

LLACAN
Centre de Linguistique Appliquée

Designing Manding spoken corpora

UMR8135 CNRS IMA/CO
Centre André-Georges Haudricourt
7 rue Guy Môquet 94801 Villejuif Cedex - France


<http://llacan.vjf.cnrs.fr/>
idiatov@vjf.cnrs.fr



Why a spoken corpus?

- primary language medium is speech (or gesturing for sign languages)
- written language differs from spoken language (often significantly)
- spoken language recordings are indispensable for phonetic and prosodic research, discourse studies...

Why not also **video** recordings?



2

Possible target groups

- phoneticians, phonologists
- research on discourse, information structure...
- usage-based grammatical and lexical research
- applied linguistics
- gesture studies

3

Corresponding design issues

- different "genres" of spoken data

Components of the Dutch Spoken Corpus

a. Spontaneous conversations ('face-to-face')	i. Live (e.g. sports) commentaries (broadcast)
b. Interviews with teachers of Dutch	j. Newsreports/reportages (broadcast)
c. Spontaneous telephone dialogues (recorded via a switchboard)	k. News (broadcast)
d. Spontaneous telephone dialogues (recorded on MD with local interface)	l. Commentaries/columns/reviews (broadcast)
e. Simulated business negotiations	m. Ceremonious speeches/sermons
f. Interviews/discussions/debates (broadcast)	n. Lectures/seminars
g. (political) Discussions/debates/ meetings (non-broadcast)	o. Read speech
h. Lessons recorded in the classroom	i. Live (e.g. sports) commentaries (broadcast)

4


Corresponding design issues

- different "genres" of spoken data
- different (physical) quality of the recorded audio signal
- different levels of depth of transcription and annotation

5

Scope & Size

- What is achievable?
- Focus on Bamana only?
- Size?
 - The Dutch Spoken Corpus: 9 million words / 1000 hours ≈ 150 words/minute
 - Le Corpus de Référence du Français Parlé: 440.000 words / 36 hours ≈ 200 words/minute
 - The spoken part of the British National Corpus: ca. 10 million words
 - The Corpus of Spoken Israeli: ca. 5 million words
 - C-Oral Corpus (Italian, French, Spanish, Portuguese): under 1 million each
- orthographic transcription of 1 min of audio by a trained native speaker ≈ 30 min (on average)
- 1 million words at 175 words/hour ≈ 95 hours
- orthographic transcription of 95 hours ≈ 2.850 hours (= 475 days of 6 hours/day)



6

Representativeness

- What is achievable?

Čermák (2009) "Spoken Corpora Design"

- demographic parameters

- gender
- age
- educational level
- occupation
- village/city
- geographic origin (implies dialectal variant)
- native/non-native speaker?



7

Representativeness

- situational parameters

PLUS (+)	MINUS (-)
<i>a. Origin of the text:</i>	
1 spoken (i.e. original)	- read (but written)
2 dialogue (i.e. original, typical)	- monologue
<i>b. Interpersonal, sociological relationship of partners and physical situation:</i>	
3 proximity of partners (friends, family)	- no proximity
4 equality of partners	- inequality
5 private (non-public)	- public
6 informal	- formal
7 interactive	- unidirectional
8 present	- distant (e.g. phone)
9 non-multiple (one-to-one)	- multiple (one-to-many)
<i>c. Topic/situation approach:</i>	
10 spontaneous (unscripted)	- prepared (beforehand, more or less scripted)
11 casual (informal)	- regular/official
<i>d. Awareness of the recording:</i>	
12 not aware	- aware

"prototypical spoken text"

+spoken(1), +dialogue(2), +proximity(3), +equality(4), +private(5),
+informal(6), +interactive(7), +present(8), +non-multiple(9),
+spontaneous(10), +casual(11), +not aware(12).

"interview for radio"

?+1, +2, -3, -4, -5, -6, +7, +8, ?-9, ?+10, ?-11, +12

8

Representativeness

- topic parameters

- topic
- professional field
- genre
- register
- etc.

"...rather **problematic** because no generally acknowledged taxonomy (except for those used by libraries etc.) is available to be followed." (Čermák 2009:118)

- OLAC Discourse Type Vocabulary (<http://www.language-archives.org/REC/discourse.html>)

- | | |
|-------------------------|-------------------------|
| • drama | • narrative |
| • formulaic_discourse | • procedural_discourse |
| • interactive_discourse | • report |
| • language_play | • singing |
| • oratory | • unintelligible_speech |

9

Copyright and privacy issues

- "informed consent" (for the field recordings)
- privacy and the possible need for **anonymisation**
- copyright (French/German laws & Malian/Ivorian/etc. laws)

Baude et al. 2005. *Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux*. Ministère de la Culture et de la Communication. http://www.culture.gouv.fr/culture/dgfi/Guide_Corpus_Oraux_2005.pdf

Newman, Paul. 2007. Copyright essentials for linguists. *Language Documentation & Conservation* 1(1). 28-43. <http://nflrc.hawaii.edu/ldc/>

- recordings of radio or television programs
- field recordings made before the start of the project with no proofs of informed consent

Corpus du français parlé (<http://sites.univ-provence.fr/delic/corpus/index.html>):

"Nous savons depuis 1999 que l'on doit obtenir, par écrit et avec signature, le 'droit de reproduction et de représentation' pour tous les enregistrements. **La partie ancienne des corpus n'a pas cette garantie.**"

10

Q20. While I was in the field I collected quite a lot of oral literatures, especially from two remarkable people. The first was an old woman who seemed to know an endless number of folktales, which she told in energetic fashion. Both she and the village elders explained to me that she was the personal custodian of the folktales, but that the tales as such were the property of the community. The second was a blind man who was admired in the village because of his linguistically expressive poetry. I recorded both of these people and with the help of a local school teacher assistant transcribed everything in the local language and translated everything into English. From a copyright point of view, who owns what?

(Newman 2007:40-41)

A20. Nothing on my copyright to the folktales. The old woman doesn't because, although she related them, she was not the author. Similarly, the community has no copyright interest because of the lack of identifiable authorship. When the elders told you that the community owned the folktales, you may have acquired certain contractual or ethical obligations regarding your use of the tales, but this would be outside of copyright law. Finally, neither you nor the teacher has any copyright interest in the folktales as such. Transcription of a recording does not constitute authorship.

As long as the poet's poetry was entirely oral, there would have been no copyright. However, once you recorded the poetry, you thereby "reduced it to tangible form"—the transcription wasn't required—whereupon copyright automatically attached. (For sake of discussion, I am assuming that the copyright laws in the country in which you were doing your research are similar to U.S. law.) The poet is now fully invested with the copyright on his poetry that you took down, and so anything that you intend to do with the poetry will require his approval.

Unless the teacher could be considered your employee and not just someone who did special tasks for you from time to time, you and the teacher jointly own the copyright to the translations. Remember, even if you paid him well for the translation work, and even if your understanding was that you would then be free to use the translation as you wanted, if you didn't get an agreement in writing saying that his free-lance work would constitute Work for Hire, the teacher obtained a 50% interest in the translation. As a joint holder of the copyright, you would be free to use the translations for your purposes and even issue non-exclusive licenses—each co-holder has that right. However, you couldn't transfer the copyright as such without the teacher's approval, and you would owe him 50% of any royalties or other income that might ensue from your combined efforts.

11

Archiving

- long-term archiving (archivage pérenne)

TGE ADONIS (« Très Grand Equipement » « Accès unifié aux données et documents numériques des sciences humaines et sociales »)



- also deposit the recordings with a local institution (e.g., the National Museum)

12

Practicalities

- good **models** to follow:
 - The Dutch Spoken Corpus / Corpus Gesproken Nederlands** (http://tst.inl.nl/cgndocs/doc_English/start.htm)
 - CorpAfroAs** (<http://web.me.com/aminamettouchi/CORPAFROAS/Abstract.html>)
- formats** (Baude et al. 2005; Adonis: http://www.tge-adonis.fr/sites/default/files/ressourcesdoc/GuideFormatsAdonis-04a_V3.pdf)

audio

- CD-Audio: 44100 Hz, 16 bits, mono/stereo (uncompressed PCM/WAV or lossless FLAC)
- a compressed format for streaming (MP3, lossless FLAC, etc.)

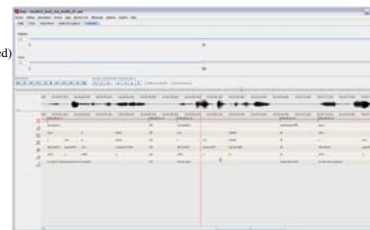
video

- container OGG: video codec THEORA + audio lossless codec FLAC
- container MPEG-4: video codec mpeg4-AVC(H.264) + audio codec mpeg4-AAC

13

Software

- ELAN** (<http://www.lat-mpi.eu/tools/elan/>)
- PRAAT** (<http://www.fon.hum.uva.nl/praat/>)
 - open source & free
 - dedicated to the creation of complex (hierarchically connected and time-aligned) annotations for audio & video
 - becoming a de facto standard
 - constant development
 - mutually convertible with many other software
 - XML, Unicode, etc. compliant
 - apparently, rather easy to learn for transcribers
 - numerous search and export options
 - possibility for adding constrained vocabularies
 - etc.



14

Tiers: transcription & annotation

- The Dutch Spoken Corpus** (ca. 9 mln words)
 - orthographic transcription**: 9 mln words (100%)
 - broad phonetic transcription & word segmentation**: 1 mln words (11%)
 - prosodic annotation**: 250.000 words (3%)
 - POS-tagging & lemmatization**: 9 mln words (100%)
 - syntactic annotation**: 1 mln words (11%)
- possible additional tiers:
 - morpheme-by-morpheme** segmentation
 - glossing**
 - free translation?**

15

Orthographic transcription

A2. Definition of the orthographic transcription

An orthographic transcription is a *literal* reflection of what was said. The transcription corresponds closely to *written* Dutch. However, on certain points the *Dutch spelling rules* are deviated from or additional information is added. This ensures that the transcription is as consistent as possible and applicable for a range of scientific purposes.

The orthographic transcription also contains *alignment*¹ at *chunk level*² and an indication of the different speakers. The transcription also allows for the registration of background sounds and there is a field for comments.

¹ *alignment*: linking a part of the transcription to a part of a sound signal by inserting boundaries. These boundaries determine the start and end times of the parts of the transcription. See the manual supplied with the Praat program for further information over inserting boundaries.

² *a chunk*: a fragment of speech of approximately 2 to 3 seconds which is bordered on both sides by a (short) visible and audible pause. Chunks do not need to correspond to sentences or parts of sentences; they are determined solely by the pauses in the speech signal. You are encouraged to keep the chunks as short as possible, (where the sound signal permits).

- A **protocol** (Goedertier & Goddijn 2000, http://tst.inl.nl/cgndocs/doc_English/topics/annot/orthography/ort_prot_en.pdf) has been developed which describes in detail what to transcribe and how to deal with new words, dialect, mispronunciations, and so on. Background noises are not represented in the transcript.

16

B2. Work as accurately as possible, but **don't waste time** on sections which are difficult to understand or where people are talking over one another; mark these with **xxx** or **Xxx** (see C9 and C10).

B3. **Do not use a capital letter** at the beginning of a sentence. **Do use capital letters** for proper names (e.g.: Bert Jansens) and for titles of books, records, etc. (e.g.: De Naam Van De Roos) (see also C1 and C2).

B4. Use the following codes where necessary or logical:

- use **'v** for **foreign words** (e.g.: tomorrow^v) (see also C3);
- use **'d** for **dialect words** (e.g.: kortewagen^d; keuje^d) (see also D4);
- use **'z** for **standard words** which are pronounced with a **strong dialect** (see D5);
- use **'n** for **new words** (e.g.: zero-toerentienⁿ) (see also C4);
- use **'t** for **new interjections** (e.g.: amaa!^t; hoppa!^t) (see also C5);
- use **'a** for **words cut short or interrupted** (e.g.: uitges^a; verpr^a) (see also C6);
- use **'u** to show **onomatopoeia** and **slips of the tongue** (e.g.: boink^u; oesoreken^u) (see also C7);
- use **'k** for words where you are **not sure** if you have **understood them correctly** (see C8).

Never use more than one of the above codes per word. Choose the most suitable code.

B5. Check if the transcription only includes words which are in the CGN lexicon³. This can be done using the special CGN spelling checker.

B6. *Only* use the punctuation marks **full stop**, **question mark** and **ellipsis** (. / ? / ...). Do not use any other punctuation marks such as comma, exclamation mark, quotes etc. (see also section E).

If in doubt: Where the protocol is insufficient, choose the spelling which best fits the conventional spelling.

B7. Use **ggg** for **clearly audible speaker's sounds** (see also C11).

17

D. How should spoken language be written down?

D1. **Write down what is said literally**. This means that you should change nothing of what is said; this also applies to where the speaker uses **incorrect sentence constructions** or makes **grammatical mistakes**:

you hear and write	do NOT correct to
de koe dat graast.	de koe die graast.
de open venster.	het open venster.
hun zijn er al.	zij zijn er al.

D2. **Only use words from the CGN lexicon**. In principle this lexicon should include all Standard Dutch words in the *written* language. Here we mention explicitly a number of forms which are in the CGN lexicon (although it could possibly be argued that they are not part of Standard Dutch). It is compulsory to use these forms if the user does.

D3. **Do not attempt to record how words are pronounced**. After all, this would quickly lead to a sort of phonetic transcription⁴. This is **not** the idea behind an **orthographic transcription**; this may only contain words from the CGN lexicon (with the exception of *v, *d, *n, *t, *z, *x, *a, *u, or where a capital letter is written).

- A decision needs to be made on **what should be reflected in the orthographic transcription** for Bamana

18

- elisions: *te a* > *t'a*
- syncope: *fila* > *fla*
- "mobile" nasalization & *ŋ/m/r. fila(n), nininkali* > /*niŋiŋgali*/
- consonant alternations & contractions: *an ka* > /*ãŋ(g)a'*/, *bagan be* > /*baga me*/, *n ye* > /*p(n)e*/
- vowel alternations in PM *be/te*
- dominant vowel assimilation: *bir'a* > /*ber'a*/
- *ns-/z/-s-*: *nson, zon, son*
- *h-|Ø-*: (*h*)*adamaden*
- *s/sh*: *seeġin / sheeġin*
- labialization: *g(w), k(w)*
- labiovelars: *gb, kp*
- palatalization: *b(y), f(y), sh(y)*
- ???
- **lexical tone** (for a more efficient subsequent parsing/tagging)
- **tonal article** (only where it can be heard, e.g. before a H or pause in Standard Bamana, also before a L in Segu)



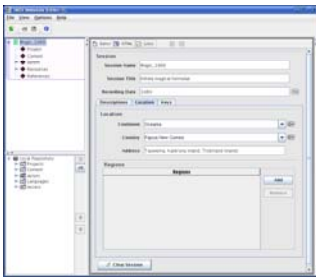
To dos...

- create a **reference lexicon** (based on the one developed for the written corpus)
- create a **list of interjections & hesitation markers** (based on the one developed for the written corpus)
- create **protocols** for all the workpackages (orthographic transcription, broad phonetic transcription, etc.)

Metadata

- follow the existing standards (requirements of OLAC, Baude et al. 2005)

IMDI (<http://www.lat-mpi.eu/tools/imdi/editor/>)



CorpAfroAs



To dos...

- work out **metadata parameters**
- **prioritize** types of data to collect
- **who** should **collect** the data (preferably, **native speakers** themselves)
- work out the **workflow**, such as:
 - data recording
 - fill in the metadata
 - clean the audio file from long silences
 - orthographic transcription
 - selective cross-check of transcriptions
 - archive the recordings
- clarify the question of **archiving**

Money matters...

- recording equipment (decent audio, video, memory cards, etc.; several sets)
- computers for transcribers
- salary for transcribers (who would also collect a part of the data)
- remuneration of the participants
- ? copyright for radio/tv recordings
- trips for training & supervising the transcribers
- how many transcribers? for how long?