Scientific Research Publishing

Open Journal of Modern Linguistics, 2022, 12, 504-509
https://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

# Guideline for the Construction of a Formal Grammar for the Malagasy Language

**Nivo Randriambololona, Aina Nambinintsoa Rakotondrafara**

Laboratoire de Recherche en Sciences Cognitives et Applications (LRSCA), Ecole Supérieure Polytechnique d'Antananarivo, Université d'Antananarivo, Antananarivo, Madagascar
Email: nivoran@gmail.com, nambiprofessionnel@gmail.com

**How to cite this paper:** Randriambololona, N., & Rakotondrafara, N. A. (2022). Guideline for the Construction of a Formal Grammar for the Malagasy Language. *Open Journal of Modern Linguistics, 12,* 504-509.
https://doi.org/10.4236/ojml.2022.124036

**Received:** July 23, 2022
**Accepted:** August 19, 2022
**Published:** August 22, 2022



## Abstract

This paper presents a guideline for the construction of formal grammar for the Malagasy language. The used method is based on a deterministic approach given that reliable corpora are not yet available for Malagasy. The main purpose of formal grammar is language recognition which will be the keystone of an automatic grammar checker. Jointly used with an existing part-of-speech-tagger, a grammar checker will bring us further by facilitating the automatic creation of reliable corpora which in turn will boost the Malagasy automatic processing.

## Keywords

NLP, Malagasy, Grammar, Language, Recognition

## 1. Introduction

A natural language may or may not be ruled through grammar, which is established by an Academy. Grammars are supposed to govern as well language production as language recognition. Structured natural languages get more and more involved in most emerging technologies. Great progress has already been achieved in terms of Natural Language Processing (NLP). Unfortunately, only a small handful of languages can really take advantage of these technological advances because they are most of the time tailored for English.

In terms of NLP, the Malagasy language, the mother language of Madagascar, is 70 years behind the English language, even if from a purely literary perspective, it is not less qualified. Promoted by the respectable Malagasy Academy and taught in all schools in Madagascar, even at some Universities, it has its own grammar, its proper vocabulary as well as a significant heritage of literary works,

which attest to the intellectual capital of the Malagasy civilization. But a civilization rarely stagnates, either it thrives or it fades slowly away under the cultural influence of other civilizations. Survival of the fittest is a law of nature. As native Malagasy citizens, the authors care for the survival of their culture and have chosen to tackle the problem by the promotion of their mother language in the NLP world.

Malagasy presents all the advantages to deserve a place in the research fields of NLP. The work presented in (Rakotondrafara et al., 2019), which is focused on the creation of a part-of-speech-tagger for the Malagasy language, is the first step in this direction. The present paper is the first of a series of guidelines, the common purpose of which is to put the automatic processing of the Malagasy language on the right track through the modelization of fundamental resources such as formal grammar, parsers, ontologies, and corpora which apparently are sorely lacking in the current state of the art.

In this paper, we will first highlight the main specificities of Malagasy grammar. Then we will present a guideline for the creation of formal grammar for the Malagasy language.

## 2. Specificities of the Malagasy Grammar

Like most grammars, Malagasy grammar, as defined in Ramik (2013), is determined by concepts like word classes, proposition types, proposition forms, and proposition structure among others. But In many points, the Malagasy grammar is easier to check compared to other languages like English or French, thanks to the following specificities:

- Nouns and determiners have no gender.
- The conjugation of a verb remains invariable for all personal pronouns.
- There are only 3 verb tenses: past, present and future and they differ from each other only through their prefix.
- The verb *to be* does not exist. In special cases, it is replaced by an emphasizing word.

Nevertheless, it has also its specificities that make the syntactic analysis a bit more complicated than in French or in English. Effectively, the structure of a proposition is determined by the respective positions of its 3 fundamental elements: the verb: *Enti-milaza* (EM), the subject: *Lazaina* (L), the object: *Fameno* (F), where "verb" is not actually a complete translation of EM since it may also be an adjective, a pronoun or a noun. Idem for the subject and the object.

## 3. A Formal Grammar for the Malagasy Language

### 3.1. Definition

There are two types of grammar (Clément et al., 2009): positive grammars for generating all grammatical sentences and negative grammars for describing ungrammaticalities. The grammar we want to construct for Malagasy is a positive context-free grammar.

According to the definition given in (Loeckx et al., 1986), a formal grammar is a quadruple G (Σ, ℕ, ℝ, S), where Σ is the alphabet that contains the terminal words, ℕ the alphabet of the non-terminal symbols, ℝ a set of rewriting rules of the form $w \to w'$ where $w$ and $w'$ are in $(\Sigma \cup \mathbb{N})^*$ and S ∈ ℕ the germ of the grammar. All words of the language are directly or indirectly derived from S through successive applications of rules.

### 3.2. The Terminal Alphabet of the Malagasy Language

It contains all possible words (gathering canonical words that we can find in conventional lexicons and slang or dialect versions of the same words) in Malagasy. Since each word is assigned to at least one word class, we will define the terminal alphabet as the union of all existing classes.

### 3.3. The Non-Terminal Alphabet of the Malagasy Language

It contains non-terminal words, which means symbols that are not parts of the language itself but that are solely used as intermediate placeholders in the application of some rules along the derivation process.

They are organized in layers:

- Layer 1: The germ {<S>}.
- Layer 2: The sentence types.
- Layer 3: The sentence forms.
- Layer 4: The structures.
- Layer 5: The part of speech sequences.
- Layer 6: The word classes.

### 3.4. The Rewriting Rules for Language Production

Our goal is to produce sentences (or propositions) in Malagasy by the repeated and subsequent applications of rewriting rules beginning with the germ <S> until we get a sentence with exclusively terminal words in it. This is a quite straightforward process. For a better understanding of the production rules for Malagasy, let us proceed step by step. To express rules, we use the formalism: $w \to w'$ (meaning: derive $w'$ from $w$ or rewrite w as $w'$), where $w$ and $w'$ are in $(\Sigma \cup \mathbb{N})^*$.

First, we must set, which kind of sentence we want to produce: **a single proposition sentence** or **a multiple proposition sentence**.

#### 3.4.1. Single Proposition Sentences

- According to Layer 2, there are 5 types of propositions. For each one, we define a rule. For example, the rule <S> → <decl> serves to initiate the production of a declarative proposition (Layer 1 → Layer 2).
- Once the proposition type is known, we have to choose the proposition form among the 16 possible combinations (Layer 2 → Layer 3).
- Next, the structure must be defined where we have the choice between 4 possibilities (Layer 3 → Layer 4).
- Finally, each structure in layer 5 is synthesized as a parse-tree whose leaves

are the symbols in Layer 6.

The language production can be viewed as the top-down construction of a tree with the germ as the root. Figure 1 shows an example.

#### 3.4.2. Multiple Proposition Sentences

For this kind of sentences, we must introduce cycles in rules, so that we can put as much propositions as needed inside a single sentence:

- <S> → <S><conj-sub><S>

We can also put the conjunction at the beginning of the first proposition, in which case an emphasizing word  must be added between the two propositions.

- <S> → <conj-sub><S><S>

Each appearance of the germ <S> in these rules can be treated like in Section 3.4.1.

### 3.5. Language Recognition Using K-N-N Algorithm

Language recognition is by far more complex than language production. Given a sentence, the following steps are globally required:

1) Segmentation

2) Lexical analysis:

a) Check for each word if it is a valid Malagasy word.

b) Proceed to a part-of-speech-tagging (pos-tagging).

c) For the whole sentence, create a corresponding pos-vector.

3) Check if the pos-vector corresponds to a legal parse tree according to Section 3.4.1. The vector is just a raw structure and has yet to be compiled into a concrete parse tree, what turns to be a classification problem. Hence, a training set is required which could be easily build if we define a representative vector and a unique number for each legal parse-tree.

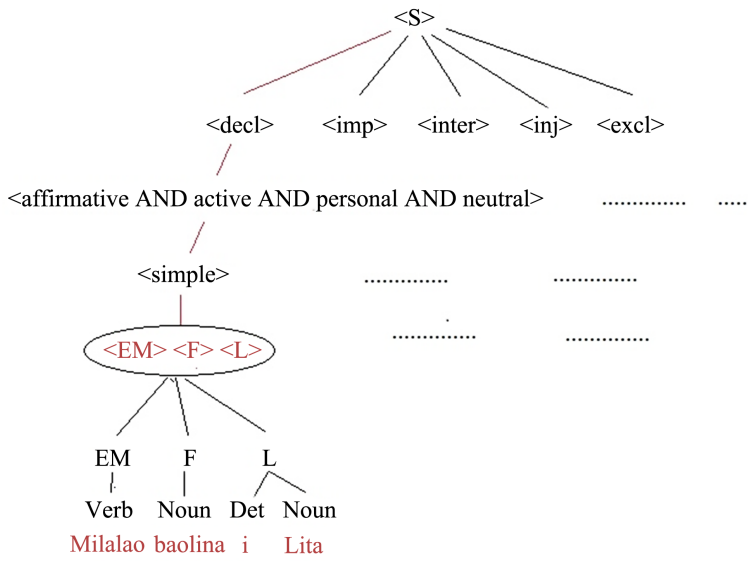


Figure 1. The rewriting rules, viewed as a tree.

4) Instead of using a forest-based algorithm as suggested in (Huang, 2008), the K-nearest-neighbors algorithm (Sun et al., 2018) seems to be more appropriate to identify an appropriated parse tree for our input vector. This algorithm would evaluate the Euclidian distances between the input-vector and each pos-vector in the training set, after which the k vectors with the minimal distances will be identified. Each vector (under the k vectors) will vote for its class. The class that obtains the best result indicates the appropriate parse tree.

Note: A threshold should be set for the Euclidian distance. If for all k vectors, the threshold is exceeded, then a non-existing class-number (−1) is assigned to the input-vector, what would mean that the input sentence is ungrammaticality.

## 4. Conclusion

The goal of this paper is to present a guideline for the creation of a formal grammar for Malagasy, not an exhaustive listing of rules, nor a detailed algorithm for language recognition.

We have shown that it is absolutely possible, for a formal grammar for the Malagasy language to model, which could be used as well for language production as for language recognition. Since we are now living in an age where machine learning-based methods are the references for solving any kind of complex problems, one might ask why we define the grammar rules manually. The answer is quite simple. Learning-based methods are appropriate and straightforward for grammar checking and rule definition if *reliable* (*grammatically speaking*) training corpora are overwhelmingly available. We are yet far from this situation regarding the existing literary work in the Malagasy language. Even if nowadays, an important quantity of texts is daily produced via numerous social networks and journalistic articles, they are riddled with errors and inaccuracies. Letting the machine learn from errors is not a good idea.

One aspect that we haven't yet considered in this work is the problem of ambiguity, which can already happen in the lexical analysis. To solve ambiguity problems, the semantic dimension should be taken into account. Therefore, semantic-based concepts like ontologies should also be created for the Malagasy language, in parallel with grammar. This will be the topic of our future research.

The soon we can implement resources like formal grammars, ontologies, and annotated corpora among other things, the sooner the time will arrive where Malagasy could also benefit from automatic learning techniques and methods.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

Loeckx, J., Mehlhorn, K., & Wilhelm, R. (1986). *Grundlagen der Programmiersprachen*. Teubner Verlag. https://doi.org/10.1007/978-3-322-94706-2

Rakotondrafara, A. N., Randimbindrainibe, F., Randriambololona, N. H., & Robinson, M.

(2019). Natural Language Processing: Malagasy Part-of-Speech Tagging. *International Journal of Advance Research and Innovative Ideas in Education, 5,* 705-717.

Ramik, D. M. (2013). Grammaire Malgache. http://dominicweb.eu/fr/malagasy/grammar/

Huang, L. (2008). *Forest-Based Algorithms in Natural Language Processing*. Scientific Research, University of Pennsylvania.

Clément, L., Gerdes, K., & Marlet, R. (2009). A Grammar Correction Algorithm—Deep Parsing and Minimal Corrections for a Grammar Checker. *Proceedings of the 14th International Conference on Formal Grammar* (pp. 1-16). https://www.researchgate.net/publication/29600813_A_Grammar_Correction_Algorithm_-_Deep_Parsing_and_Minimal_Corrections_for_a_Grammar_Checker

Sun, J. W., Du, W. X., & Shi, N. C. (2018). A Survey of kNN Algorithm. *Information Engineering and Applied Computing, 1,* Article ID: 770. https://doi.org/10.18063/ieac.v1i1.770