SpeechReporting Corpus: annotated corpora of West African traditional narratives

Katya Aplonova
LLACAN, CNRS
aplooon@gmail.com

Izabela Jordanoska
LACITO, CNRS
izabela.jordanoska@cnrs.fr

Timofey Arkhangelsky
Universität Hamburg
timarkh@gmail.com

Tatiana Nikitina
LACITO, CNRS
tavnik@gmail.com

Abstract

This paper describes the SpeechReporting Corpus, an online collection of corpora annotated for a range of discourse phenomena. The corpora contain folktales from 7 lesser-studied West African languages. Apart from its value for theoretical linguistics, especially for the study of reported speech, the database is an important resource for the preservation of intangible cultural heritage of minority languages and the development and testing of cross-linguistically applicable computational tools.

1 Introduction

Recent decades have seen an upsurge of interest in issues of language extinction, leading to increased efforts to describe and document the world's endangered languages. The major adverse effects of language endangerment are also associated with loss of different forms of traditional knowledge (Hale, 1992).

The SpeechReporting Corpus (Nikitina et al., 2022) explores the relationship between specific discourse practices that represent the nucleus of the transmission of traditional knowledge and the linguistic strategies associated with it, centering on one particular problem: discourse reporting in traditional oral storytelling in West Africa, the "oral continent par excellence" and the homeland of a rich and vibrant oral tradition (Scheub, 1985; Finnegan, 2007, inter alia).

The article is structured as follows: in Section 2 we discuss why the SpeechReporting database is particularly relevant for West Africa. Section 3 is dedicated to database composition. Our workflow and tools are described in Section 4, while Section 5 shows some basic principles of annotation of reported discourse. In Section 6, we show how the online interface of the corpus works. In Section 7, we illustrate how the corpus can be used for dissemination among linguistic communities in order to archive and help preserve intangible cultural heritage. Section 8 concludes the paper.

2 West African storytelling traditions

In traditional rural societies of West Africa, the acute feeling of loss is related to the diminishing role played by culturally significant discourse practices: even in communities that retain traditional social organization and economy, modern Western cultural practices seep into daily life with new forms of entertainment (television, radio broadcasts) and education (compulsory Western-style schooling). Under the pressure from these new practices, traditional forms of knowledge transmission — including techniques of storytelling and instruction — become unimportant, and may eventually go out of use.

In many local communities across Africa, storytelling is more than a favorite pastime. Viewed as a vital part of cultural heritage, it serves as a central medium for the transmission of cultural knowledge. Storytelling traditions have accumulated special linguistic techniques that respond to the needs of specific practices of textual production and performance. As storytelling traditions vanish with older generations of speakers, they take along with them an array of linguistic tools on which such specialized techniques relied (Nikitina, 2018).

While oral traditions of West Africa have received considerable attention from anthropologists (Finnegan, 1970, 2012), their linguistic aspects have not been subject to systematic investigation. Our knowledge of the special ways in which language is used in traditional genres is largely limited to observations of frequent use of special vocabulary and opening/closing formulae (Cosentino, 1980), singing (Innes, 1965; Burnim, 1976; Azuonye, 1999), and various forms of repetition (Finnegan, 1967, 1977). The goal of our database is to start filling this gap using an interdisciplinary approach combining rigorous analysis

Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023), pages 26–31

of primary data with meticulous attention to genre characteristics and culture-specific contexts of textual production.

3 Database composition

The SpeechReporting Corpus contains multiple sub-corpora of traditional folk stories, annotated for a number of discourse phenomena using the ELAN-CorpA software and tools (Chanard, 2015; Nikitina et al., 2019). It is updated regularly with newly available data, including data from new languages. The project currently involves 11 different languages, of which 7 are spoken in West Africa. All texts are transcribed, glossed, translated, and annotated. Table 1 lists the West African languages[1] in the database, their genetic affiliation and country where they are spoken.

Language	Affiliation	Place
Bandial	Atlantic	Senegal
Gizey	Chadic	Cameroon
Guro	Mande	Côte d'Ivoire
Kafire	Senufo	Côte d'Ivoire
Mwan	Mande	Côte d'Ivoire
Ut-Ma'in	Kainji	Nigeria
Wan	Mande	Côte d'Ivoire

Table 1: West African languages in the SpeechReporting database

In the project, we work with texts (both oral and written) and not with elicited data. This helps to avoid the influence of the working language, and speakers' potential judgment about 'proper language use'. For example, logophoric pronouns, repetitions and some interjections and ideophones are very hard to elicit, though they do occur frequently in narratives.

We also restricted the genre of the texts we work with. The corpus is annotated for reported discourse (see Section 6), and thus, we chose fairy tales as a main data source, since in most fairy tales the driving force of the narration is the communication among characters.

Despite trying to keep the genre consistent across languages, the data are still very diverse. For example, it includes archived transcriptions, data from the field, recordings of professional storytellers as well as of regular people, one or multiple participants. Table 2 contains information about the composition of the corpus.

Language	Data format	Tokens	Phrases	Texts
Bandial	text, audio, video	10,378	1260	28
Gizey	text, audio, video	5184	700	10
Guro	text, audio, video	7346	1129	2
Kafire	text, audio, video	14,921	2769	17
Mwan	text	24,949	1797	33
Ut-Ma'in	text	1159	246	7
Wan	text	48,195	5370	82

Table 2: Composition of the Discourse Reporting database (West African languages only)

In the table, 'Data format' refers to the modality of the data. For some languages, we have audio files, video files and corresponding written transcriptions, while for others, we only have the written transcriptions. While 'Tokens' refers to the number of tokens, 'Phrases' refers to the total number of intonational units into which the texts of that language are segmented. 'Texts' refers to the number of separate ELAN files per language, each corresponding to one narrative.

We transcribe texts using orthographies based on the International Phonetic Alphabet. The African languages in the database do not have a standardized orthography which is widely used by native speakers. Published materials are scarce and, in the majority of cases, were developed by the authors of the corpora and rely on the same or similar orthography.

4 Workflow and tools

The project unites multiple collaborators that work in different frameworks and use different tools for data documentation and analysis. As a result, we had two basic workflows. In one, segmentation and transcription are done in SayMore (Hatton, 2013) or ELAN (n.a., 2022; Sloetjes and Wittenburg, 2008). Segmented and transcribed texts are glossed in Toolbox or Flex and then imported to ELAN in order to add annotations of reported discourse. The other workflow allows researchers to use only one software product, ELAN-CorpA for segmentation, transcription, glossing and annotation of reported discourse.[2]

[1] In addition to describing strategies for reporting discourse employed in oral traditions of selected West African cultures, the project sets out to compare them to their functional counterparts from a geographically and historically unrelated area. Therefore, the database contains some languages spoken in Eurasia that will not be discussed in this article.

[2] ELAN-CorpA is a fork-version of ELAN, developed by Christian Chanard, check this link https://llacan.cnrs.fr/res_ELAN-CorpA_en.php for more information.

Annotated files are checked manually and by using ELAN Tools (Chanard, 2019), a collection of scripts that checks the consistency of labels and the structure of the ELAN files. The manual checking consists of, among other things, proofreading the free translations and looking out for irregularities in the glosses and the morphemic analysis. Double-checked files are uploaded to Tsakorpus.

Collaborators could contribute to the project in various ways. Since 2019, we have had 4 post-docs, 2 PhD students, 6 research assistants and 8 non-contractual academic visitors working on the corpus.

5 Annotation of reported discourse

Annotation of reported discourse consists of four levels: the function of the construction's elements; the construction's syntactic type; the semantic type of the discourse report; and the encoding of participants within the discourse report. They correspond to four additional ELAN tiers (in our template, `qt`, `rp`, `typ` and `par`, respectively). Figure 1 is an example of our annotation of a Gizey sentence in an ELAN file.

Explained below are the basic principles of annotation that are relevant to searching in the corpus interface.[3]

A reported speech construction consists of different elements; for example, in *John said: Hello!* the reported utterance (*Hello!*) is introduced by a clause describing the speech event (*John said*). In the Gizey example in Figure 1, "she says" is expressed by a Quotative, while "give millet; this red mare of mine..." is a Discourse report. The semantic type of the Discourse report is Command.[4]

Different syntactic types of reported discourse constructions are visually represented by different frames. The types are defined by the elements the construction consists of. The syntactic type in the Gizey example is Quotative + Discourse report.

The elements referring to participants in the current or reported speech event are annotated in the Participant tier. The Gizey example contains a reference to the Reported Speaker (RS).

In the Tsakorpus interface, these annotations are reflected by background colors, frames, and pop-up windows. This is illustrated in Figure 2 for a sentence in Bandial (also known as Jóola Eegima), where the reported segment is in green and the speech event is in red.

6 Searching in Tsakorpus

Equipping the annotated corpora with a web-based search interface makes them more accessible both to linguists and to language communities. We made our corpora available online with the help of the *Tsakorpus* platform.[5] The platform was mostly developed independently of the project. However, a number of features were added specifically to accommodate the needs of the SpeechReporting database.

Search queries in the online interface are formed by clicking on buttons and filling out text fields. A single-word query can include constraints on the word, its lemma, its part of speech and/or its glosses. All fields can handle Boolean functions (`,` for `AND`, `|` for `OR`, `~` for `NOT`). Word and lemma search can include regular expressions and provide instant suggestions when the user starts typing. Multi-word queries consist of several single-word queries with additional distance constraints.

When clicking "Search sentences", the user gets randomly ordered search hits, split into pages. The sound associated with a particular search hit, if any, can be played by clicking on that hit.

One limitation of Tsakorpus is that its basic search unit is a sentence (or any sentence-like segment of text). It is not possible to search for units that are either larger than a "sentence", or smaller than a "sentence" but larger than a word. ELAN segments (which normally represent intonational units) were reinterpreted as "sentences" in Tsakorpus. However, our discourse annotation often consists of multi-word spans that are either smaller than a sentence or transcend the segment sentence. In order to make them searchable, we add values of all discourse annotations that appear anywhere within a sentence as sentence-level metadata. This way, a query like "Quotative AND a word glossed as *say*" will return all sentences that contain both a Quotative span and a word glossed as "say", but they will not necessarily overlap. This option was added to Tsakorpus in the course of the project. Nevertheless, the exact spans inside sentences that have discourse annotations are highlighted with

[3] A detailed description of the annotation principles can be found on the project website `http://discoursereporting.huma-num.fr/annotation.pdf`

[4] The terminology used in the annotation of the syntactic and semantic elements in speech reporting comes from Spronck and Nikitina (2019).

[5] `http://discoursereporting.huma-num.fr/corpus/search`

Figure 1: Example of an annotated Gizey sentence in ELAN

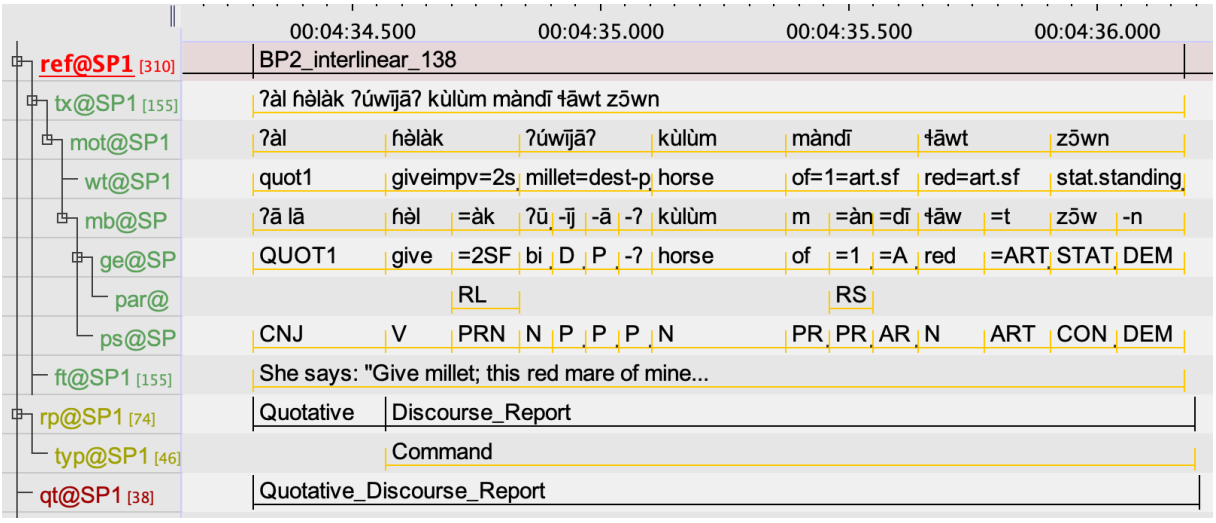
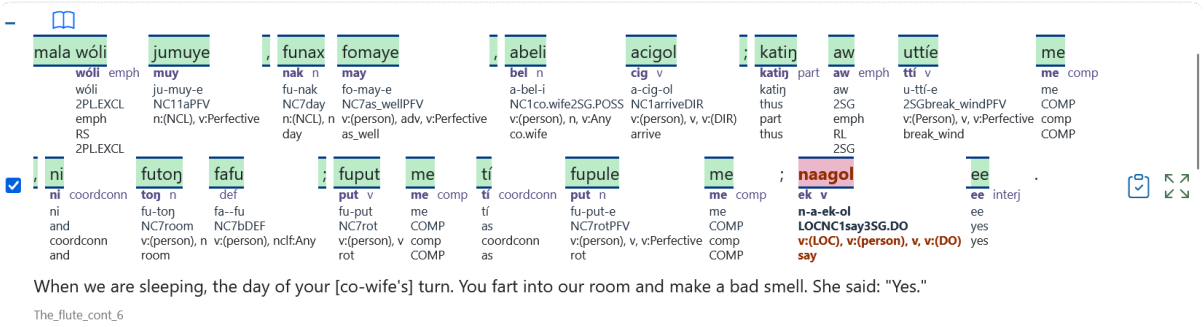


Figure 2: Example of search within Tsakorpus, a sentence in Bandial (Jóola Eegima)



different colors in the search hits.

The online corpus contains all languages present in the Discourse reporting database. When searching in the corpora, the user can choose between selecting a specific language and searching in all language subcorpora at once. In the latter case, the search query must include annotation that is uniform across the subcorpora. This includes annotation of reported discourse (see Section 5) and part-of-speech tags in the UD format (de Marneffe et al., 2021).

Currently, the corpus interface is available in English and Russian. The French interface is under construction.

7 Dissemination

Target users of the SpeechReporting Corpus are linguists and anthropologists who are interested in traditional narratives.

Besides academic uses, this corpus is a valuable source of materials for language communities to keep their languages and linguistic traditions alive; first of all, by simply having online access to recorded narration sessions of some of their folktales. In addition, we make materials, such as storybooks, that the communities can use for educational purposes. Moreover, our project has received additional funding from the *Humanités Numériques et Science Ouverte* program of the Sorbonne Nouvelle in Paris for producing animated YouTube videos of the recorded folktales and spreading them among the linguistic communities and wider audiences.

Furthermore, the availability of open-access annotated linguistic data of minority African languages is important for the development of machine learning based technologies, which currently under-represent these languages.

8 Concluding remarks

The SpeechReporting Corpus provides meticulously annotated corpora of low-resourced indigenous languages spoken in West Africa. It also offers a digital representation of reported speech constructions on different levels of analysis (morphology, syntax and semantics), which opens a potential of a new understanding of a range of discursive phenomena. The SpeechReporting Corpus offers open access tools for comparable annotation of data from different languages. It contributes to the accessibility of previously unpublished traditional narratives in indigenous languages spoken in West Africa.

Limitations

One limitation of this corpus is the difficulty of comparing between the different languages. It is hard to identify typologically applicable categories based on limited amounts of data. For example, when deciding on which discourse categories to annotate, we had to make sure that we could use the same vocabulary for all the languages in our sample.

An additional limitation is the possible lack of consistency between the different subcorpora. The cross-checking of the data is done manually. This is a tedious task that is susceptible to human error, but it is necessary to improve the quality of individual data sets.

Furthermore, we have discovered that it is challenging to bring together a perfect team for a project that is both linguistic and technological in nature.

Another possible limitation is related to transparency. Considering possible future uses of our corpora, we have tried to make the annotations as transparent as possible and have documented them all on our website.

Interdisciplinarity is another challenge: the kind of data that is suitable for dissemination in the communities is slightly different from the kind of data that is of primary interest from the point of view of linguistic theory.

References

Chukwuma Azuonye. 1999. Igbo stories and storytelling. *Traditional Storytelling Today: An International Sourcebook*, pages 33–40.

Mellonee Victoria Burnim. 1976. *Songs in Mende Folktales*. Madison: University of Wisconsin.

Christian Chanard. 2015. ELAN-CorpA: Lexicon-aided annotation in ELAN. In *Corpus-based Studies of Lesser-described Languages*, pages 311–332. John Benjamins.

Christian Chanard. 2019. ELAN Tools: Python tools for ELAN. Online access: https://llacan.cnrs.fr/res_manuels_en.php, last accessed 13/02/2023.

Donald J Cosentino. 1980. Lele Gbomba and the style of Mende baroque. *African Arts*, 13(3):54–55.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Ruth Finnegan. 1967. *Limba stories and story-telling*. Oxford: Clarendon Press.

Ruth Finnegan. 1970. A note on oral tradition and historical evidence. *History and Theory*, 9(2):195–201.

Ruth Finnegan. 1977. Oral poetry: Its nature, significance and social context.

Ruth Finnegan. 2007. *The oral and beyond: doing things with words in Africa*. James Currey/University of Chicago Press.

Ruth Finnegan. 2012. *Oral literature in Africa*. Open Book Publishers.

Ken Hale. 1992. Endangered languages: On endangered languages and the safeguarding of diversity. *language*, 68(1):1–42.

John Hatton. 2013. Saymore: Language documentation productivity [Computer software].

Gordon Innes. 1965. The function of the song in Mende folktales. *Sierra Leone Language Review*, 4:54–63.

n.a. 2022. ELAN (Version 6.4) [Computer software].

Tatiana Nikitina. 2018. When linguists and speakers do not agree: The endangered grammar of verbal art in West Africa. *Journal of Linguistic Anthropology*, pages 197–220.

Tatiana Nikitina, Ekaterina Aplonova, Abbie Jordanoska, Izabela andHantgan-Sonko, Guillaume Guitang, Olga Kuznetsova, Elena Perekhvalskaya, and Lacina Silué. 2022. The speechreporting corpus: Discourse reporting in storytelling.

Tatiana Nikitina, Abbie Hantgan, and Christian Chanard. 2019. Reported speech annotation template for ELAN. (The SpeechReporting Corpus). Online access: http://discoursereporting.huma-num.fr/annotation.pdf, last accesed 16/03/2023.

Harold Scheub. 1985. A review of African oral traditions and literature. *African Studies Review*, 28(2-3):1–72.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

Stef Spronck and Tatiana Nikitina. 2019. Reported speech forms a dedicated syntactic domain. *Linguistic Typology*, 23:119–159.