segmentation ok	Segmentation error rate	Transcription ok	Chunk error rate
Dutch	5374	5337	0.7%	5073	5.5%
Flemish	3695	3684	0.2%	3426	7.3%

Table 4 Results for segmentation and transcription verification

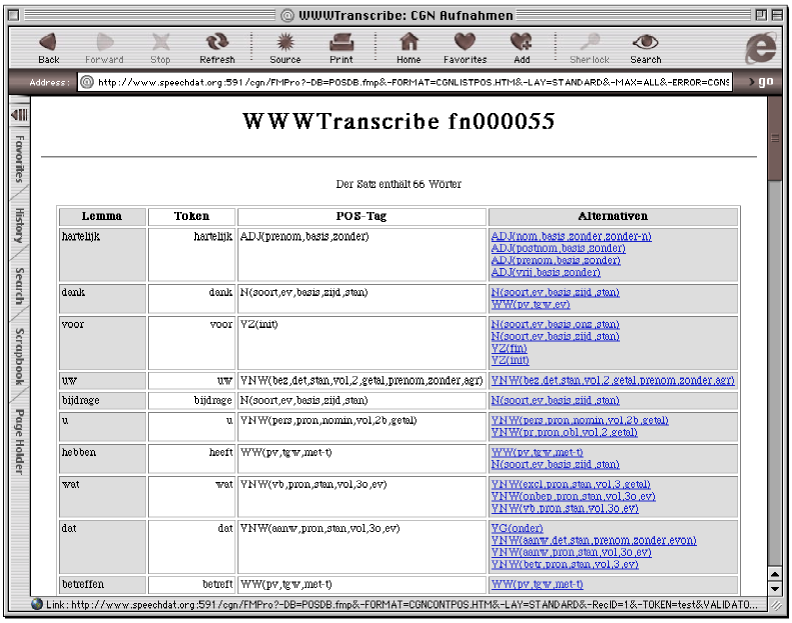
Foreign language words are not always marked as such, e.g. "*cash flow*".

The overall quality of the orthographic transcription is very good. It is debatable whether all transcription errors are really errors: in many cases our validators tried to identify words in utterance fragments that were marked as incomprehensible in the original transcription. The original transcribers were explicitly requested not to waste time on utterances difficult to understand.

The transcription validation log file is in the appendix.

3.3.2 POS-Tagging

For the verification of the POS-tagging the individual sentences of the selected recordings were presented in a table with four columns: lemma, token, POS-tag, and potential POS-tags as found in the lexicon. The validator was asked to check the lemma and the chosen POS tag; if the POS tag was incorrect, then he or she could click on the correct POS tag which was then submitted to the DBMS.


WWWTranscribe: CGN Aufnahmen
Back
Forward
Stop
Refresh
Source
Print
Home
Favorites
Add
Sherlock
Search
Address: http://www.speechdat.org:591/cgn/FMPro?-DB=POSDB.fmp&-FORMAT=CGNLISTPOS.HTM&-LAY=STANDARD&-MAX=ALL&-ERROR=CGNS
go
Favorites
History
Search
Scrapbook
Page Holder
Link: http://www.speechdat.org:591/cgn/FMPro?-DB=POSDB.fmp&-FORMAT=CGNCONTPOS.HTM&-LAY=STANDARD&-RecID=1&-TOKEN=test&VALIDATO...

WWWTranscribe fn000055

Der Satz enthält 66 Wörter

Lemma	Token	POS-Tag	Alternativen

hartelijk	hartelijk	ADJ(prenom,basis,zonder)	ADJ(nom,basis,zonder,zonder-n)
			ADJ(postnom,basis,zonder)
			ADJ(prenom,basis,zonder)
			ADJ(vrij,basis,zonder)

dank	dank	N(soort,ev,basis,zijd,stan)	N(soort,ev,basis,zijd,stan)
			WW(pv,tgw,ev)

voor	voor	VZ(init)	N(soort,ev,basis,onz,stan)
			N(soort,ev,basis,zijd,stan)
			VZ(fin)
			VZ(init)

uw	uw	VNW(bez,det,stan,vol,2,getal,prenom,zonder,agr)	VNW(bez,det,stan,vol,2,getal,prenom,zonder,agr)
bijdrage	bijdrage	N(soort,ev,basis,zijd,stan)	N(soort,ev,basis,zijd,stan)

u	u	VNW(pers,pron,nomin,vol,2b,getal)	VNW(pers,pron,nomin,vol,2b,getal)
			VNW(pr,pron,obl,vol,2,getal)

hebben	heeft	WW(pv,tgw,met-t)	WW(pv,tgw,met-t)
			N(soort,ev,basis,zijd,stan)

wat	wat	VNW(vb,pron,stan,vol,3o,ev)	VNW(excl,pron,stan,vol,3,getal)
			VNW(onbep,pron,stan,vol,3o,ev)
			VNW(vb,pron,stan,vol,3o,ev)

dat	dat	VNW(aanw,pron,stan,vol,3o,ev)	VG(onder)
			VNW(aanw,det,stan,prenom,zonder,evon)
			VNW(aanw,pron,stan,vol,3o,ev)
			VNW(betr,pron,stan,vol,3,ev)

betreffen	betreft	WW(pv,tgw,met-t)	WW(pv,tgw,met-t)

A total of 808 POS tagged sentences were checked.

Language	POS tagged sentences	POS tags	Incorrect tags	Error rate
Dutch	348	7086	19	0.3%
Flemish	460	5829	11	0.2%

Table 5 Summary of POS tag verification

Note that not all recordings selected by the randomized selection algorithm were POS tagged. Intermediate POS tag files were provided upon request by CGN (Antal van den Bosch) for all files, and they were used to implement the verification procedure. However, because they were intermediate files, they were not validated.

The POS validation log files are in the appendix.

4 Summary

The CGN project is an ambitious task which when finished will definetely boost the development of speech and language technology and encourage basic and applied research for Dutch and Flemish.

Compared to validation results in SpeechDat projects and other comparable speech data collection efforts, the third release of the CGN corpus shows in both Dutch and Flemish components good to very good results.

Most of the errors found in the formal validation can be easily corrected; no errors were found that exclude the usage of certain recordings.

The quality of the manual annotations is within the usually expected and accepted limits.

More work is needed in these areas:

1) Documentation is not sufficient. The HTML structure is very useful, but does not include corpus specifications, recording protocols, etc.

2) The consistency of frequency count files vs. lexicon vs. annotation files has to be improved.

In this context, a new calculation of the frequency lists is needed.

5 Recommendations

The following recommendations range from general to more specific issues:

Publish all specifications, including a priori corpus specifications, recording protocols and a posteriori log files describing deviations from the plan.

Consider future extensions to the corpus. The experience from Verbmobil and SmartKom shows that the value of a corpus multiplies with the number of additional representation or annotation levels. Therefore, provide a flexible multi-tier annotation interchange format that also allows non time-aligned representations that can be linked to time-aligned representations.

Make releases available to partners really working with the data immediately to find errors that cannot be detected by automatic checkers. Install a bug reporting mechanism and encourage reporting bugs, e.g. with a reward for every bug found.

Annotations change, signals don't. Hence it makes sense to provide the annotations on-line in the most recent version. It is not clear whether a version control system is implementable for corpus data, but every result obtained must be accompanied by the version (version number, date released, source, etc.) of the data used.

Perform automatic procedures and checks wherever possible, e.g. format checkers for signal files, site analysis tools for HTML structures, a validating parser for XML files, perl scripts for the generation of frequency lists, etc. These tools and scripts ideally should be platform independent and be provided together with the corpus.

The consistency issue is of particular importance for files that are used in further processing steps, e.g. the orthographic annotation in Praat files which are the basis of all frequency counts, phonetic segmentations, and POS-tagging.

Subcorpus (and supercorpus) data extraction is becoming increasingly important as the size of corpora and their number grows. Although this seems to be an issue that concerns the distributor only this is not the case: data model, access structures, tools, and documentation have to support this type of data extraction. Eudico is a good starting point, but it needs to be extended to include additional representation levels, e.g. linguistic and phonetic annotations, lexicon access, etc.

WAV or NIST format? WAV certainly is easy to use in web-based tools, and it is cross-platform. However, we doubt that many standard audio tools actually can process non-standard WAV-files, e.g. multi-channel recordings, and many signal processing software packages require NIST files. A viable option is to provide platform independent software that translates WAV to NIST format.

The volume numbering should be re-considered; a release independent numbering is more usable than a numbering within releases.

The file naming problem should be addressed. We would recommend to treat all file names (eg. in links) with capital letters according to ISO 9660 (Level 1).

For a corpus of this size it might simplify the handling if the signal data is distributed on DVD-R instead of CDROM.

We miss a detailed documentation about the technical setup of the recordings. For instance in recordings of radio shows it would be interesting to know about the exact procedure of the recording:master signal of the broadcasting station, recording via digital broadcasting (which type of receiver, converter, etc.), recording via analog receivers etc.

For online recordings the equipment should be named: type of microphone, filter (if any), A/D card, system software, recording software, known problems (for instances hums, noises etc), room charcteristics, distance to microphones, possible post-processing etc. Why are many signals clipped and others not?

6 Appendix: List of log files

File	Description
`AnnotValidation.log{NL|VL}`	Overview of annotation files required to be on CD according to `FRG` database
`FormalValidation.log{NL|VL}total`	Formal validation of signal files
`Errors_in_INFO.pdf`	List of errors found in the HTML structure. This list is the result of an automatic analysis by the Adobe SiteMill 2.0 software applied to the R3_CGN_ANN2 CD-ROM copied to a local hard disk under MacOS 9.2.
`checkORT.txt`	log file of formal check of orthography files
`checkPOS.txt`	log file of formal check of POS-tag files
`LexiconCharSet.txt`	Character set found in the token column of the CGN lexicon (v. 8.1)
`TotalphCharSet.txt`	Character set found in the TOTALPH.FRQ file
`AnnotCharSet.txt`	Character set found in the orthographic annotation files
`check{ORT|POS|PRI}.txt`	structural validation of `ORT`, `POS`, and `PRI` files
`orthoLog.txt`	extract of log files for orthographic transcription
`posLog.txt`	extract of log files for POS tags
`checkSPKIDs.txt`	formal check of speaker IDs in the FRG files and the ORT files; a positive value means the speaker ID occurred in the ORT file but not the FRG file, a negative value means the speaker ID is in the FRG file but not the ORT file.

Table 6 List of log files