Suffix inflectional morphology generation for Amharic text

Nigusie G.
Department of Information Technology
Mizan-Tepi University
Tepi, Ethiopia
gerenigusie138@gmail.com

Abstract

The Amharic language is morphologically rich language in which single lemma can form variety of words through inflection or derivation forms. Generating such variants of words manually for second language learners and NLP applications is challenging task that needs an automatic morphology generator tool. In this study, we have developed new Amharic morphology generator tool for inflecting lemmas of nouns and verbs to possessive, gender, and number forms. In case of possessive inflection, nouns can be inflected for both singular and plural forms while verbs can be inflected only for singular forms. For number inflection both nouns and verbs can be inflected. To construct these rules, we have followed Amharic word affixation rules of linguists. Before we apply the suffixation and letter series transformation rule we have analyzed the word's root form in the sentence which helps us to accurately apply the new inflected word formation rules based on the lemmas POS. Finally, we have evaluated the performance of the tool by comparing the inflected form result generated by linguists and the tool generates 76.9% accuracy compared with linguists-generated results. So as the result shows Amharic common nouns, mass nouns, and verbs suffix inflection form is generated correctly while the tool considers some proper nouns as common nouns to generate their inflected forms that need to be optimized in further studies.

1 Introduction

Morphology generation is the process of producing new words from lemma by applying inflectional rules of the language. Amharic is a Semitic family language which is largely spoken in Ethiopia and it is one of morphologically complex language, it uses Ge'ez Ethiopic scripts (Gezmu et al., 2018). Recently Amharic has become one research area for different NLP applications (Nigusie & Tegegne, 2022), like text classification (Nigusie & Tesfa, 2022), and morphological analysis (Michael, 2011). Morphology is defined as the study of the internal structures of words, over time it has been developed into a branch of linguistics that includes formation of words, their internal structures, and derivational forms (Joaquín, 1991). The advantage of having deep understanding about the morphology of a specific language is it enables learners and NLP applications to understand the appearance of words into a language and how various inflections and derivations are formed [Booij, 2010. 5]. Morphological analysis and morphological generation are the common sub-parts of studying word morphology. Morphological generation is producing a correct word form, by inflection or re-inflection rules of lemma (Ling, 2021), while morphological analysis is learning about the morphological structure of a given word form such as morphological tagging, and segmentation. Generating morphological features for both nouns and verbs is crucial task in machine translation (Kavirajan et al., 2017). The newly formed word using morphological generation will be derivational or inflectional form of the lemma. Deriving a word that has different concrete lexical meaning than the lemma is derivational morphology whereas changing the word form considering its grammatical requirements by the language is inflec-

tional morphology (Ling, 2021) Since morphology is generated based on the target language sentence pattern. The basic principle of morphological generation is to get variant forms of lemma and a set of lexical categories. The process relies on two sources of information such as dictionary/lemma of words and a set of inflectional paradigms (Melinamath & Mallikarjunmath, 2011). The lemmas can be inflected interims of number, possession, case, and others, such inflected words have significant contribution for indexing occurrence of the words in information retrieval (Khaled et al., 2007). In this study, we have developed Amharic morphology generator tool using suffix inflection for nouns and verbs.

2 Related works

Morphological generation is often implemented by concatenating related morphemes (prefixes and suffixes) with a root or lemma (Oleg et al., 2017), furthermore consonant duplicating and vowel change are also used. These lemma words can be inflected for expressing various grammatical categories, such as mood, voice, aspect, person, gender, number, and case (Khaled et al., 2007). Considering such variety of forms helps to understand the inflectional and derivational word formation from lemma (Waad & Nilotpala, 2022). Understanding the morphological structure of words in a wider form helps for better comprehensive reading and writing abilities. Because of such reasons, teachers and linguists give attention to morphology formation to increase the lexical capacity of students and speakers (Girmay, 2011). Generating the correct inflectional and derivational morphology depending on the languages is an essential part of the research and it is fundamental task in natural language processing task (Ling, 2021). To generate these inflectional and derivational forms, the word lemma will be analyzed first using morphological analysis process (Ling, 2021), then adding different affixes for the lemma word (Violetta et al., 2000). To generate such variants of word formation the morphological rules need be specified by conditions (if and then) (Joaquín, 1991). The if part is matched against the value of the feature root and then part includes addition, deletion, and replacement of prefixes, infixes, and suffixes. The other common approach for morphological generation is uses finite state transducers (Melinamath & Mallikarjunmath, 2011), generate morphology of the word using affix strata and concatenation rule sets for nouns and verbs which is working aligned with the analysis such as Arabic words represented in the standard orthography using Arabic finite state lexical transducer (Kenneth, 1996). These rules rely on hand-crafted rules and lexicography, the machine learning model such as conditional random field (Durrett & DeNero, 2013) is also employed for such morphology generation process by extracting orthographic transformation rules the Long Short-Term Memory (LSTM) network is also used to learn the paradigms of a morphologically complex language by giving knowledge of partial structural forms of the word to generate the remaining unobserved inflected forms (Robert, 2016). Transform the input to a sequence of outputs by representing the inflected form through grouping words lemma with their inflection using neural network sequence to sequence transducer, the result of the accuracy of these neural encoder-decoder models depends on the length of the word (Manaal et al., 2015). Such morphology generation strategies and investigation help to know the properties of the languages and to make them easily learnable. To study the morphology of the Amharic language priory rule-based morphological learning methods are applied (Wondwossen & Michael, 2012) which targeted on 216 manually prepared Amharic verbs. Furthermore, the Amharic language is morphologically complex in different parts of speech words, such as nouns, adjectives adverbs verbs are common word types that follow separate affixation rules, because of this we have conducted the work for generating morphology of Amharic nouns and verbs using different combination and transformation rules of the lemma word.

3 Word Formation

Word formation is process of generating new words from an existing root by adding affixes or another word and the process involves four major rules (Kenneth, 1996), such as Affixation, the addition of prefixes, suffixes, and infixes to the root word, reduplication the repetition of all or parts of the root or stem word, symbolism change the structure of the lemma by altering it in some way or compounding rule that forms new word by combining two or more

root morphemes. Amharic language shares these list of rules to generate morphology of root words based on its part of speech (Wondwossen & Michael, 2012). The morphology generation of the Amharic lemma word are inflectional or derivational according to its part of speech class. This inflectional or derivational morphology of the lemma can be affixed to person, number, gender, tense, aspect, passive, causative, and polarity (affirmative/negative). To generate the morphology of these different parts of speech lexicons, the root or lemma word, affixes, and concatenation set of rules are the basic resources required in new word formation (Melinamath & Mallikarjunmath, 2011). The lemma of the word and its part of speech is analyzed before applying this word formation using morphological analysis process which is the sub-parts of morphology generation. One of the main objectives of the tagging task is to provide lexical information that can be employed for syntactic analysis (Gezmu et al., 2018). Finite-state transducers (FSTs) and rule-based transformations are the two commonly used methods for word formation (Khaled et al., 2007). The Finite-state transducers which is working based on letter path through the lexical trees from a legal starting state to a final leaf (Kenneth, 1996). The rule-based morphological generation is used to form the inflection form of the lemma using its specified feature list as input (Nizar & Arabic, 2004).

3.1 Noun Affixation

Amharic nouns are used to name an object and they include three subcategories (Mersie, 1948), such as Proper Nouns used to name a specific person, place or thing like **አብርሃም** (Abrham), **ኢትዮጵያ**(Ethiopia). These proper nous do not have infixes and suffixes unless they are replaced by common nouns. Common nouns are used to name the common representative name of person, place or thing like **ሰው**(human), **ከተማ**(city), **አንበሳ**(lion). These nouns can be inflected for number, gender, possessives and etc, for example, **ከተማ**(city) can be inflected as **ከተሞች**(cities) for number inflection by adding suffix **ኦች** and **ከተማችን** (our city) for first person singular feminine and masculine by adding **አችን**. The third types of noun are general nouns which are used to name the collective of common nouns or name tribe. The name **እንጨት**(wood) and **ሰብል**(crop) are some examples of general nouns in Amharic. Similar to common nouns these general nouns can be inflected to number, gender, possessive, and others.

3.2 Verb Affixation

Verbs are used to express action, state of being, or a relation between words in the sentence (Jeff, 2021). Amharic has complex verbal morphology that used to express the actions and can be inflected in different forms. The following are some of the inflectional properties of these Amharic verbs tense (past, present, or future), number (singular or plural), gender, mood (such as subjunctive), person (first, second, or third), and voice (active or passive). For example, the lemma verb **በላ**(ate) can be inflected as **በሉ**(eat) for third person plural feminine and masculine.

4 Morphological analysis

Morphological analysis of highly inflected languages like Amharic is a non-trivial task because of its complex morphological structure (Abate & Assabie, 2014). This analysis process focuses on studying the word structure and part of speech (Girmay, 2011). In this study, we have used the morphological analysis process for extracting the smallest meaningful word(lemma) with its part of speech information. This lemma extraction from the word before applying morphological generation helps us to form variants of morphologically inflected words based on their part of speech from the analyzed lemma. For instance, for the word **በላችሁ**(belachu), it is morphologically inflected word, so it is impossible to apply different morphological formation rules without finding the minimal meaning of full word or lemma. So the morphological analyzer is applied to handle such issues for splitting the word to lemma and affix such as **በላ** and suffix **ችሁ**. The root word **በላ** can be concatenated with different suffixes for inflectional as well as derivational morphology generations. The HornMorpho: a system for morphological processing of

Amharic, Oromo, and Tigrinya (Michael, 2011) is used to find the lemma of the word and to identify its part of speech. The identification of POS of the word also helps us to apply accurate morphology generation rule, because the affixation rule of Amharic verbs is different from that of Amharic nouns.

5 Morphological generation

Morphological generation is process of forming new words using different inflection and derivation rules (Ling, 2021). This new word formation process is depending on the specific language subject verb argument and word part of speech. Developing morphological generation tool is vital method for understanding the structure of the language. In this paper, we have developed tool for Amharic language morphological generation. The inflected words can be generated in different forms using rules that can be applied to the lemma (Girmay). These rules include affixation which is the addition of prefixes, suffixes, and infixes to the lemma, for example, the word **ሰው**(sew) can be inflected as **የሰው** through adding prefix **የ** and **ሰዎች** by adding the suffix **ኦች**. The second method is reduplication which is generating new word by repetition of all or parts of a stem. The word **ስብርብር**(sbrbr) can be generated from **ስብር**(sbr) by repeating parts of the root **(ብር)** word. The other method is compounding which is the process of formation of new word by combining two or more root morphemes. The morphological generator tool that we have developed in this study is forming inflected words by adding suffixes and/or changing parts of the word. Our main focus in developing the morphological generator tool is for inflecting nouns and verbs to possessive (for both singular and plural), number, and gender. Some of the rules that we have used to develop the morphology generator tool are for possessive inflection of nouns which end with the 3rd and 5th Amharic series of litters, the first person singular feminine and masculine inflection will add **የ**(ye) at the end, the first person plural feminine and masculine will add **ዎቻችን**(wochachn) at the end. Number inflection for the words end up with 2nd, 3rd, 5th, and 7th series will add **ዎች** at the end. The detailed rules that we have used for such nouns possessive inflection is described in Table 1.

Table 1: Nouns possessive affixation.

Words end with	1st p. sing	1st p. prul	2nd p.sing(m)	2nd p.sing(f)	2nd p. prul	3rd p.sing(m)	3rd p p.sing(f)	3rd p. prul
3rd & 5th letter series	+የ	+ ያችን	+ ህ	+ ሽ	+ ያችሁ	+ ው	+ ዋ	+ ያቸው
4th letter series	+ የ	+ ችን	+ ህ	+ ሽ	+ ችሁ	+ ው	+ ዋ	+ ቸው
6th letter series	6th -> 5th	6th -> 4th + ችን	+ ህ	+ ሽ	6th -> 4th + ችሁ	6th -> 2nd	6th -> 4th (ዲቃላ)	6th -> 4th + ቸው
7th letter series	+ የ	7th -> 8th (ዲቃላ)+ ችን	+ ህ	+ ሽ	7th -> 8th (ዲቃላ)+ ችሁ	+ ው	+ ዋ	7th -> 8th (ዲቃላ)+ ቸው

Similar to nouns Amharic verbs have their own structure and inflectional rules which we have applied to vary them using the tool we have developed in this study. Amharic verbs can be inflected for possessive (first person, second person, and third person). To inflect these different cases, we have used the rules such as the verbs ending with 1st series of letters are inflected as the last letter changed to 6th series and add **ክ**(ku) for second person singular masculine inflection. For example, the word **ገነጠለ** will be inflected as **ገነጠልክ**, for the second person singular masculine by changing the last letter to 6th and add **ክ**. The detailed verb case inflection rule that we have applied in this study is described in Table 2 below.

Table 2: Verb possessive affixation.

Words end with	1st p. sing	1st p. prul	2nd p.sing(m)	2nd p.sing(f)	2nd p. prul	3rd p.sing(m)	3rd p p.sing(f)	3rd p. prul
1st letter series	1st->6th +ህ	1st->6th+ን	1st->6th + ከ	1st->6th+ ሽ	1st->4th + ችሁ	-	+ ች	1st->2nd + ያቸው
4th letter series	+ ሁ	+ ን	+ ህ	+ ሽ	+ ችሁ	-	+ ች	1st->2nd

In the above Table 2 as the 7th column shows it does not take any suffixation which indicates the lemma verbs will be used for 3rd person singular masculine inflection. For example, the word we have used before, **ገነጠለ** will be used directly as it is for the 7th column types of inflection rule.

6 Result and discussion

In this study, we have developed morphological generator tool for Amharic nouns and verbs. Because morphology generation is an essential issue for natural language processing tasks and learners of the language (Girmay, 2011). The word form generation has recently gained attention in many languages to increase the understandability and sentence formation of specific language (Khaled et al., 2007; Joaquín, 1991). The morphology generation tool that we have developed in this study is based on suffixation rules. Inflecting both Amharic noun and verb lemmas to possessive, gender, case, and number. To evaluate the new word formation performance of the tool for these different inflectional types, we have used sample test sentences from such test sentences **ልጁ ከሃገር ወጣ** (he is out the country) is one example sentence used for testing the generation performance of the tool which its result is illustrated in Table 3 and 4. Amharic nouns can be inflected for both singular and plural possessives (1st person, 2nd person, and 3rd person) while verbs can be inflected for singular possessive forms. As below tables (Table 2 and 3) show before we applied the inflectional rules, we analyzed its part of speech that helps us to apply the right inflection based on its part of speech type. In Table 3 the nouns **ልጁ** and **ሀገር** are inflected for eight singular possessive forms. The formation of these inflectional rules considers letter series changing or transforming in addition to adding different inflectional suffixes, for example, when the word **ሀገር** is inflected for 3rd person both feminine and masculine the last letter**(ር)** of the word is changed to other series **(ራ)** from the sixth series to fourth series and the suffix **ቸው** is added. Based on Amharic nouns suffixation rules [13], the word end with sixth series will add vowel **ኣ** (a) and consonant**ቸው** for such 3rd forms of possessive inflection. Amharic verbs can also have inflected for such eight types of singular possessive forms, while their way of suffixation is different from the suffixation rules of nouns, for instance, the verb in Table 3 row 4 for third person singular feminine and masculine inflection, the last letter of the word **(ጣ)** is changed to second series **(ጡ)**, based on the suffixation rule of Amharic verbs end with fourth series of letter will be changed to second series of letter.

Table 3: Nouns singular possessive inflection.

Word	POS	1st M&F	1st M&F	2nd M	2nd F	2nd M&F	3rd M	3rd F	3rd M&F
ልጅ	n	ልጄ	ልጃችን	ልጅህ	ልጅሽ	ልጃችሁ	ልጁ	ልጇ	ልጃቸው
ሀገር	n	ሀገሬ	ሀገራችን	ሀገርህ	ሀገርሽ	ሀገራችሁ	ሀገሩ	ሀገሯ	ሀገራቸው
ወጣ	v	ወጣሁ	ወጣን	ወጣህ	ወጣሽ	ወጣችሁ	ወጣ	ወጣች	ወጡ

The next inflection for these Amharic nouns is plural forms of possessive inflection. This plural noun suffixation has also specific rules as Row 3 shows in Table 4 the word is inflected as **ሃገሮቻቸው** for 3rd person plural possessive form the last letter of the word is changed to seventh series and add **ቻቸው**, while verbs are not inflected for such plural possessive forms.

Table 4: Nouns plural possessive inflection.

Word	POS	1st M&F	1st M&F	2nd M	2nd F	2nd M&F	3rd M	3rd F	3rd M&F
ልጅ	n	ልጄ	ልጆችን	ልጅህ	ልጅሽ	ልጆችሁ	ልጁ	ልጇ	ልጆቸው
ሀገር	n	ሀገሬ	ሀገሮችን	ሀገርህ	ሀገርሽ	ሀገሮችሁ	ሀገሩ	ሀገሯ	ሀገሮቸው
ወጣ	v	ወጣሁ	ወጣን	ወጣህ	ወጣሽ	ወጣችሁ	ወጣ	ወጣች	ወጡ

The last inflectional rule we have considered in this study is Amharic nouns and verb number inflection. In case number both noun and verb lemas can be inflected, as Table 5 shows the noun **ልጅ**(child) is inflected to **ልጆች**(children). the rule such as nouns that end with 6th series of letter will add vowel **ኦ**(o) and consonant **ች**, while for verbs it will add vowel **ኡ**(u) such as the word **መጣ** is changed to **መጡ** (they come).

Table 5: Nouns and verbs number inflection.

	word	POS	Plural form
0	ልጅ	n	ልጆች
1	ሀገር	n	ሀገሮች
2	ወጣ	v	-

The performance of the tool is evaluated by comparing it with linguist's result (Mersie, 1948). As the result shows the tool correctly generates the inflectional form of common nouns, mass nouns, and verbs for possessive number and gender inflection however in some cases the common nouns that come as proper non in the sentence are considered as common nouns which need to be improved in further studies.

7 Conclusion

Recently the morphological generation process has gained attention in different languages because morphological knowledge helps to express the idea verbally. Amharic languages have large variants of inflectional and derivational forms for single word which is challenging for second language learners to easily generate variant forms and for NLP applications to form such variants, due to this reason we are motivated to develop new Amharic morphological generator tool. The tool is developed based on variant rules for inflecting Amharic nouns and verbs for possessive, gender, and number. We have formed the rules based on linguists' rules for forming valid inflected forms for lemma. Before we apply such inflectional suffixation rules and letter series transformation, we have applied a morphological analysis process to identify the word part of speech, which helps us to use the right inflection rule based on its POS. Finally, we have compared the generation performance of our tool with the linguist's generated result, the tool generates the correct possessive form of 10(76.9%) words from 13 words of a single sentence that matches correctly with the linguist's result.

References

Mesfin Abate and Yaregal Assabie. Development of amharic morphological analyzer using memory-based learning. In *Advances in Natural Language Processing*. Springer International Publishing, 2014.

Greg Durrett and John DeNero. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2013.

A. Mekonnen Gezmu, B. Ephrem Seyoum, Michael Gasser, and Andreas Nurnberger. Contemporary amharic corpus: Automatically morpho-syntactically tagged amharic corpus. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*. Association for Computational Linguistics, 2018.

Berhane Girmay. Word formation in amharic. pp. 1–25.

Berhane Girmay. In search of pre-service efl certificate teachers' attitudes towards technology. *Procedia Computer Science*, 3:666–671, 2011.

Valerioti Jeff. *Verbs Defined*. Liberty University Online Writing Center, 2021.

A. Domínguez Joaquín. The role of morphology in the process of language acquisition and learning. *Revista Alicantina de Estudios Ingleses*, 4:37–47, 1991.

B. Kavirajan, M. Anand Kumar, K.P. Soman, S. Rajendran, and S. Vaithehi. Improving the rule based machine translation system using sentence simplification (english to tamil). In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017.

R. Beesley Kenneth. Arabic finite-state morphological analysis and generation. In *International Conference on Computational Linguistics*. 1996.

Shaalan Khaled, Talhami Habib, and Kamel Ibrahim. Automatic morphological generation for the indexing of arabic speech recordings. *International Journal of Computer Processing of Oriental Languages*, 20(1):1–14, 2007.

Liu Ling. *Morphological Generation with Deep Learning Approaches*, volume 1. 2021.

Faruqui Manaal, Tsvetkov Yulia, Neubig Graham, and Dyer Chris. Morphological inflection generation using character sequence to sequence learning. In *North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2015.

Bhuvaneshwari C Melinamath and A G Mallikarjunmath. A morphological generator for kannada based on finite state transducers. In *2011 3rd International Conference on Electronics Computer Technology*. Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (ICST), 2011.

H. Weldeqirqos Mersie. *Amharic Sewasew*, volume 1. Artistic, 1948.

Gasser Michael. Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In *COMPUTEL*. 2011.

Gebregziabihier Nigusie and Tesfa Tegegne. Amharic text complexity classification using supervised machine learning. In *Artificial Intelligence and Digitalization for Sustainable Development (ICAST 2022)*. Springer Nature Switzerland, 2022.

Gebregziabihier Nigusie and Tegegne Tesfa. Lexical complexity detection and simplification in amharic text using machine learning approach. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*. IEEE, 2022.

Habash Nizar and Morphology Arabic. Large scale lexeme based arabic morphological generation. In *Session Traitement Automatique de l'Arabe*. JEP-TALN, 2004.

Sychev Oleg, Gurtovoy Vladislav, and Penskoy Nikita. A study of efficiency of modern inflection and lemmatization software. In *Proceedings of the IV International research conference "Information technologies in Science, Management, Social sphere and Medicine" (ITSMSSM 2017)*. Atlantis Press, 2017.

Malouf Robert. Generating morphological paradigms with a recurrent neural network. San Diego Linguistic, 2016.

Cavalli-Sforza Violetta, Soudi Abdelhadi, and Mitamura Teruko. Arabic morphology generation using a concatenative strategy. In *Applied Natural Language Processing Conference*. 2000.

D. Naser Waad and Gandhi Nilotpala. Morphological analysis on the language acquisition. *Journal La Sociale*, 03:160–165, 2022.

Mulugeta Wondwossen and Gasser Michael. Learning morphological rules for amharic verbs using inductive logic programming. In *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*. 2012.