Proceedings of the Second Workshop on African Language Technology
- AfLaT 2010 -

Corpus Building in a Predominantly Oral Culture: Notes on the Development of a Multiple-Genre Tagged Corpus of Dagbani

Tristan Michael Purvis
University of Maryland, Box 25, College Park, MD 20742
tpurvis@umd.edu

Abstract

This paper describes the development of a multiple-genre tagged corpus of texts and recordings in Dagbani, a language spoken by a newly literate speech community in an African context with unique oral traditions. The paper begins with a brief review of the target speech community and corpus collection, followed by a discussion of the problems encountered and procedures conducted for tagging the corpus, and concludes with a presentation of example applications of and future plans for this corpus.

1. Introduction

Among the many challenges of developing African language technologies is the fact that language technology development highly favors the use of written resources both in terms of data management and specific applications, whereas a majority of African cultures place a higher value on oral traditions. This paper describes the development of a multiple-genre tagged corpus of texts and recordings for a language (Dagbani) spoken by a predominantly oral, yet also newly literate, speech community in an African context.

2. Background on Dagbani Language and Corpus Collection

Dagbani is a Gur (Niger-Congo, Atlantic-Congo) language spoken in northern Ghana by approximately 800,000 people (Lewis, 2009). As one of 12 nationally sponsored languages, the range of genres available for corpus collection is reinforced by government support for radio and television broadcasts, substantial funding and coordination of mother-tongue education in public schools and adult literacy programs, and government and NGO assistance for printing publications of various sorts. An attempt was made to balance the corpus with roughly equivalent written and spoken counterparts, as illustrated in Table 1. For example, the genre of oral history can be compared directly with a written counterpart; traditional (oral) fable-telling can be compared with written stories; scripted dialogue and personal letters to some extent are paired with spoken conversation; etc.

The Dagbani corpus, originally compiled and processed for use in a multi-dimensional analysis of register variation in Dagbani (Purvis, 2008; see Biber, 1988, 1995 on MD analysis methodology in general), comprises 163 texts (or text excerpts) and transcripts of recordings averaging 880 words each (roughly 140,000 words total) and representing approximately 32 genres. In addition to the primary corpus of naturally occurring data, there is a secondary corpus of 183 experimentally collected texts and recordings averaging 200 words (37,500 words total), made up of pairs of equivalent written and spoken narratives by the same speakers on the same topic.

Written	Spoken
Short stories, fables	Salima (fable-telling)
Novel excerpts	Oral interview with novelist
News articles	Oral recounting of articles
Written history	Oral history
Poetry	Oral intro/preface; Interview
Introduction	with novelist (monologue)
Personal letters	Natural conversation
Scripted conversation	Radio call-in discussion
Written plays	Broadcast talk show
	Improvisational drama
(see News article above)	*Radio & TV News
	*Broadcast announcement
Educ.materials (grammar)	College lecture
Legal texts	Courtroom proceedings
Medical pamphlets	Technical medical lecture
Agricultural pamphlets	Technical agric. discussion
Educational stories	Political speeches
Qur'an translations	Muslim sermons
Bible translations	Christian sermons
Written sermon etc.	*Christian prayers
*oral delivery, likely read from script	

Table 1. Genres collected for the Dagbani corpus

3. Tagging

Unlike corpus development for better resourced languages, which typically starts with a substantial base of electronically available texts and transcripts, the initial processing of texts and recordings for this corpus was almost entirely manual. First of all, there are hardly any online public texts in Dagbani apart from a translation of the Universal Declaration of Human Rights and selected translations of the Qur'an; so all texts for this research would have to be entered manually.

When working collaboratively with locals in a remote, semi-rural area in developing nation, simple technological assumptions such as Unicode font compatibility and keyboard set-up cannot be taken for granted. Therefore, to avoid problems in data processing and file transfer, a simplified transcription system was adopted based on a somewhat iconic use of Arabic numerals to represent the five non-Latin symbols used in Dagbani orthography: 3 = ɛ; 0 = ɔ; 6 = ɣ; 4 = ŋ; 7 = ʒ.

The option of scanning some of the written texts was ruled out due to poor print quality and frequent typographical errors or nonconventional orthography found in many Dagbani documents. A widespread number of homonyms/homographs in the Dagbani language would have posed another challenge for the automated tagging of the Dagbani texts, especially in combination with these other constraints. In speech, the number of homophones is reduced by tonal and vowel qualities, but notation of intonation and lexical tone are not represented in Dagbani orthography nor were these phonological distinctions within the scope of the research for which this corpus was first compiled. Unlike English, for example, where homonymy is rarely found among functional categories (e.g., there is only one *and*, only one *the*, only one *we*),[1] the homonymy in Dagbani does involve all sorts of major functional linguistic categories, which are often essential anchoring points in tagging programs or algorithms. Table 2 provides two different exemplary sets of such homographs and/or polysemes.

(a) *di*		(b) *ni*	
Grammatical Category	**Pre-tag Code**	**Grammatical Category**	**Pre-tag Code**
verb 'eat'	`di`	COMP; QUOtative	`ni` `ni:`

3S/INAM Subj pro; 3S/INAM Poss pro; and 3S/INAM Obj. as proclitic (rare)	`di+`
	`di$`
	`di@`
		SUBordinator, used in relative clause constructions & temporal clause	`ni*1`
			`ni*2`
			`ni-as`

NEG. IMPERative	`di*`	FUTure marker	`ni%`
preverbal particle (time depth marker)	`di%`	NP conjunction; instrumental use	`ni&` `ni&-w`
		LOCative	`ni#`

Table 2: Examples of Dagbani homography/homophony

For example, the sequence *di di di zaa* could be read as 'don't eat it all' (NEG/IMPER eat it all) or 'it eats it all' (3S/INAM eat it all). As another example, what is typed as 'bɛ pan bo (ma)' in one written text and heard as [bɛ pa m bɔ (o)] in a separate recording ends up being transcribed in standard orthography as *bɛ pa ni bɔ* ('they now will look for . . .', 3P now FUT seek). 'bɛ pan bo' is meaningless, there being no word transcribed as 'pan' in Dagbani orthography. The phonetic representation [bɛ pa m bɔ] could have been interpreted literally (but incorrectly in this case) as 'many of them look for' (3P many EMPH seek). The correct transcription *bɛ pa ni bɔ* contains homographs that could lead to misinterpretation.

Consequently, a system of notation (exemplified in the "Pre-tag Code" column of Table 2) was devised in which each word was coded upon initial data entry for part of speech and other linguistic information relevant to a given word class (noun class, verb type, inflectional categories, syntactic role, etc.). In many cases the pretag symbols were either iconic or productively used with multiple lexical items; so the system could be learned easily and transcription could be completed rapidly. For example, all pronouns are tagged upon initial data entry for syntactic role of subject (`+`), object (`@`), or possessive (`$`). Or, compared to the example of `ni&` shown in Table 2 for nominal conjunction, the ampersand is also used to distinguish VP conjunction (`ka&`) and IP conjunction (`ka&&`) from homophonous forms such as the complement FRONting/focus particle `ka!` where we find the same exclamation point as used to distinguish the subject focus/EMPHasis particle `n!` from homophonous forms like the first person singular proclitic (`n+`, `n$`).[2]

Thus, the majority of features to be tagged for potential linguistic analysis were previously coded in each lexical item or somewhere in the text or transcript and only needed to be converted to a more formalized tagging system. Various discourse categories such as addressive and left-dislocated topic were also notated immediately as the text was entered into electronic form.

A sample of the formal tagging system is presented in Table 3. The first column of the tag indicates the basic part-of-speech category (**N**oun, **V**erb, ad**J**ective, adve**R**b, **D**eterminer, pronoun (**1-5**, **W**h), **C**onjunction, **P**article, or interjection/e**X**clamation). The remaining columns account for morphological, syntactic, or lexical information specific to a given part of speech, such as valence and aspect for verbs.

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6
Noun	**S**ingular	**A**nim.	po$s.	**C**ompound Lexicalized	**E**nglish
	Plural	**I**nanim.	**L**oc.	**K**ompound	**A**rabic
	#:denom	**P**roper	**D**ir.	**V**erbal compound	**H**ausa
	Dummy	***Bu***			**T**wi
		Gu			**M**ampruli
		Lana			**G**a
		Redup.			
		Quantity			
Verb	**T**rans.	**P**erf.	**L**oc.	**1**st in serial	**E**nglish
	Intrans.	**I**mperf.	**D**ir.	**2**nd in serial	**A**rabic
	Ditrans.	i**M**per.	**N**eg.	**3**rd in serial	**H**ausa
	Modal	**R**edup.		...	**T**wi
	Sentential	***-Gi*** form		**A**-1st in n-serial	
	Caus.			**B**-2nd in n-serial	
	Adjectiv.			**C**-3rd in n-serial	
	Pluraction			...	
	#:deverbal	lexical			
		Serial			
	e**Q**uative				
	e**X**istential				

[1] Cases do exist, e.g. *to* (preposition) versus *to* (infinitive), but nowhere near as much as in Dagbani, as illustrated in Table 2.

[2] The use of symbols such as `+`, `@`, `$` of course requires adding escape characters (or other work-around strategies) in later stages of data processing, but this is outweighed by the rapid notation of linguistic information during data entry.

Col 1	Col 2	Col 3	Col 4	Col 5
Part-icle	**V**erbal	**S**entent. (*yi*,*ni*,*ti*)	**T**i-sequence marker	**P**urposive **E**vent
			Ni	**1**-herc
		Infinitive		**2**-hirc
		Negative		**3**-as/when
			Conditional	**C**ondition
				Result
		Time		
		e**M**phatic		
		Future	fut w/**N**eg	e**M**phatic
		adve**R**bial	**M**anner	
			Time	
			Purpose	**I**nf w/purp
	su**F**fix	**I**nfostatus	**I**mperfective	
			Focus	
		Nominal	**P**lural	
			Monger	

Table 3: Sample of formal tagging system

A few categories apply to multiple parts of speech - such as reduplication found with nouns, verbs, adjectives, adverbs, and interjections or language origin to document lexical borrowing or code-switching with nouns, verbs, adjectives, adverbs, determiners, conjunctions, and interjections. Conventional gloss categories and/or lemma gloss are combined with this tagging system. Perl script was used to convert pretagged texts and transcripts to fully tagged texts en masse.

As an illustration, the sentence presented earlier, *bɛ pa ni bɔ*, first encountered with nonconventional orthography as shown in Figure 1, would be entered as `b3+ pa* ni% b0VT` following the pretag data entry stage and later converted to `b3<3PSAW-3P> pa<PVRT-now> ni<PVF-FUT> b0<VT-seek>`.

"Man dii ycli ka nyin bi lan nya? Bɛ pan bɔ ma bɔ n
jc." Din ka Dagbamba booni Banguman nɔ. To Zolugu

Pretag: `b3+ pa* ni% b0VT`
Tagged: `b3<3PSAW-3P> pa<PVRT-now> ni<PVF-FUT> b0<VT-seek>`
Orthography: *bɛ pa ni bɔ*
Interlinear: 3P now FUT seek
Phonetic: [bɛ pa m bɔ]
Translation: 'they now will look for . . .'

Figure 1: Sample of nonconventional orthography (source: Sulemana, 1970)

In this case, *bɛ* (3P) is indicated as a third person plural inanimate weak personal pronoun acting as syntactic subject[3]; the word *pa* ('now') is indicated as a member of the large class of preverbal clitic particles (`<PV`) in general and more specifically as providing temporal adverbial information; the word *ni* (FUT), also of the preverbal clitic particle class, is indicated as the (affirmative) future marker; and *bɔ* ('seek') is indicated as a transitive verb with no further aspectual inflection.

Apart from time efficiency, a primary concern against using manual tagging as opposed to automated tagging algorithms is that of consistency and reliability. While automated tagging procedures are prone to mistags, at least the inaccuracy of the algorithm is consistent. Biber (1988) claims that spot checks reveal his tagging algorithms to be 90 percent accurate or better. Considering the relatively low frequency of certain key features such as "that-complements" and "that-relatives" which were reported as exemplary instances of mistagging in Biber's corpus, however, the overall accuracy rating of 90 percent may be misleading. A hand-tagged corpus as in the present study arguably benefits from a higher level of accuracy, with the potential for inaccuracy mainly dependent upon data entry errors and misjudgments on the part of the data analyst or native-speaker research assistants. The accuracy of coding linguistic features for this corpus was verified through periodic retagging of selected text excerpts and confirming that these matched up to the original tags.

4. Corpus Applications

The Dagbani corpus has been used in a number of research projects for theoretical and descriptive linguistic purposes. Immediate plans are to prepare the data for more advanced computational applications. A variety of completed, ongoing, and projected applications are reviewed below.

4.1 Factor analysis of register variation

As noted in Section 2, this corpus was originally collected for use in a study of register variation of Dagbani following what is known as the multi-dimensional analysis protocol (Purvis, 2008). For this study, factor analysis was conducted based on normalized frequency counts of targeted linguistic features in order to identity clusters of co-occurring variables and interpret the underlying communicative function. A Perl script was again used to compute frequencies of the targeted features for each text or transcript en masse. Some features were tallied directly (e.g. `FUT` for a unique future marker); while others involved variable expressions such as `<[1-4]...S` for all strong (emphatic/disjunctive) personal pronouns. Automated algorithms using Regular Expressions were also employed to track linguistic features involving multiple lexical items—e.g., `<VS*[A-Z1-3|]> ka<CSS2_ COMP2-and_COMP>` for identifying instances where *ka* as opposed to the more standard *ni* is used as a complementizer for verbs taking sentential complements. Some targeted linguistic features required additional manipulations of the texts and/or tags. For example, tallying of token-type ratio (TTR) for the first 300 words of text involved first running a Perl script to extract the

[3] Pronominal allomorphs are actually distinguished by clitic position not syntactic role, but syntactic role was tagged in the interest of theoretical and descriptive analysis (Purvis, 2007).

first 300 words and then a systematic trimming of lexical data to a subset of the tag categories and/or gloss tags so that the TTR was based on word lemmas as opposed to variably inflected word forms. Similarly, the tags were manipulated to ensure that a tally of lexical borrowings (of nouns and verbs mostly) treated inflected forms such as plural and possessive nouns as instances of the same lexical target.

4.2 Corpus analysis of pronoun allomorphs

The Dagbani corpus has also been exploited to efficiently track patterns of the use of allomorphs of weak personal pronouns (Purvis, 2007). These were formerly described as varying according to syntactic role (see, e.g., Olawsky, 1999; Wilson, 1972). Based on evidence from this corpus, however, they are better understood as clitics whose variable forms depend on their position in relation to their lexical host.

4.3 Corpus analysis of left dislocation

In another study inspired by seemingly anomalous patterns in the factor output of the original study on register variation, the corpus has provided a means to track the patterns of occurrence of and analyze the motivations behind the use of left dislocation constructions in Dagbani (Purvis, 2009).

4.4 Corpus analysis of relative constructions

Presently, the tagging of relative pronouns and other lexical items used in variable relative clause constructions in Dagbani is being expanded to indicate a number of contextual factors deemed useful in analyzing variation in the appearance of these constructions (e.g., placement in relation to verb of the main clause; placement in relation to verb of the relative clause, etc.)—in order to address unanswered questions and track previously undocumented variants (cf. Wilson, 1963). This pursuit has been facilitated by the legacy tagging system in combination with linguistic concordancing software.

4.5 Current developments

In its current state, the Dagbani corpus already constitutes a useful tool for computational applications. For example, with word by word tagging of part of speech and various categories of inflectional and derivational morphology and syntactic role in some cases, this modest corpus may be used as training data for morphological and syntactic parsing and machine translation of novel Dagbani texts and possibly adapted to other Gur languages. (The Gur cluster has been identified as one of 15 core languages and language clusters which in combination reportedly provide communicative coverage for up to 85% of Africa's population (Prah, 2002).)

Currently, we are in the process of aligning the tagged transcriptions with digitized recordings for the spoken portion of the Dagbani corpus to be available for use in applications such as speech recognition. As for the written texts, we are investigating methods for associating the tagged transcriptions with OCR-processed versions of the original documents which may contribute towards efforts to train software to correctly recognize nonstandard orthography and accurately deal with the widespread homography found in this language.

5. Summary

This paper has described the development of a multiple-genre tagged corpus of texts and recordings for a language (Dagbani) spoken by a newly literate speech community with an otherwise predominantly oral culture. Following a review of the target speech community and corpus collection, we have discussed problems encountered in data collection and presented the procedures followed for tagging the corpus. The corpus has benefitted a number of descriptive and theoretical research projects to date and is undergoing further processing for future computational applications.

6. Acknowledgements

The bulk of corpus collection and processing was funded by a U.S. Department of Education Fulbright-Hays doctoral dissertation grant in 2003-2004 and could not have been completed without the help of the numerous collaborators in the Dagbani-speaking community.

7. References

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation*. Cambridge: Cambridge University Press.

Lewis, M.P. (Ed.). (2009). *Ethnologue: Languages of the World, 15th edn*. Dallas, Tex.: SIL International.

Olawsky, K. J. (1999). *Aspects of Dagbani Grammar*. München: Lincom Europa.

Prah, K. (2002). Language, Neo-colonialism, and the African Development Challenge. TRIcontinental, No. 150. Cape Town: CASAS.

Purvis, T.M. (2007). A Reanalysis of Nonemphatic Pronouns in Dagbani. In F. Hoyt, N. Seifert, A. Teodorescu, & J. White (Eds.), *Proceedings of the Texas Linguistics Society IX Conference: The Morphosyntax of Underrepresented Languages*. Stanford: CSLI Publications, pp. 239–263.

Purvis, T.M. (2008). A Linguistic and Discursive Analysis of Register Variation in Dagbani. Doctoral dissertation. Indiana University.

Purvis, T.M. (2009). Left Dislocation in Dagbani. Paper presented at the Sixth World Conference on African Linguistics (WOCAL6), Cologne, Germany.

Sulemana, T. (1970). *Naa Luro*. Accra: Bureau of Ghana Languages.

Wilson, W.A.A. (1963). Relative constructions in Dagbani. *Journal of West African Languages*, 2 (Part 2).139–144.

Wilson, W.A.A. (1972). *Dagbani: An Introductory Course*. Tamale, Ghana: GILLBT.