International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Hybrid, Theory-Driven Transcription System for the Study of Code-Switching

Mihiretu Wakwoya
PhD Candidate, Department of Applied linguistics, School of Multilingualism, University of Pannonia, Veszprem, Hungary
Corresponding Author: Mihiretu Wakwoya**, E-mail**: mihiretu2005@gmail.com

| ABSTRACT

Transcription plays a critical role in linguistic research, especially in studies of bilingual speech and code-switching. This paper introduces a custom-built transcription protocol designed to meet the theoretical and practical demands of analysing code-switching between typologically distinct languages—Afaan Oromoo and English—under the Matrix Language Frame (MLF) and 4-M models. The proposed system incorporates morpheme-level labelling, language-specific glossing, and clause segmentation in a layered and adaptable framework, drawing on principles from existing systems such as CHAT, ELAN, and the Leipzig Glossing Rules. The integrated system facilitates the identification of matrix and embedded languages while simultaneously capturing essential morphosyntactic features. It enables the empirical monitoring of code-switching phenomena through spreadsheet-based structuring, thereby supporting both qualitative and quantitative analyses. This transcription method improves current practices by connecting with theoretical ideas, addressing complex language issues, and offering a model that can be repeated for less-studied language pairs. The approach holds promise for future expansion, automation, and adaptation across diverse multilingual contexts.

| KEYWORDS

Bilingual, code-switching, morphology, matrix language, transcription

| ARTICLE INFORMATION

ACCEPTED: 01 August 2025 **PUBLISHED:** 30 August 2025 **DOI:** 10.32996/ijllt.2025.8.9.12

1. Introduction

In linguistic research—especially in areas like sociolinguistics, corpus linguistics, and bilingual code-switching analysis—transcription is not just a way to write things down; it is a key part of the research process ([1]; [2]). Transcription serves as a crucial bridge between ephemeral speech and enduring analysis in qualitative research, especially in studies involving spoken language, interaction, or narrative. As such, the choice of a transcription system is far from a neutral or mechanical task. Instead, it is a methodological choice that shapes the representation, interpretation, and ultimate understanding of data. Research in areas like conversation analysis (CA), sociolinguistics, and ethnography have shown that transcription is not just a simple process of turning speech into text; it involves interpretation influenced by different theories, language choices, and analysis goals (e.g., [2]; [3]; [4]; [5]).

This study explores bilingual code-switching between Afaan Oromoo and English using the Matrix Language Frame (MLF) and its extension, the 4-M model, as its central analytical frameworks. The MLF model, developed by Myers-Scotton, posits that in bilingual utterances, one language—the Matrix Language (ML)—provides the grammatical structure of the clause, including system morphemes such as tense, agreement markers, and function words([6];[7]). The other language—the Embedded Language (EL)—contributes lexical items or phrases that are inserted into this grammatical frame. The 4-M model by Myers-Scotton & Jakerefined on this by categorizing morphemes into content morphemes, early system morphemes, late system morphemes, and outsider system morphemes, allowing for a more nuanced analysis of how elements from each language participate in clause formation([8]).

The central research question addressed in this study is, how is the matrix language (ML) and embedded language (EL) distribution identified in code-switching between Afaan Oromoo and English? To answer this question within the MLF/4-M theoretical framework, a transcription system was required that could support fine-grained morphosyntactic analysis. The transcription had to clearly show where sentences start and end, break down words into their smallest parts, label them with the right grammar terms, and importantly, show which language each part comes from. This level of linguistic detail was essential to determine which language functions as the grammatical host in each clause and how lexical insertions from the other language are structurally integrated.

Therefore, the transcription method adopted for this study prioritized clause-level segmentation, manual morpheme glossing using standardized grammatical labels, and consistent language tagging. This method is different from transcription systems that focus on sounds, rhythm, or conversation because it was specifically created to match the structural ideas of MLF and 4-M models. It offers the detailed analysis needed to observe language dominance, the order of morphemes, and how grammatical functions are used in different languages, allowing for a structured and theory-based study of bilingual speech.

Moreover, the choice of transcription system has implications beyond data analysis. It affects how findings are communicated to others, including other researchers, practitioners, and participants themselves. Transparency, consistency, and accessibility are all concerns that hinge on transcription practices. Inaccurate or inappropriate transcription can introduce bias, misrepresent speakers, or obscure important features of the discourse ([9]; [2]).

To meet this need, Deuchar et al. (2018) emphasize the importance of using detailed and structured transcription tools that go beyond verbatim representation to incorporate morphological, language-origin, and pragmatic information. One widely recognized system for such transcription is the CHILDES (Child Language Data Exchange System), a comprehensive database and toolkit developed for the study of child language acquisition and bilingualism ([10];[11]). Transcriptions in CHILDES follow the CHAT (Codes for the Human Analysis of Transcripts) format, which is a set of rules that allows researchers to add detailed information about grammar, word choice, and context to spoken data. A CHILDES-compatible format ensures that transcripts can be processed using specialised software such as CLAN, facilitating the systematic coding of each word's grammatical function and language of origin (see [12]). This approach is particularly effective for analysing bilingual speech, as it reveals the grammatical structure of mixed-language clauses and supports theory-driven analysis. Consequently, transcription systems must be designed to reflect the structural realities of bilingual speech to yield valid and reliable theoretical conclusions.

Studies of language contact involving typologically distinct languages—such as Afaan Oromoo and English—require particular attention to structural and grammatical features in transcription and analysis. Afaan Oromoo, a Cushitic language, is typologically distinct from English in several keyways. It is an agglutinative language characterised by rich verbal inflection, extensive suffix stacking, and a Subject-Object-Verb (S)OV) word order ([13]). Grammatical functions such as tense, aspect, mood, agreement, and case are marked morphologically, often through a series of affixed morphemes ([14]). In contrast, English is an analytic language with comparatively minimal inflection, a reliance on word order (typically SVO), and the use of function words rather than bound morphology to express grammatical relationships.

These typological differences point out the need for a transcription system capable of capturing such structural contrasts. A CHILDES-compatible system, such as CHAT, enables researchers to annotate not just lexical items but also the morphological and syntactic features that are crucial for bilingual data analysis. For example, a transcription system made for this type of data needs to show verbs at the end of clauses, agreement markers, and case markings—important parts of the Matrix Language Frame (MLF) model. To support valid cross-linguistic analysis, the system must also be sufficiently flexible and robust to accommodate languages with diverse morphosyntactic profiles.

I argues that I should take into account grammatical structures under investigation when choosing, adapting, or designing the transcription system. It needs to be aware of different languages and meet the needs of the theoretical framework being used, allowing for both in-depth analysis and, when necessary, statistical analysis of the data.

2. Existing Transcription Systems

2.1 ELAN

ELAN (EUDICO Linguistic Annotator), developed by the Max Planck Institute, is a highly flexible tool for transcription and annotation of spoken and signed language data (https://archive.mpi.nl/tla/elan). It allows for multi-tiered, time-aligned annotation, making it particularly suitable for naturalistic data involving overlapping speech, gestures, or prosody ([15];[16]).

One major advantage of ELAN is its ability to facilitate the analysis of multiple modes of communication simultaneously—such as speech, gesture, and prosody—which is particularly important in research on indigenous or under-described languages, where

non-verbal cues like emphasis, pauses, or gestures may signal clause boundaries or shifts in meaning (Himmelmann, 2006). This multimodal capability makes ELAN especially suitable for research that integrates interactional or discourse-based theories of language, including models such as the Matrix Language Frame (MLF) model (Myers-Scotton, 2002), where features like intonation and prosodic breaks can assist in identifying clause boundaries and shifts in the matrix or embedded language. As a result, ELAN is increasingly used in conjunction with theoretical frameworks that emphasize naturalistic, real-time communication ([15]; [17]).

Despite its strengths, ELAN has limitations. Its reliance on Java for cross-platform compatibility results in limited media handling performance, as Java lacks high-performance native media frameworks. Earlier versions also suffered timing inaccuracies due to dependencies on QuickTime and Java Media Framework, which affected the precision of media playback—critical in multimodal research. Media format support was initially limited, and while expanded later, performance still depends on codec stability. Additionally, ELAN's search functionality is restricted, with structured searches confined to single files and multi-file support still under development ([15]). Researchers must therefore design their own glossing conventions and export data for use in external software such as Excel, Toolbox, or FLEx, especially when dealing with complex morphosyntactic phenomena in multilingual contexts. To mitigate this, researchers working with frameworks like the MLF model often pair ELAN with structured glossing systems such as the Leipzig Glossing Rules ([18];[19]), enabling them to analyse grammatical integration across languages more systematically.

In contrast, the CHAT transcription format—part of the CHILDES project—is explicitly designed to support morphosyntactic annotation, language tagging, and computational processing([10];[11]). CHAT uses a rule-governed structure to mark each word's language origin, grammatical category, and interactional function, and it is fully integrated with CLAN software for automated analysis (Deuchar et al, 2018). This makes CHAT highly effective for theory-driven research that requires detailed and replicable tracking of grammatical features in bilingual speech, such as identifying matrix language assignment, code-switching patterns, or frequency of morpheme use ([10]; [8]). In this way, ELAN and CHAT reflect different typological emphases: ELAN prioritises multimodal, interaction-sensitive flexibility, while CHAT emphasises structured morphosyntactic annotation and computational traceability. This distinction is crucial for researchers who investigate code-switching between typologically distinct languages, such as Afaan Oromoo and English. While ELAN offers unparalleled flexibility for capturing interactional and prosodic features, CHAT provides greater analytical rigor for grammatical and corpus-based analysis. So, the transcription system must match the study's theoretical models, analytical goals, and linguistic data ([20]).

2.2 Leipzig Glossing Rules

The Leipzig Glossing Rules provide a standardised set of conventions for interlinear morpheme-by-morpheme glossing, a method used to represent the internal grammatical structure of utterances from morphologically complex languages ([21]). Interlinear glossing involves aligning each word in a sentence with one or more lines of annotation that explain its morphological components—such as case, tense, aspect, negation, or verb agreement—followed by a free translation. This method allows researchers to present primary linguistic data in a transparent and analyzable form, particularly when working with understudied or typologically diverse languages.

Descriptive linguistics and typological research rely heavily on these conventions to document and compare grammatical structures across languages. They are especially valuable in field linguistics, language documentation, and syntax-morphology interface studies, as they make structural patterns explicit for both human readers and computational analysis ([21]). Furthermore, interlinear glossing plays a crucial role in multilingual and code-switching studies—such as those involving Afaan Oromoo and English—by highlighting points of grammatical integration or boundary between the matrix and embedded languages (Myers-Scotton & Jake, 2000). See example (1). Key to glosses: 1/2/3PL, First/second/third Person Plural; 1/2/3SL, First/second/third Person Singular; POSS, possessive pronoun; DET, determiner; ACCO, Accusative; Q, Question mark; ART, Article; NEG, negative/negative particle; IMPV, imperfect verb; IMP, Imperative verb; PRV, Perfective Verb; CNV, Converb; FOC, Focus marker; COP, copula; NOM, Nominative marker; LOC, Locative marker; M, male; F, female; Pass, passive marker. Additionally, Afaan Oromoo words appear in standard font, English words in bold with @eng, and Amharic words in italics with @amh.

(1) Xaafii nam-ni export@eng hin godh -u.
Teff-ACCO man-NOM export NEG do -IMPV
"Man does not export 'teff."

(Maccaa-OC12 EYS-326)

Interlinear Gloss:

Clause	Xaafii	Name	-ni	export	him	godh	-u
Gloss	teff	Man	NOM	export	not	do	IMPV
Grammatical	ACC	N	Case marker	(inserted @eng)	NEG	V	Aspect

This gloss highlights key features of Afaan Oromoo, like how objects are marked (Xaafii, accusative), how subjects agree (nam-ni, nominative), the inclusion of an English word (export), and the complex verb forms that show negation and commands (hin-godh-u, negation with an imperfective verb). Each element is segmented and annotated in line with the Leipzig Glossing Rules, which are designed to ensure cross-linguistic comparability and clarity.

For Matrix Language Frame (MLF) analysis, using Leipzig-style interlinear glossing is important to clearly show where word parts begin and end, especially when figuring out the source and role of grammar elements in sentences that mix two languages. One critical category within the MLF model is "late outsider system morphemes"—commonly referred to as late outsider morphemes. According to Myers-Scotton, these are functional morphemes that are not conceptually activated by individual lexical items (i.e., content words), but are required to mark grammatical relations at the clause level([6]). They are “late” because they are inserted late in the production process, after the morphosyntactic frame is established, and “outsider” because they rely on grammatical information external to the embedded language content morphemes.

In bilingual speech, the language that provides these late outsider morphemes typically functions as the Matrix Language (ML)—the dominant grammatical frame into which elements from another language (the Embedded Language) are inserted. Identifying such morphemes is crucial for determining matrix language assignment, especially in structurally mixed clauses. For instance, in example(1):
"hin godh-u"
NEG do-IMPV
“ does not do (it).”

The negation prefix "hin-" and the imperative suffix "-u" are functional morphemes required by the grammar of Afaan Oromoo. If an English-origin verb stem such as export is inserted—e.g., "hin export-u"—the surrounding negation and aspect markers from Afaan Oromoo act as late outsider morphemes, indicating that Afaan Oromoo is the Matrix Language. These morphemes are not dependent on the English verb's semantics but are necessary for sentence well-formedness in Afaan Oromoo syntax. Thus, interlinear glossing that clearly separates and labels such morphemes is indispensable for rigorous MLF-based analysis, as it enables researchers to trace how structural integration occurs and which language governs the grammatical frame of the utterance.

Although not a transcription system per se, Leipzig glossing plays a foundational role in this study by enabling transparent analysis of clause structure in bilingual speech. It provides a systematic method for representing each morpheme within a clause along with its grammatical function, using standardized abbreviations ([21]). In the context of this study, Leipzig glossing was used not only to annotate the grammatical function of each morpheme (e.g., PST, 3SG, NEG, ACC), but also to indicate its language of origin using explicit markers such as @orm for Afaan Oromoo, @eng for English, and @amh for Amharic. This dual-layered annotation—grammatical glossing alongside language tagging—facilitates a clear visualization of how elements from different languages interact within a single clause. Such precision is essential for operationalizing the Matrix Language Frame (MLF) and 4-M models, which require identification of morpheme types and their alignment with either the Matrix or Embedded Language (Haspelmath ,2014).

Deuchar and her colleagues demonstrate that using Leipzig-style interlinear glosses in conjunction with the CHAT transcription framework enhances grammatical transparency in multilingual corpora, allowing researchers to trace structural patterns and test theoretical claims about code-switching ([12]) . Similarly, Deuchar (2006) applies Leipzig glossing in her Welsh–English bilingual data to disentangle the syntactic contributions of each language and clarify how grammatical structures are shared or diverge across languages([22]) . These examples underscore that Leipzig glossing is not merely a descriptive convenience, but a critical tool for theory-driven bilingual analysis.

Beyond individual studies, broader computational-linguistic research supports Leipzig glossing as a standard for cross-linguistic interoperability. For example, Nordhoff highlights its use in large-scale projects involving endangered and typologically diverse languages, where consistent glossing and language annotation facilitate data sharing, machine readability, and comparative analysis([23]). In this study, Leipzig glossing, improved with language marking, was crucial for figuring out the grammatical role and language source of each morpheme, which allowed for a detailed analysis of code-switched clauses within the MLF framework.

However, published examples often use Leipzig glosses instead of full corpus transcription. They require additional effort in formatting and are typically applied post hoc rather than during initial transcription. To maximise their benefit, Leipzig glossing should be embedded in a structured, tiered system like ELAN or CHAT.

2.3. CHAT

The CHAT transcription system, developed within the CHILDES project is one of the comprehensive systems available for morphosyntactic analysis of spoken language. It provides a system with different levels, letting researchers write down the exact words, break down the parts of words, identify the language used, and understand the grammar—all of which are important for MLF-based analysis([10]).

In bilingual corpus research, CHAT has been adapted for detailed language tagging at the morpheme level (e.g., in my data @eng (for English), @amh (Amharic), allowing researchers to differentiate matrix from embedded languages across the entire corpora. Deucher et al. applied a combined approach in their work on Building and Using the Siarad Corpus Bilingual conversations in Welsh and English, using the CHAT transcription framework for structuring bilingual data and integrating Leipzig-style interlinear glosses to capture grammatical morphemes and code-switch points indicated through language-specific tags (e.g., @eng, @amh) applied at the morpheme level within the transcription tiers([12]) . It was this incorporation of Leipzig glossing—rather than CHAT alone—that enabled fine-grained morphological analysis and clear identification of language boundaries within bilingual clauses. Their tier structure included separate lines for speaker ID, utterance segmentation, morpheme glosses, and metadata, enabling automated analysis using CLAN software (see also [12]). This technique was essential for verifying matrix language hypotheses with statistical rigor.

Deuchar et al. show that the CHAT transcription system is particularly effective for studying code-switching because it is designed to include explicit morpheme-level language tags, grammatical information, and structured tiers (e.g., speaker ID, utterance, gloss, translation [12]. This detailed formatting allows for fine-grained descriptions of bilingual utterances — including which morphemes belong to which language, and how they function syntactically. Importantly, CHAT's standardized structure enables researchers to compare data across large corpora and to automate searches for specific linguistic patterns using tools like CLAN. This makes it well-suited not only for small-scale qualitative analysis but also for large-scale, data-driven investigations of code-switching behavior across different speakers, contexts, or language pairs.

Despite its robust theoretical grounding and widespread use in language documentation, the strengths of the CHAT transcription framework come at a cost. It is highly labor-intensive, requires specialized training, and its formal markup conventions can create challenges—particularly for novice transcribers or fieldwork involving under-documented and structurally complex languages ([2]; cf. [24]). These challenges are especially acute in contexts where computational tools and language technologies are not well developed or readily available.

3. Methodological Considérations for Data Transcription and Analyses

3.1 Choice of Transcription System

The selection of a transcription and annotation framework in this study was guided by both theoretical and practical considerations, drawing specifically on the methodological principles proposed by Ochs , Lapadat and Lindsay ([2] ;[25]). Ochs emphasizes that transcription is not a neutral or mechanical act but a theory-laden process in which choices about what to represent and how to represent it reflect the researcher's analytical goals. Similarly, Lapadat and Lindsay advocate for transcription methods that are aligned with the interpretive framework of the study, highlighting the importance of making methodological decisions transparent, context-sensitive, and purposeful. Following these approaches, this study adopted a transcription system that prioritized the morphosyntactic features relevant to the Matrix Language Frame model, rather than prosodic or conversational features, and was structured to support clause-level analysis, language attribution, and manual glossing.

The choice to study code-switching between Afaan Oromoo and English using the MLF and 4-M frameworks required a system that could accurately break down words into their smallest parts, identify which language is being used, and show how the grammar works. Existing systems offer partial but insufficient solutions. CHAT offered solid support for studying language structure, particularly because it includes an optional glossing tier that allows researchers to label grammatical parts of utterances. This feature, when used with CLAN software, helps researchers analyze language structure in detail and allows for the computational processing of bilingual corpora ([10]; [26]). ELAN is a versatile tool that lets researchers add notes to different layers of spoken and signed language recordings, helping them line up various types of data—like sound, video, gestures, and speech—on the same timeline ([15]). This setup makes ELAN especially suitable for studying complicated communication situations such as when people talk over each other, take breaks, use tone, and show body language, while the Leipzig Glossing Rules help clarify word forms ([15]). However, none of these systems could completely handle the specific needs of this research, which included bilingual morphosyntactic coding, matching sounds with text, and writing styles specific to each language.

CHAT was selected as the foundational model primarily because it includes an optional glossing tier, which allows researchers to annotate grammatical morphemes and align them with lexical material. This feature helps researchers analyse the structure of language in detail, especially when paired with language tagging and the CLAN software tool, which is essential for testing theories like the System Morpheme Principle and the Morpheme Order Principle ([24]; [6]). However, given the computational demands and steep learning curve of full CHAT implementation, I chose to extract and adapt its most relevant analytical layers for manual use.

In parallel, I incorporated conventions from the Leipzig Glossing Rules. These rules enabled consistent morphological annotation using established grammatical abbreviations (e.g., PST.3SG for past tense third person singular, NOM for nominative case).The resulting transcription and analysis system integrates selected features from both CHAT and Leipzig glossing conventions. First, I used the language tagging system from CHAT to keep things consistent with the best methods in corpus linguistics and to make it easier to compare with earlier studies that used the Matrix Language Frame model (e.g., [26] ;[24]). In this process, each morpheme was manually segmented and assigned a language origin marker—such as @orm for Afaan Oromoo, @eng for English, and @amh for Amharic—reflecting language marking. In a separate annotation step, I applied Leipzig-style glossing to each morpheme following Leipz Rule-2 to indicate grammatical categories (e.g., IMPV, 3SG, NEG), enabling morphosyntactic analysis of code-switched clauses in line with the MLF and 4-M frameworks.

All transcription and glossing in this study were conducted entirely manually. This approach was necessary due to the under-resourced status of Afaan Oromoo in computational linguistics. Unlike widely studied languages, Afaan Oromoo lacks core digital tools such as morphological analyzers, part-of-speech taggers, or glossing automation systems that could support automatic annotation. While platforms such as ELAN, CLAN, and Toolbox offer semi-automatic functions—such as tier-based templates, searchability, and reapplication of tags—these features require structured lexicons, pre-defined glossing rules, or integrated parsers, none of which are currently available for Afaan Oromoo.

The linguistic analysis in this study draws on both my own linguistic competence as a native speaker of Afaan Oromoo supported with formal training at the tertiary level (including a Bachelor of Education with a minor in Afaan Oromoo), and on established grammatical descriptions, such as([27] ;[28] ;[14]). These sources were important for understanding verb forms, adding parts to words, and sentence structure, which helped keep the process of breaking down and labelling word parts consistent.

As argued by scholars such as Himmelmann (2006), manual transcription is supported as a methodologically sound and linguistically sensitive approach, particularly when working with under-documented or under-resourced languages([29]). Their support is grounded in the view that in the absence of computational tools, the most reliable way to capture the structural and functional aspects of a language is through detailed, hands-on methods. These scholars emphasize that manual transcription allows for greater control over linguistic decisions—such as morpheme segmentation, language attribution, and glossing—ensuring that the transcription aligns with the unique grammatical and typological features of the language being documented.
Therefore, my adapted Excel-based system represents both a practical response to technological limitations and a commitment to analytical depth. It allowed us to create a language-accurate collection of data that works well with existing theories and meets the needs of analyzing speech in Afaan Oromoo, English, and Amharic.

3.2 Suitability for Language-Specific Structures

Afaan Oromoo has agglutinative morphology with rich verbal inflection and noun case marking, while English relies more heavily on function words and fixed word order ([30] ;[14]). To address this difference in language structure, the transcription system was designed to keep the Qubee writing system of Afaan Oromoo intact ([27] ;[31]). English was transcribed using Oxford Dictionary

standards. Since the MLF framework sees the source and order of morphemes as important for identifying the main language, care was taken to clearly break down morphemes in Afaan Oromoo.

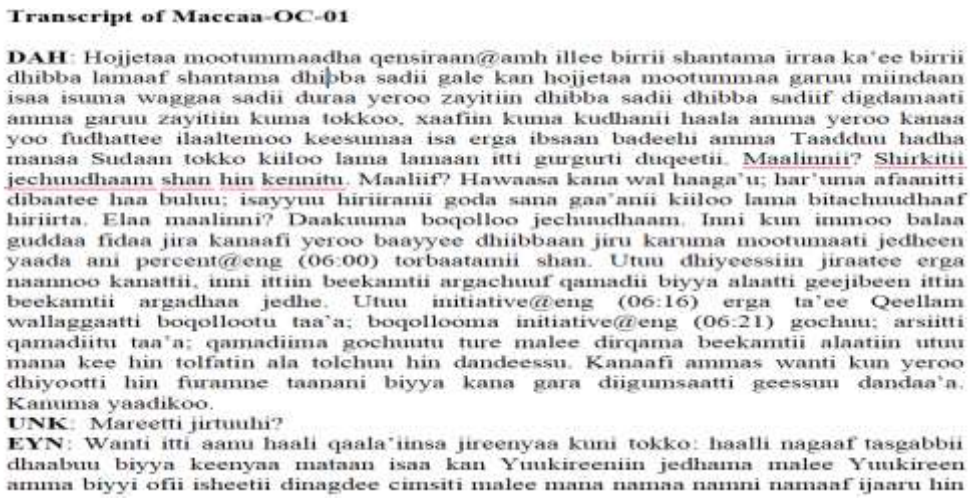
Additionally, due to the morphological complexity of many Afaan Oromoo words, each was manually segmented into its constituent morphemes—roots and affixes. These morphemes were annotated with both their grammatical role (e.g., content or system morpheme) and language of origin (e.g., @orm, @eng, @amh). This morpheme-level glossing and language tagging enabled the systematic identification of the Matrix and Embedded Languages within clauses. It also allowed for the manual tracking of language distribution, based on the frequency and positioning of morphemes. Such detailed annotation was essential for applying the Matrix Language Frame model, which determines the matrix language primarily through morpheme order and the origin of system morphemes.

3.3 Format and Data Presentation

This study employs a composite transcription framework, drawing on selected features from both the CHAT transcription system and the Leipzig Glossing Rules. The framework follows a structured three-stage process—initial transcription, clause segmentation, and morpheme-level annotation—each designed to serve a specific analytical function. The system integrates CHAT-style language tagging (e.g., @eng, @orm) to track code-switching, alongside Leipzig-style grammatical glossing (e.g., 1SG.IMPF, NEG) to annotate morphosyntactic structure. This combined approach enables both fine-grained grammatical analysis and systematic identification of Matrix and Embedded Languages, in alignment with the analytical goals of the Matrix Language Frame model. The initial data source consists of naturalistic audio recordings of bilingual speakers engaged in conversation involving Afaan Oromoo, English, and Amharic. I developed two subsequent formats from these recordings: a readable text transcription and a clause-segmented spreadsheet. These three forms—audio, transcript, and analytical tablework together in a step-by-step way to create a clear system for studying code-switching based on the Matrix Language Frame (MLF) and 4-M models ([7] ;[8]).

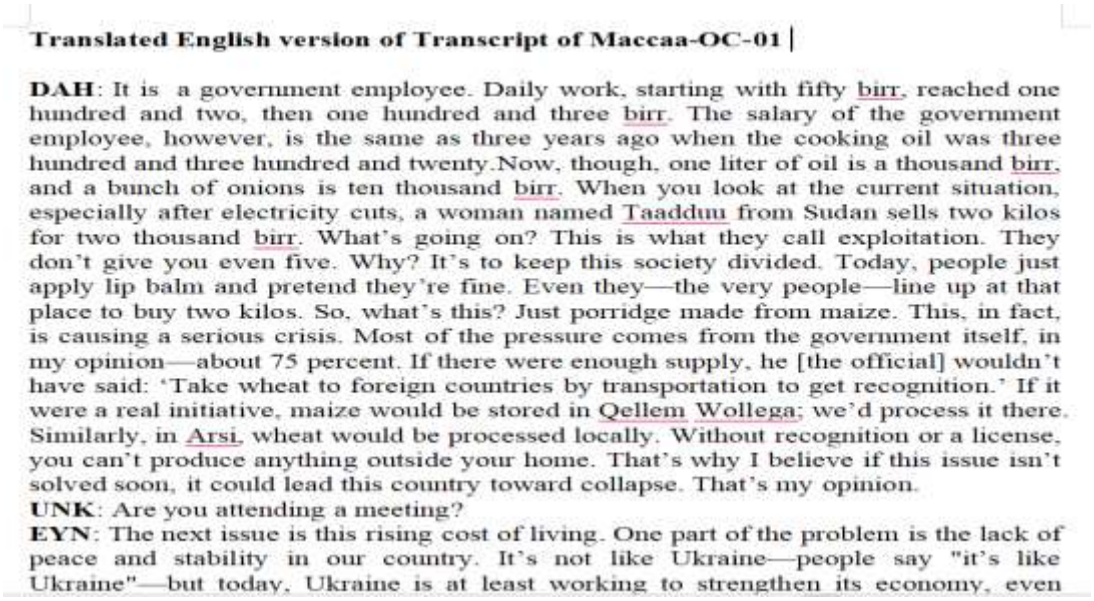
In the second stage, I transcribed the audio into a readable text format using the Qube system capturing only audible verbal interaction. This version did not attempt to represent pauses, non-verbal cues, or overlapping speech but focused on lexical and morphosyntactic content. Customized conventions were used to highlight other language elements, specifically English and Amharic: all English words were tagged with Afaan Oromoo with @orm, @eng and Amharic with @amh. I added time stamps directly after each occurrence of a non-Afaan Oromoo word, indicating the minute and second it appeared (e.g., English word "percent" [6:00]. This format allowed for rapid identification of code-mixed segments and provided a bridge between the audio data and deeper theoretical coding. See Figure 1 for an illustrative sample. Then, the transcription was segmented into clauses, consistent with the analytical requirements of the MLF model. The clause was selected as the unit of analysis because the model's predictions hinge on identifying the matrix and embedded languages at the clause level ([8]). Segmentation mainly relied on finding finite verbs, but subject-verb agreement and markers at the end of clauses were also used to check the boundaries.

Figure 1. Transcript of Maccaa-OC-01: An excerpt from informal conversation data illustrating code-switching between Afaan Oromoo, English, and Amharic.



Transcript of Maccaa-OC-01

DAH: Hojjetaa mootummaadha qensiraan@amh illee birrii shantama irraa ka'ee birrii dhibba lamaaf shantama dhibba sadii gale kan hojjetaa mootummaa garuu miindaan isaa isuma waggaa sadii duraa yeroo zayitiin dhibba sadii dhibba sadiif digdamaati amma garuu zayitiin kuma tokkoo, xaafiin kuma kudhanii haala amma yeroo kanaa yoo fudhattee ilaaltemoo keesumaa isa erga ibsaan badeehi amma Taadduu hadha manaa Sudaan tokko kiiloo lama lamaan itti gurgurti duqeetii. Maalinnii? Shirkitii jechuudhaam shan hin kennitu. Maaliif? Hawaasa kana wal haaga'u; har'uma afaanitti dibaatee haa buluu; isayyuu hiriiranii goda sana gaa'anii kiiloo lama bitachuudhaaf hiriirta. Elaa maalinni? Daakuuma boqolloo jechuudhaam. Inni kun immoo balaa guddaa fidaa jira kanaafi yeroo baayyee dhiibbaan jiru karuma mootumaati jedheen yaada ani percent@eng (06:00) torbaatamii shan. Utuu dhiyeessiin jiraatee erga naannoo kanattii, inni ittiin beekamtii argachuuf qamadii biyya alaatti geejibeen ittin beekamtii argadhaa jedhe. Utuu initiative@eng (06:16) erga ta'ee Qeellam wallaggaatti boqollootu taa'a; boqollooma initiative@eng (06:21) gochuu; arsiitti qamadiitu taa'a; qamadiima gochuutu ture malee dirqama beekamtii alaatiin utuu mana kee hin tolfatin ala tolchuu hin dandeessu. Kanaafi ammas wanti kun yeroo dhiyootti hin furamne taanani biyya kana gara diigumsaatti geessuu dandaa'a. Kanuma yaadikoo.
UNK: Mareetti jirtuuhi?
EYN: Wanti itti aanu haali qaala'iinsa jireenyaa kuni tokko: haalli nagaaf tasgabbii dhaabuu biyya keenyaa mataan isaa kan Yuukireeniin jedhama malee Yuukireen amma biyyi ofii isheetii dinagdee cimsiti malee mana namaa namni namaaf ijaaru hin



Translated English version of Transcript of Maccaa-OC-01 |

DAH: It is a government employee. Daily work, starting with fifty birr, reached one hundred and two, then one hundred and three birr. The salary of the government employee, however, is the same as three years ago when the cooking oil was three hundred and three hundred and twenty.Now, though, one liter of oil is a thousand birr, and a bunch of onions is ten thousand birr. When you look at the current situation, especially after electricity cuts, a woman named Taadduu from Sudan sells two kilos for two thousand birr. What's going on? This is what they call exploitation. They don't give you even five. Why? It's to keep this society divided. Today, people just apply lip balm and pretend they're fine. Even they—the very people—line up at that place to buy two kilos. So, what's this? Just porridge made from maize. This, in fact, is causing a serious crisis. Most of the pressure comes from the government itself, in my opinion—about 75 percent. If there were enough supply, he [the official] wouldn't have said: 'Take wheat to foreign countries by transportation to get recognition.' If it were a real initiative, maize would be stored in Qellem Wollega; we'd process it there. Similarly, in Arsi, wheat would be processed locally. Without recognition or a license, you can't produce anything outside your home. That's why I believe if this issue isn't solved soon, it could lead this country toward collapse. That's my opinion.
UNK: Are you attending a meeting?
EYN: The next issue is this rising cost of living. One part of the problem is the lack of peace and stability in our country. It's not like Ukraine—people say "it's like Ukraine"—but today, Ukraine is at least working to strengthen its economy, even

Figure 1 presents a segment of naturally occurring informal conversation that demonstrates multilingual code-switching involving Afaan Oromoo, English, and Amharic. The transcript reveals the predominance of Afaan Oromoo as the matrix language, providing the grammatical frame within which lexical items from English and Amharic are embedded. For instance, English lexical insertions such as "percent@eng (06:00)" and "initiative@eng (06:16)" occur within otherwise fully grammatical Afaan Oromoo clauses. These instances suggest the insertion of content words without structural disruption, aligning with the assumptions of the Matrix Language Frame (MLF) model. Additionally, the Amharic noun "qensiraan@amh" appears in the speaker DAH's utterance, functioning as a content word integrated into the host language discourse. The majority of the morphosyntactic structure remains consistent with the grammatical rules of Afaan Oromoo. This example provides a compelling case for studying how words are inserted and integrated in different languages, especially in the social language use of western Oromia.

The third stage involved converting the transcribed text into a structured spreadsheet to facilitate detailed morphosyntactic analysis. In line with the CHAT framework, the transcription retains utterance-level structure, speaker identification, and a tiered format that allows for linguistic annotation. However, instead of using the usual CHAT transcription that works in the CLAN software and allows for automatic translations and sound matching, this study changed the CHAT rules to fit a custom spreadsheet format. This adaptation allowed for direct integration with Leipzig glossing rules, making the system better suited for analysing morphologically rich languages like Afaan Oromoo, and enabling both qualitative coding and quantitative comparison within a single, accessible platform.

Thus, while structurally informed by CHAT, the transcription differs in its practical implementation, prioritising flexibility and cross-linguistic applicability over full CHAT protocol compliance, which is informed by prior implementations in bilingual corpora (e.g., [26];[12]The spreadsheet started with a metadata header that included details such as the speaker's pseudonym, gender, and age. Each speaker's turn was segmented into individual clauses. I arranged each clause in a vertical column format, one above the other, and assigned a row number. The other columns included the transcription, Leipzig-style interlinear glosses, English translation, clause type (e.g., monolingual, bilingual), and matrix/embedded language classification. This format made it easier to analyze the data and run statistical tests, helping to check the MLF model's predictions about the order of morphemes and which language was dominant.

3.4 Analytical Scope and Exclusion Criteria

Crucially, I excluded the first five minutes of each recording from detailed clause segmentation to ensure the data reflects naturalistic speech. Initially, when speakers become aware of the recording, they may alter their behavior or speech style. They may modify their behavior or speech style. However, as the conversation continues, they typically become more relaxed and shift toward more spontaneous, natural speech. For this reason, I believe data beyond the five-minute mark is more representative of authentic language use. I only included the conversation beyond this point in the spreadsheet for detailed analysis. The first five minutes were transcribed in a readable format but excluded from further linguistic coding. This methodological decision ensured consistency and focus on the dataset while preserving the integrity of the initial interaction in the transcript. Recent studies have

underscored the importance of theory-driven transcription in bilingual corpora, moving beyond surface-level representation to capture the structural and functional aspects of code-switching more accurately e.g. [12] ;[26]

4 Summary

The transcription system adopted in this study is an important improvement for research on bilingual speech, especially for studies based on the Matrix Language Frame (MLF) and 4-M models. Its design responds directly to the limitations of existing transcription tools when applied to typologically distinct language pairs, such as Afaan Oromoo and English.

The transcription system used in this study takes elements from existing methods—CHAT, ELAN, and the Leipzig glossing rules—to effectively analyze bilingual speech that includes and the Leipzig glossing rules to meet the needs of analyzing bilingual speech that involves different language structures, in this case Afaan Oromoo and English. The study employs CHAT, ELAN, and the Leipzig glossing rules to meet the requirements of transcribing and analyzing bilingual speech in structurally divergent languages like Afaan Oromoo and English. Leipzig glossing rules—to address the specific demands of bilingual speech analysis involving structurally divergent languages, Afaan Oromoo and English. From the CHAT system (MacWhinney, 2000), this study adopts the use of speaker-identified utterance lines and basic tier structures but applies them within a customised spreadsheet format rather than the CLAN environment. This method is different from what Deuchar et al. (2018) did, where CHAT was employed in full compliance with its XML-based protocol and used alongside CLAN and ELAN tools for comprehensive annotation—including automated glossing, disfluency markers, and prosodic features. In this study, the CHAT format was instead adapted for focused linguistic analysis, particularly morpheme-level tracking in bilingual speech, and was integrated with Leipzig glossing conventions to make morphosyntactic structures—especially the agglutinative Afaan Oromoo—analytically transparent.

From ELAN, this system takes inspiration from the feature of time-aligned annotation. However, unlike ELAN's automated synchronization, in this study time-stamping was performed manually, focusing specifically on instances where English morphemes occurred within Afaan Oromoo utterances. This mixed annotation helps analyze the detailed structure of embedded morphemes needed for the Matrix Language Frame (MLF) and 4-M models, without the extra complexity of fully using ELAN, since Afaan Oromoo is not very compatible with technology.

The use of Leipzig glossing rules provides a critical enhancement, allowing for consistent morpheme-by-morpheme glossing that reveals the grammatical and lexical makeup of bilingual utterances. This is particularly necessary given the morphological richness of Afaan Oromoo, where single words can encode tense, aspect, mood, agreement, and case through affixation. Combining Leipzig glossing with CHAT-style transcription in a spreadsheet makes it easier to see the structure clearly and allows for flexible number-based analysis, which fits the study's goals. This approach provides structural clarity and flexible quantitative analysis that is tailored to the theoretical goals of the study.

This system is important because it works as both a way to write down language and a way to analyze it, letting us capture detailed language information during transcription that can be used for both in-depth analysis and finding patterns in numbers. In this way, it helps combine different research methods, which is what bilingualism studies need right now: tools that can grow with the research while still being accurate in language use. In summary, the system differs from previous models not by introducing entirely new tools, but by recombining and adapting existing features in a streamlined, theory-driven framework. It is optimised for the structural and resource constraints of languages like Afaan Oromoo, facilitating reproducible, fine-grained, and flexible bilingual data analysis in under-resourced language contexts.

A major contribution of this system lies in its ability to make the internal architecture of utterances visible which is especially important in MLF research, where the assignment of matrix and embedded language roles depends on detailed information about morpheme origin, grammatical function, and clause structure. The present system meets these requirements by providing tiered annotation of each utterance, identifying content morphemes, system morphemes, and their language sources. This transparency is essential for both theoretical validation and corpus comparability.

While not in itself a novel methodological innovation, the use of spreadsheets to organise and analyse transcribed bilingual data remains an effective and widely adopted practice in code-switching research (e.g. Deuchar et al., 2018). In this study, using spreadsheets helps organize the analysis after transcription, making it easier for researchers to count how often code-switching happens, look at patterns among different speakers and situations, and test ideas about language use and switching. This method helps researchers understand both the details and the numbers behind code-switching, showing a trend in bilingualism research that combines different types of transcription and analysis, reflecting current methodological trends toward mixed-methods inquiry in bilingualism research. It is important to note, however, that transcription and analysis serve distinct purposes:

transcription aims to represent the linguistic data faithfully, whereas the analytical layer interprets this data in light of specific theoretical frameworks.

The system's adaptability is particularly valuable when working with under-resourced languages such as Afaan Oromoo, whose morphologically rich structure poses specific transcriptional challenges. Since there aren't complete automated glossing systems for these languages, using Leipzig glossing conventions helps researchers clearly and consistently show internal changes in word structure (like adding prefixes or suffixes). This procedure ensures that important grammatical information is preserved during transcription and made visible for analysis. If there had been many digital language tools—like part-of-speech taggers, morphological parsers, or annotated corpora—for Afaan Oromoo, the process of writing down the language could have been easier and faster, making it possible to analyze more data. If linguistic resources—such as part-of-speech taggers, morphological parsers, or annotated corpora—had been available for Afaan Oromoo, the transcription process could have been more automated and less labor-intensive, potentially accelerating data preparation and allowing for more extensive corpus-based analysis. However, in the current resource-limited context, a manual but theoretically informed transcription protocol remains essential for maintaining analytical rigour.

The approach also reinforces the view, articulated by Bucholtz (2007), that transcription is never merely technical—it is theoretical. Every transcription system embodies assumptions about what linguistic features matter, what constitutes meaningful structure, and how language should be represented. The system created here puts the ideas of the MLF and 4-M models into practice, making it a real extension of the theory instead of just a simple first step before analysis. As such, it contributes not only to better transcription practices but also to the refinement of theory itself, especially where it is tested against previously under-examined language pairs.

Nonetheless, limitations remain. The system is relatively manual and linguistically informed, requiring transcribers with basic linguistic awareness and familiarity with standard spreadsheet tools. While not heavily reliant on specialised software beyond standard applications such as Microsoft Excel or Google Sheets, the transcription process does require attention to morpheme-level detail and consistency in glossing conventions. This design choice increases accessibility, especially in under-resourced research settings where high-end software like ELAN or CLAN may be unavailable or impractical to use. Although it enables high-precision analysis, it may not yet be feasible for enormous datasets without additional support—such as automated morpheme tagging or semi-structured annotation tools. Furthermore, its applicability to discourse-level features such as turn-taking, hesitation phenomena, or non-verbal cues would require additional tiers and conventions.

In summary, the transcription protocol proposed here addresses a methodological gap in bilingual code-switching research, especially for under-resourced languages like Afaan Oromoo. It does so by linking transcription explicitly to theoretical categories, adapting to linguistic complexity, and enabling both structural and statistical analysis. In doing so, it lays the groundwork for more robust, reproducible, and theory-sensitive research in language contact studies.

5 Future Directions

This study has argued that transcription, particularly in bilingual speech research, must be understood not as a mechanical procedure but as a methodological and theoretical act. The transcription system created in this study, which uses CHAT, ELAN, and Leipzig glossing conventions, were specifically made to help analyse bilingual data from Afaan Oromoo and English using the Matrix Language Framework (MLF) and 4-M models. This combined system offers a strong way to identify main languages, sort morphemes by their role and source, and connect language transcription with morphosyntactic theory.

The examples given show how the system handles the complicated structure of sentences that mix languages, clearly showing the grammar and the basic parts of words that help determine the main language used. Unlike simple text formats—like regular transcripts or translations that don’t show grammar—the proposed system allows researchers to examine bilingual speech in a detailed and reliable way with both qualitative insight and empirical rigor. It works especially well for language pairs that are very different from each other, where regular English-based models can't fully represent the details of languages that combine words or have complex structures.

Furthermore, this transcription approach makes a contribution to broader debates in linguistic methodology. Looking forward, several avenues for further research emerge. Firstly, we could adapt the transcription tool for other language pairs, particularly those that involve minority or under-documented languages. Second, the system’s structure lends itself well to semi-automated annotation, suggesting future integration with natural language processing tools for scaling corpus analysis. Third, we could conduct inter-rater reliability studies to assess the consistency of the system and enhance its coding conventions.

In summary, the transcription system proposed in this article contributes to advancing both the practice and the philosophy of linguistic analysis. It demonstrates how careful, theory-driven design can transform transcription from a descriptive tool into a scientific method, one capable of revealing the deep structural dynamics of bilingual speech.

Funding: This research received no external funding
Conflicts of Interest: The authors declare no conflict of interest.
ORCID ID: https://orcid.org/0000-0003-4001-9139

Reference

[1] McMullin, "Transcription and qualitative methods: Implications for third sector research," *VOLUNTAS: International journal of voluntary and nonprofit organizations*, vol. 34, no. 1, Art. no. 1, 2023.
[2] Ochs, "Transcription as theory," *Developmental pragmatics*, vol. 10, no. 1, Art. no. 1, 1979.
[3] Edwards, "The transcription of discourse," *The handbook of discourse analysis*, pp. 321–348, 2005.
[4] Bucholtz, "Variation in transcription," *Discourse studies*, vol. 9, no. 6, Art. no. 6, 2007.
[5] Bezemer and D. Mavers, "Multimodal transcription as academic practice: A social semiotic perspective," *International journal of social research methodology*, vol. 14, no. 3, Art. no. 3, 2011.
[6] Myers-Scotton, *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press, USA, 2002.
[7] Myers-Scotton, *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press, 1993.
[8] Myers-Scotton and J. Jake, "Four types of morpheme: Evidence from aphasia, code switching, and second-language acquisition," 2000.
[9] Green, M. Franquiz, and C. Dixon, "The myth of the objective transcript: Transcribing as a situated act," *TESOL quarterly*, vol. 31, no. 1, Art. no. 1, 1997.
[10] MacWhinney, "The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database," 2000.
[11] MacWhinney and C. Snow, "The child language data exchange system," *Journal of child language*, vol. 12, no. 2, Art. no. 2, 1985.
[12] Deuchar, K. Donnelly, and P. Webb-Davies, "Building and using the Siarad Corpus," 2018.
[13] Wakweya, "Inflectional morphology in Mecha Oromo," *Journal of Languages and Culture*, vol. 8, no. 8, Art. no. 8, 2017.
[14] Wakweya, "Inflectional morphology in Mecha Oromo," *Journal of Languages and Culture*, vol. 8, no. 8, Art. no. 8, 2017.
[15] Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: A professional framework for multimodality research," presented at the 5th international conference on language resources and evaluation (LREC 2006), 2006, pp. 1556–1559.
[16] Tacchetti, "User's Guide for ELAN Linguistic Annotator," *The Language Archive, MPI for Psycholinguistics, Nijmegen, The Netherlands.[Google Scholar]*, 2017.
[17] Crasborn and H. Sloetjes, "Using ELAN for annotating sign language corpora in a team setting," 2010.
[18] Comrie, "From the Leipzig Glossing Rules to the GE and RX lines," in *Corpus-based Studies of Lesser-described Languages*, John Benjamins Publishing Company, 2015, pp. 207–219.
[19] Comrie, M. Dryer, and Y. Matras, "Empirical Approaches to Language Typology".
[20] Borovanský, C. Kirchner, H. Kirchner, P.-E. Moreau, and C. Ringeissen, "An overview of ELAN," *Electronic Notes in Theoretical Computer Science*, vol. 15, pp. 55–70, 1998.
[21] Haspelmath, "The Leipzig style rules for linguistics," *Max Planck Institute for Evolutionary Anthropology, Leipzig, URL http://www. uni-regensburg. de/sprache-literatur-kultur/sprache-literatur-kultur/allgemeine-vergleichende-sprachwissenschaft/medien/pdfs/haspelmath_2014_style_rules_ linguistics. pdf*, 2014.
[22] MDeuchar, "Welsh-English code-switching and the Matrix Language Frame model," *Lingua*, vol. 116, no. 11, Art. no. 11, 2006.
[23] Nordhoff, "Modelling and annotating interlinear glossed text from 280 different endangered languages as linked data with LIGT," presented at the Proceedings of the 14th Linguistic Annotation Workshop, 2020, pp. 93–104.
[24] Deuchar, K. Donnelly, and P. Webb-Davies, "Building and using the Siarad Corpus," 2018.
[25] JLapadat and A. C. Lindsay, "Transcription in research and practice: From standardization of technique to interpretive positionings," *Qualitative inquiry*, vol. 5, no. 1, Art. no. 1, 1999.
[26] Carter, P. Davies, M. Deuchar, and M. del C. P. Couto, "A systematic comparison of factors affecting the choice of matrix language in three bilingual communities," *Journal of Language Contact*, vol. 4, no. 2, Art. no. 2, 2011.
[27] Alemayehu Ayanso and A. Mawadza, *Oromo: dictionary & phrasebook*. New York: Hippocrene Books, Inc, 2017.
[28] Gragg, "Oromo of Wellegga," *The non-Semitic languages of Ethiopia*, vol. 166, p. 195, 1976.
[29] Himmelmann, "Language documentation: What is it and what is it good for," *Essentials of language documentation*, vol. 178, no. 1, Art. no. 1, 2006.
[30] Crystal, *The Cambridge encyclopedia of the English language*. Cambridge university press, 2018.
[31] Griefenow-Mewis, "A grammatical sketch of written Oromo," *(No Title)*, 2001.