Annette Gerstenberg, Valerie Hekkel, Julie Kairet

LangAge Corpora: Transcription guide

Potsdam University, May 2018

1 Introduction[1]

LangAge corpus data represent spoken language, as both a transcription and the original audio recording.

The transcription of spoken language is a unique challenge: a process which inevitably reduces the rich inventory of gestures, intonational features, and variants of linguistic expressions used in an individual interaction. The result of a transcription is a linear, letter-based representation of this multifaceted event. On the one hand, this can be considered as a loss of complexity; on the other hand, the transformation to a machine-readable form allows a fine grained, computationally supported linguistic analysis and annotation. Additionally, time-aligned transcription, as used in LangAge corpora, makes the dimensions of acoustic realisations immediately accessible and provides all users with the opportunity to compare the transcription offered by the corpus database with the own auditory impression.

The LangAge corpus aims at representing the spoken word with respect to its very properties. However, transcription is never neutral,[2] corpus data are always the product of abstraction, reduction and transformation – and they are never objective or authentic in a genuine way. Each corpus is meant to provide data for a series of research questions which guides the choices in the complex process of corpus building, which includes the many decisions made in the transcription of audio data, the choice of events accompanying the interaction and finally the linguistic annotation. All of these decisions define the possibilities, but also the limits of the corpus exploration. Instead of aiming to establish a complete transcription fulfilling the needs of every possible research question, in what follows, we want to explain the transcription's rationale in a transparent way.[3]

The LangAge corpus provides machine readable transcriptions which include the most important interactional features. In what follows, we illustrate the decisions made in order to build a sufficiently sizable corpus prepared for further linguistic annotation. For a fine grained interactional analysis or prosodic labelling, Transcriber files can be transformed in Praat format in order to enrich the orthographic transcription.

The corpus of LangAge contains uncompressed audio data (*.wav) that has been transcribed using the software Transcriber 1.5.1. The output of the software is an xml file (*.trs) which –

[1] A sincere thank you to Freya Hewett for her diligent proofreading of this guide.

[2] Among the many publications concerning transcription of spoken language, we just mention Ochs 1979 and Blanche-Benveniste & Jeanjean 1987.

[3] These principles are formulated, in a previous version, in Gerstenberg 2011.

depending on individual requirements – can be edited either within Transcriber or any other text editor.

1.1 Architecture of a transcription

A transcription within Transcriber is structured hierarchically. It contains sections that are composed of turns that are again composed of segments.

Sections are used to indicate if the text contained in it is to be transcribed or not. Their attribute type respectively takes the value “report” or “nontrans”.

Turns arrange the text according to the speaker uttering it. A turn by another speaker marks a new turn.

Segments structure the text within a turn. The segmentation of LangAge follows syntactic guidelines, so that one segment consists of one main clause (if given), elliptic or not completed sentences, *yes/no*-answers, or pragmatically independent utterances.

1.2 Transcription principles

The transcription follows the standard of modern orthography according to *Le Petit Robert* unless deviations are indicated in the present guide. Spaces, hyphens and underscores are used in a consequent way to facilitate the tokenisation of the interviewee’s speech.

The interviews are anonymised,[4] replacing all names and precise indications that might lead to the identification of an interviewee, as well as in some cases for ethical reasons, with a neutral label; placeholders such as NPR for ‘proper noun’ are used.

In order to capture the particularities of spoken language, all false starts, interruptions and hesitation phenomena are included in the transcription (2.2.1) while punctuation, except for the question mark, is omitted. The question mark is only used if the question is not marked syntactically (1).

```
(1)  c’ est à six heures ?
```

1.3 Mark-up

The "mark-up" event in transcriber is used to enrich the transcription with phenomena that are not contained in the plain text or by metalinguistic comments.

[4] See Baude 2006 for legal issues in corpus compilation.

Only perceptively long pauses up to 1 second are annotated with the "mark-up" event (longer pauses are placed in a separate segment, 2.2.4). As Transcriber allows the transcription of two people at the same time, the mark-up *overlapping speech* is used when a syntactic unit (main clause) uttered by one speaker goes beyond one turn.

2 Going into details

2.1 The architecture of a transcription

Sections[5] are used in LangAge to classify text which is transcribed or not. The values of the attribute *section* are "report" and "nontrans". While the value "report" refers to transcribed text, text that is, for various reasons (e.g. confidentiality, beginning of the transcription, interruptions of the interview), not transcribed, takes the value "nontrans".

Turns group the segments uttered by one speaker. A change of the active speaker is marked by the start of a new turn. An exception, however, concerns back channel signals, such as *m-hm* or *oui*, which are not segmented and not transcribed if they don't introduce a new turn. In an interactional context, those signals are normal and frequent, but according to the transcription's rationale, they are not included in the basic transcription.

Overlapping speech is transcribed[6] when two speakers speak at the same time, unless the speech is too dense. In the latter case a "nontrans" section and the event "chv" are created for the segment. First, this decision can be justified by the fact that LangAge focuses more on monologic speech. Second, if wanted, this text can be edited further using different tools.

The segmentation into *segments* follows syntactic principles. One syntactic unit corresponds to one main clause as in (2) and in (3).

```
(2)  il l' a été pendant deux mois
(3)  il avait soixante-dix-huit ans à l' époque
```

There are special cases, according to the needs of a transcription of spoken language:

- Short (intrasentential) insertions are not segmented (4)

```
(4)  vous voyez ben elle avait quel âge euh
```

- *oui*, *non* are segmented if they correspond to independent utterances
- Enumerations (of main clauses) are segmented unless they are prosodically dense.

[5] Menu: Segmentation > Create section > Report / Nontrans.

[6] Menu: Create Turn > Overlapping speech. For technical reasons, a segment of overlapping speech always starts with spk2 (the main interviewee), irrespective of whether the other speaker is the interviewer or another person present in the interview.

- *hein*, *voilà*, *bon ben* at the end of an utterance are attached to the previous utterance if they have a conclusive value.

Exceptions:

- Prosodically dense utterances are not segmented.
- Repetition/ imitation/ echo at the end of an utterance are considered to be part of the segment and thus are not taken apart, as in (5) and in (6).

```
(5) ça ça n' est pas comparable hein ça n' est pas
    comparable
(6) ç() ça fait juste dix ans il est mort le six janvier
    quatre-vingt-quinze ça fait juste dix ans
```

In overlapping speech, priority is given to timing. This means that the segmentation according to syntactic units becomes secondary. If the overlapping speech concerns more than two interlocutors, a fine-grained segmentation is the aim: as far as possible, new turns are created.

2.2 Transcription principles

2.2.1 Spoken language

Repetitions, interruptions, incomplete or incomprehensible syllables are transcribed as far as possible.

Incomprehensible syllables in the beginning or in the ending of a word are transcribed using brackets as in (7) and in (8): if content is reconstructed, it is also written in brackets.

```
(7) je s()
(8) (par)ce que
```

Incomprehensible syllables are marked by up to three *X* (*XXX)*. This has the advantage that a transcription does not need to be forced whenever a word sequence is incomprehensible, and the risk of misunderstanding is minimized.

2.2.2 Lexical units

The transcription facilitates further computational analysis. Thus, hyphens and underscores are used to link multi word expressions and letters that form lexical units.

- Numbers: Words are used instead of numbers (*mille-neuf-cent-quarante-deux*)
- Proper nouns: capital letters and underscores link the single constituents (*Jeanne_d'Arc, Rue_Royale* ; *Place_du_Martroi*, *Parole_et_écrit*)
- Underscores are also applied to multi word slogans
- For spelled out words: Letters are linked by an underscore.
- Acronyms such as *SNCF* are transcribed without underscores

- Countries, institutions, associations, organisations, terminologies and titles are treated in the same way as names: *Diplôme_d'études_supérieures*, *Jeunesse_Ouvrière_Chrétienne*, *Certificat_d'études*, *Cœurs_veillants, Oscar_et_la_Dame_rose*, *École_normale*, *Arte*, *Brevet_élémentaire*, *Université_du_temps_libre*
- *Etcetera* is transcribed “etc”

Clitics attached with hyphens to the preceding word are used as in standard French; a full inventory of these clitics is used in the tokenisation (*celui-là*, *dites-moi*).

Apostrophes alone are not to be counted as word boundaries, for they also appear in words such as *aujourd’hui* or *d’abord*. Therefore, a horizontal space is added if the apostrophe is located at a word boundary.

Spaces are omitted for the following expressions:

- *aujourd’hui*
- *c’est-à-dire*
- *d’abord*
- *d’accord*
- *d’ailleurs*
- *d’autant*
- *d’habitude*
- *d’œuvre*
- *n’est-ce-pas*
- *quelqu’un*

2.2.3 Pronunciation

Following the orthographic standard of *Le Petit Robert*, pronunciations that deviate from the standard form are transcribed, nonetheless, according to the standard. In fact, in spoken language, between the full form of, e.g., *mais enfin* and *m’enfin*, a variety of shortened forms can be realised. Thus, the decision to transcribe the short form or not is highly subjective, and for the analysis of these variants, more sophisticated and time consuming operations are needed, such as those provided in Praat. For the basic transcription, these variants are reduced to the standard form.

Accordingly, shortened forms are transcribed as follows:

- *t’écoutes ? > tu écoutes ?*
- *j’ viens > je viens*
- *i i pense > il il pense*
- *d’ mère > de mère*
- *m’enfin > mais enfin*

Slips of the tongue or erroneously deviant pronunciations are annotated using the *event* "mark-up"[7] and a phonetic transcription of the realized string, according to the SAMPA standard. In (9), the usually silent final *-s* is pronounced, so the SAMPA-transcription includes it.

```
(9) je faisais mes cours <Event desc="kuRs"
    type="pronounce" extent="previous"/>
```

The same process is applied for variation concerning *muta cum liquida*: The standard form is transcribed, even in the case of deviant pronunciation (*quatre*, even if pronounced [kat]). The event “pro”, applied to the previous word, hints at the deviation.

2.2.4 Pauses

Pauses are differentiated according to whether they appear intra- or intersententially. Perceivable intrasentential pauses don’t lead to segmentation, they are instead annotated using the *event* "mark-up [pau]"[8]. Intersentential pauses and pauses between turns which are longer than 1 second are transcribed by an empty segment and attributed to the previous speaker.

2.2.5 Interjections

An inventory of interjections occurring in LangAge corpora was established in order to reduce the number of variants originating in, e.g., lengthening an interjection.

These variants are highly subjective, and the inventory facilitates the automatic treatment.

- *ah*
- *bah*
- *beh*
- *ben*
- *chh*
- *eh*
- *euh*
- *ha*
- *hé*
- *hein*
- *hop*
- *hum*
- *m-hm* (back-channel)
- *mmh* (back-channel)
- *mm* (hesitation)
- *oh*
- *ouf*
- *pff*
- *youh*

[7] Menu: Insert event / CTRL+d > Pronounce, Apply to previous word.

[8] Menu: Insert event / CTRL+d.

2.2.6 Events

The signs used in "mark-up" of *events* consist of three letters.[9] A space is added before and after an event. The metalanguage of the abbreviations is French.

Deviant pronunciation, dense overlapping speech and direct speech are annotated as events. Events are applied to a previous word (+*[event]*), a selection (*[event-][-event]*) or they can annotate an independent instantaneous event (*[event]*).

- [chv] dense overlapping speech (French: *chevauchement*)
- [dir-]/[-dir] direct speech: Select what is uttered as a direct speech (DS; reported speech). If DS includes several main clauses/ segments, the annotation is applied to each segment individually

Events are also used to annotate verbal and nonverbal sounds.

- [bru] noise
- [bou] *pt*
- [exp] expiration
- [ges] audible gesture
- [ins] inspiration
- [ono] onomatopoeia (even if an invented instance)[10]
- [pau] pause (inside a segment)[11]
- [rac] clearing throat
- [rir] laughter; if the pause lasts longer than one word, apply to the whole segment.
- [tou] coughing
- [tss] clicking with tongue
- [sou] sighing

Additionally, free descriptions can be added using the *event* "mark-up". In this case, square brackets are used: (*[boit du thé]*).

Words pronounced in other languages are labelled using the *event* "mark-up language", and a value set to a short code indicating the language according to ISO 639-1:

- de: German
- ru: Russian
- en: English
- di: Dialect

The event is applied to the previous word or to the selection if more than one word is concerned; multi word slogans in a foreign language are united by an underscore.

[9] The mark-up of comments like [buzz], [lang], [pron] have a different form. Events are inserted via, in the menu, Edit > Insert event" or CTRL+d.

[10] Check "apply to previous word".

[11] Pauses that are longer than 1s. are visible in the segmentation (empty segment).

3 References

Baude, Olivier. 2006. *Corpus Oraux*: *Guide des bonnes pratiques*. Orléans: Presses Universitaires d'Orléans; CNRS Éditions.

Blanche-Benveniste, Claire & Colette Jeanjean. 1987. *Le français parlé : transcription et édition*. Paris : Éditions Interco.

Gerstenberg, Annette. 2011. *Generation und Sprachprofile im höheren Lebensalter*: *Untersuchungen zum Französischen auf der Basis eines Korpus biographischer Interviews* (Analecta Romanica 76). Frankfurt am Main: Vittorio Klostermann.

Hirst, Daniel. 2013. Anonymising Long Sounds for Prosodic Research. In Brigitte Bigi & Daniel Hirst (eds.), *Tools and Resources for the Analysis of Speech Prosody*. TRASP 2013, 36–37. Aix-en-Provence: Laboratoire Parole et Langage.

Ochs, Elinor. 1979. Transcription as theory. *Developmental pragmatics*. 43–72.

SAMPA = Wells, John C. 2005. Speech Assessment Methods Phonetic Alphabet. SAMPA. http://www.phon.ucl.ac.uk/home/sampa/index.html (20180415).

4 Internal section

A transcription always starts with a999 or b999 or c999 (spk1).

4.1 Anonymisation

In the original versions (e.g., a006o), critical sections are marked with buzz, but transcribed. Only in the very last, anonymized versions (e.g., a006a), the content is replaced by codes.

Following rules are established for anonymization, The content to be anonymised is categorised (*NPx*, *NUM*, *anon*), selected (Ctrl + d, apply to selection) and annotated using the event *buzz*.

After the event *buzz*, continue with the same speaker.

4.2 Anonymisation: abbreviations

Text that should be anonymised because of proper nouns is framed by the *buzz* event and the code according to the type of the proper noun:

▪	anon	personal or ethical reasons
▪	NPp	proper noun, given name (*prénom*)
▪	NPf	proper noun, family name (*nom de famille*)
▪	NPl	proper noun, locality (village, city) (*localité)*
▪	NPc	proper noun, country
▪	NPe	proper noun, company (*entreprise*)

Word formations based on proper nouns are also anonymised:

▪	A	Adjective	fond familial fleuryssois -> NPl(A)
▪	N	Noun	Délégués de l' Association XY -> NPe(N)
▪	N	Noun	Il était fleuryssois -> NPl(N)

For years/dates with high personal relevance, NUM code is used.

In order to protect the privacy of the speakers, locations which could potentially help to identify them, are anonymised with the code NPl if they have less than 5000 inhabitants.

4.3 Anonymisation for ethical reasons

Text that should be anonymised because of personal or ethical reasons (for example when the interviewee asks the interviewer to be discrete about something) is framed by the *buzz* event and the code "anon"; STRG+d, event *buzz*.

If necessary, a short description of the content is included with STR+d event "instantaneous event".

5 Technical details

5.1 Metadata inside the xml document

File – edit episode attributes

- audio file name -> x000a
- Transcriber's name -> LangAge corpora
- Principal language -> fr
- Program -> LangAge corpora (2005|2012|2015|XXX)
- Recording date (automatically)

Options – general

- Default scribe's name -> LangAge corpora
- Encoding -> UTF8

Make sure that interviewer is spk1, interviewee spk2, others spk3 or spk4 as in (11) and (12):

```
(10) <Speaker id="spk1" name="a999" check="no"
     dialect="nonnative" accent="fr” scope="local"/>
(11) <Speaker id="spk2" name="A48" check="no"
     dialect="native" accent="" scope="local"/>
```

5.2 Protocol file

A protocol file is used to document the development of the transcription and the correction of the data. Linguistic phenomena worth consideration are also referred to in the protocol file (État des lieux.xlsx dans documentation). Workflow manager: Julie.

5.3 Corrections and conventions

Following elements should be controlled systematically:

- the use of question marks
- the use of events ([dir])
- c'est > c' est except for c'est-à-dire
- attention *qu' est -> qui est; qu' était -> qui étaient* (if applicable)

Lower-case letters

- *bac*
- *sms* (lower-case, without dots)
- *papa et maman* : lower case letters
- *nord, sud, est, ouest* : lower case letters
  exception : proper names such as *Amérique_du_Sud*, *Afrique_du_Nord*
- *exode*
- *messe*

Upper-case letters

- Proper names : *Google, Renault*
- Nationalities used with capital letters: *les Allemands*
- (le) *Daesh*, (la) *Gestapo*
- Journaux: *Le_Monde*, *L'Express*, *L'Histoire*, *La_République_du_Centre*

List of words starting with a capital letter that aren't names :

- Pâques, Noël, Christ, Jésus, Dieu, Bible, Homme
- (la) Résistance, (la) Libération

5.4 Foire aux questions : Particular cases

For the revision of the transcriptions, an individual code is used to mark questions or uncertainties:

- § for VH
- for JMK

Questions that arose during the transcription or correction were collected and discussed. The results of the discussions can be found here:

List of onomatopoeia:

- papapapapapa (A10)
- toudoudoum toudoudoum (A10)
- psout (A10)
- Han (B18)
- Paf
- ch (c018)

Orthographic doubts:

- *Noël* has a plural form ➔ *Noëls*
- *Conflit de générations* : with -s
- *Schnock*
- *boche*

5.5 Preparing for tokenisation

Chose speakers individually (only spk1 or only spk2).

All xml-elements are deleted and the plain text is transformed to one word per line (word wrap after space and apostrophe). Exceptions (use one line only):

- m'enfin+mais enfin+
- ç' +ça +
- aujourd' hui+aujourd'hui+
- c' est-à-dire+c'est-à-dire+

- d' abord+d'abord+
- d' accord+d'accord+
- d' Arc+d'Arc+
- d' ailleurs+d'ailleurs+
- d' autant+d'autant+
- d' oeuvre+-d'oeuvre+
- n' est-ce pas+n'est-ce pas+
- quelqu' un+quelqu'un+

Clitics after hyphen treated as an individual word (segmented in a new line):

- +ce+ci+elle+elles+en+il+ils+je+j'+la+là+les+leur+lui+même+mêmes+m'+moi+nous+on+t'+toi+tu+vous+y+