From Fieldwork to Annotated Corpora: The CorpAfroAs project

Amina Mettouchi° & Christian Chanard°°
°(University of Nantes & Institut Universitaire de France)
°°(CNRS-LLACAN, Villejuif)*

Introduction

In the first years of this new century, in the domain of linguistics, much emphasis is being put on language diversity, as well as on language technologies. Not so long ago, grammatical theories were content to rely on a small number of well-described European and Asian languages, and corpora-design was limited to some of those well-known languages[1]. With the development of typology, and the growing concern about the fast disappearance of hundreds of the estimated 6000 languages currently spoken on our planet, language descriptions are now given more and more importance. In the meantime, language technologies have become more and more accessible to the linguist, through the generalization of the use of computers, and the availability of high-quality portable recording devices. The first result of this technological revolution was the development of language archives aiming at preserving the work of fieldwork linguists through the digitalization of recordings and transcripts. Such initiatives as the LACITO Archive[2], the CRDO[3],

* amina.mettouchi@univ-nantes.fr, chanard@vjf.cnrs.fr.
[1] See for instance such initiatives as the London-Lund Corpus of spoken English, the British National Corpus, C-Oral Rom, etc.
[2] http://lacito.vjf.cnrs.fr/archivage/presentation.htm
[3] http://crdo.risc.cnrs.fr/exist/crdo/

or DOBES[4] or other centers for the preservation of language diversity and endangered languages have emerged. A number of texts in a great variety of languages have thus been digitalized. However, annotations are not always provided, and when they are, they are not standardized and/or do not allow complex queries in the database. CorpAfroAs[5], a project funded by the Agence Nationale de la Recherche (ANR) in France, has emerged in this context, as a pilot corpus aiming at providing a structured database of spontaneous recordings of Afroasiatic languages, transcribed, translated, and annotated in view of allowing complex queries.

The ultimate goal of CorpAfroAs is to trigger a number of similar endeavors for various language families. This is why the design of the corpus, and the scientific decisions made, must be brought to the knowledge of the community, and proposed for discussion and implementation. Hence this paper, which is an update on a preceding paper (Mettouchi et al. to appear) presenting the main lines and goals of CorpAfroAs.

Our aim here is to focus on the theoretical and technical developments of the project. In part 1, we present the motivations for the choice of software made for the project, in part 2, we focus on the design of annotation tiers in relation to some queries relevant for our language family, and in part 3 we briefly present our metadata form.

1. General design and annotation procedure

[4] http://www.mpi.nl/DOBES

[5] The CorpAfroAs project is conducted by three French research laboratories, and associate French and International researchers. The principal coordinator is A. Mettouchi, the associate coordinators are M. Vanhove and D. Caubet. Two experts are following the project and providing feedback and guidelines: B. Comrie (MPI Leipzig and UCSB), and S. Izre'el (University of tel-Aviv). The complete list of members can be found on http://www.univ-nantes.fr, keyword 'CORPAFROAS', or on http://web.me.com/aminamettouchi/CORPAFROAS/Abstract.html.

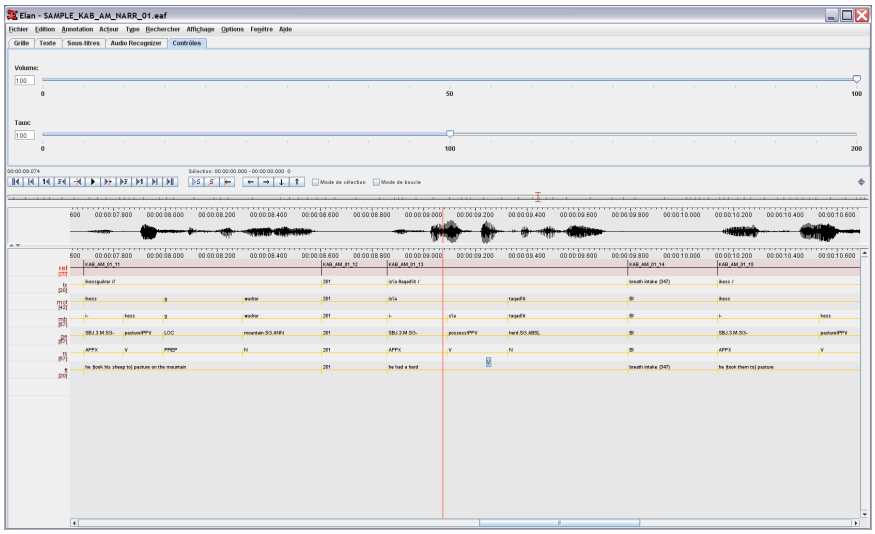
CorpAfroAs is organized along two axes, linked to the nature of the materials and to the aim of the project, which is typological comparability among languages: prosodic analysis, and morphosyntactic glossing.
The body of data is spoken, and we have decided to fully take into account this oral dimension by working on segmentation. We do not use the punctuation system of written texts, because it is not adapted to the specificities of the spoken language (Wichmann 2000). Instead, we are adapting the widely accepted system of boundary-marking used for instance in the C-ORAL-Rom developed by Cresti & Moneglia[6].
We therefore analyze the prosodic units of our languages into minor (non-terminal) and major (terminal) units, using the software Praat[7]. No other specification (tones, contours etc.) is given to those boundaries, but the fact that the transcription is indexed to the sound, itself available in .wav format, will allow more in-depth prosodic studies on the available data.
Segmentation by native speakers provides the basis for analysis of the major (terminal) intonation-units, which turn out to be based on cues used in a wide variety of languages, namely pitch reset, lengthening, anacrusis, and pauses. Minor intonation unit are typically more difficult to define. The fact that the sound file will be linked to the segmented transcription will facilitate alternative proposals by other researchers.

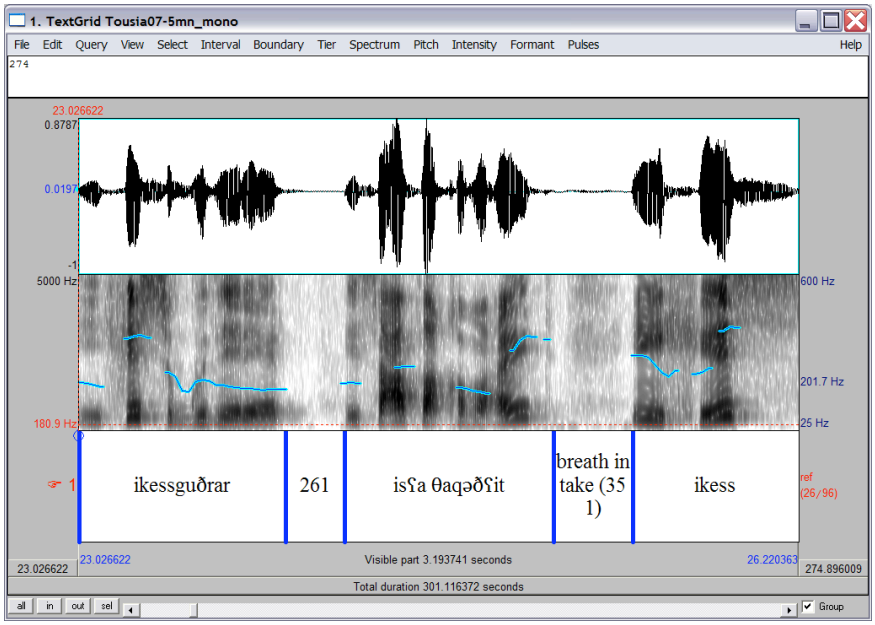
The software in which CorpAfroAs is designed and will ultimately be put online is ELAN[8], developed by the Max Planck Institut in Nijmegen. This software was chosen for a number of reasons: it is dedicated to the creation of complex annotations on video and

[6] http://lablita.dit.unifi.it/coralrom/
[7] Paul Boersma & David Weenink, http://www.fon.hum.uva.nl/praat/
[8] http://www.lat-mpi.eu/tools/elan/

audio resources, it is open-source and free, it is regularly updated, and it is widely used among linguists. Annotations are created on multiple layers, called tiers, which can be hierarchically connected, and time-aligned to the media.

In Elan, the sound is only available through the online visualization of the waveform. Although it is theoretically possible to align sound and transcription using this visualization, the results are actually too inaccurate: a segmentation made into Elan and checked under Praat showed systematic misalignment of the segments. This is the reason why we decided to start the segmentation process in Praat. This software allows visualization of spectrum, pitch and intensity, on top of the waveform, with important zooming effects.





However, Praat has no hierarchical structure allowing complex annotations and queries. The segmentation process achieved, we therefore use the capacity of Elan to import the Praat segmented file. At this point we have a segmented text correctly synchonized with the sound. It would be possible to segment the words of the text into morphemes and annotate them into Elan, but there would be no consistency guaranteed in this hard work.

For that reason we just prepare the text into Elan, by adding a *reference* tier and a *word* tier for each speaker, to allow morphosyntactic annotation into another sofware, Toolbox.

The *reference* tier displays a unique numbered label for each segmented unit, to identify it for later referencing. This labelization can be automatically generated by Elan.

The *word* tier contains each word of the text tier in a separated cell. It can be automatically generated by Elan text tokenizer, provided

the text in the *tx* tier is transcribed without sandhis, and normalized to some extent. If not, that is if *tx* is transcribed with assimilations, then an additional intermediary tier must be provided for tokenization to be successful.

When the text is segmented into words, we can export it into Toolbox[9]. Toolbox is a software dedicated to the management of textual databases such as lexicon and/or phrase databases. In addition, it can annotate a text with the contents of a lexicon. This is an interactive process in which the software searches the lexicon for each word of the text to interlinearize, and proposes the glosses it finds, each one on a line, vertically aligned under the word. If the actual word doesn't exist in the lexicon, Toolbox tries to isolate possible affixes (which may be listed in the same lexicon or in a special one), glosses them if they exist, until it finds the root in the lexicon, or, if not, outputs a failure mark for the rest of the word.

The user has to interactively choose between gloss ambiguities, correct wrong segmentation or add new lexemes into the lexicon with their glosses. This ensures a high level of consistency in the morphosyntatic annotation process.

2. The tiers in ELAN and their technical and theoretical motivations

After much discussion and a number of tests, the CorpAfroAs team decided to adopt a format containing six linguistic annotation tiers. As we will see, other tiers are added for technical reasons.

2.1. The technical organizing principles

The **ref** line references each segment by a numbered label. It is the only one which is synchronized to time, the *tx* tier being in

[9] http://www.sil.org/computIng/toolbox/

symbolic association to it, that is to say they share the same time segmentation. This *ref* tier is the ultimate reference that subsumes the other tiers. So, any cell, in the end, refers to a *ref* segment parent, and this allows Elan to jump to that main segment when asked to.
tx: is the tier in which the text is transcribed in broad phonetics, into 'phonological' words (with assimilations, sandhis etc.). Major and minor boundaries are indicated (/ & //), and pauses over 200 ms appear in a separate unit.
mot: is the tier in which the text is transcribed into grammatical words, with no morphemic separators (- =), and using a phonological (i.e. 'regularized' as compared to the broad phonetics one) transcription.
mb: is the tier in which the text is segmented into morphemes (one cell per morpheme); - goes in the cell that contains the affix, = goes in the cell that contains the clitic.
ge: is the tier in which a gloss is provided for each morpheme cell. The glossing is into grammatical category labels, and is based on the Leipzig Glossing Rules[10]. Other relevant information (parts of speech, verb class, syncretism phenomena, etc.) goes into tier **rx**.
rx: is the tier in which all information relevant and necessary for retrieval purposes is entered. If there is more than one label per cell, we separate them with a slash.
ft: is the tier where the text is translated (free translation). The translation is indexed to minor or major units depending on the syntax of the language.

The principle of Elan is to document a media resource (audio or video signal). The signal is displayed on a horizontal timeline, and tiers can be created under that line to synchronically annotate the

[10] http://www.eva.mpg.de/lingua/tools-at-lingboard/glossing_rules.php

signal: there is a vertical correspondence between the annotation lines and the signal line.

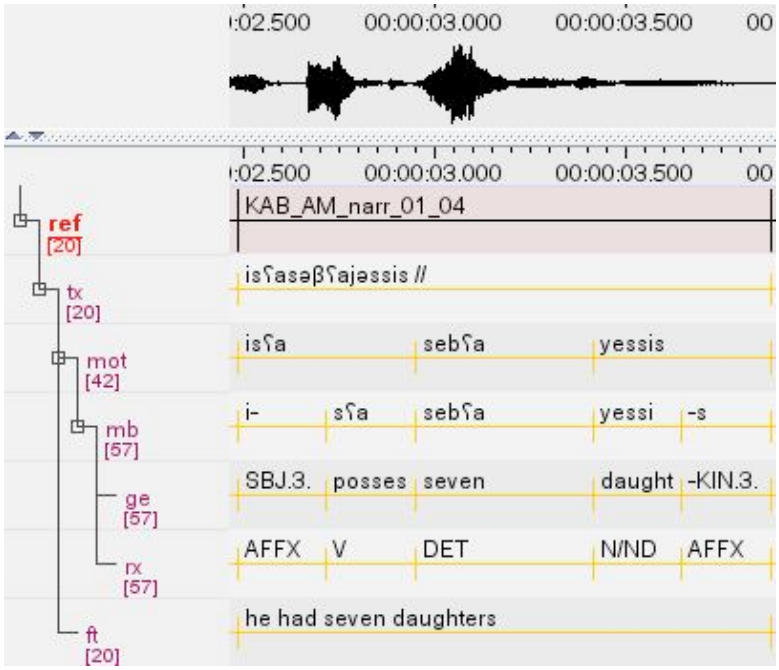
All the annotation lines do not need to be directly synchronized to the signal. In the CorpAfroAs project, only the first tier corresponding to the segmentation into minimal prosodic units is synchonized. The other tiers are indirectly indexed to time by dependency relationship among them.

The *word* tier has a *symbolic subdivision* dependency with the *text* tier. This means that the time duration of a text segment unit is divided (equally[11]), at the *word* tier level, between the words that belong to that segment. These words are not synchronized to sound, but they share the same time segment than the text segment which they belong to. They therefore inherit the time boundaries of their parent segment.

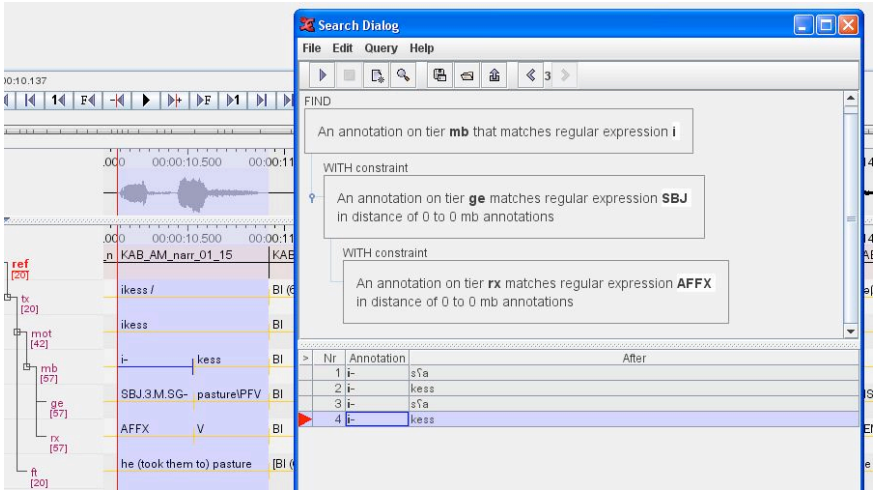
The *morpheme* tier is a *symbolic subdivision* of the *word* tier, i.e the different morphemes of a word share the time segment of the word they belong to.
The *gloss* and *rx* tiers have a *symbolic association* dependency with the *morpheme* tier. That means there is a correspondance term to term between morpheme and glose or extra glose.

These dependencies from tier to tier - also called child and parent relations - makes it possible to vertically align the elements in the way linguists usually present interlinear texts.

[11] The space occupied by a word with regard to its parent text segment does not represent its actual duration in the signal. This is why these are called *symbolic* subdivisions in Elan.



Elan has a retrieval engine allowing to look for a sequence (a word, a morpheme...) in a specific tier, with possible additional constraints (another sequence in another tier). The correspondence between the two (or more) tiers may be just direct, i.e sharing the same time segment, or the second one may be searched within a certain distance from the segment of the first. In the example below, we are looking for all the morphemes (tier *mb*) ' i ' that have 'SBJ' in their direct (distance 0) corresponding glossing tier (*ge*) and that are affixes ('AFFX' in the *rx* tier).



Thanks to the hierarchical structure of the tiers, when such a search is made, Elan will display all the occurrences of this morpheme, one per line, with the left and right context. From any occurrence, a jump is possible to the time segment to which the morpheme belongs, since Elan is able to look back in the hierarchy from child to parent. This time segment will display all the tiers depending on it, and clicking on the play button will allow listening to the sequence.

In the same way, a concordance can be made for a sequence (morpheme, word, gloss...), which will display occurrences centered in the line, with the left and right context within a selected distance. Statistics can also be displayed.

2.2. The theoretical organizing principles

The **tx** line is the one that holds the transcription of minor prosodic units. Its purpose is to reflect as closely as possible the sound file, including false starts and other phenomena found in spontaneous speech. As the phonology of the language is known, the

transcription is not completely phonetic, although it includes word-boundary phenomena (sandhi etc.), as those may be interesting for the phonology-syntax interface.

The **mot** line is mainly an intermediary tier that allows the subsequent segmentation into morphemes. It contains grammatical words, the definition of those words being language-dependent, therefore, this tier may not reflect exactly the word segmentation of the *tx* tier. The **mb** line is segmented into morphemes, allowing for allomorphs and all such variation desirable for a varied morpheme inventory.

The **ge** line is the morpheme-by-morpheme gloss of the *mb* line. Its syntax is based on the Leipzig Glossing Rules:

- When a single object-language element is rendered by several metalanguage elements (category labels), these are separated by periods. Ex: 3.M.SG (Rule 4)
- When a single object-language element is rendered by several metalanguage elements (words), these are separated by underscores. Ex: be_tall (Rule 4A)
- If a grammatical property in the object-language is signalled by a morphophonological alternation (ablaut, mutation, tone alternation, etc.), the backslash is used to separate the category label and the rest of the gloss. Ex: write\PFV (Rule 4D)

The list of abbreviations provided by the LGR is incomplete, and therefore one of the tasks we have completed is the creation and unification of all the proposed glosses for the languages of our pilot-corpus, with the assistance of Bernard Comrie, one of the creators of the LGR. A number of problems arose, to which solutions were proposed. Those solutions have been implemented within the Afroasiatic phylum, but are exportable to other language families, and will be listed and published at the end of the project. Here are some examples of the issues that were discussed:

- Traditional labels: for each language family of the phylum, descriptive traditions going back sometimes to more than a

century ago, have consecrated the use of some labels, such as ‘suffixal conjugation’ in Arabic, ‘free state’ in Berber. Those labels, although they have their motivation and are grounded in decades of analyses, make little or no sense to linguists that do not work within those traditions. We have decided to use more widespread labels whenever appropriate. Thus, the ‘suffixal conjugation’ of our Arabic varieties was labelled ‘perfective’, and the ‘free state’ of Berber was labelled ‘absolute’.

- Aprioristic vs nonaprioristic categorization of morphemes: however, this unification may have undesired side-effects, in that it may erase the language-specific function of those forms. For instance, the use of the label ‘marked nominative’ for the ‘annexed state’ of Berber might at first sight be desirable, because it is currently widespread among typologists. But the function of the annexed state of Berber is far more complex than the definition of the marked nominative implies. Therefore, the traditional label was retained, and the reference to case avoided.
- Use of labels covering different phenomena in different languages: in Berber a special form of the verb, invariable with respect to person, number and gender (but marked for aspect and mood) appears in relative clauses when the antecedent has the same referent as the subject of the subordinate clause. This form is traditionally called “participle”. However, due to the fact that the label “participle” in Indo-European languages refers to a different notion, the label SSREL, for “same subject relative form” was retained.

The **rx** line was originally a part-of-speech line. But when we started thinking about the queries that such an online corpus was supposed to allow, we realized that parts of speech were only just a small part of the necessary information. We therefore started with the queries themselves, and implemented the **rx** line with all relevant information, regardless of their linguistic domain. We thus also provide complementary morphological information

(neutralization or syncretism, morphological verb-class, etc.), as well as syntactic (word-order, etc.) and semantic (stative verb, etc.) information. We are currently testing the **rx** line for all those types of information. If the information load were too high, we might create an additional tier, but this in turn would imply more constraints on the computer programme that will treat the queries and provide the results.

The labels used in **rx** are sometimes the same as those used in **ge**. But they cover a different domain. For instance PREP in **ge** is a special prepositional paradigm of affixes, that is found in Berber, Semitic and Chadic. The prepositions in **ge** are glossed by their value only (either grammatically, e.g. LOC, or semantically, e.g. BETWEEN). In **rx**, PREP means that the morpheme is a preposition. This is useful for specific queries, because sometimes, the same morpheme can be a preposition, or a conjunction.

Here is an example of query: “search ANN in ge & ND in rx” will give us a concordance listing all the examples (with context) where a noun which does not morphologically mark the distinction between the two states (ND= no distinction) is (covertly) in the annexed state (ANN). The usefulness of the query lies in the fact that the distinction in Kabyle is covert for half the nouns in texts, therefore it may be interesting to retrieve all those cases, and see what their statistical distribution is: as postverbal subject, nominal modifier, complement of prepositions, etc.

Finally, the **ft** line was apparently unproblematic, but eventually raised some questions since it appeared that indexation to the minor units was only possible in some languages, while others were better translated within broader units (major ones). It also appeared that translating a text was in no way an easy task, since contrary to the translation of isolated examples for grammatical purposes, text

translations must also provide equivalences for pragmatic dimensions.

We are also planning on adding another tier synchronized to time which will fuse minor units into major units. This would allow to listen to a major unit instead of only minor units, when, for example, the latter is too short to understand the meaning of the sequence. Another free translation tier corresponding to those longer units could be added too.

3. The metadata

In relation to the previous point, translations often contain a certain amount of implicit information, which might be difficult to retrieve for a linguist who did not participate in the recording. This type of information, as well as other types, should be contained in the metadata accompanying the corpus.

There are two types of metadata: one is linked to the technical characteristics and status of the audio recording, the other to the texts themselves. The latter must at the same time provide all the necessary information for the texts to be anchored and understandable to an outsider who was not present during the recording, and protect the recorded speakers from any prejudice. In that view, as the data is to be made available online to the community, a thorough reflection process was engaged before data collection, concerning the deontological aspects of the project. Thus, anonymization procedures, as well as control over sensitive data (restricted access), have been implemented. In this process, we followed international recommendations, stated in *Corpus Oraux, Guide des Bonnes Pratiques*[12] (Baude 2006).

[12] http://www.culture.gouv.fr/culture/dglf/Guide_Corpus_Oraux_2005.pdf

At the same time, all the relevant information was listed, in order to provide rich metadata on the recordings. These metadata follow the requirements of OLAC[13] (Open Language Archives Community). We provide in annex the metadata form we have devised for each recording.

Conclusion

Two years after the beginning of the CorpAfroAs project, we are able to present a layout (the "CorpAfroAs format"), with a series of organized tiers, and a number of transcription and glossing rules, as well as a list of glosses for the *ge* and *rx* tiers, and a metadata form. Minor alterations will be made in the next two years, but the format is bound to remain quite similar to what it is now. The remaining work consists in finishing the annotation of the data, and working on the queries, theoretically as well as technically. The development of the software for the queries and for end-user visualization is also part of the remaining tasks.

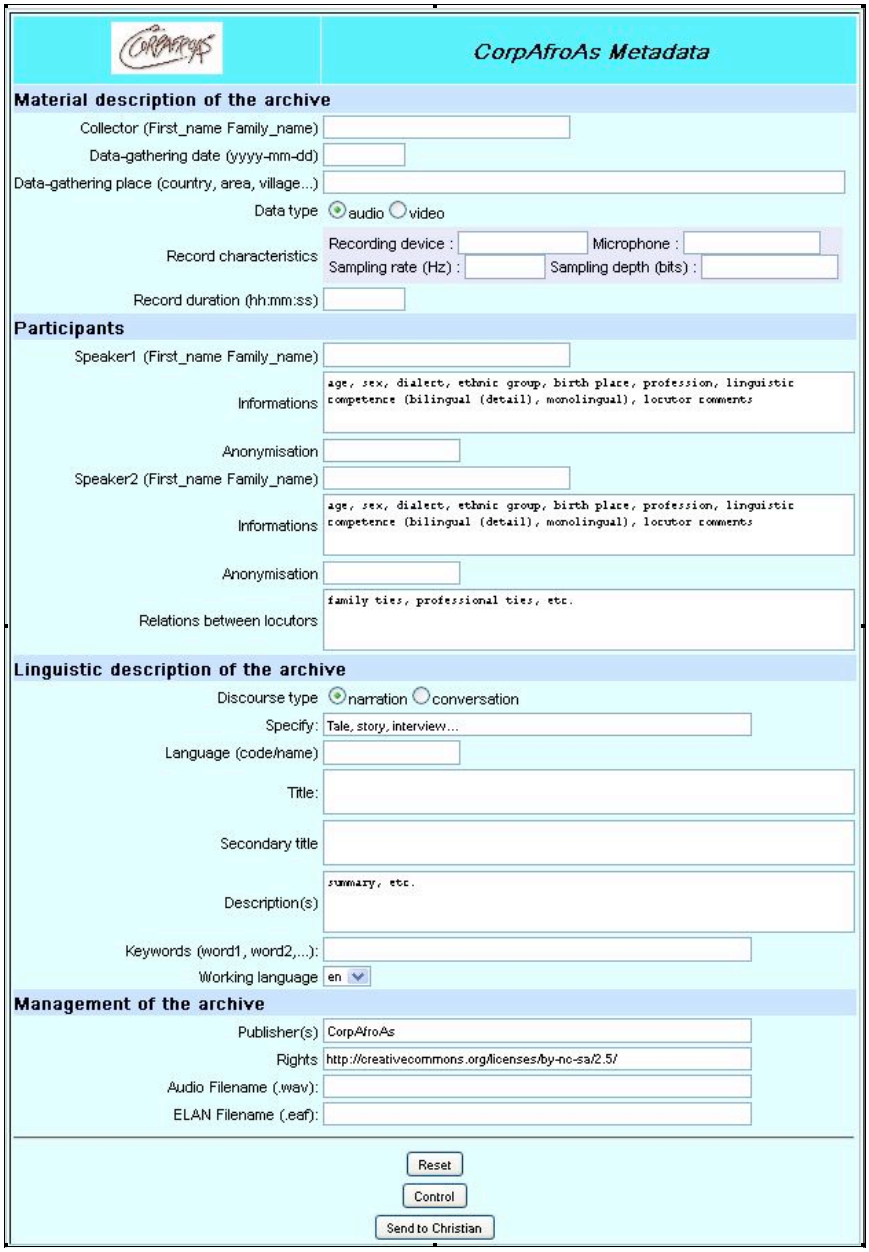
References

Baude, O. (ed). 2006. Corpus Oraux, Guide des bonnes pratiques. CNRS: Paris.

Mettouchi, A., D. Caubet, M. Vanhove, M. Tosco, B. Comrie & S. Izre'el. To appear. "CORPAFROAS, A Corpus for Spoken Afroasiatic Languages: Morphosyntactic and Prosodic analysis", Proceedings of the XIII Incontro Italiano di Linguistica Afroasiatica, M.F. Fales & G.F. Grassi (eds), Udine.

Wichmann, A. 2000. Intonation in Text & Discourse: Beginnings, Middles and Ends. Longman: Harlow.

[13] http://www.language-archives.org



CorpAfroAs Metadata

Material description of the archive

Collector (First_name Family_name)

Data-gathering date (yyyy-mm-dd)

Data-gathering place (country, area, village...)

Data type: (•) audio ( ) video

Record characteristics: Recording device : Microphone : Sampling rate (Hz) : Sampling depth (bits) :

Record duration (hh:mm:ss)

Participants

Speaker1 (First_name Family_name)

Informations: age, sex, dialect, ethnic group, birth place, profession, linguistic competence (bilingual (detail), monolingual), locutor comments

Anonymisation

Speaker2 (First_name Family_name)

Informations: age, sex, dialect, ethnic group, birth place, profession, linguistic competence (bilingual (detail), monolingual), locutor comments

Anonymisation

Relations between locutors: family ties, professional ties, etc.

Linguistic description of the archive

Discourse type: (•) narration ( ) conversation

Specify: Tale, story, interview...

Language (code/name)

Title:

Secondary title

Description(s): summary, etc.

Keywords (word1, word2,...):

Working language: en

Management of the archive

Publisher(s): CorpAfroAs

Rights: http://creativecommons.org/licenses/by-nc-sa/2.5/

Audio Filename (.wav):

ELAN Filename (.eaf):

Reset

Control

Send to Christian