Annotations in the LANCHART corpus

Philip Diderichsen, Torben Juel Jensen, Marie Maegaard, Janus Spindler Møller, Natalie Carmen Hau Sørensen, Peter Agerlin Trolle, Aviaja Klarskov Skotte-Nielsen, Lukas Christian Backhausen.

Translated to English by Lukas Christian Backhausen with help from ChatGPT.

LANCHART Centre, early 2023

Last update: 12.11.2024

This is an overview of the annotations of the LANCHART corpus, to be freely used by all who need detailed information about what those annotations mean.

There are two main chapters in this overview: 'Tiers and tags' and 'Metadata'. Chapter 2, 'Tiers and tags', describes annotational layers (tiers) of the corpus. The chapter 'Metadata' describes the conversational and informant-based metadata, transferred from the LANCHART informant-base, which are searchable categories in the corpus.

See chapter 1 below for a short description of the most important basic principles and terms.

Table of Contents

1 Important principles and terms

Tiers, tokens and tags

A **tier** is an annotational layer where researchers have annotated transcriptions with phonetical and grammatical categories, among others. Annotations are typically made in the phonetical annotation and analysis program Praat, where an annotational layer is called a tier. In this context, a word is called a **token**. Tokens can in certain tiers be connected to categories called **tags**. When a single token is connected to a tag, it is called **word coding**. When a tag spans several tokens, it is called **sequence coding**.

Tier names in parentheses

Evert tier is described through a user-friendly name, which is also shown in the corpus. If this name differs from the name of the Praat tier, the Praat tier name is written in parentheses. For example, ”Simpel ordklasse (RedPos)” (lit. ”simple word-class”). The name in parentheses is also what will occupy column titles in data exports from Korp.

BIO-labeling

Sequence coded Praat-tiers are recoded into a variety of the BIO-format (’Begin, In, Out’, also often called the IOB-format). This format makes it easy to find the first, last and possible middle tokens in an annotated sequence. In practice, this is done by using the following prefixes and suffixes:

- Prefix: Number in sequence (1_, 2_, 3_ and so on).
- Suffix: ”_I” for ‘in’, ”_E” for ‘end’.

When for example a sequence of three tokens is annotated with the category *dialektcitat* (lit. dialect quote), those three tokens will be annotated: 1_dialektcitat_I, 2_dialektcitat_I, 3_dialektcitat_E. The number before the tag shows the position of the tag in the sequence and the letter after the tag shows whether it is the last tag in the sequence or not.

Empty values

If a token is missing an annotation in a tier (which it commonly will), it will have the searchable value “__UNDEF__” (‘undefined’) instead. If you want to search for words that specifically AREN’T annotated in a certain tier, for example “AN-markering”, search for tokens where “AN-markering” is “__UNDEF__”.

Underscore replaces space

Space (’ ’) in annotations will result in errors in the corpus database. Therefore, spaces in the original Praat-annotations er replaced with underscore instead. For example, the generic pronoun-code “G AS” will be replaced with “G_AS”

Status level

There are three different status levels for tiers:

- **Open** means that this tier can be accessed freely without the need for a password.
- **Researcher level** means that this tier can be accessed at lanchartkorp.ku.dk, which is password protected with individual usernames and passwords.

- **Private** means that this tier is protected by a project password that is only given to employees of that project.

2 Tiers and tags

In this chapter are documentation for all annotational layers (tiers) in the LANCHART corpus. The corpus is divided into a number of project corpuses with names such as LANCHART_AMAGER, LANCHART_BYSOC and so on. Some tiers are present in all project corpuses, others only in some corpuses. The “general tiers” present in all corpuses are described below.

2.1 General tiers

- Contact: Torben Juel Jensen, tjuelj@hum.ku.dk
- Status: Åben. All tiers are freely available.

2.1.1 Ordform (ortografi) – Orthographical wordform

- Word-coded

Print out in common ortography of what informants, interviewers or passers-by say in the interview/group conversation/self-recording. There is a tier for every speaker in the recording (informant code follows in parentheses after *ortografi*.

For additional information, refer to the article Udskrivningsmanual (currently only in Danish) and the transcribing guide itself. The newest version is always available on DGCSS’ website under ‘Publikation\Manualer’ and ‘rapporter\Manualer’.

2.1.2 Kommentarer – Non-speaker Comments

- Sequence-coded (BIO-labeled). Original annotation sequence ”X X X” recoded to ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

This tier contains non-speaker-specific comments. Note: It includes recordings of errors to some extent.

Example of sequence coding: When a token is annotated as "spring i optagelsen," this token will receive the following tag: "1_ Spring_i_optagelsen_E." (translated: recording skips) The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.1.3 Comments

- Sequence-coded (BIO-labeled). Original annotation sequences "X X X" are recoded as "1_X_I 2_X_I 3_X_E" (as described in the BIO-labeling description above).

Speaker-specific comments - the tier is associated with a specific speaker. In the original tier name in Praat files (Textgrids), the speaker code appears in parentheses after "Comments."

Example of sequence coding: When a token is annotated as "siges leende," this token will receive the following tag: "1_siges_leende_E." (translated: said while laughing) The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.1.4 uncertain transcription

- Sequence-coded (BIO-labeling). The original annotation sequences "X X X" are recoded as "1_X_I 2_X_I 3_X_E" (as described in the BIO-labeling description above).

The tier is associated with a specific speaker - the speaker code appears in parentheses after 'uncertain transcription'.

The tier is generated automatically and creates boundaries at the words in the orthography that are transcribed as uncertain. In the Praat file, they are marked with '?'.

In BIO-tagging, a sequence of uncertain transcriptions looks like this: "1_?_I 2_?_I 3_?_E".

In some cases, it is a series of individual words in a row that are marked as uncertain. In such cases, it looks like this: "1_?_E 1_?_E 1_?_E".

2.1.5 IPA-lydskrift (real_IPA)

- Word-coded

The annotational layer IPA-lydskrift contains a transliteration into IPA characters (represented as numeric HTML codes) of the Praat IPA codes in the IPA tier. The annotations do not appear as a tier in the underlying TextGrids, only in the Corpus and the alternate display as an aid in deciphering the otherwise somewhat obscure phonetic transcription codes.

2.1.6 IPA

- Word-coded

This tier contains an automatically generated phonetic transcription of the words in the Orthography tier using codes that Praat can interpret as IPA phonetic transcription symbols. There is an IPA tier for each speaker. The transcription is generated based on a standard pronunciation of a given word, similar to what one might find in a dictionary. This means that the transcription may not necessarily match the actual pronunciation.

The Praat IPA codes can be seen in the "Code in Praat Documentation" column in the table below. In some cases, they have been modified in the corpus (see the "Praat Code in Corpus" column). Instances where the Praat code has been modified are indicated in bold in the tables below. [1]

[1] For further information on IPA and TtT notation in the Corpus, please refer to the manual for the Language Change Center's program Phonix (currently only available in Danish), which is a part-of-speech and phonetic transcription tagger. The Phonix manual is not publicly available but can be provided upon request. However, please be aware that there

An overview of Praats IPA-coding can be found here: https://www.fon.hum.uva.nl/praat/manual/Phonetic_symbols.html

Vowels

IPA	Code in Praat Documentation	Praat Code in Corpus	TtT	Example (ortography)
i	i	i	i	vi, hvile, hvil
e	e	e	e	det, dele, del
ɛ	\ef	\ef	E	gæs, næse
æ	\ae	\ae	z	være, vane
a	a	a	a	vand, mand
ɑ	\as	\as	A	bark, bak
y	y	y	y	lys, tys, nyse
ø	**\o/**	**\o** (cannot be shown in Praat)	**q**	**øl, pølse**
œ	\oe	\oe	Q	søn, høne, løn
œ	\Oe	\Oe	x	grøn, trøje
u	u	u	u	guld, gule, fuld
o	o	o	o	tone, rod
ɔ	\ct	\ct	c	kul, kål, kåbe
ʌ	\vt	\vt	C	tom, som

---

are errors in some of the tables in the Phonix manual regarding the notation of Praat codes as outlined in the table below.

Errors in the Phonix manual	**Incorrect**	**Correction**
Table 1	[ø] is written as \yc, but this Praat code represents the sound [ʏ].	The Praat code for [ø] is \o/ (in the Corpus, it is \o, see the table above).
	[œ] is written as \o/, but this Praat code represents the sound [ø].	The Praat code for [œ] is \oe.
	[œ] is written as \oe, but this Praat code represents the sound [œ].	The Praat code for [œ] is \Oe.
	[ə] is written as \ic, but this Praat code represents the sound [ɪ].	The Praat code for [ə] is \sw.
Table 2	[ɕ] is written as \sh, but this Praat code represents the sound [ʃ].	The Praat code for [ɕ] is \cc.
Table 3	It says /cc.	It should be \cc.
	It says /rc.	It should be \rc.

ɒ	**\ab**	**O**	**X**	**vores, båre**
ə	\sw	\sw	0	-e (”schwa”)

Consonants

IPA	Code in Praat Documentation	Praat Code in Corpus	TtT	Example (ortography)
b	b	b	b	by, op
d	d	d	d	dø, at
g	**\gs**	**g**	**g**	**gå, tak**
v	v	v	v	ve, ulv
f	f	f	f	fe, hof
s	s	s	s	sø, es
ɕ	\cc	\cc	S	sjat, tusch
h	h	h	h	hest, hø
m	m	m	m	mand, mus
n	n	n	n	næse, nisse
l	l	l	l	lå, æsel
ʁ	**\ri**	**\rc** (becomes ʀ in Praat)	**r**	**rat, rå**
p	p	p	p	på, pølse
t	t	t	t	tå, taske
k	k	k	k	ko, kat
j	j	j	ᴊ	jul, haj
ʀ*	**\rc**	**\ri** (becomes ʁ in Praat)	**R**	**mor, ser**
w	w	w	w	hav, av
ð	\dh	\dh	D	hid, tid
ŋ	\ng	\ng	N	gang, hænge

*in final position, it should be \at\nv for a non-syllabic a-schwa [ə̯].

Diacritics

IPA	Code in Praat Documentation	Praat Code in Corpus	TtT	explanation
ˈ	\'1	\'1	2	stress
ː	\:f	\:f	:	extension
ːˀ	\?g	\?g		stød

2.1.7 TtT

• Word-coded

TtT is a way to write IPA using more common characters. There is a one-to-one correspondence between IPA tiers and TtT tiers. The tables above show which TtT characters correspond to which IPA codes.

For further information on TtT notation and more, please refer to the Phonix manual (see footnote on page 5-6).

Examples

familiemenneske

IPA: {fam\'1il\?gj\swm\efn\swsg\sw}
TtT: [fam2il!J0mEn0sg0]

døgnberedskab
IPA: {d\'1\vtjnbe\rc\ae\dh\?gsg\ae\:f\?gb}
TtT: [d2CJnberzD!sgz:!b]

øveweekend
IPA: {\1\o\\:fw\swWi\:fg\efn\?gd\vt}
TtT: [2q:w0Wi:gEn!dC]

etagebyggeri
IPA: {et\1\ae\:f\?g\cc\swbyg\vt\rci\:f\?g}
TtT: [et2z:!S0bygCri:!]

2.1.8 Ordklasse (PoS) – Part of Speech

- Word-coded

In this tier, each token is tagged based on its part of speech (PoS). The PAROLE tags are used for this purpose. They consist of up to 11 characters, which function as positions for a range of fixed categories, with positions 3-11 varying depending on the higher-level part of speech. A brief explanation of PAROLE tags can be found below, while a detailed description can be accessed here (currently only available in Danish):

https://korpus.dsl.dk/documentation/PAROLE-dokumentation/paroledoc_da.pdf

Note: The part-of-speech tier should be used with caution; the part-of-speech categories in the PAROLE tag set are very detailed, but the tier has been annotated automatically, so one must be cautious about potential annotation errors.

Different parts of speech and their corresponding categories.

Part of Speech	**Sub-category**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Noun (N)	C, P	Genus: C, N, ., -	Numerus: S, P, ., -	Casus: G, U	=	=	Definite-ness: D, I, ., -			
Verb (V)	A, E	Modus: D, M, F, G, P	Tempus: R, A, -	=	Numerus: S, P, ., -	Genus: C, ., -	Definite-ness: D, I, ., -	Trans-cat-egoriza-tion: R, A, ., -	Diathe-sis: A, P, -	Casus: G, U, ., -

Adjective (A)	N, C, O	Comparative: P, C, S, A, -	Genus: C, N, ., -	Numerus: S, P, ., -	Casus: G, U, -	=	Definite-ness: D, I, ., -	Transcate-gorizing: R, U, -		
Adverb (R)	G	Comparative: U, P, C, S, A								
Preposition (S)	P									
Conjunction (C)	C, S									
Pronoun (P)	D, I, T, P, O, C	Person: 1, 2, 3, -	Genus: C, N, ., -	Numerus: S, P, .	Kasus: G, U, N	Owner nu-merus: S, P, ., -	Reflexive: N, Y, ., -	Posture: U, O, F, P, -		
Interjection (I)	=									
Unique (U)	=									
Other (X)	A, F, P, R, S, X									

The first position in a PAROLE tag indicates the overall part of speech.

The second position indicates a subcategory of the overall part of speech. The possible values will vary for each part of speech. For example, when dealing with an adjective/numeral, the second position can take one of the following values: N (Normal), C (Cardinal), or O (Ordinal). If it's a conjunction, the second position may take the following values: C (Coordinating) or S (Subordinating).

Values for positions 3-11, as mentioned earlier, vary for each part of speech (see the table above or Figure 4 on page 8 in the guide linked above). Additionally, in some cases, a value may not be relevant or clearly defined for a specific word or part of speech. In such cases, the following placeholders are used:

=	Used when a value is not relevant for an entire part of speech.
-	Used when a value is not relevant for the specific word.
.	Used when a value is ambiguous for the specific word.

An example of the use of "." is in nouns that can appear in both common and neuter gender. Words like *virus* and *parameter* are examples of such words, and they would be described with a period in the slot that specifies grammatical gender (unless there is an explicit specification like "et parameter").

Below is a table of all the different PAROLE tags that can be used in the PoS tier. The examples are primarily taken from the corpus. In cases where there were no suitable examples in the corpus, new examples have been constructed and are enclosed in square brackets.

The different PoS tags in the PAROLE tag set

PoS (PAROLE)	Description	RedPoS	Example
AC---G=--	adjective cardinal genitive. Alt: AC---G	NUM_GEN	[*Det er de <u>tos</u> pro-jekt*]
AC---U=--	adjective cardinal unmarked-case. Alt: AC---U	NUM	*<u>to</u> personer*
ANA---=-R	adjective normal absolute-superlative adverbial-use. Findes kun som ANA---	ADJ	*en af de klasser der er <u>allermindst</u> sam-menhold i*

ANA..U=DU	adjective normal absolute-superlative unmarked-case definite unmarked-use. Alt: ANA..U	ADJ	*den allerstørste virksomhed*
ANC---=-R	adjective normal comparative adverbial-use. Alt: ANC--- eller ANC	ADJ	*som nu ikke eksisterer mere*
ANC.PU=.U	adjective normal comparative plural unmarked-case unmarked-use. Alt ANC.PU	ADJ	*der var gået flere år*
ANC.SU=IU	adjective normal comparative singular unmarked-case indefinite unmarked-use. Alt ANC.SU	ADJ	*der var ikke mere plads derude*
ANC..G=.U	adjective normal comparative genitive unmarked-use	ADJ_GEN	[*Nu er det den/de ældres tur*]
ANC..U=.U	adjective normal comparative unmarked-case unmarked-use. Alt: ANC..U	ADJ	*deres regning bliver ikke større end min*
ANP---=-R	adjective normal positive adverbial-use. Alt: ANP---	ADJ	*det kunne godt være*
ANPCSU=IU	adjective normal positive common singular unmarkedcase indefinite unmarked-use. Alt: ANPCSU	ADJ	*en halv dag*
ANPCSU=.U	adjective normal positive common singular unmarkedcase unmarked-use. Alt: ANPCSU	ADJ	*være religiøs på sin egen måde*
ANPNSU=IU	adjective normal positive neuter singular unmarkedcase indefinite unmarked-use. Alt: ANPNSU	ADJ	*det er sgu rimelig sjovt*
ANPNSU=.U	adjective normal positive neuter singular unmarkedcase unmarked-use. Alt: ANPNSU	ADJ	*det er jeres eget problem*
ANP.PG=.U	adjective normal positive plural genitive unmarked-use. Alt: ANP.PG	ADJ_GEN	*for manges vedkommende*
ANP.PU=.U	adjective normal positive plural unmarked-case unmarked-use. Alt: ANP.PU	ADJ	*de/nogle små børn*
ANP.SG=DU	adjective normal positive singular genitive definite unmarked-use. Alt: ANP.SG	ADJ_GEN	*den enkeltes ideologi*
ANP.SU=DU	adjective normal positive singular unmarked-case definite unmarked-use. Alt: ANP.SU	ADJ	*det var for hele skolen*
ANP.SU=IU	adjective normal positive singular unmarked-case indefinite unmarked-use. Alt: ANP.SU	ADJ	*det er en amerikansk rapper*
ANP.SU=.U	adjective normal positive singular unmarked-case unmarked-use. Alt: ANP.SU	ADJ	*da jeg var lille*
ANP..G=.U	adjective normal positive genitive unmarked-use. Alt ANP..G	ADJ_GEN	[*Nu er det den/de stores tur*]
ANP..U=.U	adjective normal positive unmarked-case unmarkeduse. Alt: ANP..U	ADJ	*det var i sidste uge*
ANS---=-R	adjective normal superlative adverbial-use. Alt: ANS---	ADJ	*det jeg var mest interesseret i*

ANS.PU=DU	adjective normal superlative plural unmarked-case definite unmarked-use. Alt: ANS.PU	ADJ	*sådan er de fleste faktisk*
ANS.PU=.U	adjective normal superlative plural unmarked-case unmarked-use. Alt: ANS.PU	ADJ	*dem der får flest point*
ANS.SU=DU	adjective normal superlative singular unmarked-case definite unmarked-use. Alt: ANS.SU	ADJ	*for det meste sidder vi og snakker*
ANS.SU=IU	adjective normal superlative singular unmarked-case indefinite unmarked-use	ADJ	-
ANS..U=DU	adjective normal superlative unmarked-case definite unmarked-use. Alt: ANS..U	ADJ	*de/den bedste koncerter/koncert*
ANS..U=.U	adjective normal superlative unmarked-case unmarkeduse. Alt: ANS..U	ADJ	*det var lettest synes jeg*
AO---G=--	adjective ordinal genitive. Alt: AO---G	NUM_ORD_GEN	*Christian den fjerdes oprindelige byggestil*
AO---U=--	adjective ordinal unmarked-case. Alt AO---U	NUM_ORD	*syvende klasse*
CC	conjunction coordinative	SKONJ	*karakterer og årskarakterer*
CS	conjunction subordinative	UKONJ	*sjovere end syvende klasse*
I=	interjection. Alt: I	INTERJ	*ja*
NCCPG==D	noun common common-gender plural genitive definite. Alt: NCCPG	N_GEN	*det er meget forældrenes valg*
NCCPG==I	noun common common-gender plural genitive indefinite	N_GEN	*forældres skænderier*
NCCPU==D	noun common common-gender plural unmarked-case definite. Alt: NCCPU	N	*i ferierne plejer du at være sammen med dem*
NCCPU==I	noun common common-gender plural unmarked-case indefinite	N	*vi skal optræde tre gange*
NCCPU==.	noun common common-gender plural unmarked-case	N	*de/nogle pårørende*
NCCSG==D	noun common common-gender singular genitive definite. Alt: NCCSG	N_GEN	*byens historie*
NCCSG==I	noun common common-gender singular genitive indefinite	N_GEN	*en kammerats mor*
NCCSU==D	noun common common-gender singular unmarkedcase definite. Alt: NCCSU	N	*så skal vi i biografen*
NCCSU==I	noun common common-gender singular unmarkedcase indefinite	N	*de stod på række*
NCNPG==D	noun common neuter-gender plural genitive definite. Alt: NCNPG	N_GEN	*i årenes løb*

NCNPG==I	noun common neuter-gender plural genitive indefinite	N_GEN	*to minutters udvisning*
NCNPU==D	noun common neuter-gender plural unmarked-case definite. Alt: NCNPU	N	*ordene har fået en anden værdi*
NCNPU==I	noun common neuter-gender plural unmarked-case indefinite	N	*i femten år*
NCNSG==D	noun common neuter-gender singular genitive definite. Alt: NCNSG	N_GEN	*alle landets skoler*
NCNSG==I	noun common neuter-gender singular genitive indefinite	N_GEN	*dit livs big time vilde experience*
NCNSU==D	noun common neuter-gender singular unmarked-case definite. Alt: NCNSU	N	*jeg tager toget til Glumsø*
NCNSU==I	noun common neuter-gender singular unmarked-case indefinite	N	*der er et problem*
NC.PU==D	noun common plural unmarked-case definite	N	*de her virus*
NC.PU==I	noun common plural unmarked-case indefinite. Alt: NC.PU	N	*noget virus*
NC.SU==I	noun common singular unmarked-case indefinite. Alt: NC.SU	N	*det er til gavn og glæde*
NC..G==.	noun common genitive	N_GEN	-
NC..U==I	noun common unmarked-case indefinite. Alt: NC..U	N	*Vi har snakket lidt i øst og i vest*
NC..U==.	noun common unmarked-case	N	*lad dem da få lov*
NP--G==-	noun proper genitive. Alt: NP--G	EGEN_GEN	*Danmarks bedste jord*
NP--U==-	noun proper unmarked-case. Alt: NP--U	EGEN	*jeg er født i Århus*
PC--PG---	pronoun coordinative plural genitive	PRON_REC_GEN	*vi skrev hinandens navne op*
PC--PU---	pronoun coordinative plural unmarked-case	PRON_REC	*vi kender hinanden*
PD-CSG--U	pronoun demonstrative common singular genitive unmarked-style	PRON_DEMO_GEN	[*beskriv dennes form*]
PD-CSU--O	pronoun demonstrative common singular unmarkedcase obselete	PRON_DEMO	[*hvad skete der egentlig siden hin aften, da telefonen kimede*]
PD-CSU--U	pronoun demonstrative common singular unmarkedcase unmarked-style	PRON_DEMO	*lejligheder som denne her*
PD-NSU--U	pronoun demonstrative neuter singular unmarked-case unmarked-style	PRON_DEMO	*jeg skulle have ordnet det kød derude*
PD-.PG--U	pronoun demonstrative plural genitive unmarked-style	PRON_DEMO_GEN	[*beskriv disses former*] (søgt)

PD-.PU--U	pronoun demonstrative plural unmarked-case unmarked-style	PRON_DEMO	*under de/disse forhold*
PD-..U--U	pronoun demonstrative unmarked-case unmarked-style	PRON_DEMO	[*det må du/de selv om*]
PI-CSG--U	pronoun indefinite common singular genitive unmarked-style	PRON_UBST_GEN	*når det er ens egne børn*
PI-CSU--U	pronoun indefinite common singular unmarked-case unmarked-style	PRON_UBST	*det var en jeg fik*
PI-C.N--U	pronoun indefinite common nominative unmarkedstyle	PRON_UBST	*alle de her idrætsskader man hører om*
PI-NSU--U	pronoun indefinite neuter singular unmarked-case unmarked-style	PRON_UBST	*du må ikke forandre noget*
PI-.PG--U	pronoun indefinite plural genitive unmarked-style	PRON_UBST_GEN	*nogen andres forældre*
PI-.PU--O	pronoun indefinite plural unmarked-case obselete	PRON_UBST	*det er somme tider hårdt*
PI-.PU--U	pronoun indefinite plural unmarked-case unmarkedstyle	PRON_UBST	*nogle dejlige naboer*
PI-..G--U	pronoun indefinite genitive unmarked-style	PRON_UBST_GEN	*for nogens vedkommende*
PO1CSUPNF	pronoun possessive 1st-person common singular unmarked-case plural nonreflexive formal	PRON_POSS	*for hun er en af vor egne*
PO1CSUSNU	pronoun possessive 1st-person common singular unmarked-case singular nonreflexive unmarked-style	PRON_POSS	*deres regning er mindre end min*
PO1NSUPNF	pronoun possessive 1st-person neuter singular unmarked-case plural nonreflexive formal	PRON_POSS	*billeder fra vort bryllup*
PO1NSUSNU	pronoun possessive 1st-person neuter singular unmarked-case singular nonreflexive unmarked-style	PRON_POSS	*det er mit arbejde*
PO1.PUPNF	pronoun possessive 1st-person plural unmarked-case plural nonreflexive formal	PRON_POSS	*i vore dage*
PO1.PUSNU	pronoun possessive 1st-person plural unmarked-case singular nonreflexive unmarked-style	PRON_POSS	*det står i mine papirer*
PO1..UPNU	pronoun possessive 1st-person unmarked-case plural nonreflexive unmarked-style	PRON_POSS	*de skulle ud i vores køkken*
PO2CSUSNU	pronoun possessive 2nd-person common singular unmarked-case singular nonreflexive unmarked-style	PRON_POSS	*han er din mand*
PO2NSUSNU	pronoun possessive 2nd-person neuter singular unmarked-case singular nonreflexive unmarked-style	PRON_POSS	*du passer dit arbejde*

PO2.PUSNU	pronoun possessive 2nd-person plural unmarked-case singular nonreflexive unmarked-style	PRON_POSS	*har du fået lavet <u>dine</u> lektier*
PO2..UPNU	pronoun possessive 2nd-person unmarked-case plural nonreflexive unmarked-style	PRON_POSS	*en forskel i <u>jeres</u> sprog*
PO2..U.NP	pronoun possessive 2nd-person unmarked-case nonreflexive polite	PRON_POSS	*De lader <u>Deres</u> mand om det*
PO3CSUSYU	pronoun possessive 3rd-person common singular unmarked-case singular reflexive unmarked-style	PRON_POSS	*på <u>sin</u> vis*
PO3NSUSYU	pronoun possessive 3rd-person neuter singular unmarked-case singular reflexive unmarked-style	PRON_POSS	*hver passer <u>sit</u>*
PO3.PUSYU	pronoun possessive 3rd-person plural unmarked-case singular reflexive unmarked-style	PRON_POSS	*sørge for <u>sine</u> børn*
PO3..UPNU	pronoun possessive 3rd-person unmarked-case plural nonreflexive unmarked-style	PRON_POSS	*<u>deres</u> regning bliver ikke større*
PO3..USNU	pronoun possessive 3rd-person unmarked-case singular nonreflexive unmarked-style	PRON_POSS	*<u>hans/hendes</u> søn kom hjem fra Oslo*
PP1CPN-NU	pronoun personal 1st-person common plural nominative nonreflexive unmarked-style	PRON_PERS	*sådan noget havde <u>vi</u> da*
PP1CPU-.U	pronoun personal 1st-person common plural unmarked-case unmarked-style	PRON_PERS	*så kigger han på <u>os</u>*
PP1CSN-NU	pronoun personal 1st-person common singular nominative nonreflexive unmarked-style	PRON_PERS	*<u>jeg</u> har fået lov*
PP1CSU-.U	pronoun personal 1st-person common singular unmarked-case unmarked-style	PRON_PERS	*en million kroner til <u>mig</u>*
PP2CPN-NU	pronoun personal 2nd-person common plural nominative nonreflexive unmarked-style	PRON_PERS	*hvad siger <u>I</u> til det*
PP2CPU-.U	pronoun personal 2nd-person common plural unmarked-case unmarked-style	PRON_PERS	*men hvad så med <u>jer</u>*
PP2CSN-NU	pronoun personal 2nd-person common singular nominative nonreflexive unmarked-style	PRON_PERS	*bruger <u>du</u> sukker*
PP2CSU-.U	pronoun personal 2nd-person common singular unmarked-case unmarked-style	PRON_PERS	*det er fuldstændig op til <u>dig</u> selv*
PP2C.N-NP	pronoun personal 2nd-person common nominative nonreflexive polite	PRON_PERS	*gør <u>De</u> bare det*
PP2C.U-.P	pronoun personal 2nd-person common unmarked-case polite	PRON_PERS	*jeg lagde det på bordet for <u>Dem</u>*
PP3CSN-NU	pronoun personal 3rd-person common singular nominative nonreflexive unmarked-style	PRON_PERS	*og <u>hun/han</u> var ydermere flink*
PP3CSU-NU	pronoun personal 3rd-person common singular unmarked-case nonreflexive unmarked-style	PRON_PERS	*<u>den</u> er fra Færøerne*
PP3NSU-NU	pronoun personal 3rd-person neuter singular unmarked-case nonreflexive unmarked-style	PRON_PERS	*<u>det</u> er rigtigt*

PP3.PN-NU	pronoun personal 3rd-person plural nominative nonreflexive unmarked-style	PRON_PERS	*hvis de kan skaffe mig en lejlighed*
PP3.PU-NU	pronoun personal 3rd-person plural unmarked-case nonreflexive unmarked-style	PRON_PERS	*dem jeg har kendt*
PP3..U-YU	pronoun personal 3rd-person unmarked-case reflexive unmarked-style	PRON_PERS	*man brokker sig over huslejen*
PT-CSU--U	pronoun interrogative-relative common singular unmarked-case unmarked-style	PRON_INTER_REL	*hvilken skole gik du på*
PT-C.U--U	pronoun interrogative-relative common unmarked-case unmarked-style	PRON_INTER_REL	*hvem skal så bo her*
PT-NSU--U	pronoun interrogative-relative neuter singular unmarked-case unmarked-style	PRON_INTER_REL	*hvilket år er du født*
PT-.PU--U	pronoun interrogative-relative plural unmarked-case unmarked-style	PRON_INTER_REL	*hvilke lærere synes du er de fede*
PT-.SU--U	pronoun interrogative-relative singular unmarked-case unmarked-style	PRON_INTER_REL	*heller ikke fløde eller hvad*
PT-..G--U	pronoun interrogative-relative genitive unmarked-style	PRON_INTER_REL_GEN	[*hvis er det*] (alle forekomster af *hvis* er markeret som konjunktioner)
RGA	adverb general absolute-superlative	ADV	*hvor vil du allerhelst bo*
RGC	adverb general comparative	ADV	*det var længere nede*
RGP	adverb general positive	ADV	*så længe du bor i stuen*
RGS	adverb general superlative	ADV	*dem der bor længst væk*
RGU	adverb general unmarked-comparison	ADV	*det fik jeg ikke gjort*
SP	adposition preposition	PRÆP	*ude ved Strandvejen*
U=	unique.	UNIK	*dem der ikke kunne*
VADA=----A-	verb main indicative past active. Alt: VADA	V_PAST	*det vidste jeg ikke*
VADA=----P-	verb main indicative past passive. Alt: VADA	V_PAST	*vi skiftedes lidt*
VADR=----A-	verb main indicative present active. Alt: VADR	V_PRES	*jeg er født i Århus*
VADR=----P-	verb main indicative present passive. Alt: VADR	V_PRES	*det staves ikke som det siges*
VAF-=----A-	verb main infinitive active. Alt: VAF, VAF-	V_INF	*jeg skal lige have fat i dem*
VAF-=----P-	verb main infinitive passive. Alt: VAF, VAF-	V_INF	*den skal hjælpes*
VAG-=SCI--U	verb main gerund singular common indefinite unmarked-case. Alt: VAG-	V_GERUND	*den fysiske formåen*

VAM-=------	verb main imperative. Alt: VAM-	V_IMP	*lad dem da få lov*
VAPA (VAPA=....-.)	verb main participle past	V_PARTC_PAST	*jeg har fået lov*
VAPA=---R--	verb main participle past adverbial-use. Alt: VAPA	V_PARTC_PAST	[*han sprang forskrækket op*]
VAPA=P..A-G	verb main participle past plural genitive	V_PARTC_PAST	[*de/nogle skræmtes historie*]
VAPA=P..A-U	verb main participle past plural unmarked-case. Alt: VAPA	V_PARTC_PAST	[*de/nogle røgede sild*]
VAPA=SCDA-U	verb main participle past singular common definite unmarked-case	V_PARTC_PAST	[*alle de andre er gåen nu*]
VAPA=S.DA-G	verb main participle past singular definite genitive	V_PARTC_PAST	[*den skræmtes historie*]
VAPA=S.DA-U	verb main participle past singular definite unmarkedcase. Alt: VAPA	V_PARTC_PAST	[*den røgede sild*]
VAPA=S.IA-U	verb main participle past singular indefinite unmarkedcase. Alt: VAPA	V_PARTC_PAST	[*en røget sild*]
VAPA=S.I.-U	verb main participle past singular indefinite unmarkedcase	V_PARTC_PAST	[*en forskrækket mand*] [*han sprang forskrækket op*]
VAPR=---R--	verb main participle present adverbial-use. Alt: VAPR	V_PARTC_PRES	*det har været forbløffende nemt*
VAPR=...A-U	verb main participle present unmarked-case. Alt: VAPR	V_PARTC_PRES	*det flyvende tæppe*
VAPR=....-U	verb main participle present unmarked-case. Alt: VAPR	V_PARTC_PRES	*man blev bare stående* [*han sov stående*]
VEDA=----A-	verb medial indicative past active. Alt: VEDA	V_MED_PAST	*det lykkedes*
VEDR=----A-	verb medial indicative present active. Alt VEDR	V_MED_PRES	*det synes jeg er meget*
VEF-=----A-	verb medial infinitive active. Alt: VEF-	V_MED_INF	*vi mødes på et tidspunkt*
VEPA=....-U	verb medial participle past unmarked-case. Alt: VEPA	V_MED_PARTC_PAST	*du har skændtes med din mand*

PoS tags from the *Residual* category (code X in the PAROLE tag set) are described separately in the table below because these tags in the corpus are not used as described in the manuals. According to the manuals, they should cover abbreviations (XA), foreign words (XF), formulas (XR), and others (XX). However, they are not used consistently, and in addition, the category XZ, which is not described in the manuals, has been used extensively. The table below provides an overview of how they are actually used in the corpus.

Table of PoS tags from the *Residual* category

PoS (PAROLE)	Description	RedPoS	Example
XA	(Egl. "residual abbreviation") Letters in material ("A", "B", "C"), exclamations ("uh"), compounds, and the like involving a letter material (e.g., snabel_A), as well as a series of seemingly random words (e.g., nærigt, steppet, and grovt).	FORK	*ottende B*
XF	(Egl. "residual foreign") This tag is used for foreign words (though not all), some Danish words that orthographically resemble foreign words (e.g., and), and a series of seemingly random words.	UL	*det er sådan what*
XR	(Egl. "residual formulae") This tag is used for foreign words (though not all), some Danish words that orthographically resemble foreign words (e.g., and), and a series of seemingly random words.	FORM	*Bilka og og OBS*
XX	(Egl. "residual other") This tag appears to be used more or less randomly.	XX	*tysk fremfor fransk*
XZ	(Not in the PAROLE tag set) Proper nouns (words that begin with a capital letter), self-interruptions, words with underscores, xxx (incomprehensible speech), oe (American Danish Øh).	[diverse]	*i ha- i starten af halvfjerserne*
U	(Not in the PAROLE tag set) Ordene *at*, *som* og *der*.	UNIK	*at lære at stave*
U	Words with underscores – but not all words with underscores are annotated with this tag (there are currently only about 100 hits in total).	*U*	*den der T_shirt*

2.1.9 Simpel ordklasse (RedPoS) - Reduced Parts of Speech

- Word-coded

This tier contains simplified Reduced Parts of Speech (RedPoS) PAROLE tags. In the table above, you can find the simplified part of speech for each PAROLE tag in the RedPoS column. The table below provides descriptions and examples for each of these reduced tags. The examples are primarily taken from the corpus. In cases where there were no suitable examples in the corpus, new examples have been constructed and are enclosed in square brackets. Ambiguous tags (as described in the Residual category on page 15) are highlighted in yellow.

Table of Reduced PAROLE Tags (RedPoS)

RedPoS	Description	Example
ADJ	Adjective	*en halv dag*
ADJ_GEN	Adjective in genitive	*for manges vedkommende*
ADV	Adverb	*det fik jeg ikke gjort*
EGEN	Proper noun	*jeg er født i Århus*
EGEN_GEN	Proper noun in genitive	*Danmarks bedste jord*
FORK	Abbreviation	
FORM	Formula	
INTERJ	Interjection	*ja*
N	Noun	*så skal vi i biografen*
N_GEN	Noun in genitive	*to minutters udvisning*
NUM	Numerals	*to personer*
NUM_GEN	Numerals in genitive	[*Det er de tos projekt*]
NUM_ORD	Ordinal numbers	*syvende klasse*
NUM_ORD_GEN	Ordinal numbers in genitive	*Christian den fjerdes oprindelige byggestil*
PRON_DEMO	Demonstrative pronoun	*lejligheder som denne her*
PRON_DEMO_GEN	Demonstrative pronoun in genitive	[*beskriv dennes form*]
PRON_INTER_REL	Interrogative pronoun	[*hvis er det*]
PRON_INTER_REL_GEN	Interrogative pronoun in genitive	*hvem skal så bo her*
PRON_PERS	Personal pronoun	*jeg har fået lov*
PRON_POSS	Possessive pronoun	*det er mit arbejde*
PRON_REC	Reciprocal pronoun	*vi kender hinanden*
PRON_REC_GEN	Reciprocal pronoun in genitive	*vi skrev hinandens navne op*
PRON_UBST	Indefinite pronoun	*nogle dejlige naboer*
PRON_UBST_GEN	Indefinite pronoun in genitive	*for nogens vedkommende*
PRÆP	Preposition	*ude ved Strandvejen*
SKONJ	Coordinating conjunction	*og, eller, men*
UKONJ	Subordinating conjunction	*fordi, hvis, mens*
UL	[Foreign word]	*det er sådan what*
UNIK	Unique – different particles	*som, der, at*
V_GERUND	Verb, gerundium	*den fysiske formåen*
V_IMP	Verb in imperative	*lad dem da få lov*

V_INF	Verb in infinitive	*jeg skal lige <u>have</u> fat i dem*
V_MED_INF	Medial verb in infinitive	*vi <u>mødes</u> på et tidspunkt*
V_MED_PARTC_PAST	Medial verb in perfect (past)	*du har <u>skændtes</u> med din mand*
V_MED_PAST	Medial verb in preterite (past)	*det <u>lykkedes</u>*
V_MED_PRES	Medial verb in present (present)	*det <u>synes</u> jeg er meget*
V_PARTC_PAST	Verb in perfect	*jeg har <u>fået</u> lov*
V_PARTC_PRES	Verb in present participle (pluperfect or past perfect)	*det <u>flyvende</u> tæppe*
V_PAST	Verb in preterite	*det <u>vidste</u> jeg ikke*
V_PRES	Verb in present	*jeg <u>er</u> født i Århus*
XX	Other	
XZ	Unclear function	

2.1.10 Events (events)

- Sequence-coded (BIO-labeled). The original annotation sequences "X X X" have been recoded as "1_X_I 2_X_I 3_X_E" (as described in the BIO-labeling description above).

Speaker-produced, meaningful sounds, such as coughs, sneezes, smacking, or clearing the throat. Note: Some of the annotations instead indicate errors and deficiencies.

Example of sequence coding: For instance, when a token is annotated with "rømmen," (Eng. Clearing the throat) that token will receive the following tag: 1_rømmen_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.1.11 Globale events (global events)

- Word-coded

Used for noise that cannot be attributed to a specific speaker, such as "The dog barks."

Note: This should be sequence-coded since there are instances of multi-word sequences marked with this code.

2.1.12 Emfase (emphasis)

• Word-coded

A word pronounced with emphatic stress is marked with an exclamation point, <!>. This tier contains only exclamation points in the same time intervals as the words that are pronounced with emphatic stress.

2.1.13 Phonetic (phonetic)

• Word-coded

Contact person: Marie Maegaard, mamae@hum.ku.dk

In this tier, both emphasis and elongation are marked. Emphasis is marked with an exclamation point, <!>, after the word, and elongation is marked with a colon, <:>, after the symbol representing the prolonged sound. The tier contains only the words that are pronounced with emphatic stress or elongation.

Examples:

afprøver!

der:

du!

ps:ykotisk

2.2 IIV-tiers

• Contact persons: Torben Juel Jensen, tjuelj@hum.ku.dk; Astrid Ag, astridag@hum.ku.dk; Frans Gregersen, fg@hum.ku.dk

• Status: Open. All tiers are open to everyone.

IIV stands for Intra Individual Variation. At the levels of Activity Type, Conversation Type, and Macro-speech Act, full coding has been performed, meaning everything is coded, and no part of a conversation falls outside a category. At the levels of Interaction Structure, Genre, and Utterance, partial coding has been performed, meaning only certain structures, genre sequences, phrases, or words are coded, and therefore, there are parts of the conversation that do not have a category. The individual codes are not specific to speakers; all speakers are coded. This means that each level within IIV only receives one tier in each conversation, regardless of the number of participants in the conversation.

2.2.1 IIV Aktivitetstype (Aktivitetstype) - IIV Activity Type

- Sequence-coded (BIO-labeled). The original annotation sequences "X X X" have been recoded as "1_X_I 2_X_I 3_X_E" (as described in the BIO-labeling description above).
- Fully coded.

All conversations are categorized into activity types based on the activity or "phase" in which the participants are engaged. The six activity types we categorize conversations into are:

Abi	Background interview
Asa	Conversation
Asi	Conversation with non-participant
Ael	Elicited speech
Asp	Language attitude survey
Ati	Declaration of consent

These categories can be combined if desired.

Example of sequence coding: For instance, when 70 tokens are annotated with Abi, those 70 tokens will receive these 70 tags: 1_Abi_I, 2_Abi_I, 3_Abi_I, [...], 69_Abi_I, 70_Abi_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.2.2 IIV Interaktionsstruktur (Interaktionsstruktur) - Interaction structure

- • Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Partial coding

Interaction structure is defined based on the degree of control in the conversation, determined by the symmetry or asymmetry in the local relationship between the participants' interactional roles. In individual interviews and group interviews, five different interaction structures are coded: I4, I5, I6, I7, and I8.

Table of different types of interaction structure:

I1	Interviewer takes strong initiative, and the informant responds briefly
I2	Interviewer takes strong initiative, and the informant responds at length
I3	Interviewer takes the initiative and formulates extensively, and the informant responds briefly
I4	No control
I5	Informant takes strong initiative, and the interviewer responds
I6	Fighting for the chance to speak
I7	Informant takes strong initiative, and another informant responds
I8	Monologue
I9	Other

Group conversations without an interviewer are coded for I6, I7, and I8.

According to footnote 28 in the coding manual (page 26), I1, I2, I3, and I9 are no longer used. However, they appear in the corpus, and their descriptions are therefore included in this document.

Example of sequence coding: For instance, when 39 tokens are annotated with I8, those 39 tokens will receive these 39 tags: 1_I8_I, 2_I8_I, 3_I8_I, [...], 38_I8_I, 39_I8_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.2.3 IIV Genre (Genre)

- • Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Partial coding

There are eight different genres distinguished:

Gna	Narrative
Gsr	Specific account
Ggr	Generel account

Gsb	Soapbox
Gsl	Gossip and derogatory statements
Gbe	Confession
Gre	Reflection
Gvi	Joke

Some of these have a clear specific function and can be seen as falling under one of the categories at the macro-speech act level. Still, we have chosen to place them collectively at the genre level for the sake of systematics. Other spoken language genres, such as quarreling and counseling, which may often appear in other conversation types, are very rare, approaching non-existent, in the sociolinguistic interviews and group conversations in our data, and therefore, they are not included here.

The eight genres can be combined in coding if desired.

Example of sequence coding: For instance, when 97 tokens are annotated with Gsl, those 97 tokens will receive these 97 tags: 1_Gsl_I, 2_Gsl_I, 3_Gsl_I, [...], 96_Gsl_I, 97_Gsl_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.2.4 IIV Samtaletype (Samtaletype) - Conversation type

- • Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Fully coded.

Each conversation is assigned an overarching code depending on the composition of participants. There are five conversation types, and these five types cover the spread in the corpus concerning interaction type (interview or non-interview), number of participants (one, two, or more), and familiarity between interviewer(s) and informant(s). If there are prolonged shifts in participant composition during the conversation, the specific sequence will be given a different code than the overarching one. The five conversation types are:

Siek	Single interview, familiar
Sieu	Single interview, unfamiliar
Sifk	Group interview, familiar
Sifu	Group interview, unfamiliar
Sgfk	Group conversation without an interviewer

Example of sequence coding: For instance, when 136 tokens are annotated with Sifk, those 136 tokens will receive these 136 tags: 1_Sifk_I, 2_Sifk_I, 3_Sifk_I, [...], 135_Sifk_I, 136_Sifk_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.2.5 IIV Makro-sproghandling (Makro-sproghandling) - Macro-Speech Act

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Fully coded.

Conversations are categorized at this level based on the pragmatic function of language in the conversation at a given time – in other words, the "macro-speech act" taking place. This is defined based on the type of content exchanged between conversation participants. The following five overarching macro-speech acts are distinguished:

Mvi	Exchange of information
Mho	Exchange of opinions
Mfø	Exchange of emotions
Mha	Action-bound speech
Mfi	Fiction

These five categories can be combined in coding if desired.

Example of sequence coding: For instance, when 682 tokens are annotated with Mvi, those 682 tokens will receive these 682 tags: 1_Mvi_I, 2_Mvi_I, 3_Mvi_I, […], 681_Mvi_I, 682_Mvi_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.2.6 IIV Udsigelse (Udsigelse) - Utterance

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Partial coding.

We collectively code for quotes, imitations, and illustrative sounds, whether it is the speaker themselves or another person, animal, or object being quoted, imitated, or illustrated.

The code for this is Ucil.

Example of sequence coding: For instance, when six tokens are annotated with Ucil, those six tokens will receive these six tags: 1_Ucil_I, 2_Ucil_I, 3_Ucil_I, 4_Ucil_I, 5_Ucil_I, 6_Ucil_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.2.7 Leksis og fraser – Lexis and Phrases

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

Found in BySoc, Køge, ModSjæl, Næstved, and Odder.

Note: A comprehensive description of this category has not yet been obtained. However, "Ltot/ltot" and "Like/like" should probably be harmonized regarding uppercase/lowercase, and the following inventory should be checked, supplemented, and explained:

Ltot/ltot/Â¨Ltot	"Contact appeals" (ikke, vel, hallo)
Like	(hvad pokker, javel, aha, er du rigtig klog, herregud, det må jeg love dig, uha, hov, uh)
like	(fucked up, eller sådan et eller andet, eller hvordan det, skidt med det, saftsus-, det er helt vildt, gud)
ludl	(yes yes yes)
Lsl	(ikke en junk)
s#	(halvtreds)
lslk	(funker)
Lsk	(tøsen)

Example of sequence coding: For instance, when a token is annotated with Ltot, this token will receive the following tag: 1_Ltot_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.2.8 IIV AIG-kommentarer (IIV AIG Kommentarer) – AIG Comments

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

Any comments on Activity Type, Interaction Structure, and Genre coding. Note: At least partially contains records of errors.

Example of sequence coding: For instance, when a token is annotated with "Afbrud i interview pga. båndskift" (Eng. Interruption due to tape change), this token will receive the following tags 1_ Afbrud_i_interview_pga._båndskift_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.2.9 IIV SMU-kommentarer (IIV SMU Kommentarer) – SMU Comments

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

Any comments related to Conversation Type, Macro Speech Acts, and Utterance coding. Note: At least partially contains records of errors.

Example of sequence coding: For instance, when a token is annotated with "Ucil attributed to HCM", this token will receive the following tags 1_Ucil_attributed_to_HCM_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.2.10 IIV SMUL-kommentarer (IIV SMUL Kommentarer) - SMUL Comments

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

Any comments related to Conversation Type, Macro Speech Acts, Utterance, and Lexis and Phrases coding. Note: The tier is currently unavailable for an unknown reason. However, it is also nearly empty, so not much has been lost.

2.2.11 IIV-kommentarer (IIV Kommentaerer) – IIV Comments

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

Any comments related to IIV coding. Note: This tier mainly contains records of errors.

2.3 Phonetic Tiers

- Contact person: Marie Maegaard, mamae@hum.ku.dk
- Status: Open. All tiers are accessible to everyone.

More information can be found in the article “Fonetisk variantsætning” in the DGCSS Wiki and manuals in the Language Change Center's phonetics folder on the N-drive – reach out to Marie if this seems relevant.

These tiers contain codings of pronunciation of individual sounds in different contexts. Generally, each pronunciation is encoded with a number that has a specific meaning for the given category. In cases of uncertainty, various methods have been used to indicate this. Either the general symbols are used (see below) or, for example, "12" is written if there is uncertainty about whether the sound is of quality 1 or 2.

Generelle symboler:

*	Doubt, impossible decision due to following segment (=schwaD and participles)
* K	Uncertainty about quality
* T	Uncertainty about stress
* L	Uncertainty about lengthening
/	Separation of different suggestions for a decision
¤	Quality that falls outside the conventional variants of the variable
.&	Lengthening due to hesitation
.§	Lengthening due to emphasis
:	Previous marking of lengthening
.!	Previous marking of lengthening due to laughter or other extralinguistic factors
.	Prolonged vowel. <.> (= period) is written after the number symbols (see the categories below)

2.3.1 AN-markering (ANmarkering) – AN Marking

- Word-coded

AN is coded alongside AM, but they are in separate tiers. AM/AN symbolizes stressed short A in contexts other than before or after R and before a semivowel. This refers to phonologically assumed short A, which is not part of falling diphthongs and is not a neighbor sound to R.

The AN category is used for a-sounds that are expected to be pronounced as /a/ before /n/.

The following codes are used:

1	The usual vowel quality in Copenhagen long 'a' in words like 'mase'.
2	The usual vowel quality in standard Danish short 'a' in words like 'kast'.
3	The usual vowel quality in standard Danish short 'a' in words like 'gammel'.
4	The vowel quality, which is more retracted than 3 and is sometimes heard in Copenhagen dialect in words like 'kaffe'.
0	The vowel quality that is higher than 1 and therefore sounds like 'e' in 'meld' (or even higher).

2.3.2 AM-markering (AMmarkering) – AM Marking

- Word-coded

AM is coded alongside AN, but they are in separate tiers. AM/AN symbolizes stressed short A in contexts other than before or after R and before a semivowel. This refers to phonologically assumed short A, which is not part of falling diphthongs and is not a neighbor sound to R.

The same codes as for ANmarking are used.

2.3.3 ENG-markering (ENGmarkering) – ENG Marking

- Word-coded

The ENG variable refers to the variation in the pronunciation of the stressed vowel in words like 'tænke' (think), 'engelsk' (English), 'enkelt' (single), 'penge' (money), etc.

The following codes are used:

1	signifies an e-pronunciation, which is common in words like 'vinde' (win), 'dele' (share).
2	signifies an æ-pronunciation, which is common in words like 'æske' (box), 'sæbe' (soap).

2.3.4 AJ-markering (AJmarkering) – AJ Marking

- Word-coded

AJ symbolizes the falling diphthong found in words such as 'dig' (you), 'vej' (road), 'leg' (game) in contexts other than after R. AJ is coded alongside ANR. Any omission of the semivowel 'J' in AJ is denoted using 0 - (for example, 30), when the quality of 'A' is judged as 3, and there is subsequent omission of the semi-vowel.

The following codes are used:

1	The usual vowel quality in Copenhagen long 'a' in words like 'mase'.
2	The usual vowel quality in standard Danish short 'a' in words like 'kast'.
3	The usual vowel quality in standard Danish short 'a' in words like 'gammel'.
4	The vowel quality, which is more retracted than 3 and is sometimes heard in Copenhagen dialect in words like 'kaffe'.
0	The vowel quality that is higher than 1 and therefore sounds like 'e' in 'meld' (or even higher).
x0	Semivowel omission (after a vowel with quality 'x' – there is a number instead of 'x').

2.3.5 ANR-markering (ANRmarkering) – ANR Marking

- Word-coded

ANR signifies specific special words with the monophthong A, where an R is present in the immediate context, for example, 'andre' (other) and 'anderledes' (different). ANR is coded alongside AJ.

The same codes used in AN-marking are applied.

2.3.6 RU-markering (RUmarkering) – RU Marking

- Word-coded

RU symbolizes the pronunciation variation between "ru/ro" in words such as 'rulle' (roll), 'rugbrød' (rye bread), 'russisk' (Russian), and 'bruge' (use), 'rude' (window), 'rune' (rune).

The following codes are used:

1	for 'u' pronunciation
2	for 'o' pronunciation

2.3.7 W-markering (Wmarkering) – W Marking

- Word-coded
- W refers to the pronunciation of words like 'skrevet' (written), 'prøvet' (tried), and 'lovet' (promised), which can be pronounced without the W-diphthong - as "skredd," "prødd," and "lådd."

The following codes are used:

1	for 'w' pronunciation

0	for omission of 'w'

2.3.8 D-schwa-D-markering (DschwaDmarkering) – D Schwa D Marking

- Word-coded

SchwaD refers to the ending -et, which in Danish is usually realized either as "schwa-soft d" or as "schwa-hard d". It is the variation found in words like 'boet' (lived), 'løbet' (run), 'siddet' (sat), 'huset' (house), 'vandet' (water), 'brødet' (bread). The variable is divided into two and categorized into two tiers: one tier called XschwaDmarkering and one called DschwaDmarkering. The point is that some Danish language researchers have argued that the variable is often (some say always) pronounced with a hard 'd' when it follows a soft 'd'. That is, words like 'siddet' and 'brødet' typically have a hard 'd', even if the speaker otherwise uses a soft 'd'. This division allows examination of this phenomenon.

The following codes are used:

1	for soft 'd'
2	for hard 'd'
0	if the ending is not pronounced

2.3.9 X-schwa-D-markering (XschwaDmarking) - X Schwa D Marking

- Word-coded

SchwaD refers to the ending -et, which in Danish is usually realized either as "schwa-soft d" or as "schwa-hard d". It is the variation found in words like 'boet' (lived), 'løbet' (run), 'siddet' (sat), 'huset' (house), 'vandet' (water), 'brødet' (bread). The variable is divided into two and categorized into two tiers: one tier called XschwaDmarkering and one called DschwaDmarkering. The point is that some Danish language researchers have argued that the variable is often (some say always) pronounced with a hard 'd' when it follows a soft 'd'. That is, words like 'siddet' and 'brødet' typically have a hard 'd', even if the speaker otherwise uses a soft 'd'. This division allows examination of this phenomenon.

The following codes are used:

1	for soft 'd'
2	for hard 'd'
0	if the ending is not pronounced

2.3.10 Kontekst fonetik (kontekst fonetik) - Context Phonetics

- Word-coded

A comprehensive description of this tier has not been provided yet.

2.3.11 Variant fonetik (variant fonetik) – Variant Phonetics

- Word-coded

This tier is only found in the exploratory files (cf. section "Metadata at the conversation level").

In this tier, the vowel quality for A, E, and U is indicated. The tier is closely associated with the tiers *variant fonetik R*, *variant fonetik kontekst forventet* and *variant fonetik kontekst realiseret*.

In variant phonetics R, it is indicated whether an /r/ comes before or after the vowel. In *variant fonetik kontekst forventet* and *variant fonetik kontekst realiseret*, what comes after the vowel is transcribed - both the expected and the realized.

A:

Her benyttes fire varianter:

A1	flat “katte” a
A2	“Katte” a
A3	“Kappe” a
A4	retracted “kappe” a
A0	an a raised so much that it sounds like [æ]

If the vowel is long, a colon is added after the number; if it is overlong due to hesitation, an "&" is added after the colon.

AJ:

The diphthong /aj/ is classified for the quality of the vowel with the same values as A, i.e., AJ2 for the slightly old-fashioned, conservative pronunciation, and AJ3 for the pronunciation it has in modern Copenhagen standard Danish (and AJ1 and AJ4 if necessary, for example in 'bage' and 'meget').

Additionally, a 0 is added if the [j] is completely absent, for example in words like 'jeg, mig, sig, dig'. For instance, AJ30 for an A with quality 3 and a completely missing [j].

EN:

Vokalkvaliteten for E er fordelt på 2 kategorier:

EN0	“Æble” æ
EN1	“Mele” e

R:

The lowering from [a] to [ɑ] after /r/ as in 'ret, frem, græde'. Here, three qualities are distinguished:

RA0	occurrence of narrow [a] after /r/.
RA3	same quality as with A
RA4	same quality as with A

The vowel quality represented in the orthography with <u> when it occurs immediately after /r/, as in 'rulle, rugbrød' is indicated here. Two qualities are distinguished:

RU0	vowel quality [u]
RU1	vowel quality [o]

Examples:

Tier	*bagved*	*Fælledparken*	*ja*	*grund*	*tænkt*
variant phonetics	A1:	A3	A2	RU1	EN0
variant phonetics R		_R		R_	
expected context variant phonetics	?+veD#	+g@n#	#	n?#	N?gd#
realized context variant phonetics	+ve#	g#	#	n#	N#

2.3.12 Variant fonetik R (variant fonetik R) - Variant Phonetics R

- Word-coded

This tier is only found in the explorative files (refer to the section "Metadata at the conversation level").

Variant Phonetics R indicates whether an /r/ occurs before or after the occurrence of the variant described in variant phonetics. If there is no /r/ in the immediate vicinity, the interval is empty.

Inventory:

_R	/r/ occurs after the variant
R_	/r/ occurs before the variant
	/r/ does not occur

2.3.13 Variant fonetik kontekst forventet (variant fonetik kontekst forventet) - expected context variant phonetics

- Word-coded

This tier is closely related to the tiers *variant fonetik* and *variant fonetik kontekst realiseret*.

In this tier, what is expected to come after the variant indicated in variant phonetics is transcribed. It's a phonetic transcription of the distinct pronunciation of the given word.

The following phonetic transcription convention is used (middle column):

IPA (rough)	DGCSS convention	example
i	i	mit
e	e	lidt
ɛ	E	mæt
æ	æ	bade
a	a	ladt, dreng
ɑ	A	lak
y	y	tyst
ø	ø	øst
œ	Ø	skøn
œ̟	rØ	grynt
ɶ	ö	grønt
U	u	husk
O	o	foto
ɔ	å	bund
ɒ	År	orne
ʌ	Å	bånd
ə	@	sætte
p^h	p	pas
t^s	t	tal
k^h	k	kat
b	b	bip
d	d	dit
g	g	gik
m	m	mit

n	n	nat
ŋ	N	lang
l	l	lidt
ʁ	r	rod
w	w	hav
ð	D	med
j	j	jeg

Additionally, < # > is used for word boundaries and < + > for syllable boundaries.

2.3.14 Variant fonetik kontekst realiseret (variant fonetik kontekst realiseret) - realized context variant phonetics

- Word-coded

This tier is closely related to the tiers *variant fonetik* and *variant fonetik kontekst forventet*.

In this tier, what comes after the variant indicated in *variant fonetik* is phonetically transcribed. It's essentially a phonetic representation of the pronunciation of the word.

The same phonetic transcription convention used in the tier variant fonetik kontekst forventet is employed here (see above).

2.4 Grammatical tiers

- Contact person: Varied, see individual tiers.
- Status: Varied, see individual tiers.

2.4.1 Grammatik generisk pronomen (grammatik) – Generic Pronoun Grammar

- Contact person: Torben Juel Jensen: tjuelj@hum.ku.dk.
- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Status: Researcher level.

In this tier, coding is done for generic pronouns in nominative form (du, man) without taking into account generic occurrences of "en" and "den." This is done using the analysis tool "generisk" (see the "Generisk" tier on page 33). All codes from the *Generisk* tier are also found in the *Grammatik* tier. The *Grammatik* tier also includes coding for all reflexive pronouns, coded according to the analysis tool "refleksiv", which is described below.

Refleksiv

Additional information can be found in the document "Analyseapparat (refleksiv)" located at the directory Korpusgruppe/_oversigter/Oversigt over tiers og tags/Kildedokumenter/Grammatiske analyseapparater fra TJJ/.

All occurrences are categorized concerning "usage"; reflexively used occurrences are additionally categorized concerning "domain" and "reference".

Coding:

The coding involves marking all instances of the search strings representing pronouns with an "R" along with the smallest sub-code for "usage," such as "R_AR," "R_AN," "R_AA," or "R_AI." In connection with occurrences categorized as reflexive, additional sub-codes for "domain" and "reference" are added, for instance, "R_AR_DSE_RP."

When uncertain about a coding, the letter "X" is used in the respective position, such as "R_AX" or "R_AR DX_RP." If an occurrence is deemed ambiguous, two codes can be used (in alphabetical order), such as "R_AR_DSE_RPT." Note: this option of coding should be minimized as much as possible.

Here is an overview of the codes used:

Usage (**Anvendelse**)		**Domain**		**Reference**	
Tag	Description	Tag*	Description	Tag	Description
AR	**Reflexive** usage	DSE	Pronoun occurs after (**efter**) subject	RP	**Person** (singular), including positions/roles that can be

					assumed by people, and animals or objects that can be referred to with 'han' or 'hun'.
AN	non-reflexive **anaforic**, kataforic or exoforic (deictic) usage	DSS	The pronoun occurs in a non-clausal comparative (**sammenlignende**) element introduced by "end" or "som."	RT	Animals or objects (**ting**) (singular).
AA	Other (**anden**) usage	DSF	Pronoun occurs before (**før**) subjects	RG	**Generic**
AI	**Incomplete** construction	DLA	Coreference within **accusative** with "(at-)infinitive"	RU	Indefinite (**Ubestemt**)
		DLI	Coreference between direct and **indirect** object	RF	Pluralis (**Flertal**)
		DLO	Coreference between direct object og **predicative**	RA	Other (**Anden**)
		DLP	Coreference between direct object and a with the sentence paired **prepositional construction**		
		DLN	Coreference between a dependent element and the core element in a **nominal subordination** or between nominal elements in apposition.		
		DLØ	Coreference between a pronoun and another form of "logical subject"		
		DIF	The pronoun is part of a fixed expression.		
		DII	The pronoun is co-referential with an **implicit** (most often generic) "subject" in an **infinitive**.		
		DBO	The pronoun is co-referential with both (**både**) the grammatical subject and (**og**) one of the types of logical subjects mentioned above.		

*Domaintags can be separated into three categories: grammatical subject (S), logical subject (L), implicit (I)

Example of sequence coding: When a token, for instance, is annotated with G_AI, this token will receive the following tag: 1_G_AI_E. The number preceding the tag indicates the position, and the letter following it indicates whether it is the last tag in the sequence or not.

2.4.2 Gramma_II (gramma_II)

- Contact person: Torben Juel Jensen: tjuelj@hum.ku.dk.
- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Status: Researcher level.

This tier encodes for generic pronouns such as "den" (it), "en" (one), and all instances of "ens" (one's), "dens" (its), "dig" (you), "din" (your), "dit" (your - neuter), and "dine" (your - plural). This coding is done according to the "generisk” analysis tool described in the section "Generisk" below. All generic codes from this tier are also found in the *Generisk* tier.

Example of sequence coding: For instance, when a token is annotated with G_AB_DI, this token will have the following tag: 1_G_AB_DI_E. The number preceding the tag indicates the position, and the letter following it indicates whether it is the last tag in the sequence or not.

Generisk (Generic)

- Contact person: Torben Juel Jensen: tjuelj@hum.ku.dk.
- Word-coded
- Status: Researcher Level.

This tier contains all generisk-codes from the tiers ”grammatik” and ”gramma_II”.

Analysis apparatus for occurrences of pronouns with generic meaning:

All instances of search strings representing pronouns are marked with a "G" and the smallest sub-code for "usage", such as "G_AS", "G_AA", "G_AI", and so on. In connection with the occurrences categorized as other than "Other usage" and "Incomplete", the partial code for "discourse function", i.e., "G_AS_DM", "G_AS_DS", or "G_AS_DI", is also added. For the first two mentioned types, the partial codes for "syntax" and "reference" are then added, for instance, "G_AS_DS_SB_RA".

In cases of doubt, the letter "X" is used in the respective position, for instance, "G_AX" or "G_AS_DX_SA_RX". For ambiguous occurrences, two codes are used (in alphabetical order), for example, "G_AS_DS_SA_RAJ" or "G_AS_DMS_SA_RA".

Categories:

Usage (**Anvendelse**)		**Discourse Function**	
Tag	Description	Tag	Description
AS	Pronoun functions as grammatical **subject**	DM	Truism or **moral**
AO	Pronoun functions as grammatical **object** (direct or indirect) or predicative	DS	Generalization of **situation**

AP	Pronoun functions as governed in a **prepositional phrase** (including indirect objects)	DI	Non-generalizing (**Ikke-generaliserende**) usage
AB	Pronoun functions as determinant (**bestemmerled**) in a nominal phrase		
AA	Other (**anden**) usage		
AI	**Incomplete** construction		

Syntax		**Reference**	
Tag	Description	Tag	Description
SB	Conditional construction (**betingelseskonstruktion**)	RA	**All** people or a specific group of people including both speaker and addressee
SP	In complement clause (often a nominal clause) to a **projecting** clause	RJ	Specific group of people including the speaker (**Jeg**) but not the addressee
SA	Other (**andet**)	RD	Specific group of people including the addressee (**Du**) but not the speaker
		RH	Specific group of people that neither (**Hverken**) includes the speaker nor the addressee

2.4.3 Ordstilling (ordstil) – Word Order

- Contact person: Torben Juel Jensen: tjuelj@hum.ku.dk.
- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Status: Researcher Level.

Analysis apparatus for subordination word order (subordinate clauses)

L: Subordinate clause, F: Function, O: Word order, M: Matrix clause, _: Other.

Coding:

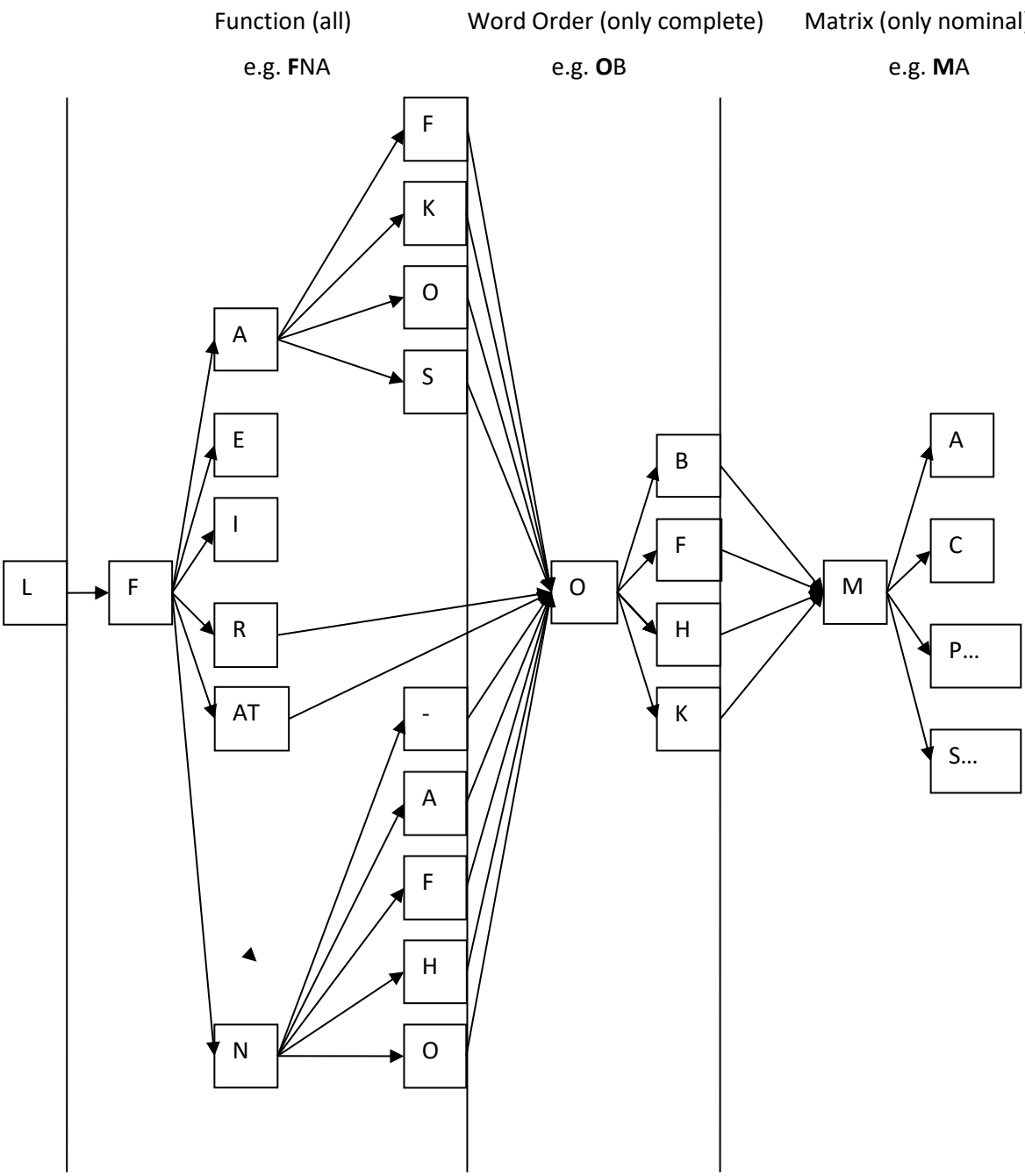
All occurrences of subordinations with sentence adverbs are coded with an "L" and at least the code for Function, that is at least "L F**E**" or "L F**I**". The functions "ellipsis" and "Incomplete" are not further coded.

In cases of doubt, the letter "X" is used in the respective place, for example, "L F**X**", "L FNA O**X** MA PI Aikke", or "L FNA OH M**X** PI Aikke". However, X-codes should not be found in the finished files. If an occurrence expresses two things at once, two codes are used in alphabetical order, for example, "L FAO OBF PI Aikke". However, double codes are only used in connection with word order and matrix clauses.

For subordinations embedded in other subordinations, parentheses are used around the codes, including the code for the "superordinate" subordination, for example, "(L FAK OB PI Aikke(L FNA OB MA PI Aikke) (LFN- OB MA PI Aikke))".

A complete description of the codes used can be found in the document "Analyseapparat(ordstilling)" (not yet translated into English).

Here is an overview of the codes used:



Examples:

(L FATNN\at OB A\ikke SL3P\de PRL\glemme FPS PF D0 L9 II FVN SPN))

(L FNA OH MA PF Aaltså\bare PRhave\brug\for TN S3Pde AMrentfaktisk SYST

L FAK\fordi OH A\jo SL3P\alle\maskiner PRL\være\styre FPS PF D0 L25 II FVF SPN

L FAK\fordi\at OH A\egentlig SL1S PRL\tænke FPS PF DI L44 IU FVF SPN

L FAO\hvis OB A\bare\lige SL1S PRL\blive\sætte\ind\i FPS PF D0 L18 II FVN SPN

L FAO\nÃ¥r OB A\ikke SL3S\hun PRL\vÃ¦re FPS PU D0 L6 II FVF SPN

L FI

L FN- OH MA PF Abare PRbestemme TTFD S1P AMegentlig SYO

L FN- OH MCA PF Aikke

L FN- OH MCA PF Ajo\ikke

L FN- OH MSA PF Ajo

L FN- OH MSA PF Asgu\da\fÃ¸rst

L FN- OK MSA PK Anok

L FNA OH MA PF Agerne PRkigge\på TN S3Sman AMegentlig SYST

L FNA OH MA PF Agodt PRhave\ngt TTD S1P AM0 SYO

L FNA OH MA PF Aikke PRsige TTD S1S AMbare SYO

L FNH OB MA PF Aikke PRsige TN S3Pde AM0 SYO

Examples of sequence coding: When for example nine tokens are annotated with *L FAK\fordi OH A\jo SL3P\de PRL\have FPS PF PR\0 S\0 AM\0 DD L9 KU IU FVF SPN*, the nine tokens will have these tags:
1_L_FAK\fordi_OH_A\jo_SL3P\de_PRL\have_FPS_PF_PR\0_S\0_AM\0_DD_L9_KU_IU_FVF_SPN_I,
2_L_FAK\fordi_OH_A\jo_SL3P\de_PRL\have_FPS_PF_PR\0_S\0_AM\0_DD_L9_KU_IU_FVF_SPN_I,
3_L_FAK\fordi_OH_A\jo_SL3P\de_PRL\have_FPS_PF_PR\0_S\0_AM\0_DD_L9_KU_IU_FVF_SPN_I, […],
8_L_FAK\fordi_OH_A\jo_SL3P\de_PRL\have_FPS_PF_PR\0_S\0_AM\0_DD_L9_KU_IU_FVF_SPN_I,
9_L_FAK\fordi_OH_A\jo_SL3P\de_PRL\have_FPS_PF_PR\0_S\0_AM\0_DD_L9_KU_IU_FVF_SPN_E.

The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.4.4 Participium (Part) - Participial

- Contact person: Torben Juel Jensen: tjuelj@hum.ku.dk.
- Word-coded
- Status: Researcher Level.

The study is a collaboration between the grammar researchers and the phonetics researchers. The focus of investigation is the use of '-en' in the past participle.

For the phonetic part of the study: refer to the article "Phonetic Variants Section" for further information. For the grammatical part of the study: refer to the article "Analyseapparat(participier)" (not yet translated into English).

The following codes are used:

Each code consists of a combination of two letters describing the function and a number (or U) describing the pronunciation:

FV	Verbal function
FU	Indeterminate (It cannot be conclusively determined whether the participle has a verbal function, i.e., whether it denotes an action/event or a quality)
FA	Other function than verbal (i.e., adjectival or nominal)
FI	Occurrences in unfinished (i.e., unanalyzable) constructions

1	Represents 'soft d'
2	Represents 'hard d'
3	Represents '-en'
0	Represents schwa or no suffix
U	Not specified

Examples:

FA U

FI U

FU 0

FU 1

2.4.5 Epistemisk sætning (epistsætn) – Epistemic sentence

- Word-coded
- Contact person: Tanya Karoli Christensen: tkaroli@hum.ku.dk
- Status: Researcher level

In this tier, epistemic (including evidential) verbs and predicative constructions with epistemic adjectives are described: *tro, tænke, mene, vide* og *givet, klart, sikkert, tydeligt, indlysende, oplagt, muligt, sandsynligt, uklart, usikkert, umuligt, usandsynligt*.

For each of these words, the occurrence (F), construction type (K), clause constituent position (L), placement of the epistemic clause (P), tense (T), subject (S), sentence adverbial (A), meaning (B), and relations in the associated clause (C) are coded.

All codings start with "ES."

Occurrence:

The specific word is written here. For fixed expressions, the entire phrase is written. For example, *Ftro, Fdet/vil/sige*

Construction type:

The following codes are used:

KN	Predicate with NP-object
KS	Predicate with sentence object
KSK	Predicate with sentence object in the form of a sentence node
KSE	Predicate with elliptical sentence object
K0	Predicate that does not have a sentence object
KX	Uncertainty

Word order

The following codes are used:

LOVS	Object (NP), finite verb, subject
LSVO	Subject, finite verb, object (NP)
LSV	Subject, finite verb (without NP-object and in cases of KX)
LVS	Finite verb, subject (without NP-object and in cases of KX)
LA	Other
LX	Uncertainty

Placement of the epistemic clause

The following codes are used:

PI*	Placement initially with respect to the controlled/modified clause
PM	Placement medially with respect to the controlled/modified clause
PF	Placement finally with respect to the controlled/modified clause
PA	Placement alone (i.e., not controlling/modifying another clause)
PX	Uncertainty

*PI has several subcodes:

PIk	initial placement and subordinating conjunction
PIØ	initial placement and omitted subordinating conjunction
PI0	initial placement without possibility of subordinating conjunction
PIX	uncertainty

Tense

The following codes are used:

TN	present tense
TD	past tense
TFN	present perfect
TFD	past per-fect
TI	infinitive
TX	uncertainty

Subject

The following codes are used:

SP[form]	Pronouns, e.g., SPJeg and SPman
SL	Appellative
SA	Other, e.g., infinitive as *At rejse er at leve*.
S0	None
SX	Uncertainty

Sentence adverbial (in the sentence with the epistemic predicate)

The following codes are used:

A[form]	Write the form, e.g., Aikke or Aslet/ikke
A0	No sentence adverbial
AX	Uncertainty

Meaning

The following codes are used:

BU	epistemic uncertainty
BS	epistemic certainty
BA	Other kinds of meaning than epistemic
BX	Used in cases of KSE where it is not possible to determine if it's BU or BS.

Relationships in possible associated sentences

The following codes are used:

CP	Predicate in the associated sentence, CP[lemma]. If there is both a subordinate clause and a main clause, write the main clause's predicate first and then the subordinate clause's predicate in parentheses.
CT	Tense, with subcodes TN, TD, TFN, TFD. Any tense of the subordinate clause is written in parentheses. TI + S0= imperative T0 + verbal CP + S0 = infinitive
CS	Subject, with subcodes SP, SL, SA, S0, SX (e.g., CSPjeg). Any subject of the subordinate clause is written in parentheses.
CA	Any adverbials in the associated sentence, e.g., CAjo/ikke. Any adverbials of the subordinate clause are written in parentheses.
CL	Word order in the associated sentence (the first two to three positions in the sentence template), e.g., CLSV
	CLØVS: When the forfelt is empty
	CLHVS: HV-word in the forfelt
	CLXVS: Coded when there is something other than the subject in the forfelt
	CLOVS: In nodes, the part of word order that is not already indicated by the L-code for the epistemic sentence

For further information, refer to the coding manual *Kodningsmanual for epistemiske sætninger 23 02 15.doc* (not yet translated into English)

Examples:

Det ved jeg ikke hvad han havde set

ES Fvide KS LOVS PIk TN SPjeg Aikke BU CPse/have CTFD CSPhan CA0 CLSV

(double representation of argument)

Majbritt tror jeg hun hed.

ES Ftro KSK LVS PIØ TN SPjeg A0 BU CPhedde CTD CSPhun CA0 CLXSV

Det er svært at sige, fordi ..

ES Fsvært/at/sige K0 LSV PA TN SPdet A0 BU

2.4.6 semvar epistemicitet (semvar) – Semvar Epistimicity

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Contact person: Tanya Karoli Christensen: tkaroli@hum.ku.dk
- Status: Researcher Level.

In this tier, sentence adverbs for epistemicity are encoded. Some of these epistemic sentence adverbs are polysemous and/or homographs with other words. Such words are encoded with the tag "LA." The sentence adverbs are encoded for the categories: Occurrence (F), Tense (T), Verb (V), Subject (S), Negation (N), Clause Type (L), Meaning (B), and Placement (P).

All codes start with "E."

Occurrence

Here, the retrieved word is written, for example, *Fbestemt, Fgangskegivet*

Tense

The following codes are used:

TN	present tense
TD	past tense
TFN	present perfect
TFD	past perfect
TX	uncertainty

Verb

The lemma of the main verb is written after the code V. In coordinate constructions (e.g., 'går og synger' meaning 'walks and sings'), both lemmas are written with 'og' (and) in between (*Vgåogsynge*). Any modal verbs are written after the main verb, for example, *Vsekunne* (from the sentence *det kan de vel godt se*).

Subject

The subject in the sentence where the search word occurs. The following codes are used:

SP[form]	Pronouns. The form of the pronoun is indicated.
SL	Appellatives (proper nouns and nouns)
SA	Other
S0	No subject
SX	Uncertainty

Negation

The following codes are used:

N[form]	The form of negation is indicated.
N0	No negation

Clause Type

The following codes are used:

LP	Modal particle in the form of 'nok', 'vel', or 'vist'
LS	Sentence adverbial
LA	Other (all non-epistemic sentence adverbs)

Meaning

The following codes are used:

BS	Epistemic certainty
BU	Epistemic uncertainty
BX	Uncertainty

Placement

The following codes are used:

PF	Foundation field
PA	Sentence adverbial position
PH	Extraposition to the right

PV	Extraposition to the left
PU	Without finite sentence
PI	Infinitive sentence
PX	Uncertainty

Examples of coding:

ja **klart nok** (bysoc1ny-05-JKS)

E Fklartnok LP BS PU

nu er det godt **nok** noget siden jeg sidst har været derovre

E Fnok LA (koncessiv betydning)

men altså min mor havde **tilsyneladende** planlagt det

E Ftilsyneladende TFD Vplanlægge SA N0 LS BU PA

det er sgu ikke **sikkert**

E Fsikkert TN Vvære SPdet Nikke LS BU PA

og **sandsynligvis** så er der ikke noget om det

E Fsandsynligvis TN Vvære SA Nikke LS BU PV

Example of sequence coding: When, for instance, a token is annotated as E Foverhovedet V T S N LA B P, this token will have the following tag: 1_E_Foverhovedet_V_T_S_N_LA_B_P_E. The number before the tag indicates the position, and the letter after indicates whether it's the last tag in the sequence or not.

2.4.7 Ledsætning (ledsaet) – Subordinate clause

- Contact person: Torben Juel Jensen: tjuelj@hum.ku.dk.
- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Status: Researcher Level.

Coding:

All instances of subordination are coded with an ”L” and at least the code for Function, i.e., at least ”L F**E**” (for "ellipse") or ”L F**I**” (for "Incomplete"). The functions of ”ellipsis” and ”incomplete” are not coded further.

or occurrences of all other functions, i.e., ”adverbiel”: ”L F**A**”; ”relativ”: ”L F**R**”, ”samordning”: ”F**S**” or ”nominal”: ”L F**N**”, their sub-codes are also marked, for example, ”L FA**S**”, ”L FR**E**”, ”L FN**H**”.

Additionally, the code for Word **Order** is added, for example, ”L FRE **ON**”.

In connection with the subordinations categorized as ”nominal”, the code for Matrix Clause is also added, for example, ”L FNA OH **MP**”.

It is important to put spaces between the sub-codes as indicated above!

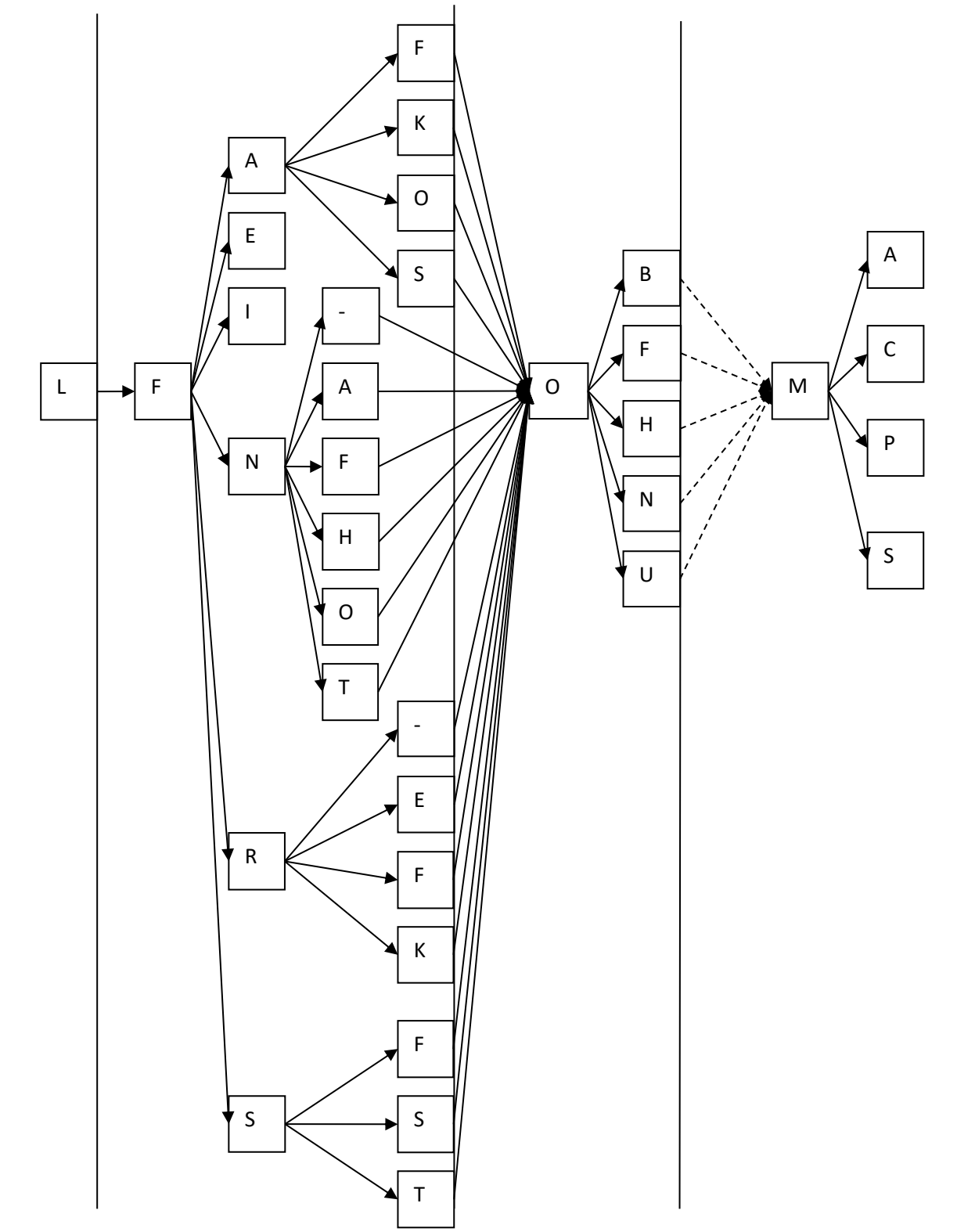
In case of doubt, the letter ”X” is used in the respective place, for example, ”L F**X**”, ”L FNT O**X** MA” or ”L FNA OH M**X**”. However, X-codes should not be found in the finished files. If an occurrence expresses two things at once, two codes are used (in alphabetical order), for example, ”L FAO **OHU**”.

For subordinations embedded in other subordinations, parentheses are put around the codes, including the code for the ”superior” subordination, for example, ”(L FAK ON (L FNA ON MA) (LFN- ON MA))”.

The tier is only found in the exploratory files (see the section ”Metadata på samtaleniveau”).

A complete overview of the different codes can be found in the document ”Analyseapparat(ledsætninger) 18 04 07” (not yet translated into English)

The following is an overview of the codes used:



Analysis apparatus for occurrences of generic pronouns.

All occurrences of search strings representing pronouns are marked with a 'G' along with the smallest subcode for 'usage,' i.e., 'G AS', 'GAA', 'G AI', and so on. In connection with occurrences categorized as other than 'Other usage' and 'Not completed,' the subcode for 'discourse function,' i.e., 'G AS DM', 'G AS DS,' or 'G AS DI,' is also added. For the first two types mentioned, subcodes for 'syntax' and 'reference' are then added, for example, 'G AS DS SB RA'.

In cases of uncertainty, the letter 'X' is used in the respective position, for example, 'G AX' or 'G AS DX SA RX.' For ambiguous occurrences, two codes are used (in alphabetical order), for instance, 'G AS DS SA RAJ' or 'G AS DMS SA RA'.

Categories:

Usage		**Discourse function**	
Tag	Description	Tag	Description
AS	Pronoun functions as a grammatical **subject**	DM	Truism or moral.
AO	Pronoun functions as a grammatical object (indirect or direct) or predicative	DS	Generalization of situation.
AP	Pronoun functions as a controller in a prepositional connection (also intermediate objects)	DI	Non-generalizing usage.
AB	Pronoun functions as a determiner in a nominal phrase.		
AA	Other usage.		
AI	Incomplete construction.		

Syntax		**Reference**	
Tag	Description	Tag	Description
SB	In conditional construction	RA	All people or a delimited group of people, including both the speaker and the addressee
SP	In a complement clause (often a nominal clause) to a projecting sentence	RJ	Delimited group of people including the speaker (Jeg), but not the addressee
SA	Other	RD	Delimited group of people including the addressee (Du), but not the speaker
		RH	Delimited group of people that includes neither the speaker nor the addressee

Example of sequence coding: For instance, when ten tokens are annotated (L FNA ON MA), these ten tokens will have these ten tags: 1_(L_FNA_ON_MA_I, 2_(L_FNA_ON_MA_I, 3_(L_FNA_ON_MA_I, […], 9_(L_FNA_ON_MA_I, 10_(L_FNA_ON_MA_E). The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.4.8 GEX Generel udvider (GEX) – General Extender

- Contact person: Torben Juel Jensen: tjuelj@hum.ku.dk.
- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Status: Researcher Level.

Refer to the document ”Analyseapparat for general extenders” (not yet translated into English) for further information.

General extenders are expressions that typically consist of a coordinating conjunction (*og*/*eller*) and a generic nominal phrase that refers to something immediately preceding it. In a broader sense, these are expressions that serve the function of invoking the notion of a larger quantity/category based on the presumed knowledge of the referent. In any case, the expression is postposed.

Coding:

All occurrences of expressions that function as general extenders are labeled with an interval in the GEX tier. The label indicates, at a minimum, the code for 'expression,' such as 'UI' or 'UNN\and\or\something\like\that.' Unless the occurrence is categorized as 'UI,' the codes for 'other elements in the phrase (L*),' 'positioning in the utterance (Y*),' and 'turn shift (T*)' are also added, for example, 'UNA\and\the\whole\mess LNIF\2 YF T0.'

In cases of uncertainty, the letter 'X' is used in the respective position, for instance, 'UX' or 'UNA\and\the\whole\mess LNIF\2 YX T0.

Categories:

U: Expression (type and orthographic representation of words with’ \’ instead of spaces)

L: ther elements in the phrase (type and the number of preceding elements separated by ”\”)

Y: Positioning in the utterance

T: Change in Turn

Expression		**Other elements in the phrase**	
Tag	Description	Tag	Description
UNN	Nominal phrase with noget/nogen/nogle as the core	LNP(E\F)	The other elements in the phrase are nominals or nominal phrases referring to persons or other agentive entities
UNT	Nominal phrase with ting/alting as the core	LNI(E\F)	The other elements in the phrase are nominals or nominal phrases referring to individualized entities
UNA	Nominal phrase with another core	LNM	The other elements in the phrase are nominals or nominal phrases referring to non-individualized masses

UA	Adverbial phrase	LPAP	The other elements in the phrase are other forms of predicates than nominal predicates
UP	Prepositional construction	LPØ	The other elements in the phrase consist of other types of predicates than predicative constructions
UØ	Other expressions	LC	The other elements in the phrase consist of independent utterances that function as actual or inferred quotes
UI	Not completed or occurs in an interrupted utterance before expressing a meaning	LB	The other elements in the phrase consist of several different types mentioned above (mixed)
		LA	The other elements in the phrase are of a different type than those mentioned above
		LU	The other elements in the phrase are unidentifiable or cannot be distinctly delineated

Placering i ytring		**Turskifte**	
Tag	Beskrivelse	Tag	Beskrivelse
Y0	The phrase and GEX in itself constitute an utterance (zero)	T0	The speaker continues without the listener's contribution, i.e., zero turn shift
YI	The GEX is placed **initially** in the utterance, before the finite verb but after the phrase it attaches to	TP	The speaker continues without the listener's contribution but after a **pause**
YM	he GEX is placed **medially** in the utterance	TM	The speaker retains the turn, but there is a **minimal** contribution from the listener
YF	The GEX is placed **finally** in the utterance	TOV	The speaker continues, but is **overlapped** for a while by the listener
YU	The placement in the utterance is **undecided**	TS	Turn **shift**
		TU	The phrase's placement in relation to turn-taking is **undecided**

Examples:

UA\eller\hvad LNPE\1 Y0 T0

UA\og\så\videre LPAP\1 YF T0

UA\og\så\videre LPØ\1 YF TM

UA\og\sådan LPØ\1 YFI TP

UNA\eller\et\eller\andet LB\2 YF T0

UNA\eller\sådan\et\eller\andet LNPE\1 YF T0

UNN\eller\noget LA\1 YF TM

UNN\eller\sådan\noget LA\2 YF TM

UNN\eller\sådan\noget LC\1 YF TP

UNN\og\alt\sådan\noget\underligt\noget LPØ\1 YF TS

UNN\og\sådan\noget LA\1 YI T0

UNN\og\sådan\noget LNIF\1 YF T0

UNN\og\sådan\noget LNIF\1 YF TM

Example of sequence coding: When, for instance, two tokens are annotated UA\og\sådan LPØ\1 YF TP, those two tokens will receive these two tags: 1_UA\og\sådan_LPØ\1_YF_TP_I, 2_UA\og\sådan_LPØ\1_YF_TP_E. The number before the tag indicates the position, and the letter after indicates whether it's the last tag in the sequence or not.

2.4.9 at-tab-…

- Contact person: Kasper Boye: boye@hum.ku.dk

There are a number of different tiers related to the loss of the subordinating conjunction 'at.' So far, it has not been possible to provide a comprehensive description of these tiers. The tiers are:

- at-tab-ekspl
- at-tab-ekspl-komm
- at-tab-vide
- at-tab-vide-komm
- at-tab-mene
- at-tab-sige
- at-tab-sige-komm
- at-tab-kunnehuske
- at-tab-adj
- at-tab-nogetmed
- at-tab-tro
- at-tab-tro-komm

2.4.10 Case

- Word-coded
- Status: Open
- Contact person: Jeffrey Parrott

Concerns case variation in personal pronouns occurring in coordinated noun phrases. Both pronouns and conjunctions are coded.

It is coded for person (P), number (N), gender (G), case (Ca), conjunct (Co), and coordinate head (C0) (the conjunction itself):

Person

The following codes are used:

P1	first person
P2	Second person
P3	Third person

Number

The following codes are used:

Nsg	Singular
Npl	Plural

Gender

The following codes are used:

Gfe	female
Gma	male

Case

The following codes are used:

CaSF	Subject form
CaOF	Oblique form

Conjunct

The following codes are used:

Co1	First element in the coordinated noun phrase
Co2	Second element in the coordinated noun phrase
Co3	Third element in the coordinated noun phrase
Co*n*	*n*-th element in the coordinated noun phrase

Coordinate head

This indicates the noun phrase's function in the sentence. This category has several subcategories. The complete list can be found in the coding manual *Skelton key 2.doc*.

Example:

hende	*og*	*jeg*	*drikker*	*øl*
P3 Nsg Gfe CaOF Co1	C0 SM	P1 Nsg CaSF Co2		

2.4.11 GIDDY Gender in Danish and Dutch (GIDDY)

- Word-coded
- Contact person: Contact person: Torben Juel Jensen: tjuelj@hum.ku.dk (originally Frans Gregersen)
- Project: GIDDY (Gender in Danish and Dutch)

In this tier, the grammatical gender used in nouns, adjectives, pronouns, and constructions is described. Each tag displays the word class along with the given word(s). Additionally, values for relevant categories are provided - for a large portion of the tag set, it entails the expected and realized gender.

The following tags have been used:

Nouns

- **D**eterminers (indefinite, singularis)
- DUU[lemma]: standard **U**trum, usus **U**trum — fx DUUstol = *en stol*
- DNN[lemma]: std. **N**eutrum, us. **N**eutrum — fx DNNbord = *et bord*
- DUN[lemma]: std. U, us. N — fx DUNstol = *et stol*
- DNU[lemma]: std. N, us. U — fx DNUbord = *en bord*
- DT[evt. lemma]: uncertainty

- Suffixes **(B**øjningssuffikser) (definit, singularis)
- BUU[lemma]: std. U, us. U — fx BUUstol = *stolen*
- BNN[lemma]: std. N, us. N — fx BNNbord = *bordet*
- BUN[lemma]: std. U, us. N — fx BUNstol = *stolet*
- BNU[lemma]: std. N, us. U — fx BNUbord = *borden*
- BI0[lemma]: naked form, +expected genus — fx BI0måde = *[på sjov] måde*
- BI1[lemma]: naked form, ÷expected genus — fx BI1skole = *[gå i] skole*
- BT[evt. lemma]: uncertainty

Adjectives (singularis)

- AUUA*(attributive)*[lemma_referent]: std. U, us. U — fx AUUAgod_stol = *[en] god [stol]*
- AUUAA[lemma_referent]: std. U, us. U, deviant (afv.) article — fx *et god stol*
- AUUUA[lemma_referent]: std. U, us. U, without (u) art — fx *gå med rød vest*
- AUUP*(predicative)*[lemma_referent]: std. U, us. U — fx AUUPgod_stol = *[stolen er] god*

- ANNA[lemma_referent]: std. N, us. N — fx ANNAgod_bord = *[et] godt [bord]*
- ANNAA[lemma_referent]: std. N, us. N, afv. Art. — fx en godt bord
- ANNUA[lemma_referent]: std. N, us. N, u art — fx gå i sjovt tøj
- ANNP[lemma_referent]: std. N, us. N — fx ANNPgod_bord = *[bordet er] godt*
- AUNA[lemma_referent]: std. U, us. N — fx AUNAgod_stol = *[en] godt [stol]*
- AUNAA[lemma_referent]: std. U, us. N, afv. Art. — fx et godt stol
- AUNUA[lemma_referent]: std. U, us. N, u art — fx gå med rødt vest
- AUNP[lemma_referent]: std. U, us. N — fx AUNPgod_stol = *[stolen er] godt*
- ANUA[lemma_referent]: std. N, us. U — fx ANUAgod_bord = *[et] god [bord]*
- ANUAA[lemma_referent]: std. N, us. U afv. Art. — fx en god bord
- ANUUA[lemma_referent]: std. N, us. U u art — fx gå med sjov tøj
- ANUP[lemma_referent]: std. N, us. U — fx ANUPgod_bord = *[bordet er] god*
- AIA [item_referent]: attributive, ÷genusmarker — fx AIAlette_bord = *[det] lette [bord]*
- AIAA: attributive, -genusmarker, deviant article
- AIP[item_referent]: predicative, ÷genusmarker — fx AIPhele_ABSTRAKT= *[det] hele*
- AØA[lemma_referent]: indeterminate due to context — fx en halv time
- AT[evt. lemma]: uncertainty

Pronouns (singularis)

- Pronominal references to nominals
- PUUA[lemma_referent]: std. U, us. U — fx PUUAmin_stol = *min [stol]*
- PUUP[lemma_referent]: std. U, us. U — fx PUUPmin_stol = *[stolen er] min*
- PNNA[lemma_referent]: std. N, us. N — fx PNNAmin_bord = *mit [bord]*
- PNNP[lemma_referent]: std. N, us. N — fx PNNPmin_bord = *[bordet er] mit*
- PUNA[lemma_referent]: std. U, us. N — fx PUNAmin_stol = *mit [stol]*
- PUNP[lemma_referent]: std. U, us. N — fx PUNPmin_stol = *[stolen er] mit*
- PNUA[lemma_referent]: std. N, us. U — fx PNUAmin_bord = *min [bord]*
- PNUP[lemma_referent]: std. N, us. U — fx PNUPmin_bord = *[bordet er] min*
- Special pronominal references
- PF[item]: placeholder — fx PFdet = *det [er i orden]*
- PHU[lemma]: higher-order reference, us. U — fx PHUden = *den [er fin med dig]*
- PHN[lemma]: higher-order reference, us. N — fx PHNden = [*jeg skal nok gøre*] *det*
- PIU[lemma]: conditional reference, us. U — fx PIUdenne = *[brug] denne [her]*
- PIN[lemma]: conditional reference, us. N — fx PINdet = *[giv mig] det [der]*
- PVU[lemma]: vague/unclear reference, us. U
- PVN[lemma]: vague/unclear reference, us. N
- PT[evt. lemma]: uncertainty

Construktions (connections of multiple words) with genusmarking

- K[entire construction/all words] — fx Kenafvennerne = *en af vennerne*
  fx Ketellerandet = *et eller andet*

Examples:

ANNPnormal_arabisk_SPROG

BI1ingeniør

BI1lejrskole

BNNhold

Knogetaf

PNNPden_tørklæde

PUUAden_liga

PVNden

2.4.12 Hjælpeverber (AUX) – Auxillary Verbs

- Word-coded
- Contact Person: Anu Laenemets
- Status: Open

In this tier, the last past participle in verbal connections (main verbs) is encoded. These verbs are encoded for the categories Auxiliary verb (AU), Voice (D), Argument structure (R), Adverbials (A), Type of event (E), Subject (S), Coordination and ellipsis (K), Distance (O), and Pronunciation (U).

Auxiliary verb:

Inventory:

AUH/_	The auxiliary verb is 'have'. The form is written after the forward slash. Multi-word connections are written with forward slashes between each word, e.g.: AUH/ville/have
AUV/_	The auxiliary verb is ’være’. The form is written after the forward slash as described above.
AUB/_	The auxiliary verb is ’blive’. The form is written after the forward slash as described above.
AUF/_	The auxiliary verb is ’få’. The form is written after the forward slash as described above.
AUX	Uncertainty
AUZ	This code indicates that it could not be determined whether ’har’ or ’er’ was said
AUQ	Incomplete construction. If it is obvious which word is missing, the relevant word is coded with Q.
AUY	If the word is tagged incorrectly (in the PoS tier).

Diathesis:

Inventory:

DP	Passive

DA	Active

Argument structure:

Inventory:

RI	Intransitive
RT	Transitive
RO	Prepositional object
RP	Predicative Construction
RX	Uncertainty

Adverbials:

Note: Sentence adverbials are not noted except for negations (*ikke, aldrig*)

For each tag, all subcategories for adverbials should be written. If a particular type of adverbial does not appear, <Q> is added to the tag to indicate this. If an adverbial consists of multiple words, they are written without spaces between them.

Inventory:

AK/_	Incorporation and transferred/metaphorical meaning. The form is written after the forward slash.
AR/_	Directional adverbials. The form is written after the forward slash.
AS/_	Locational adverbials. The form is written after the forward slash.
AT/_	Temporal adverbials. The form is written after the forward slash.
AM/_	Manner adverbials. The form is written after the forward slash.
AZ/_	Other types of adverbials. The form is written after the forward slash.
AX	Uncertainty
A_/Q	Unfinished adverbials where you can see the type but not which word. In this case, Q stands in place of the word. For example, AK/iQ. Note that this code differs from AKQ, which indicates that there are no adverbials of the AK type.

Type of event:

Inventory:

ET	Telic
EA	Atelic
ES	State
EX	Uncertainty

Subject:

Subject code consists of two positions: S12. Position 1 describes if the subject is animate (A) or inanimate (I), and position 2 describes the subject's form (noun or pronoun). Thus, S will be followed by either A or I, and then a code describing the subject's form. Nouns are coded with S, and pronouns with P. Personal pronouns are coded as follows: *jeg* = P1, *du* = P2, *han/hun* = P3, *vi* = P4, *I* = P5, *de* = P6.
Impersonal pronouns (*man, nogen*) are coded with U, and relative pronouns are coded with R.

Formal subject is coded with F in both positions (*SFF*), while placeholder and real subject are coded with F in the second position (*SIF*). X is used in cases of uncertainty.

SQ indicates that a subject is missing.

Inventory:

First position	
SA_	animate subject
SI_	inanimate subject
SX_	uncertainty
SQ	No subject
Second position	
S_S	subject = noun
S_P1	subject = *jeg*
S_P2	subject = *du*
S_P3	subject = *han/hun*
S_P4	subject = *vi*
S_P5	subject = *I*
S_P6	subject = *de*
S_U	impersonal pronouns
S_R	relative pronouns
SFF	formal subject
S_F	Placeholder/real subject
S_P	subject = other pronoun

Coordination and ellipsis:

n coordination, main verbs are marked with K1 and K2 (and possibly K3) to show that they belong together. For ellipsis, the codes used are KE1 and KE2 (and possibly KE3) instead. If it's not about coordination/ellipsis, KQ is used.

Distance:

This code indicates the number of words between the auxiliary verb and the main verb.

O0	Zero words between the auxiliary verb and the main verb
O1	One word between the auxiliary verb and the main verb
O2	Two words between the auxiliary verb and the main verb
On	N words between the auxiliary verb and the main verb

Pronunciation:

Here, it's noted which word precedes *er/har*

U/_	After the slash, the word preceding *er/har* is indicated
UP	This indicates there's a pause before *er/har*
UQ	This indicates that the auxiliary verb is something other than *er/har*

Examples:

AUB/blev D R AK AR AS AT AM AZ E S K O U

AUH/har DA RT AK AR AS AT AM AZ E S K O U

AUV/er DA RI AK/påefterløn ARQ ASQ ATQ AMQ AZQ ET SAP3 KQ O3 U/han

AUV/var DA RI AK/nediomsætning ARQ ASQ ATQ AM/fyrreprocent AZQ ET SIS KQ O1 UQ

2.4.13 Kløvning 1 (KL)

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Contact person: Marie Herget Christensen: wzr818@hum.ku.dk
- Project: Kløvninger

Note: It has not been possible to provide a fully comprehensive description of this tier so far. Also, please note that it is only found in files with encoding errors.

- KLdet Fdet KPA LA0 KON- FOK0 PRÆS4 K-
- KLdet Fdet KPA LAikke KONim FOK1 PRÆS4 Khvor
- KLdet Fdet KPO LAjo KON- FOKal PRÆS4 Kat
- KLdet FsÃ¥ KPA LAfÃ¸rst KONim FOKal PRÆS4 Kat

1. ((L, (L, F, KL0, KLder, KLdet, KLel, KLo, KLpeu, KLpseu, KLpseudo, L
2. FAK, FAO, Fdengang, Fder, Fdet, Fhv-, Fhvad, FhvDad, Fhvem, Fhvor, Fhvornår, FI, Fmåske, FN-, FNA, Fnogle, Fnu, FØ, FRE, FRF, FRK, Fså, FSS
3. KMO, KP0, KPA, KPMO, KPO, KPS, KPSP, LA, meget, OB, OH, OHU, ON, ON)))
4. LA, LA0, LA1, LAalligevel, LAaltid, LAaltså, LAaltså\ikke, LAbare, LAda, LAda\bare, LAda\også, LAegentlig, LAegetnlig, LAellers, LAfaktisk, LAfaktisk\ikke, LAførst, LAgodt, osv.

5. KOim-, KON, KON-, KONek, KONex, KONi, KONim
6. FOK, FOK0, FOK1, FOK2, FOKal, FOKO, FOKØ
7. PRÆS, PRÆS2, PRÆS3, PRÆS3S, PRÆS4
8. K, K-, K., K0, Kat, Ksom, Ksom\der

Example of sequence coding: If, for instance, eight tokens are annotated *KLdet Fdet KPMO LA0 KON- FOK1 PRÆS2 K0*, those eight tokens will get these eight tags: 1_KLdet_Fdet_KPMO_LA0_KON-_FOK1_PRÆS2_K0_I, 2_KLdet_Fdet_KPMO_LA0_KON-_FOK1_PRÆS2_K0_I, 3_KLdet_Fdet_KPMO_LA0_KON-_FOK1_PRÆS2_K0_I, […], 7_KLdet_Fdet_KPMO_LA0_KON-_FOK1_PRÆS2_K0_I, 8_KLdet_Fdet_KPMO_LA0_KON-_FOK1_PRÆS2_K0_E . The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.4.14 Kløvning 2 (kl)

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Contact person: Marie Herget Christensen: wzr818@hum.ku.dk
- Project: Kløvninger

Note: It has not been possible to provide a fully comprehensive description of this tier so far. Also, please note that it is only found in files with encoding errors.

KL0X
KLel
KLdet_Fdet_KPMO_LAjo_KON-_FOK1_PRÆS4_(KLdet_TMP_TLSP_Fdet_KPO_NP_AN0_DEFI_PAT_TING_Vvære_KØ_PRÆS1_OU)
KLdet_Fdet_KPO_LA0_KON-_FOK2_PRÆS4(KLdet_TMPR_TLSPR_Fdet_KPO_AP_Vvære_KØ_PRÆS1_OU)__
KLder
KLdet_Fdet_KPO_LA0_KON-_FOK1_PRÆS4(KLdet_TMPR_TLSPR_Fdet_KPO_NP_AN0_DEFI_PAT_TING_Vvære_KØ_PRÆS1_OU)
KLdet_Fdet_KPMO_LA0_KON-_FOK0_PRÆS4(KLdet_TMPART_TLSPART_Fdet_KPMO_NP_AN0_DEFI_ØV_TING_Vvære_KØ_PRÆS1_OU)
KL0an
KLdet_Fhv-_KP0_LA0_KON-_FOKØ_PRÆS2(KLdet_TMP_TLSP_Fhv_KPO_XP_Vvære_KØ_Præs1_OU)
KLdet_Fdet_KPMO_LA0_KONex_FOK1_PRÆS3(KLdet_TMP_TLSP_Fdet_KPMO_NP_AN0_DEFD_PAT_AB_Vvære_KØ_PRÆS1_OU)_
KLdet_Fhv-_KPO_LA0_KON-_FOKØ_PRÆS2(KLdet_TMP_TLSP_Fhv_KPO_XP_Vvære_KØ_Præs1_OU)
KLdet_Fdet_KPO_LA0_KON-_FOK1_PRÆS4(KLdet_TMPR_TLSPR_Fdet_KPO_NP_AN0_DEFI_PAT_IND_Vvære_KØ_PRÆS1_OU)

Annotation positions:

1. KL (kløvningstype?): 0X, el, det, der, 0an
2. F (fokus?): det, hv-
3. KP (komplement?): MO, O, 0
4. LA: jo, 0
5. KON-:
6. FOK: 0, 1, 2, Ø
7. PRÆS: 2, 3, 4

In parentheses:

1. KL: det
2. TM: P, PR, PART,
3. TLS: P, PR, PART
4. F: det, hv
5. KP: O_NP, O_AP, O_XP, MO_NP ..
6. V: være
7. K: Ø
8. PRÆS: 1
9. O: U

2.5 Dialect tiers

- Contact person: Malene Monka: monka@hum.ku.dk
- Status: Researcher level

Note: In Odder, the tiers where 32 occurrences are marked are called "endelser_32," "ordmarkering_32," and "blodt_d_32," as for some informants, fully coded interviews are also included.

2.5.1 Bylderup artikel (Byl_ARTIKEL) – Bylderup Article

- Word-coded

Coding for the dialect's anterior and the standard language posterior article.

Inventory:

- 1 = dialect
- 2 = standard
- 3 = indeterminable

2.5.2 Bylderup IKKE (Byl_IKKE) – Bylderup IKKE

- Word-coded

Study of dialect versus standard variant of the negation *ikke.*

Inventory:

- 1 = dialect
- 2 = standard
- 3 = indeterminable

2.5.3 Bylderup JEG (Byl_JEG) – Bylderup JEG

- Word-coded

Study of the dialect versus standard pronunciation of *jeg.*

Inventory:

- 1 = dialect
- 2 = standard
- 3 = indeterminable

2.5.4 Bylderup omlyd (Byl_OMLYD) – Bylderup umlaut

- Word-coded

Study of possible dialect variance in the words *har, går, får, står, slår, foreslår*.

Inventory:

- 1 = dialect
- 2 = standard
- 3 = indeterminable

2.5.5 Vinderup artikel (artikel) – Vinderup article

- Word-coded

Investigation of anterior articles.

Inventory:

- 1 = dialect
- 2 = standard
- 3 = indeterminable

2.5.6 Blødt D Odder (blodt_d) – Soft d Odder

- Word-coded

Investigation of dialectal vs standard pronunciation of soft d.

Inventory:

- D0 = standard
- D1 = dialect
- D2 = neutral
- D3 = regional
- D4 = indeterminable
- D = an automatic coding script has been established. Some of the words are not suitable or fall outside the ratio coded within.

2.5.7 Blødt_D_32 Odder (blodt _d_32) – Soft d Odder, 32

- Word-coded

Investigation of dialectal vs standard pronunciation of soft d, where only the first 32 occurences after 900 seconds are annotated.

Inventory:

- D = an automatic coding script has been established. Some of the words are not suitable or fall outside the ratio coded within.

- D0 = standard
- D1 = dialect
- D4 = indeterminable

2.5.8 Vinderup IKKE (ikke) – Vinderup IKKE

- Word-coded

Investigation of dialectal vs standard pronunciation of *ikke*.

Inventory:

- 1 = dialect
- 2 = standard
- 3 = indeterminable

2.5.9 Udtale JEG (jeg) – JEG pronunciation

- Word-coded

Investigation of dialectal vs standard pronunciation of *jeg*.

Inventory:

- JA = an automatic coding script has been established. Some of the words are not suitable or fall outside the ratio coded within.
- JA0 = standard
- JA1 = dialect

2.5.10 Endelser (endelser) – Ending inflections

- Word-coded

1. Dialect or standard pronunciation of -et in the participle, definite form, etc. (the word "meget" is marked separately). The feature is defined by Nielsen and Nyberg as follows: R = [E6], -R = [Et, Ed] (Nielsen and Nyberg 1988:46).

2. Dialect or standard pronunciation of -ede in the past tense (here marked are the first 32 occurrences after 900 seconds)

Inventory:

- END0 = standard
- END1 = dialect
- END2 = regional
- END3 = -en inflection of strong verbs
- END4 = indeterminate

2.5.11 Endelser_32 (endelser_32) – Ending Inflections 32

- Word-coded (each token (word) is annotated separately).

Check the endings. The first 32 occurrences after 900 seconds are marked.

Inventory:

- END0 = standard
- END1 = dialect
- END2 = regional
- END3 = -en inflection of strong verbs
- END4 = indeterminate

2.5.12 Skriftens OR (skriftens_or) – OR in ortografy

- Word-coded

Dialect or standard pronunciation of -or in the following words: gjorde, gjort, fjorten, historie, horn, jordbær, lort, skjorte, sort, torden, fjortende, fjortenårig, lortet, and tror.

Inventory:

- OR0 = standard
- OR1 = dialect
- OR4 = indeterminate

2.5.13 Langt O-vokalisme (OOvokalisme) – Long O vocalisms

- Word-coded (each token (word) is annotated separately).

Investigation of diphthongization of long 'o'.

NOTE: In certain codings, the code (e.g., OO) is not deleted in front of the number, meaning long 'o' pronounced with dialect is OO1.

Inventory:

- OO = an automatic coding script has been established. Some of the words are not suitable or fall outside the ratio coded within.
- 0 = standard
- 1 = dialect
- 3 = regional
- 4 = indeterminate
- 5 = *gjort* pronounced without final t

2.5.14 Langt E-vokalisme (EEvokalisme) – Long E vocalisms

- Word-coded (each token (word) is annotated separately).

Investigation of diphthongization of old long 'e'.

Inventory:

- EE = an automatic coding script has been established. Some of the words are not suitable or fall outside the ratio coded within.
- 1 = standard
- 2 = dialect
- 3 = indeterminable

2.5.15 Ordmarkering (ordmarkering) – Word Marking

- Word-coded (each token (word) is annotated separately).

Marking of *ikke* and *jeg* in Vinderup (and also in Odder?).

Inventory:

- DS0 = standard
- DS1 = dialect
- DS2 = neutral
- DS3 = regional
- DS4 = indeterminable

2.5.16 Ordmarkering_32 (ordmarkering_32) - Word Marking 32

- Word-coded (each token (word) is annotated separately).

Marking of *ikke* and *jeg* in Vinderup (and also in Odder?). The first 32 occurrences after 900 seconds are marked.

Inventory:

- ORD = an automatic coding script has been established. Some of the words are not suitable or fall outside the ratio coded within.
- ORD0 = standard
- ORD1 = dialect
- ORD4 = indeterminable
- ORD5 = regional

2.5.17 Vestjysk stød (V) - West Jutlandic Stød

- Word-coded (each token (word) is annotated separately).

Investigation of West Jutlandic stød.

Inventory:

- V= an automatic coding script has been established. Some of the words are not suitable or fall outside the ratio coded within.

- V0 = standard
- V1 = dialect

2.5.18 Klusilspring (K) – Clusilic Sprouting

- Word-coded (each token (word) is annotated separately).

Investigation of clusilic sprouting.

Inventory:

- K = an automatic coding script has been established. Some of the words are not suitable or fall outside the ratio coded within.
- K0 = standard
- K1 = dialect

2.6 The Køge Project

- Contact person: Janus Spindler Møller: janus@hum.ku.dk
- Status: Open

2.6.1 Tyrkisk-dansk (TR-DK) – Turkish to Danish

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

Translations from Turkish to Danish.

Example of sequence coding: When, for instance, three tokens are annotated with "hun_snakkede_aldrig," the three tokens will receive these three tags: 1_hun_snakkede_aldrig_I, 2_hun_snakkede_aldrig_I, 3_hun_snakkede_aldrig_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.7 Language Attitude Tiers

- Contact person: Tore Kristiansen: tk@hum.ku.dk
- Status: Open

2.7.1 SH Sprogholdning (SH)

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

In this tier, instances related to language attitudes are marked. The used tags are simply used to identify passages relevant to language attitudes. There are three types of relevant passages:

SHsp	Passages related to the language attitudes study.
SHyt	Passages with direct expressions of or about language attitudes.
SHin	Passages interesting in connection with language attitudes (reading habits, TV habits, radio habits, etc.). (This category was not used and cannot be found in the corpus.)

Example of sequence coding: When, for instance, three tokens are annotated with SHsp, the three tokens will receive these three tags: 1_SHsp_I, 2_SHsp_I, 3_SHsp_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

Refer to the coding manual *kodningsmanualSH.doc* for further information. (Not yet translated into English).

2.8 Hanne Sæderup

- Contact person: Hanne Sæderup, sbc846@hum.ku.dk
- Status: Private (all tiers as well as audio files).

2.8.1 DP Discourse Presentation (DP)

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

See chapter 4 in the document HANNE_SÆDERUP_PHD.pdf (not yet translated into English)

Discourse Presentation

”På basis af en kodningsmanual, jeg har afstemt til mine data, opmærkes korpusset på tre niveauer: ift. gengivelsesmodus, ift. tale-, skrift- og tankegengivelseskategorier og endelig ift. grammatiske realiseringer af disse kategorier”

Plads 1: Discourse presentation

SIS	Speech presentation, indirect speech
SFIS	Speech presentation, free indirect speech
SDS	Speech presentation, direct speech
SRV	Speech presentation, representation of voice
SRSA	Speech presentation, representation of speech act
WRN	Writing presentation, representation of writing
WRWA	Writing presentation, representation of writing act
WIW	Writing presentation, indirect writing
WFIW	Writing presentation, free indirect writing
WDW	Writing presentation, direct writing
TRT	Thought presentation, representation of thought
TRTA	Thought presentation, representation of thought act
TIT	Thought presentation, indirect thought
TFIT	Thought presentation, free indirect thought
TDT	Thought presentation, direct thought
RU	Report of language use

Plads 2 (modifikatorer til plads 1):

h (hp, hn, hf)	Hypothetical (proper, negation, future)

i	Non-specific discourse presentation (generic and iterative)
s	Specific discourse presentation
e	Embedded discourse presentation
p	RV/RW/RN or RSA/RWA/RTA with topic (topic = p)
m	Metonymic discourse presentation
inel	Interactional, elicited
in	Interactional discourse presentation
u	Unfinished discourse presentation
s1sing, s3sing, etc.	Speaker Voice (relevant for all instances of discourse presentation)

Plads 3: Taler. 3 underpladser:

Person (S1, S2, S3)	Speaker 1st person, speaker 2nd person, speaker 3rd person
Tal (s, p)	Singular, plural
Rolle (d, p, dy, dg, ...)	Doctor, patient, doctor (ps**y**chiatrist), doctor (**g**eneral), ..

Plads 4: Genericity

Rs	Specific
Ri	Non-specific

Plads 5: Hypothetical discourse presentation

Hp	Hypothetical, proper
Hn	Hypothetical, negated
Hf	Hypothetical, future

2.9 Vollsmose ... (LaPUR)

- Contact person: Pia Quist: pqj@hum.ku.dk
- Status: NORS level

See the document ”Tiers til kodning af Vollsmose” for additional information (not yet translated into English).

Additonal information also available through Astrid Ravn Skovses ph.d.-thesis (Danske Talesprog 18), which concerns this data as well as the data from Bylderup.

2.9.1 Blødt D Vollsmose (Blødt d) – soft d Vollsmose

- Word-coded (each token (word) is annotated separately).

ð-elipsis inside of words.

Inventory:

u d	Short for ”uden d” (eng. without d)

2.9.2 Udtale -EDE (EDE) – pronunciation -EDE

- Word-coded (each token (word) is annotated separately).

Realization of the ending, weak verbs in the past tense with the ending -ede, respectively standard [əðə] and Funen [əd]. For example, "hoppede" (jumped).

Inventory:

0	Indeterminable
1	standard
2	dialect
EDE	General annotation to identify potentially interesting places
u d	short for ”uden d” (eng. without d)

2.9.3 Udtale -ET (ET) – pronunciation -ede

- Word-coded (each token (word) is annotated separately).

Realization of the ending in various word forms with -et, respectively standard [əð] and Funen [əd]. For example, "huset" (the house), "fjollet" (silly), (has) "hoppet" (jumped).

Inventory:

0	Indeterminable
1	standard
2	dialect
ET	General annotation to identify potentially interesting places

2.9.4 Fortsætterintonation (Fortsætterint) – Continuing intonation

- Word-coded (hvert token (ord) er annoteret separat).

Continuing intonation

fort (or f or 1)	Continuing intonation

2.9.5 Fynsk intonation (Fynsk int) – Funen intonation

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

Special intonation patterns.

The tier described in the coding manual is called "Intonation."

Inventory:

fynsk	Funen intonation
ME	Multietnolect

Example of sequence coding: When, for instance, three tokens are annotated with "fynsk," the three tokens will have these three tags: 1_fynsk_I, 2_fynsk_I, 3_fynsk_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.9.6 Grammatisk køn (Gramm. køn) – Grammatical gender

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

Variation in grammatical gender. The annotation spans a given word and its article.

Inventory:

i	Neuter

f	Common

Example of sequence coding: When, for instance, three tokens are annotated with "i," the three tokens will have these three tags: 1_i_I, 2_i_I, 3_i_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.9.7 Kommentarer - Comments

Commentary tier.

Examples:

"kraftigt""""" t""""" i tænkte og tiden;

"thorn""""-lyd; "under""""" minder om Ebru;

holder f0 næsten konstant meget længe;

ME;

ME-agtigt;

obs speciel intonation

2.9.8 Leksis - Lexis

- Word-coded (each token (word) is annotated separately).

Remarkable things in the vocabulary. For example, "walla."

Only 3 filled intervals.

2.9.9 Manglende vokalkontrast (Mgl vokalkontrast) – Missing Vowel Contrast

- Word-coded (each token (word) is annotated separately).

Abbreviation of long vowels, e.g., [sgolə] instead of [sgo:lə].

Inventory:

f	for ”forkortelse” (eng. abbreviation)

2.9.10 Palatalisering - Palatalization

- Word-coded (each token (word) is annotated separately).

Palatalization: /t/ → [tj]

Inventory:

0, 1, 2, R

0	Indeterminable
1	Standard
2	dialect
TJ	General annotation to identify potentially interesting places

2.9.11 Præposition (Præp) - Preposition

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).

Deviant use of prepositions. The annotation spans several words as otherwise, a missing preposition cannot be marked.

Inventory:

i/af/på… + d	Drop of preposition (e.g., "vi går skole" instead of "vi går i skole")
i/af/på… + a	Deviation in preposition (e.g., "jeg skal på skole" instead of "jeg skal i skole")
f	The meaning of this annotation has been lost

Example of sequence coding: When, for instance, three tokens are annotated with på+a, the three tokens will have these three tags: 1_på+a_I, 2_på+a_I, 3_på+a_E. The number before the tag indicates the position, and the letter after indicates whether it is the last tag in the sequence or not.

2.9.12 Stødafvigelse (Stødafv) – Stød Deviation

- Word-coded (each token (word) is annotated separately).

Stød deviation – either missing stød or "extra" stød where stød was not expected.

Inventory:

m s	minus stød
p s	plus stød
u s	uden stød

2.9.13 Ustemt R (Ustemt r) – Voiceless R

- Word-coded (each token (word) is annotated separately).

A particular quality of consonant + r that is difficult to distinguish from a long voiced plosive.

Inventory:

0	Indeterminable
1	standard
2	dialect
R	General annotation to identify potentially interesting places

2.9.14 V2->V3 (V2->V3)

- Word-coded (each token (word) is annotated separately).

Word order alternation where the verb ends up in the 3rd position, e.g., "men ellers jeg ved godt hun har gode venner."

Inventory:

v3	Verb in 3rd position

2.10 AMDA (Amerikadansk: AmDa, ArgDa, CanDa)

2.10.1 Language-stated

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Contact person: Jan Heegård Petersen: janhp@hum.ku.dk (orig. Gert Foget Hansen)
- Status: Researcher level

'Language-stated' describes the languages in which the transcribers have annotated the words in orthography. For example, if it says Spanish, it is because the transcriber meant that the given word is a Spanish word (spoken in Spanish). It is also noted when it is not clear which language has been used (due to pronunciation), and when a hybrid of two or more languages has been used. A given word can only have one of these tags associated with it.

Note: some language stated-data seems to have been entered in comments_x

Tags used in this tier:

Sprog (language:

- dansk (Danish)
- engelsk (English)
- norsk (Norwegian)
- spansk (Spanish)
- svensk (Swedish)
- tysk (German)

Other notations:

- ambig
    - Words that, due to their pronunciation, cannot be attributed to a language are marked with the tag 'ambig'.
- hybrid
    - Words that are composed of words or a word and an affix from more than one language are marked with the tag 'hybrid', regardless of the languages involved.
- ambig, hybrid
    - A combination of the two tags.
- non-autonom
    - If informants read aloud, sing, or recite (i.e., produce non-autonomous language use), it is marked with the tag 'non-autonom'.
- discuss
    - A note to the proofreader that there is a doubtful case. This tag only appears in combination with other tags (e.g., 'ambig, discuss').

2.10.2 Language revised

- Sequence-coded (BIO-labeled). The original annotation sequences ”X X X” have been recoded as ”1_X_I 2_X_I 3_X_E” (as described in the BIO-labeling description above).
- Contact person: Jan Heegård Petersen: janhp@hum.ku.dk (orig. Gert Foget Hansen)
- Status: Open

'Language-revised' describes the languages in which the transcribers have annotated the words in orthography. For example, if it says Spanish, it is because the transcriber meant that the given word is a Spanish word (spoken in Spanish). It also notes when it is not clear which language has been used (due to pronunciation) and when a hybrid of two or more languages has been used. In addition, there are tags for all non-language-specific sounds such as pauses, laughter, etc. A given word can only have one of these tags associated with it.

Tags used in this tier:

Sprog (language:

- dansk (Danish)
- engelsk (English)
- norsk (Norwegian)
- spansk (Spanish)
- svensk (Swedish)
- tysk (German)

Other annotations:

- ambig
    - Words that, due to their pronunciation, cannot be attributed to a language are marked with the tag 'ambig'.
- hybrid
    - Ord der er sammensat af ord eller et ord og et affiks fra mere end ét sprog, er markeret med tagget 'hybrid' – uanset hvilke sprog der er tale om.
- ambig, hybrid
    - A combination of the two tags.
- non-autonom
    - If informants read aloud, sing, or recite (i.e., produce non-autonomous language use), it is marked with the tag 'non-autonom'.
- egennavn
    - Proper nouns are not attributed to specific languages but are instead annotated with 'propernoun'.
- fejl
    - If there is a spelling error in the ortography tier, the word is not attributed to a specific language but is instead annotated with 'fejl'.
- latter
    - Laughter (the word 'ha' in orthography) is annotated with the word 'latter'.
- minimalrespons
    - In minimal responses (the word 'mm' in orthography), the word 'minimalrespons' is used.
- selvafbrydelse
    - If the speaker interrupts themselves (half-words ending with a hyphen in orthography), it is annotated with 'selvafbrydelse'.
- tøven
    - Hesitation (the word 'oe' in orthography) is annotated with the word 'tøven'.

- udbrud
    - For non-language-specific exclamations (such as 'uh,' 'arh,' or 'årh'), the word 'udbrud' is used.
- uforståelig
    - For unintelligible words (annotated with 'xxx' in orthography), the word 'uforståelig' is used.
- zzz
    - 'zzz' is used when 'zzz' is written in orthography. This tag is used when there are deliberately untranscribed passages in the interview for various reasons.

Additionally, all language names and 'ambig', 'hybrid', 'ambig, hybrid' and 'non-autonom' may have the suffix '-farm' added. This means that the word in question has the stem 'farm.' This has been done because words with the stem 'farm' are very common, and one would like to analyze these words separately.

2.10.3 AmDa Syntax (Syntax)

- Sequence-coded (BIO-labeled). The original annotation sequences "X X X" have been recoded as "1_X_I 2_X_I 3_X_E" (as described in the BIO-labeling description above).

cf. HUM-NORS-L-Danske_Stemmer_i_USA_og_Argentina/Danske Stemmer_kodning/Jessie and Ditte's manual for Amda and Argda_Syntakskorrektur_20161212.docx (all are yet to be translated into English).

Inventory:

The Danish Voices Project's syntax coding deals with full sentences. Each syntax annotation typically includes coding for the type and more of one main clause and additional coding for each subordinate clause if present. The codings for each clause (main and subordinate) are separated by "_" (originally by " ").

The coding for a single sentence consists of two parts separated by a slash: sentence type and constituent order. They are described below.

Sentence Type

If there are multiple sentences in the main clause, a serial number is added directly after the sentence type code, e.g., H/.. L1/.. L2/.. Hd3/.. Ld4/..

H	Main clause
Hkoo	Main clause with congruence construction
Hd	Main clause in direct speech
L	Subordinate clause
Lkoo	Subordinate clause with congruence construction
Ld	Subordinate clause in direct speech
Xeng	English main clause (without further coding)

Xspan	Spanish main clause (without further coding)
PU	Sentence-like adverbial phrase (always the last element in the overall annotation)

Constituent Order

The part after the slash indicates the sequence of the subject, finite verb, and any sentence adverbials.

A	Sentence adverbials
Vfin	Finite verb; any infinitive parts of the verb are not coded.
S	Subject
Sf	Formal/preliminary subject (including *som/der* in relative clauses)
Se	Elliptical subject
0	Omitted subject

2.11 Unknown Tiers

Here are tiers for which it has not been possible to find comprehensive information, and their status is therefore unclear.

2.11.1 kommentarer

Forekommer i: Odder2; BySoc1; Odder1; Næstved1.

The LANCHART corpus contains various comment tiers with very similar names. This will be cleaned up when possible.

2.11.2 Kommentar

Only 3 intervals. In Odder2. Another one of the many comment tiers, as mentioned above.

2.12 Hidden Tiers

Tiers in this section are currently removed from the corpus as they seemingly are not used by anyone. They can be reintegrated if necessary.

2.12.1 variant leksikalsk

808 filled intervals. Only found in the file "bysoc1gl-87-KJN.TextGrid." It is unclear what type of categories these are, but they seem to be some kind of word classes, although not in the traditional sense.

Inventory:

D	?
S	Slang?
EE	English words
B	Profanity
d	?
L	?
ED	?
E0	?

2.12.2 global turn

Likely indicates which speaker has the turn. Only found in Bylderup3 and Modsjæl2.

2.12.3 global sync

Only empty intervals. Found only in Bylderup3 and Modsjæl2.

2.12.4 sync

Found in: HanneSæderup; ModSjæl2; AmDa-dana; ArgDa; CanDa-ch; Køge3; Amager3; Vollsmose3; AmDa-tk; Vinderup1; Bylderup3; Næstved2; Køge25; Bornholm; AmDa-kbl; Vinderup0; CanDa; Munkebjerg3. 1.4 million intervals, none filled. It is currently unknown what it is used for. The tier matches the timestamps in the original transcriber file.

2.13 Deleted Tiers

Tiers in this section are either empty (zero results in the old search engine) or otherwise unusable. They have been completely deleted from the LANCHART corpus and are only included here for historical documentation.

2.13.1 Comments global

Presumably, according to the transcription manual: "Comments related to background noise or other things that cannot be attributed to a specific speaker should be transcribed as a global comment, for example, {global: CMP leaves the room}, {global: chat sound}" (translated from Danish). Only 2 empty intervals. Removed. Why are there only 2 empty intervals? It cannot be true that there are no global comments. Answer: Perhaps those are in "global events."

2.13.2 dialekt comments

Comments related to dialect tiers (moved into a more general comment tier).

2.13.3 grammatik2DUPLICATE

Responsible: Torben Juel Jensen.

Only 99 filled intervals. Only in Næstved3.

2.13.4 intervaltier

?

2.13.5 lydskrift

Old phonetic script tier?

2.13.6 parole_PoS

Old version of parts of speech?

2.13.7 parole_PoSRED

Old version of reduced parts of speech?

2.13.8 style

Old version of IIV?

2.13.9 Syntax edit

Syntax annotation from the Danske Stemmer project. All hits were in test files that are now deleted.

2.13.10 translation

Empty tier.

3 Metadata

In the LANCHART corpus, various metadata at the corpus, conversation, and informant levels are searchable in parallel with the annotations in the different tiers. These metadata are described below. Designations in parentheses are the raw variable names used internally in the search system and appear as column names in data export.

Note: Empty values in Korp are represented by the special value "__UNDEF__". For instance, if one wants to specify that a given word is NOT annotated in the "Socialklasse" category, a search should be conducted for tokens where "Socialklasse = __UNDEF__".

3.1 Special metadata (segmentation and duration)

The following categories are time and duration variables generated from each TextGrid and from turn segmentations. Conversations are segmented into what resembles turns of speech, although often not corresponding to entire turns. Therefore, the term "turdel" is used for individual segments.

- **Speaker's word number** (text_enum): Sequential number for the speaker's words in the conversation.
- **Word number in conversation** (ordnummer): Sequential number for the given word in the entire conversation.
- **Word number in turn** (turn_enum): Word's sequential number in the given turn.
- **Word start** (xmin): Time of the word's beginning (in seconds from TextGrid start).
- **Word end** (xmax): Time of the word's end (in seconds from TextGrid end).
- **Word length** (xlength): Duration of the word interval in seconds according to TextGrid.
- **Turn number** (turnummer): Sequential number of the turn in the conversation.
- **Speech source** (talekilde): "deltager", "baand" (= voice sample) eller "andet" (forbipasserende (eng. passerby) and so on).
- **Turdel start** (turnmin): Time of the turn's beginning (in seconds from TextGrid start).
- **Turdel end**(turnmax): Time of the turn's end (in seconds from TextGrid start).
- **Turdel length** (turnduration): Duration of the turn in seconds.

3.2 Traditional metadata

Traditional metadata includes categories such as gender, birth year, social class, etc.

Variables with the prefix "samtaler_" come from the informant database's "samtaler table, compare "informanter_" and " projekter_". If a variable does not directly come from the informant database but is generated by the code, this is indicated.

3.2.1 Metadata on Project Corpus Level

Currently, the only metadata category at the corpus level is the individual project corpus ID in the corpus database:

- **CWB-korpus-id** (label): For example, LANCHART_AMAGER, LANCHART_OKSBOEL, etc.

3.2.2 Metadata on Conversation Level

Information that applies to an entire conversation/recording.

- **Filename** (filename): Transcript file name, e.g.,"oksboel3gl-11-CMP+JLY+JRI+KAK+MLB+SJU_selvoptagelse.TextGrid"
- **Recording date** (samtaler_dato): Date of the audio recording, e.g., 2006-01-31
- **New/old recording** (oldnew, generated from filename): ”gl” (first recording with an informant, NOTE: only relevant for certain projects); ”ny” (second recording with an informant – recorded several years after the first recording, NOTE: also only relevant for certain projects).
- **Conversation type** (samtaler_samtaletype): ”Interview” (conversations between an interviewer and one informant. NOTE: In certain projects (Vissenbjerg, Odder 1, Næstved 1), the category also includes conversations with multiple informants, which should rightly be categorized as "Gruppeinterview"); ”Gruppeinterview” (conversations between an interviewer and several informants); ”Gruppesamtale” (conversations between two or more informants); ”Selvoptagelse” (conversations that the informant recorded themselves); ”Andet” (other types of conversations, such as radio interviews and advisory conversations).
- **Explorative** (Eksplorativ): Does the conversation belong to the "explorative corpus" – yes or no (1/0). The explorative corpus is a smaller sub-corpus intended for exploratory and hypothesis-generating work.
- **Proofread** (Korrektur): Has the transcript been proofread – yes or no (1/0).
- **Prioritized conversation** (Prioriteret samtale): Is the conversation part of the collection of prioritized conversations – yes or no (1/0).
- **Prioritized conversation, extra** (Prioriteret samtale, ekstra): Is the conversation one of the additional conversations categorized as prioritized – and thus not part of the original prioritized conversations. Yes or no (1/0).
- **Project** (projekter_name): Project code, e.g., "Odder3". Find more information about individual projects on the Center for Language Change's website.
- **Project-id** (samtaler_projekt): Project's sequential number.
- **Recording length** (textduration, generated): Conversation's length in seconds according to TextGrid.

3.2.3 Metadata on Informant Level

Information specific to individual informants.

- **Speaker** (speaker): Speaker code. Typically three uppercase letters, e.g., "ADA." However, there are some exceptions. S01, S02, etc., are voice samples, i.e., tape recordings that informants must respond to. There used to be a system of codes for passersby (UK1, UK2, etc.) but that has been phased out and replaced with unique codes for each passerby.
- **Gender** (informanter_koen): "m" or "f".

- **Birthyear** (informanter_foedselsaar): Four digits.
- **Generation** (informanter_generation): Read more about generation categories here:, https://www-tandfonline-com.ep.fjernadgang.kb.dk/doi/full/10.1080/03740460903364003).

Birth year	Generation code
-1879	0000
1880-1918	000
1919-1932	00
1933-1941	0
1942-63	1
1964-1980	2
1981-1986	25
1987-1996	3
>1997	4

- **Social Class** (informanter_socialklasse): "AK" for 'arbejderklasse' (eng. working class) og "MK" for 'middelklasse' (eng. middle class).
- **Role** (rolle): "meddeler" (informant), "interviewer", "forbipasserende" (passerby), "andet" (other).