Kieler Arbeiten zur skandinavistischen Linguistik 6

Iben Nyholm Debess

FADAC Hamburg 1.0

Guide to the Faroese Danish Corpus Hamburg (Version 1.0)

Kiel 2019

Christian-Albrechts-Universität zu Kiel
Institut für Skandinavistik, Frisistik und Allgemeine Sprachwissenschaft

Iben Nyholm Debess
Kiel University
Institute of Scandinavian Studies, Frisian Studies and General Linguistics

Published 18 March 2019

DOI: 10.21941/publ/kasl6

Die *Kieler Arbeiten zur skandinavistischen Linguistik* publizieren Arbeitspapiere und Forschungsergebnisse aus der skandinavistischen Linguistik am Institut für Skandinavistik, Frisistik und Allgemeine Sprachwissenschaft der Christian-Albrechts-Universität zu Kiel.

Contents

1 FADAC Hamburg description

This paper provides a short description of the *Faroese Danish Corpus Hamburg* (FADAC Hamburg), Version 1.0, which can be accessed online via the permanent identifier (PID) http://hdl.handle.net/11022/0000-0007-D0D5-D.

FADAC Hamburg is a corpus corpus of spoken, informal Faroese and Faroese Danish (469,000 words) that has been collected as part of the project *Variation in the Multilingualism on the Faroe Islands* (Project K8) at the Research Centre on Multilingualism at the University of Hamburg.[1] Earlier versions (numbered 0.2) of FADAC Hamburg subcorpora (FADAC-DAN and FADAC-FAR) were published online through the Hamburg Center for Language Corpora in cooperation with the Specialist Information Service Northern Europe at Kiel University Library. Due to technical problems, however, access was limited to smaller parts of the data. The current version was prepared as part of a project at Kiel University, which aimed at converting legacy corpus data into sustainable formats and ensuring internal consistency.[2] The corpus is intended for use with EXMARaLDA (exmaralda.org).

The corpus consists of 92 informal interviews (20–60 minutes) that were conducted with 56 speakers from three generations (70+, 40–50, 16–21) and four regions of the island, the participants being evenly distributed with regard to age and gender. All of the interviewees have Faroese as L1 and Danish as L2. For 37 of the participants, there are recordings in both Faroese and Danish, for another 15 there are recordings in Faroese only, and with four participants the interview was only conducted in Danish. Apart from questions about the speakers' sociolinguistic characteristics and language backgrounds, only few predefined questions about school, hobbies, and children's games were asked in the interviews. Otherwise, the interviews (recorded on the Faroe Islands from 2005 to 2009) were conducted as informal talk about the Faroe Islands, trips abroad, books, and the Second World War (with the oldest generation).

The corpus comprises two subcorpora based on the language used in the communications, i.e. (Faroese) Danish vs. Faroese.

Socioeconomic and sociolinguistic data (such as age, place of birth, ancestry, education, occupation(s), time spent in countries other than the Faroe Islands and contact with Danish and other languages) as well as the (linguistic) conditions pertinent to the communication (including information about the language(s) used in the interview, the first/second/third languages of the interviewee and the interviewer) are also part of the corpus.

---

1 The K8 project (2005–2011) was led by Kurt Braunmüller and funded by the German Research Foundation as part of the Research Centre on Multilingualism at the University of Hamburg. The data was collected, transcribed and annotated by Hjalmar Petersen and Karoline Kühl as well as various student assistants.

2 The project (2017–2018) was funded by the German Research Foundation (DFG), hosted by the Specialist Information Service Northern Europe at Kiel University Library, led by Ruth Sindt and supervised by Steffen Höder, in cooperation with the Hamburg Center for Language Corpora. Iben Nyholm Debess was responsible for carrying out the technical work.

The transcriptions of the Faroese interviews differ from the Danish interviews with regard to transcription conventions. There are also some differences between the interviews of the different generations within the Faroese material. See more in Section 3.3 of this document.

An overview of number and length of the interviews:

	Generation	Number of interviews	Length (h:m)
Faroese	Young	16	5:37
	Mid	21	16:29
	Old	15	9:16
	Total	*52*	*31:23*
Danish	Generation	Number of interviews	Length (h:m)
	Young	12	6:08
	Mid	17	7:41
	Old	11	10:58
	Total	*40*	*24:49*

Tab. 1: Overview of number and length of interviews (Faroese and Danish)

DOI: 10.21941/publ/kasl6

2 File names and folder structure

2.1 File names

Every informant in the project has a pseudonym, example: *MO2T*. The pseudonyms are made to indicate the informants' gender, generation and regional affiliation.

Gender	Generation	Number	Region
M/W	Y/M/O	2	T/V/SU/EYN
Man/Woman	Young/Middle/Old	(The numbers are used to differentiate informants with the same background characteristics)	Tórshavn/Vágar/ Suðuroy/Eysturoy-Norðuroyggjar

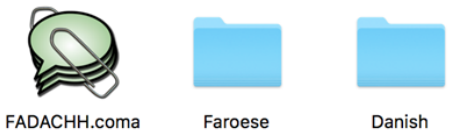
Tab. 2: Pseudonym structure

The interviews are named after the informant(s) in the interview followed by either *_f* or *_d* depending on the interview being in Faroese or in Danish, respectively, example: *MO2T_d*.

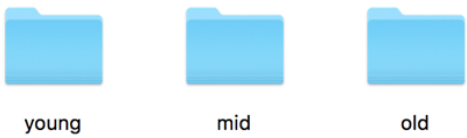
2.2 Folder structure

Important: The file names and folder structure is to be maintained at all times when working with FADAC Hamburg.

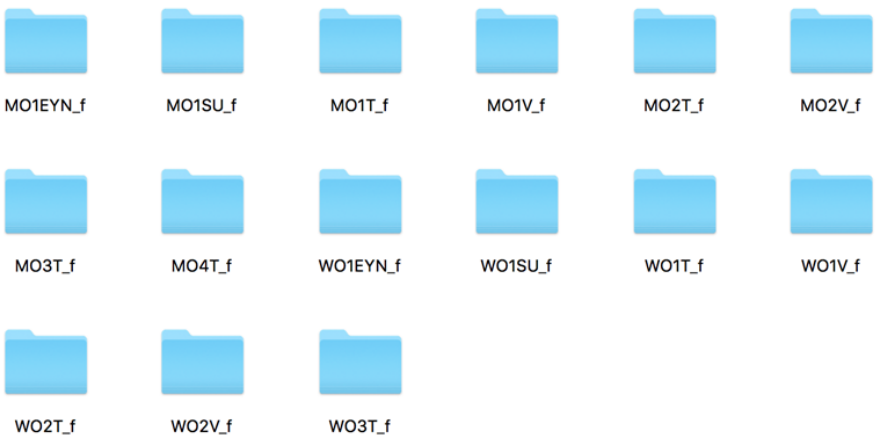
The first folder level contains one folder for the Faroese material and one folder for Danish material as well as the CoMa-file for the whole corpus. This file contains all the metadata (see also Section 3.2).



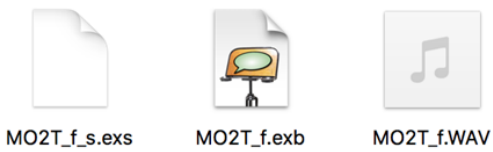
The second folder level contains three folders, one for each generation (young, mid and old).



The third folder level contains a folder for each interview. The folders have the same name as the interviews. The example shows the content of the folder *Faroese\old*.



The last folder level contains three files for each interview: a WAV-sound file, a basic transcription file and a segmented transcription file. The sound file and basic transcription file have the same name as the interview (e.g. MO2T_F) and the segmented transcription file has the same name as the interview followed by *_s* (e.g. MO2T_F_*s*).



DOI: 10.21941/publ/kasl6

3 How to use FADAC Hamburg

To work with FADAC Hamburg you need to download the EXMARaLDA system at ⟨exmaralda.org⟩. You will need the three tools:

a. PartiturEditor (transcription and annotation tool);
b. CoMa (Corpus Manager, a tool for managing the corpora);
c. EXAKT (query and analysis tool).

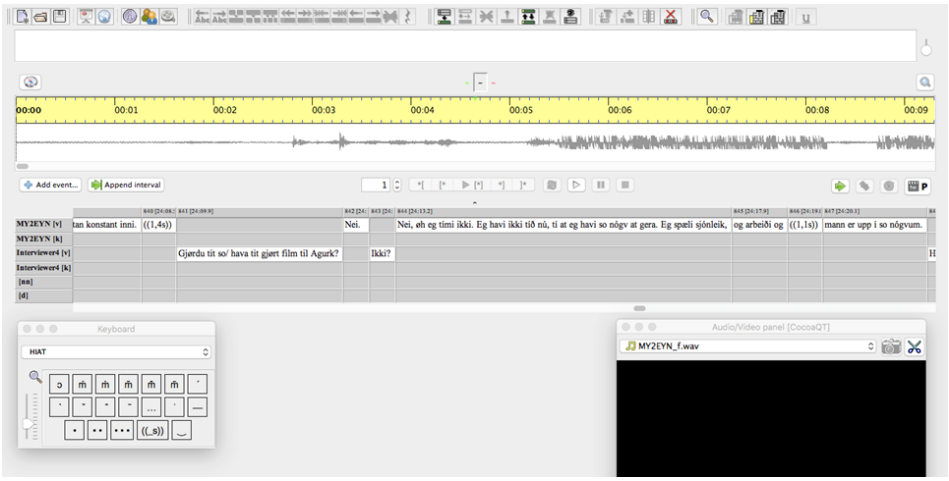
This document will give a brief introduction to the functions of each tool and how they are relevant for FADAC Hamburg . Too see extensive user manuals, see ⟨exmaralda.org/en/manuals-and-tutorials⟩.

3.1 PartiturEditor

PartiturEditor is a transcription and annotation tool. All the interviews are transcribed with this tool. See more about the transcription conventions in section 5 of this document.

PartiturEditor has many features, and can be used for e.g. annotation and exporting the transcriptions into other formats like FOLKER, Praat TextGrid, TEI (XML) and other.

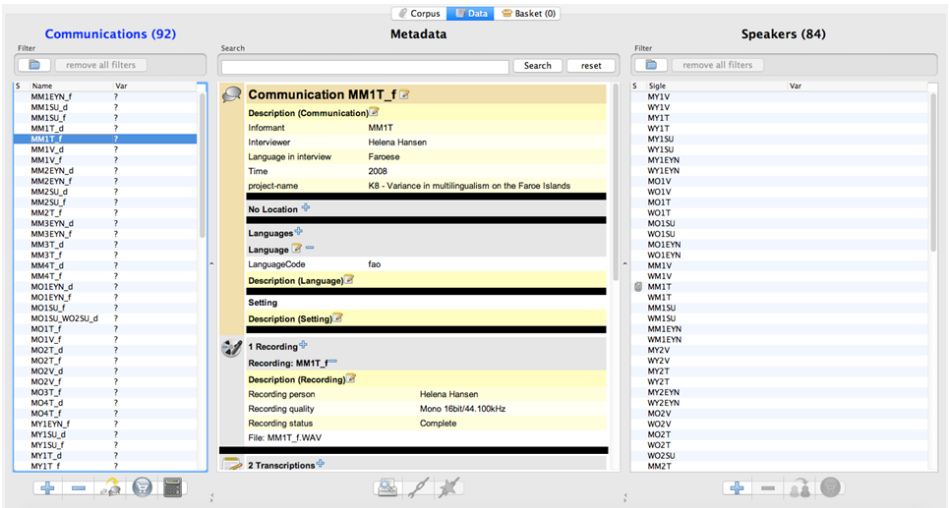
Tip: When working with the files of FADAC Hamburg, it is preferable to use the CocoaQT Audio player. This will usually be the default player, as seen in the example below.



3.2 CoMa (Corpus Manager)

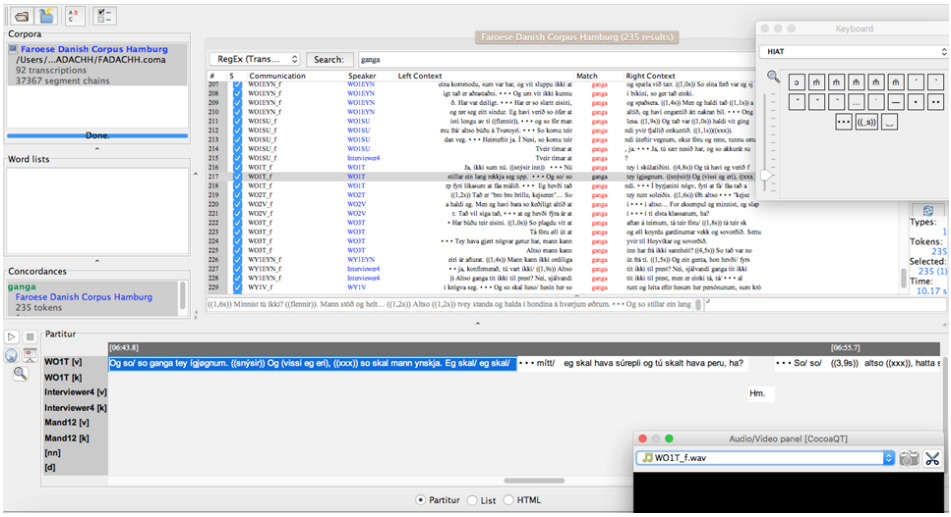
Corpus-Manager is a tool for managing the files and metadata of the corpus. All the information about the corpus, the interviews and the informants is stored in the CoMa file FADACHH in the first folder level. The file has information that links interviews and informants, it shows all the accessible information on the informants, and links sound files and transcription files to the metadata. The file contains details on the status of every interview. Working with the CoMa file is a great way to get an overview of the material in FADAC Hamburg.

DOI: 10.21941/publ/kasl6



3.3 EXAKT

EXAKT is a query and analysis tool. Opening the previously mentioned CoMa file FADACHH with EXAKT enables the search possibilities. This is probably the most usable tool for working with FADAC. You can make your own concordances and save your own annotations. By double clicking any example in the search results, you can listen to the aligned audio. Make sure to read the user manual for EXAKT to get the full potential of the tool.



If you experience difficulty and errors with these tools when working with the corpus, make sure the file names and folder structure is intact as described in Section 2.

DOI: 10.21941/publ/kasl6

4 Sound files

The sound files in FADAC Hamburg vary in length and quality. All the sound files in the corpus are 16 bit and have a 44.1 kHz sample rate. The number of audio channels varies, some sound files are in stereo and some are in mono. For information about the individual sound file quality and length, see the FADACHH CoMa file. Each interview has a descriptive section for the recording.

The surrounding settings for the interviews and sound files varies and some of the sound files may have significant background noise.

5 Transcription information

As FADAC Hamburg was part of the K8 project, the transcription conventions are based on the K8 conventions. In general, the K8 transcriptions follow the HIAT conventions as described in the handbook by Rehbein et al. (2004) (available in a PDF version at ⟨exmaralda.org/de/hiat⟩). In some points, the transcriptions of FADAC Hamburg deviate from HIAT, though. These are the following:

a. The utterance end sign and superscript period are not part of the K8 conventions.
b. Words are transcribed according to Danish or Faroese Standard orthography. If the pronunciation deviates strongly or the speaker produces a 'wrong' form of a word, this is noted in the comments tier in double round brackets (e.g. *((opvokst))* instead of *opvokset*).
c. Three (HIAT) bullets are used for pauses with a length below 1 second, not specifying the length of the pause further.
d. Emphatic stress is represented by capitals, used on syllable basis.
e. Lengthening is always represented by three repetitions of the letter (not distinguishing between the lengthening by two and three repetitions).
f. Unintelligible parts are indicated by *((xxx))* – not exceeding three repetitions of *x*, notwithstanding the length of the unintelligible part.

5.1 Danish

The Danish transcriptions mostly comply with the standard mentioned above. Additional information is to be found in the CoMa file FADACHH.

Personal information might not be anonymized in the Danish transcriptions.

Known deviations from the HIAT conventions and K8-specifications are:

a. Lengthening might be represented with more or less than three repetitions.
b. Pauses below 1 s. may be indicated with less than three HIAT bullets.
c. The spelling might deviate from standard orthography.
d. Pauses in speech are not systematically assigned to speakers, and may be transcribed in a non-verbal tier.

Examples from the Danish transcriptions:

27 [01:36.8]	28 [01:42.5]	29 [01:44.6]
Var du den æh/ den/ den æhh/ nummer hvad i rækken af børnene var du?		I midten altså.
	Jeg er nummer fire.	

30 [01:45.8]	31 [01:46.4]	32 [01:47.5]	33 [01:52.4]	34 [01:56.7]
	Okay.		Blev de alle sammen på Færøerne, eller hvad?	
	((sagte))			
I midten, ja.		Vi var fem drenge og tre piger.		Nej, • • • ikke

120 [04:04.6]	121 [04:06.2]	122 [04:0	123 [04:07.3]
Har du SELV været på sådan en skole?		Nej.	
	Jeg har ikke været der endnu.		Det er min æøh/ det er min plan?

124 [04:09.2]	125 [04:09.7]	126 [04:13.9]
Okej.		
((smækker med tungen))	Jeg æææhmm om måske fire fem år, så ønsker jeg	går til California.

5.2 Faroese

The Faroese transcriptions conventions are not the same for all generation: the mid generation differs from the young and the old generation.

5.2.1 Mid generation

The Faroese mid generation mostly complies with the standard mentioned above. Known deviations from the HIAT conventions and K8 specifications, that have been described above, are:

DOI: 10.21941/publ/kasl6

a. Utterances might lack utterance end signs.
b. Utterance end signs might have been used for other purposes ('…' to indicate pause, instead of aborted utterance).
c. Utterances do not always start with capital letter.
d. The spelling is generally based on standard orthography, although there are known deviations.

The transcriptions for the mid generation have anonymized personal names in the transcription and in the sound file, but other personal information is not anonymized.

The status of the transcriptions with regard to HIAT conventions and K8 specifications has been noted in the metadata in the CoMa file as *K8 compliant* or *K8 with deviations*.

Examples from Faroese – mid generation:

	1581 [31	1582 [31	1583 [31:56.9]	1584 [31:58.2]
altso, av foreldrum		• • •	so tað havi eg nógv…	Tað er einki at siga, tað er meira tað, at mann hevur sálina við
	ja			

	1585 [32	1586 [32:01.5]	1587 [32:0	1588 [32:02.8]	1589 [32	159
ella	• • •	gongur høgt uppí tað ha?		Eg trúgvi tað er betur úti á bygd í nógvum tilfeldi		ja
			hm hm		ja	

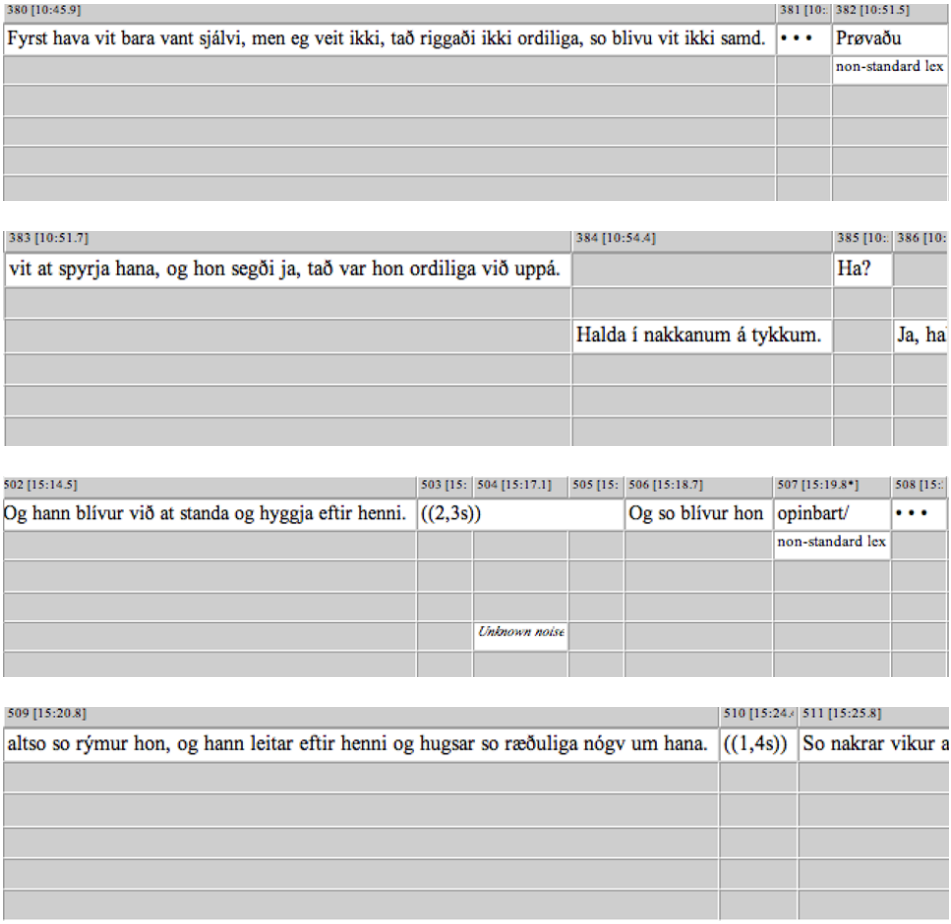
5.2.2 Young and Old generation

The transcriptions in the young and the old generation all comply with the standards mentioned above. Additionally, there are some conventions that have been made specifically for the Faroese interviews, which are described below.

All spelling in the transcriptions is based on standard orthography. Some words occur, that do not have standard Faroese spelling (loan words, hybrids etc.). These words are noted with the comment *non-standard lex* for non-standard lexical item in the annotation (k)-tier.

Inflections that differ very much from the standard are noted with the comment *non-standard infl* for non-standard inflection.

Examples:



Online Faroese dictionaries are available at ⟨sprotin.fo/dictionaries⟩. All linguistic transcription is in Faroese, but comments and extralinguistic description is written in English.

5.2.3 Additional information

5.2.3.1 Interjections

Standardized interjections and minimal responses in the Faroese transcriptions are:

øh	for all hesitation sounds
áh	

DOI: 10.21941/publ/kasl6

ja	
ná	
hm	

5.2.3.2 Anonymization

Names, telephone numbers, addresses and birth dates are anonymized both in the transcriptions and in the sound file. In the sound file the content is replaced with brown noise. In the transcription the anonymized content will be noted as follows:

the speaker's name	the speaker's pseudonym (e.g. *WY1V*)
the speaker's address	*((bústaður/adresse))*
the speaker's telephone number	*((telefonnummar/telefonnummer))*
the speaker's date of birth (not year)	*((føðidagur/fødselsdato))*
names of the speaker's immediate family members and relatives	*((navn familjulim/navn familiemedlem))*
	possibly specifying the relative like *((navn á pápa/navn på far))*
	or not *((navn))*

The following information is contained in anonymized form in the transcription:

the speaker's home town	*((bygdarnavn/býarnavn/bynavn))*
the speaker's school	*((navn á skúla/navn på skole))*
the speaker's work place	*((arbeiðspláss/arbejdsplads))*

Examples from anonymized content (the *X* in the last tier (d) indicates, that the content has also been anonymized in the aligned sound file):

28 [00:40.0]	29 [00:41.2]	30 [00:42.2*]	31 [00:42.9]	32 [00:4	33 [00:44.4]
((1,3s))	Ja. • • • Eg eiti	WY2EYN	WY2EYN.		
				• • •	Og so má eg vita, hvar tú býrt, vissi eg skal hava fatur í tær aftur?
		X	X		

34 [00:48.3]	35 [00:4	36 [00:5	37 [00:52.3]	38 [00:53.0]	39 [00:54.0]
Ja, eg búgvi á	((bústaður)).				• • • Mítt telefonnummar, ella heima?
			• • •	Og telefonnummar?	
			((xxx)).		
	X	X			

5.2.3.3 Dialect features (especially regarding the Southern dialect of Faroese)

If the dialect has specific lexical items, they will be written as they are listed in the dictionary (e.g. standard words: *vit*, *tit*, *bæði*, *skulu*; Southern dialectal words: *okur*, *tykur*, *báði*, *skulja*). These dialectal words do not follow the general pronunciation pattern distinguishing this dialect from the other dialects, and therefore the words are written as they are pronounced.

One exemption to this rule is the 1. person singular pronoun *eg/jeg*. It is difficult to distinguish between the pronunciation of the two dialectal pronouns in spontaneous speech, and therefore all 1. person singular pronouns are written as *eg*, regardless of pronunciation or known dialect.

Pronunciations that differ from the other dialects on a morphological level will not be noted, and will be written as the orthographic standard (e.g. Standard pronunciation: *lærarin*, *skrivaðu* (pl.); Southern dialect pronunciation: *leraron*, *skrivaði* (pl.)). Pronunciation differences like these are systematic and are therefore not noted in the transcription.

5.2.3.4 Standard tier structure

Presented below is the standard tier structure. Not all tiers are present in all transcriptions.

DOI: 10.21941/publ/kasl6

speaker *x*	transcription tier (v)	
	annotation tier (k)	
	description tier (nv)	only for non-verbal actions for individual speaker
interviewer	transcription tier (v)	
	annotation tier (k)	
	description tier (nv)	only for non-verbal actions for individual speaker
no speaker	description (nn)	for all non verbal descriptions of background noises, recording info etc.
	description (d)	only used for anonymizing sequences