American Journal of Artificial Intelligence
2025, Vol. 9, No. 2, pp. 110-121
https://doi.org/10.11648/j.ajai.20250902.12

SciencePG
Science Publishing Group

Research Article

Towards a Set of Morphosyntactic Labels for the Fulani Language: An Approach Inspired by the EAGLES Recommendations and Fulani Grammar

Zouleiha Alhadji Ibrahima[1], Charles Moudina Varmantchaonala[2], Dayang Paul[1, *], Kolyang[3]

[1]Department of Mathematics and Computer Science, Faculty of Science, University of Ngaoundere, Ngaoundere, Cameroon

[2]Institute for Physics, Faculty V, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

[3]Department of Computer Science, University of Maroua, Maroua, Cameroon

Abstract

This paper details the development of a morphosyntactic label set for the Adamawa dialect of the Fulani language (Fulfulde), addressing the critical lack of digital resources and automatic processing tools for this significant African language. The primary objective is to facilitate the creation of a training corpus for morphosyntactic tagging, there by aiding linguists and advancing Natural Language Processing (NLP) applications for Fulani. The proposed label set is meticulously constructed based on a dual methodological approach: it draws heavily from the well-established EAGLES (Expert Advisory Group on Language Engineering Standards) recommendations to ensure corpus reuse and cross-linguistic comparability, while simultaneously incorporating an in-depth analysis of Fulani grammatical specificities. This adaptation is crucial given the morphological richness and complex grammatical structure of Fulani, including its elaborate system of approximately 25 noun classes, unique adjective derivations, and intricate verbal conjugations. The resulting tagset comprises 15 mandatory labels and 54 recommended labels. While some EAGLES categories like "article" and "residual" are not supported, new categories such as "participle," "ideophone," "determiner," and "particle" are introduced to capture the nuances of Fulani grammar. The recommended tags further detail the mandatory categories, subdividing nouns into proper, common singular, and common plural; verbs based on voice and conjugation (infinitive active, middle, passive; conjugated active affirmative/negative, middle affirmative/negative, passive affirmative/negative); and adjectives and pronouns into more specific types based on demonstrative, possessive, subject, object, relative, emphatic, interrogative, and indefinite functions. Participles are divided into singular and plural, adverbs into time, place, manner, and negation, numbers into singular and plural, and determiners into singular and plural. Particles are further broken down into dicto-modal, abdominal, interrogative, emphatic, postposed, and postposed negative. The categories of preposition, conjunction, interjection, unique, punctuation, and ideophone remain indivisible. This meticulously defined tag set was utilized to manually annotate 5,186 words from Dominique Noye's Fulfulde-French dictionary, creating a valuable, publicly accessible resource for linguistic research and NLP development. Furthermore, the paper outlines a robust workflow for automatic morphosyntactic tagging of Fulfulde sentences, leveraging a Hidden Markov Model (HMM) in conjunction with the Viterbi algorithm. This approach, which extracts transition and emission probabilities from the annotated corpus, enables the disambiguation of morphosyntactic categories within context, considering the specific syntactic and lexical patterns of the Adamawa dialect. Ultimately, this work significantly contributes to the digitization and standardization of the Fulani language, enhancing the performance of linguistic tools and fostering its

[*]Corresponding author: piusday@gmail.com (Dayang Paul)

Received: 4 June 2025; **Accepted:** 27 June 2025; **Published:** 21 July 2025

integration into digital technologies and multilingual systems.

Keywords

Fulani, Fulfulde, Adamawa, Natural Language Processing (NLP), Part-of-speech (POS) Tagging, Morphosyntactic Tagging, EAGLES, African Languages

1. Introduction

The Fulani language, an African language of the Niger-Congo family [1] and the most widely distributed in Sub-Saharan Africa, faces a lack of resources for computerization and automatic processing due to the absence of lexical or linguistic resources. In particular, the dialect of the Adamawa region is distinguished by its morphological richness and complex grammatical structure. In most states where it is spoken, Fulani is positioned as a regional language. It is thus a language scattered across several geographical zones. In Cameroon, it is geographically, and perhaps statistically, the leading African lingua franca of the country [2]. However, this language remains under-represented in digital resources and automatic analysis tools, which significantly limits advances in fields like automatic translation, information retrieval, and other Natural Language Processing (NLP) applications. Several NLP tasks exist for building applications. One of the most important tasks is morphosyntactic tagging, also known as Part-of-Speech (POS) tagging, which attempts to assign a label to each word in a sentence, indicating its function or grammatical category (Proper noun, adjective, determiner, etc.) to address this issue for the Fulani language of the Adamawa dialectal area, we propose the creation of a morphosyntactic tag set [3]. This tagset is based on a dual methodological approach: on the one hand, it is inspired by the standards established by the EAGLES project (Expert Advisory Group on Language Engineering Standards) [4], which provides recognized guidelines for the creation of linguistic resources, and on the other hand, it relies on an in-depth study and analysis of Fulani grammar. This combination allows the adaptation of universal NLP concepts to the specific morphosyntactic features of Fulani, particularly regarding verbal inflection, nominal agreement, and agglutination mechanisms. Numerous theoretical studies on the linguistic and language aspects of this language exist Stennes [5-9]; We have therefore taken advantage of the richness of these linguistic research works on Fulani to apply them to the field of Natural Language Processing (NLP), specifically to the morphosyntactic tagging of this language. The overall objective of our work is to help transition Fulani from an under-resourced language to a moderately resourced one. To this end, we will design a morphosyntactic tag set for Fulani, allowing us to annotate corpora in this language and develop various NLP applications. Ideally, a tag set should allow for:

1) Representing the richness of lexical information;
2) Representing the information necessary for contextual disambiguation of morphosyntactic tags;
3) Encoding the useful information for the linguistic processing for which morphosyntactic tagging has been deployed [10].

The choice of tag sets is particularly delicate. Thus, tag sets for languages have been developed by different groups. Faced with these divergent practices, which form an obstacle to the exchange and reuse of corpora, sometimes based on different conceptions of what a data tag set should be, significant efforts have been made to standardize and align them within international projects such as EAGLES and MULTEX [4].

In the goal of establishing a standard tag set for the Fulani language, ensuring its comparability in a multilingual system, we have created the morphosyntactic tag set for Fulani based on the recommendations provided by the EAGLES group. This standard is very useful because it differentiates between mandatory tags, recommended tags, and special extensions. This article aims to detail the steps in the process of creating this tag set, highlighting the methodological choices from the EAGLES project, the data from grammatical studies, the challenges encountered, and the solutions provided. We will also discuss the prospects offered by this tool for developing NLP resources, particularly in applications like machine translation and text analysis.

2. Theoretical Foundations of Morphosyntactic Tagging

The development of Natural Language Processing tools for an under-resourced language like Fulani requires a deep understanding of its linguistic structures and adherence to annotation standards. Therefore, this section is dedicated to the theoretical foundations, first outlining the principles of the EAGLES project for morphosyntactic tagging, and then analyzing the fundamental properties of Fulani grammar that guided our choices.

2.1. Presentation of the EAGLES Project

For the creation of tagsets for the Fulani language of the Adamawa dialect, we draw inspiration from the recommendations of the EAGLES project [11] which aim to establish solid foundations for linguistic annotation through rigorous standards, while remaining adaptable to the specificities of different languages and diverse research objectives. These guidelines ensure quality, consistency, and interoperability of linguistic resources used in natural language processing (NLP). Founded in February 1993, the EAGLES project is an initiative of the European Commission within the”Linguistic Research and Engineering” (LRE) program. Its objective is the development of standards for language resources (written and oral corpora, electronic lexicons), as well as the development of means for structuring and exploitation, while proposing evaluation procedures for resources and tools. The EAGLES model will serve as a starting point for the creation of the Fulani language tag set. Since languages differ from one another, we cannot apply the EAGLES model as it is to the Fulani language. Some tags must be removed and others added according to the grammatical structure of sentences. Table 1 presents the list of obligatory tags defined by EAGLES.

Table 1. *List of obligatory tags from the EAGLES model.*

Number	Category
1	Noun (N)
2	Verb (V)
3	Adjective (AJ)
4	Pronoun (P)
5	Article (A)(D)
6	Unique (U)
7	Adverb (AV)
8	Adposition (AD)
9	Conjunction (C)
10	Numeral (NU)
11	Interjection (I)
12	Residual (R)
13	Punctuation (PU)

The recommendations of the EAGLES have played a key role in harmonizing morphosyntactic annotation standards. By defining a set of standardized grammatical categories and morphological features, EAGLES has ensured the interoperability of linguistic resources and optimized the development of NLP tools, particularly in automatic annotation, computer-assisted translation, and information retrieval.

2.2. Linguistic Foundations of Fulani

Fulani (or Fulfulde) is an Atlantic language of the Niger-Congo family, spoken in several countries in West and Central Africa. Its morphological richness and particular syntactic structure make it a language of interest for NLP, but also a major challenge due to the lack of dedicated resources and tools. The study of the linguistic foundations of Fulani is essential for de- signing adapted NLP tools, particularly in morphosyntax and stemming. This section presents an analysis of the linguistic structure of Fulani, focusing on its morphological and syntactic particularities. The objective is to lay the theoretical foundations necessary for implementing an automatic processing system, integrating morphosyntactic tagging models that conform to the specificities of Fulani. We have used documents [12-16] to derive the linguistic foundations for Fulani processing.

3. Theoretical Foundations of Fulani

Having established the normative framework with the EAGLES recommendations, we now turn our attention to the Fulani language itself. This section delves into the linguistic foundations of Fulani, highlighting its unique characteristics that directly influence the design of our tag set.

3.1. Noun Classes

Fulani uses an elaborate system of noun classes, similar to grammatical genders in other languages but much more complex. There are approximately 25 noun classes, marked by prefixes or suffixes indicating number (singular/plural) and sometimes semantic features (humans, objects, animals, abstract concepts).

Table 2. *Examples of noun classes in Fulani.*

Singular	Plural	Example (Meaning)
gó-	ɗi-	góru (man) → ɗi-góru (men)
ya-	ɓe-	yáaɓe (chief) → ɓe-yáaɓe (chiefs)
ma-	ñi-	máaco (person) → ñi-máaco (people)
bu-	ji-	buɗɗo (sorcerer) → ji-buɗɗo (sorcerers)

Noun classes influence morphological agreement with adjectives, pronouns, and verbs, which poses a challenge for NLP.

3.2. Adjectives

Adjectives are words that provide details and descriptions to nouns, thereby enriching the language and allowing the expression of relationships such as possession, proximity, or distance. They are di- vided into several categories, including possessive adjectives, proximal demonstratives, distal demonstratives, anaphoric demonstratives, and deictic demonstratives. Fulani has a limited number of independent adjectives, with most qualities expressed through stative verbs or derived forms.

1. Base adjectives (Independent): ndiyam (”clear”), neɗɗo (”red”).
2. Adjectives Derived from Stative Verbs:

1) ŋarii → ŋari (”to be tall” → ”tall”)
2) ɓalii → ɓali (”to be good” → ”good”)

3. Adjectives derived by suffixation:

ɗam (”belonging to”) → fulɓeɗam (”Fulani”)

Table 3. Example of verb with prefix and suffix.

Mi mbaɗi (”I left”)	O mbaɗi (”He/She left”)	Min mbaɗen (”We left”)

Adjectives must agree with the noun class of the noun they modify:

1) g ɔ́ru balɗo (good man)
2) ɗi-g ɔ́ru balɓe (good men)

This agreement system is fundamental for morphosyntactic anno- tation in NLP.

3.3. Verbs

The verb is a variable word that can take different forms depending on the intended meaning. Like nouns, pronouns, and adjectives, it changes according to number (singular or plural). However, it also varies according to person, tense, mood, and voice.

Structure of verbs in Fulani

A verb in Fulani consists of a root, a prefix or suffix indicating the tense and aspect, and a subject marker, sometimes obligatory. An example of a verb with suffix and prefix is presented in Table 3.

Fulani distinguishes several tenses and aspects, influenced by affixes as shown in Table 4.

Table 4. Examples of tense and aspect in Fulani verbs.

Tense	Example	Meaning
Present Past	Mi ñaamii Mi ñaami	I eat I ate
Future	Mi ñaamata	I will eat

The verbal system is essential for NLP applications, particularly in machine translation and corpus annotation.

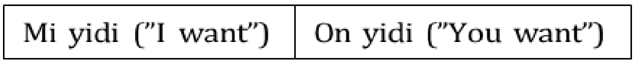
3.4. Pronouns

Pronouns are words that replace a noun or noun phrase in order to avoid repetition and make sentences more fluid. They can perform various grammatical functions, such as those of subject, object, or emphasis. Fulani has several categories of pronouns, depending on their function.

1. Personal Subject Pronouns

Table 5. Personal subject pronouns in Fulani.

Person	Singular	Plural
1st	mi	min
2nd 3rd	a o	on ɓe



Mi yidi (”I want”)	On yidi (”You want”)

Figure 1. Examples of personal subject pronouns.

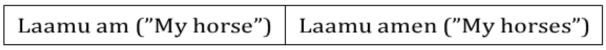
2. Possessive Pronouns

Possessive pronouns vary according to the possessor and the possessed.

Examples of possessive pronouns, see *Figure 2.*

Table 6. Possessive pronouns in Fulani.

Person	Singular (my)	Plural (my)
1st 2nd	am ma	amen onen
3rd	makko	beyen



Laamu am (”My horse”)	Laamu amen (”My horses”)

Figure 2. Examples of possessive pronouns in Fulani.

3.5. Adverbs

Adverbs are invariable words that qualitatively or quantitatively modify the meaning of a verb, an adjective, another adverb, or a noun, or that serve to affirm, negate, or ask a question. Adverbs in Fulani indicate time, place, or manner. Examples of adverbs are:

1) Jooni (”now”)
2) Heen (”there”)
3) Suka (”quickly”)

3.6. Derivations

A derived word is formed by adding one or more affixes (prefixes or suffixes) attached to a lexical morpheme called a base. The ultimate, minimal base is called the root. Derivation is the process of forming different verbs from roots to express specific modalities of action. While French uses prefixes such as're-' and 'pre-', Fulfulde uses derivational suffixes, which are added after the root to form a derived verb. This derived verb takes the same ending as simple verbs. With these derivational suffixes, it is possible to create or form various nouns, places, instruments, agents, and so on. Fulani allows the derivation of nouns, adjectives, or verbs by adding affixes.

Examples of derivation are:

1) Verb → Noun: ɗofti ("to advise") → ɗoftugol ("advice")
2) Noun → Adjective: fulɓe ("Fulani people") → fulɓeɗam ("Fulani")

Derivation is important for the automatic recognition of lexical forms.

3.7. Ideophones

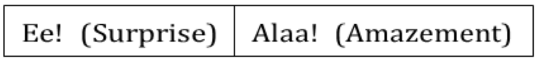
An ideophone is a word that vividly depicts sensory imagery or sensations, evoking ideas of action, sound, movement, color, or shape. They are often phonologically distinguishable from other words in a language, sometimes featuring unusual sound patterns, repetition (reduplication), or unique intonation.

Examples:

1) ɗumɗum (dull sound)
2) ŋguruŋguru (trembling)

3.8. Interjections

These are invariable words that can be used on their own to ex- press the speaker's emotional attitude or sudden feelings. They are typically short, often grammatically independent from the rest of a sentence, and frequently convey surprise, joy, pain, disgust, or other strong reactions. These words are known as interjections. Figure 3 shows some examples of interjections:



Ee! (Surprise)	Alaa! (Amazement)

Figure 3. Examples of interjections in Fulani.

3.9. Conjunctions and Prepositions

Conjunctions are terms used to join two words or groups of words. Following are some examples: Ko ("that") and Konndaa ("but").

Prepositions are grammatical words used to introduce a complement, indicating the relationship between the complement and the word it completes. Some examples of prepositions in Adamawa Fulani are: e ("with") and ko ("like"). This detailed analysis of Fulani grammatical categories:

1) Including noun classes, adjectives, verbs, pronouns, adverbs, derivations, ideophones, interjections, conjunctions, and prepositions.
2) Forms an essential foundation for the creation of a morphosyntactic tag set tailored to Fulani. By precisely identifying the various classes and their morphological interactions, we can design a robust annotation system that will facilitate the development of automatic processing tools for Fulani, particularly for morphosyntactic tagging, machine translation, and information retrieval.

4. Development of the Morphosyntactic Tag Set for Fulani

4.1. Methodology for Design

A tag can contain the following information: a grammatical category (mandatory), e.g., verb, noun, adjective, adverb, determiner, etc.; inflectional information, e.g., gender, number, person, tense, mood, etc.; and morphosyntactic information, e.g., distinction between proper nouns and common nouns. The mandatory tags include the parts of speech of a language, represented by word classes that share the same semantic and grammatical properties. Considering the specifications of each class in Fulani, the tagset we propose includes fifteen categories, unlike the EAGLES model, which proposes thirteen. The categories "article" and "residual" from the EAGLES model are not supported by Fulani, but it introduces the categories "participle," "ideophone," "determiner," and "particle." Table 7 provides the list of mandatory tags for the Adamawa dialect of Fulani.

Table 7. Proposed set of tags for Adamawa Fulani.

No.	Categories	No.	Categories
1	Noun (N)	9	Number (NU)
2	Verb (V)	10	Interjection (I)
3	Adjective (AJ)	11	Unique (U)
4	Pronoun (PR)	12	Ideophone (ID)
5	Participle (PP)	13	Punctuation (PU)
6	Adverb (AV)	14	Determiner (D)
7	Preposition (P)	15	Particle (PA)
8	Conjunction (C)		

4.2. Presentation of the Tagset

Table 7 provides the list of mandatory tags for the Adamawa dialect of Fulani. To define the recommended tags, we will detail the mandatory tags using more specific tags related to morphological changes in word forms and semantic properties.

1) Noun: Nouns can be common or proper, with common nouns varying in singular or plural number.
2) Verb: Verbs are of two types: conjugated verbs and infinitives, all represented by active, passive, and middle voice markers.
3) Adjective: Adjectives are of demonstrative and possessive types, but in Fulani, demonstrative adjectives are divided into four groups: deictic demonstratives, proximal demonstratives, distal demonstratives, and anaphoric demonstratives.
4) Pronoun: A pronoun is a word that generally replaces an object or fact previously mentioned, fulfilling the grammatical role of what it replaces.
5) Participle: Participles are verb forms that confer adjectival properties to verbs.
6) Adverb: Adverbs are of three types: time, place, and manner. They are invariable words that modify the meaning of a verb, adjective, another adverb, or a noun, or serve to affirm, negate, or interrogate.
7) Preposition: Prepositions belong to the general category of relational words. They express various semantic values, including spatiotemporal location, instrument, direction, possession, belonging, and accompaniment.
8) Conjunction: Terms that serve to join two words or groups of words. They can mark subordination or coordination.
9) Number: A word that represents a number or rank.
10) Interjection: A word or sound that expresses emotion, annoyance, or surprise. These words are generally autonomous, without syntactic relation to other words in the same sentence.
11) Unique: The unique value is assigned to terms that do not fit into usual categories, such as foreign words, mathematical formulas, and symbols. Even if these terms are not part of the lexicon of the language being processed, they occur frequently and therefore need to be tagged.
12) Ideophone: Terms that aim to convey a sensation, such as a smell, color, shape, sound, or movement.
13) Determiner: Noun classes that allow the categorization of words.
14) Particle: Often short words used to express grammatical nuances, spatial or temporal relations, or to indicate changes in meaning within a sentence.
15) Punctuation: Graphic signs used to organize written text.

Once the mandatory tags have been identified and extracted, Table 8 presents the list of recommended morphosyntactic tags for the Adamawa dialect of Fulani. We have expanded from 15 tags to 54 tags because several tags have been presented in detail.

1) The noun tag is subdivided into 3 subgroups: proper noun, common singular noun, and common plural noun.
2) The verb tag includes 9 tags: infinitive active voice, infinitive middle voice, infinitive passive voice, conjugated active affirmative, conjugated active negative, conjugated middle affirmative, conjugated middle negative, conjugated passive affirmative, conjugated passive negative.
3) The adjective tag includes 10 tags: possessive singular adjective, possessive plural adjective, proximal demonstrative sin- gular adjective, proximal demonstrative plural adjective, distal demonstrative singular adjective, distal demonstrative plural adjective, anaphoric demonstrative singular adjective, anaphoric demonstrative plural adjective, deictic demonstrative adjective, interrogative adjective.
4) The pronoun tag includes 10 tags: singular subject personal pronoun, plural subject personal pronoun, singular object personal pronoun, plural object personal pronoun, singular relative pronoun, plural relative pronoun, singular emphatic pronoun, plural emphatic pronoun, interrogative pronoun, indefinite pro- noun.
5) The participle tag includes 2 tags: singular participle, plural participle.
6) The adverb tag includes 4 tags: time adverb, place adverb, manner adverb, negation adverb.
7) The number tag is subdivided into 2 groups: singular number, plural number.
8) The determiner tag includes 2 tags: singular determiner, plural determiner.
9) The particle tag includes 6 tags: dicto-modal particle, abdominal particle, interrogative particle, emphatic particle, postposed particle, and postposed negative particle.
10) The 6 tags: preposition, conjunction, interjection, unique, punctuation, and ideophone remain indivisible.

After highlighting the mandatory and recommended tags for Fulani, Table 8 presents the proposed tagset for this language.

Table 8. List of Recommended Tags for Fulani.

Number	Title	Symbol
1	Proper Noun	NP
2	Common Singular Noun	NCS
3	Common Plural Noun	NCP
4	Infinitive Active Voice Verb	VIA
5	Infinitive Middle Voice Verb	VIM

Number	Title	Symbol
6	Infinitive Passive Voice Verb	VIP
7	Conjugated Active Affirmative Verb	VCAA
8	Conjugated Active Negative Verb	VCAN
9	Conjugated Middle Affirmative Verb	VCMA
10	Conjugated Middle Negative Verb	VCMN
11	Conjugated Passive Affirmative Verb	VCPA
12	Conjugated Passive Negative Verb	VCPN
13	Possessive Singular Adjective	AJPS
14	Possessive Plural Adjective	AJPP
15	Proximal Demonstrative Singular Adjective	AJDPS
16	Proximal Demonstrative Plural Adjective	AJDPP
17	Distal Demonstrative Singular Adjective	AJDES
18	Distal Demonstrative Plural Adjective	AJDEP
19	Anaphoric Demonstrative Singular Adjective	AJDAS
20	Anaphoric Demonstrative Plural Adjective	AJDAP
21	Deictic Demonstrative Adjective	AJDD
22	Singular Subject Personal Pronoun	PRPSS
23	Plural Subject Personal Pronoun	PRPSP
24	Singular Object Personal Pronoun	PRPCS
25	Plural Object Personal Pronoun	PRPCP
26	Singular Relative Pronoun	PRRS
27	Plural Relative Pronoun	PRRP
28	Singular Emphatic Pronoun	PRIS
29	Plural Emphatic Pronoun	PRIP
30	Singular Participle	PPS
31	Plural Participle	PPP
32	Time Adverb	AVT
33	Place Adverb	AVL
34	Manner Adverb	AVM
35	Preposition	P
36	Conjunction	C
37	Interjection	I
38	Unique	U
39	Punctuation	PU
40	Ideophone	ID
41	Singular Number	NUS
42	Plural Number	NUP
43	Singular Determiner	DS
44	Plural Determiner	DP
45	Dicto-Modal Particle	PAD
46	Abdominal Particle	PAA
47	Interrogative Particle	PAI
48	Emphatic Particle	PAIN
49	Postposed Particle	PAPPA
50	Interrogative Adjective	AJI
51	Interrogative Pronoun	PRI
52	Negation Adverb	AVN
53	Indefinite Pronoun	PRIN
54	Postposed Negative Particle	PANP

4.3. Comparison with Other Existing Tagsets (French)

To better situate the Fulani tagset within a broader linguistic framework, we compare it to the morphosyntactic categories of French based on the Universal Dependencies (UPOS) standard. This comparison allows us to identify direct correspondences, divergences, and specificities unique to each language, facilitating the integration of Fulani into multilingual NLP tools.

5. Application and Perspectives

Beyond its design, the true value of the tagset lies in its practical application and its potential to transform the NLP landscape for Fulani. This section explores its use for corpus annotation and its implications for future applications in automatic language processing.

5.1. Use of the Tagset in Annotating Fulani Corpora

After proposing a tagset for the Fulani language, comprising 54 tags, we proceeded to manually annotate 5,186 words from Dominique Noye's Fulfulde-French [17]. This work constitutes a valuable resource for research in linguistics and Natural Language Processing (NLP). In a spirit of transparency and knowledge sharing, we have made this resource available in the form of a tagged dictionary, accessible via GitHub. This dictionary can be used for various applications, including the training and evaluation of NLP tools for Fulani, such as morphosyntactic parsers, machine translators, or information retrieval systems. Furthermore, this resource can serve as a reference for enriching annotated Fulani corpora, facilitating the creation of more accurate linguistic models. We hope that it will encourage collaborative contributions and stimulate further re- search on the structuring and analysis of this language. The tagged dictionary is accessible via the following link: https://github.com/zouleihaalhadji/dicopeulh/blob/main/DICTIONNAIRE%20PEUL%20ETIQUETE.csv.

5.2. Workflow and Algorithms for HMM-based Tagging of Fulfulde Sentences

In this section, we present the method adopted for the automatic morphosyntactic tagging of Fulfulde sentences, based on a Hidden Markov Model (HMM) combined with the Viterbi algorithm. The tagging process is organized into a structured workflow that starts with the design of an appropriate tagset for the Fulfulde language, followed by the manual annotation of a training corpus. This corpus is used to extract the probabilistic parameters of the HMM, including transition and emission probabilities. Once the model is trained, input sentences are tokenized and analyzed using the Viterbi algorithm, which computes the most probable sequence of tags for the given sentence. The entire process is represented in the flowchart below, which illustrates each step from tagset creation to the generation of automatically tagged sentences. This approach allows for the disambiguation of morphosyntactic categories in context, taking into account the syntactic and lexical patterns specific to the Fulfulde language of the Adamawa dialectal area. The algorithms and work- flow described here constitute the foundation of the tagging system developed in this study.

The automatic morphosyntactic tagging process of Fulfulde sentences using a Hidden Markov Model (HMM) and the Viterbi algorithm follows a series of structured steps described below and as depicted in Figure 4:

1. Tagset Design: A tagset is developed based on the grammatical structure of Fulfulde, especially from the Adamawa dialectal area. It includes categories such as nouns, verbs, pronouns, determiners, and conjunctions, and is inspired by the EAGLES recommendations to ensure interoperability and standardization. This tagset will serve as the reference for both manual annotation and automatic tagging.
2. Annotated Corpus Creation: A Fulfulde corpus is manually an- notated using the tagset. Each word is assigned a corresponding morphosyntactic tag. This annotated data is crucial for training the statistical tagging model and serves as the basis for extracting the probabilistic parameters of the HMM.
3. Extraction of HMM Probabilities: From the annotated corpus, two types of probabilities are computed: Transition probabilities: the likelihood of a given tag following another tag in a sentence (e.g., P(VERB | PRON)). Emission probabilities: the likelihood of a word being associated with a given tag (e.g., P(yahii | VERB)). These probabilities are estimated using frequency counts from the corpus and are fundamental to the tagging model.
4. Tokenization of the Input Sentence: The sentence to be tagged is first segmented into individual words (tokens). For example:

Miɗo yahii gese → [”Miɗo”, ”yahii”, ”gese”]

Tokenization allows the model to process each word independently while considering the sequence context during tagging.

5. Application of the Viterbi Algorithm: The Viterbi algorithm is applied to the tokenized sentence. It uses the previously computed transition and emission probabilities to evaluate all possible tag sequences and determines the one with the highest over- all probability. The algorithm builds a probability matrix (trellis) and uses dynamic programming to optimize the process.
6. Selection of the Most Probable Tag Sequence: After computing all possible paths, the algorithm selects the most probable sequence of tags that best fits the input sentence. This sequence corresponds to the maximum-likelihood estimate of the true tag sequence under the HMM.
7. Output - Automatically Tagged Sentence: The final result is a version of the sentence where each word is labeled with its predicted morphosyntactic tag.

For example: Miɗo/PRON yahi- i/VERB gese/NOUN

This tagged output can be used for various downstream tasks such as syntactic analysis, machine translation, or automatic summarization.

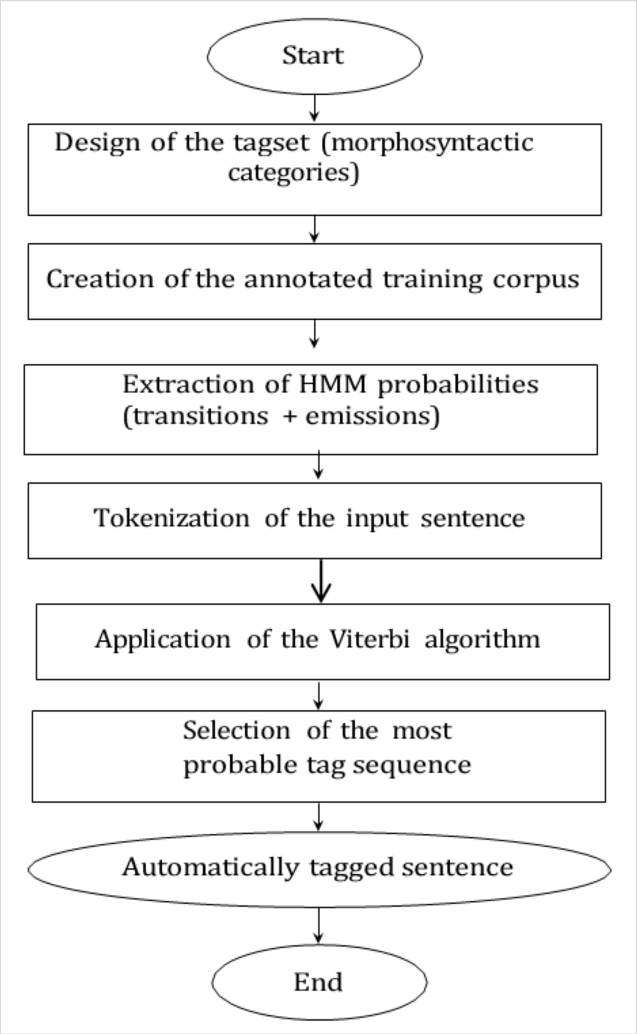

Start
Design of the tagset (morphosyntactic categories)
Creation of the annotated training corpus
Extraction of HMM probabilities (transitions + emissions)
Tokenization of the input sentence
Application of the Viterbi algorithm
Selection of the most probable tag sequence
Automatically tagged sentence
End

Figure 4. *Flowchart of the HMM-Based Tagging Process for Fulfulde.*

The Tagging Model Based on HMM and the Viterbi Algorithm: In the field of natural language processing, morphosyntactic tagging plays a crucial role in enabling machines to understand the grammatical structure of human language. For under-resourced languages like Fulfulde, the development of reliable tagging models is essential for building more advanced language technologies such as translation systems, syntactic parsers, and information retrieval tools.

This part presents the statistical model we implemented for the automatic tagging of Fulfulde sentences. The approach is based on a Hidden Markov Model (HMM), a probabilistic framework that models the sequential nature of linguistic data by associating hid- den states (tags) with observable outputs (words). To infer the most probable sequence of tags given a sentence, we use the Viterbi algorithm, a dynamic programming method that efficiently finds the optimal path through the model's state space.

We describe here the theoretical foundations of the HMM, the estimation of model parameters from an annotated corpus, and the functioning of the Viterbi decoding process. This model constitutes the core of the tagging architecture developed in this study and is well suited to the morphosyntactic characteristics of the Fulfulde language.

A Hidden Markov Model (HMM) is a probabilistic model used to represent systems where the sequence of underlying states is hidden, but the observable outputs generated by those states are known. In the context of morphosyntactic tagging, HMM assumes that:

1) Each word in a sentence is generated by a hidden tag (e.g., NOUN, VERB, PRONOUN);
2) The sequence of tags follows a Markov process, meaning each tag depends only on the previous tag;
3) Each tag emits words according to a probability distribution.

The goal of using HMM is to define the statistical model and to find the most probable sequence of hidden states given an observation sequence. This amounts to solving:

$$T\hat{} = \frac{\arg\max P\ (T|W)}{T} \qquad (1)$$

Where $T\hat{}$ is the most probable tag sequence, T represents a tag sequence, and W represents the word sequence.

The Viterbi algorithm is a dynamic programming method widely used for decoding the most likely sequence of hidden states in Hid- den Markov Models (HMMs). Given a sequence of observations, the algorithm efficiently finds the single best path — that is, the sequence of states that maximizes the joint probability of the observed data and the hidden states. This makes the Viterbi algorithm essential in many applications such as speech recognition, bioinformatics, and natural language processing, where uncovering hidden information from observed data is crucial. The algorithm operates through a series of systematic steps to ensure optimality and computational efficiency.

1. Initialization:

Initialize the probabilities for the first observation by calculating the probability of each state at time $t = 1$, considering the start probabilities and the likelihood of observing the first observation in each state.

2. Recursion:

For each subsequent observation $t = 2... T$, compute the maxi- mum probability of each state by considering the previous state probabilities, the transition probabilities between states, and the observation likelihoods. Store the path that gives the maxi- mum probability.

3. Termination:

After processing all observations, determine the final state with the highest probability at time T.

4. Backtracking:

Trace back through the stored paths to retrieve the most probable sequence of hidden states that led to the final state.

Algorithm Variables:

1) $\delta_t(i)$: Maximum probability of reaching state i at time t

2) $\psi_t(i)$: Pointer to the most likely previous state for state i at time t

3) P^*: Maximum probability of the most likely path

4) q^*: Most likely state at time t

Time Complexity: $O(N^2 T)$ where N is the number of states (tags) and T is the length of the observation sequence (sentence).

Space Complexity: $O(N T)$ for storing the dynamic programming tables.

Table 9*. Viterbi Algorithm for HMM-based POS Tagging.*

Algorithm

```
Require: Observation sequence $W = w_1, w_2... w_T$
Require: Tag set $S = \{s_1, s_2... s_N\}$
Require: Transition probabilities $A = \{a_{ij}\}$ where $a_{ij} = P(s_j|s_i)$
Require: Emission probabilities $B = \{b_j(w_t)\}$ where $b_j(w_t) = P(w_t|s_j)$
Require: Initial probabilities $\pi = \{\pi_i\}$ where $\pi_i = P(s_i)$
Ensure: Most probable tag sequence $T\hat{} = t_1, t_2... t_T$
```

Algorithm

```
Initialization:
for $i = 1$ to $N$ do
    $\delta_1(i) = \pi_i \cdot b_i(w_1)$
    $\psi_1(i) = 0$
end for
Recursion:
for $t = 2$ to $T$ do
    for $j = 1$ to $N$ do
        $\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) \cdot a_{ij}] \cdot b_j(w_t)$
        $\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) \cdot a_{ij}]$
    end for
end for
Termination:
$P^* = \max_{1 \le i \le N} \delta_T(i)$
$q^*_T = \arg\max_{1 \le i \le N} \delta_T(i)$
Backtracking:
for $t = T - 1$ down to 1 do
    $q^*_t = \psi_{t+1}(q^*_{t+1})$
end for
return Tag sequence $T\hat{} = q^*_1, q^*_2, ..., q^*_T$
```

5.3. Impact and Application Perspectives in NLP

The development of a tagset adapted to Fulani represents a major advancement for Natural Language Processing (NLP), particularly for under-resourced languages. Its impact and contributions manifest at several levels:

1. Improvement of Linguistic Annotation for Fulani:

1) Provides a standardized structure for the morphosyntactic annotation of Fulani corpora.

2) Enables automatic disambiguation of grammatical categories, improving the quality of annotated corpora.

3) Facilitates interoperability with other tagsets such as Universal Dependencies (UPOS).

2. Development of Linguistic Tools for Fulani:

1) Morphosyntactic Taggers: A well-defined tagset allows for the training of robust annotation models.

2) Syntactic Parsers: The identification of grammatical structures becomes more precise, facilitating applications such as phrase recognition and dependency parsing.

3) Enriched Lexicons: Annotation allows for the association of detailed grammatical information with entries in digital dictionaries.

3. Impact on NLP Applications:

1) Machine Translation: Enriches syntactic modeling and improves alignment quality between Fulani and other languages.
2) Information Retrieval: Optimizes search engines and question- answering systems in Fulani through better query analysis.
3) Grammatical Correction: Enables the detection and correction of morphosyntactic errors in Fulani texts.
4) Automatic Summarization and Synthesis: Improves segmentation and extraction of essential information from texts.

4. Promotion and Preservation of the Fulani Language:

1) Contributes to the digitization and standardization of the language.
2) Promotes the inclusion of Fulani in linguistic technologies, reducing the digital divide for native speakers.
3) Encourages the creation of open linguistic resources, accessible to researchers and NLP developers.

6. Conclusion

This work represents a fundamental and crucial step in the advancement of Natural Language Processing (NLP) for the Fulani language, specifically for the Adamawa dialect. By defining a comprehensive and structured set of morphosyntactic labels, inspired by EAGLES recommendations while being rigorously adapted to the grammatical specificities of Fulani, we have laid the necessary foundation for the construction of high-quality annotated corpora. The development of this tagset, comprising 15 mandatory and 54 recommended categories, has enabled the manual annotation of an initial corpus of 5,186 words, thereby making a valuable resource available to the scientific community. This corpus, combined with the implementation of a workflow based on the Hidden Markov Model (HMM) and the Viterbi algorithm, constitutes an effective preliminary tool for morphosyntactic tagging, offering invaluable assistance to linguists and a basis for future Fulani applications in the digital domain.

Beyond the advancements presented, a promising future research direction would involve exploring the application of state-of-the-art machine learning models and deep neural networks for Fulani morphosyntactic tagging. While the Hidden Markov Model (HMM) and the Viterbi algorithm provide a solid and effective foundation, modern architectures such as Recurrent Neural Networks (RNNs), notably LSTMs (Long Short-Term Memory) and GRUs (Gated Recurrent Units), as well as Transformer-based models, have demonstrated superior performance in similar tasks for other languages. These approaches could potentially better capture long-term dependencies and complex contextual nuances inherent in Fulani's rich morphology and syntax. Integrating transfer learning methods, by leveraging pre-trained models from more resource-rich languages, could also accelerate the development of high-precision tools for Fulani, despite the still limited size of available annotated corpora. This would pave the way for even more robust and adaptive Fulani language processing systems, essential for applications such as neural machine translation, speech recognition, and semantic analysis.

In summary, this work significantly contributes to the digitization and standardization of the Fulani language. The tools and resources developed here not only improve the performance of linguistic instruments but also ensure a better integration of Fulani into digital technologies and multilingual systems, thereby paving the way for a stronger and more relevant presence of this language in the global digital landscape.

Author Contributions

Zouleiha Alhadji Ibrahima: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Writing – original draft, Writing – review & editing

Charles Moudina Varmantchaonala: Investigation, Methodology, Supervision, Visualization, Writing – review & editing

Dayang Paul: Conceptualization, Project administration, Supervision, Validation, Visualization

Kolyang: Supervision, Validation, Visualization

Conflicts of Interest

The authors declare no conflicts of interest

Appendix

Table A1. Comparative Table of Mandatory Tags for Fulani and French.

Fulani (15 Tags)	French (UPOS)	Comments and Equivalences
Noun (N)	NOUN	Direct correspondence: common and proper nouns.
Verb (V)	VERB	Direct correspondence with lexical verbs.

Fulani (15 Tags)	French (UPOS)	Comments and Equivalences
Adjective (AJ)	ADJ	Qualifying adjectives in French.
Pronoun (PR)	PRON	Includes personal, possessive, demonstrative pronouns.
Participle (PP)	VERB/ADJ	French participles function as verbs or adjectives.
Adverb (AV)	ADV	Direct correspondence.
Preposition (P)	ADP	Includes prepositions and postpositions.
Conjunction (C)	CCONJ/SCONJ	Coordinating and Subordinating conjunctions.
Number (NU)	NUM	Direct correspondence with numbers.
Interjection (I)	INTJ	Direct correspondence.
Unique (U)	X	Not classified in UPOS.
Punctuation (PU)	PUNCT	Direct correspondence.
Determiner (D)	DET	Articles and demonstrative, possessive determiners.
Particle (PA)	PART	Invariable grammatical words.

References

[1] Seydou, C. (2021). Language and identity. Legends of the origin of the Fulani. Oral Literature Notebooks, (pp. 79–99). URL: https://doi.org/ 10.4000/clo.4714

[2] Seck, F. (2018). Putting Language First: An Interview with Boubacar Boris Diop. African Literary Studies, *46*, 91–105.

[3] Stat4decision (2021). French Natural Language Processing (TAL/NLP). URL: https://www.stat4decision.com/fr/traitement-langage-naturel-francais-tal-nlp consulted on January 5, 2021.

[4] Calzolari, N., & Monachini, M. (1996). Multext- common specifications and notation for lexicon encoding. *Rapport interne*.

[5] Stennes, L. H. (1967). A reference grammar of adamawa fulani.

[6] Maiga, A. et al. (2009). Fulfulde/Pulaar-French bi-grammar.

[7] Taylor, F. W. (1953a). A grammar of the adamawa dialect of the fulani language (fulfulde).

[8] Mohamadou, A. (1994). Classifiers and representations of lexical properties in Fulani: Aadamaawa dialects.

[9] Leith-Ross, S. (1925b). *Fulani Grammar: By Sylvia Leith-Ross*. Humphrey Milford Oxford University Press [Verlag bezeichnung].

[10] Allah, F. A., Boulaknadel, S., & Souifi, H. (2014). Morpho-syntactic tag game of the Amazigh language. *Asinag*, (pp. 171–184).

[11] Calzolari, N., & McNaught, J. (1996). Eagles-expert advisory group on language engineering standards. *URL: http://www.ilc.pi.cnr.it/EAGLES/home.html*

[12] Stennes, L. H. (1967b). *A Reference Grammar of Adamawa Fu- lani*. The University of Virginia, African Studies Center, Michigan State University.

[13] Maiga, A. (2009). Fulfulde/Pulaar-French dual grammar. Department of Education and Training: French language learning program in a multilingual context.

[14] Taylor, W. F. (1953b). *A Grammar of the Adamawa Dialect of the Fulani Language (Fulfulde)*. Clarendon Press.

[15] Mohamadou, A. (1991). Classifiers and representation of lexical properties in Fulani: dialects of Aadamaawa.

[16] Leith-Ross, S. (1925a). *Fulani Grammar*. Oxford University Press.

[17] Noye, D. (1989). Fulani-French Dictionary. Fulani dialect of Diamaré, North Cameroon.