NATURAL LANGUAGE PROCESSING: MALAGASY PART-OF-SPEECH TAGGING

RAKOTONDRAFARA Aina Nambinintsoa[1], RANDIMBINDRAINIBE Falimanana[2], RANDRIAMBOLOLONA Nivo Harisoa[3], ROBINSON Matio[4]

[1] *Student in Doctoral School of Science and Technic Engineering and Innovation, Laboratory of Cognitive Sciences and Application, High School of Polytechnical University in Antananarivo, Madagascar*

[2] *Professor, in Doctoral School of Science and Technic Engineering and Innovation, Laboratory of Cognitive Sciences and Application, High School of Polytechnical University in Antananarivo, Madagascar*

[3] *Doctor, in Doctoral School of Science and Technic Engineering and Innovation, Laboratory of Cognitive Sciences and Application, High School of Polytechnical University in Antananarivo, Madagascar*

[4] *Doctor, in Doctoral School of Science and Technic Engineering and Innovation, Laboratory of Cognitive Sciences and Application, High School of Polytechnical University in Antananarivo, Madagascar*

ABSTRACT

The objective of this work is the realization of a part-of-speech tagging (POS tagging) for the official Malagasy Language. The tagger uses a set of tags, a tagged learning corpus, various mathematical models, and an algorithm that implements the basic rules of the Malagasy grammar. We will describe our tagset in a two-level approach and the performance of our part of part-of-speech tagging in relation to unigram-based tagger, the bigram model and the Hidden Markov Model. We will then present detailed results of each simulation.

Keyword: *part-of-speech tagging, POS tagging, official Malagasy Language, tagset, corpus, algorithm, Malagasy grammar, unigram model, bigram model, Hidden Markov model (HMM Model)*

1. INTRODUCTION

The part-of-speech tagging consists of assigning a tag that represents a grammatical class to a word or group of words. During this work, we will create an official Malagasy tagger based on a supervised learning. Our tagger uses the NLTK tool which will permit us to independently tag a sentence or a corpus from a learning corpus.

2. PRESENTATION OF THE TAGSET

In order to obtain satisfactory results, we have decided to fix the principles of Malagasy grammar, and to avoid the usual mistakes when tagging the learning corpus, in accordance with the writer and teacher Régis RAJEMISA-RAOLISON'S work "Grammaire malgache". As far as the creation of the tagset for the Malagasy language is concerned, we relied on the tagset of the NLTK tool.

Our tagset lead us to adopt a two-level approach tagset in which the first level corresponds to the 13 tags representing the 13 grammar classes and the second level to the complete set of 42 tags.

2.2 First level of tags

Table -1: First level of tags

Tags	Description
CC	Conjunction
CD	Number
DT	Determiner
NN	Noun
JJ	Adjective
PR	Pronoun
VB	Verb
RB	Adverb
IN	Preposition
RP	Particle
FW	Foreign word
IJ	Interjection
.	Punctuation

2.3 Second level of tags: Determiner's case

Malagasy Language is composed of two types of determiner [1]:

- Definite determiner: **ny, ilay, ikala**
- Nominal determiner: **i, Ra, An, ry**

Table -2: Second level of tags for the determiner

Second level of tags	Description	Example
DTDEF	Defined Determiner	**Ny** lanitra sy **ny** tany **The** sky and **the** earth.
DTNOM	Nominal Determiner	Tsy tonga **ilay** mpandrafitra **The** mason hasn't come.

2.4 Second level of tags: noun's case

Malagasy Language is composed of two kinds of name: the proper noun and the common noun. The proper name characteristics remain the same as that of any other language such as French, English, etc. Nonetheless, the common name's have its particularity.

Table -3: Second level of tag for the noun

Second level of tags	Description	Example
NNP	Proper noun	Mahay ny **Barea** The **Barea** are talented.
NN	Common noun	Apetraho ny **penina** Put your **pen** down.
NNDVB	Common noun derived from the verb	Milalao ny **mpianatra** The **students** are playing
NNDJJ	Common noun derived from the adjective	**Hatsaram**-panahin' olona A person's **goodness** of soul.

2.5 Second level of tags: adjective's case

There are five kinds of adjectives in Malagasy Language: the adjective qualifier, the numerical adjective, the interrogative adjective, the demonstrative adjective, the indefinite adjective. The possessive adjective does not exist in Malagasy, it is supplemented by the personal pronoun or suffix: **ko (o), nao (ao)**, **ny** ...; instead of saying: my house, we say in Malagasy: *ny tranoko* (the house of mine)[1].

Table -4: Second level of tag for the adjective

Second level of tags	Description	Example
JJ	Qualifier Adjective	Trano **kely** A **little** house
JJCD	Numerical adjective	Mpianatra **roa** ihany no afaka Only **two** students succeeded
JJI	Interrogative adjective	Olona iza io ? **Who** is this man**?**
JJD	Demonstrative adjective	Omeo ahy **itsy** boky **itsy** Give me **that** book.
JJIND	Indefinite adjective	Olona **maro** no tonga **Many** persons came here.

2.6 Second level of tags: pronoun's case

There are six kinds of pronouns in Malagasy Language: the personal pronoun, the demonstrative pronoun, the relative pronoun, the interrogative pronoun and the indefinite pronoun.

Table -5: Second level of tag for the pronoun

Second level of tags	Description	Example
PRP	Personal pronoun	Nahoana **aho** no hihemotra ? Why would **I** fall back?
PRD	Demonstrative pronoun	Aza manao **izao** ! Do not do **it**!
PRR	Relative pronoun	Hosazina **izay** tsy manaraka ny lalàna We will punish **those** who violate laws
PRI	Interrogative pronoun	**Iza** no nilaza an'izany ? **Who** said that?

2.7 Second level of tags: verb's case

According to traditional grammar, the verb is a word that expresses the process, which means the action the subject does or undergoes [2].

There are three voices in Malagasy:

- The active voice where the subject is the agent of the action.
- The passive voice where the subject is the object of the action
- The relative or circumstantial voice where the subject is circumstance of the action

Table -6: Second level of tag for the verb

Second level of tags	Description	Example
VBA	Active verb	**Manarona** ny sakafo amin'ny lovia **aho**? **I cover** the meal with the plate.
VBP	Passive verb	**Saronako** amin'ny lovia ny sakafo The meal is **covered** with the plate.
VBR	Relative verb	**Anaronako** ny sakafo ny **lovia** The **plate** serves **me** to **cover** the meal
VBN	Participle	Faly aho rehefa **tafiditra** ny trano **Once back** home I feel happy.

2.8 Second level of tags: adverb's case

In Malagasy grammar, there are several adverbs which are:

- Adverbs of time which express either time, duration, or frequency;
- Adverbs of place which specify ordinary places or even demonstrative;
- Adverbs of order which express ordering;
- Adverbs of number;
- Adverbs of affirmation which express affirmation;
- Adverbs of negation;
- Adverbs of interrogation;
- Adverbs of doubt;
- Adverbs of quantity;
- Adverbs of opposition;
- Adverbs of defense

Table -7: Second level of tag for the adverb

Second level of tags	Description	Example
RBT	Adverb of time	**Taloha** mora ny fiainana **Once** life was easy
RBP	Adverb of place	**Mifanatrika** ny tsena ny tranonay Our house is **in front of** the shop.
RBO	Adverb of order	**Voalohany** mila voaloha ny karamany **First of all,** our wages must be paid.
RBN	Adverb of number	Miditra **tsiroroa** ny mpianatra Student enter **two by two**
RBM	Adverb of manner	Mila ovaina **tsikelikely** ny toetsaintsika ratsy We haveto change our way of thinking **little by little.**
RBA	Adverb of affirmation	**Eny** ramose ! **Yes** sir **!**
RBC	Adverb of negation	**Tsy mbola** tonga ny ekipa nationaly National team has**n't** arrived **yet**.
RBI	Adverb of interrogation	**Aiza** no nametrahany ny peninany? **Where** did he put his pen?
RBD	Adverb of doubt	Mety tara **angamba** izy He **might** be late.
RBQ	Adverb of quantity	Raha dinihina **kely** ny nataony If we take a look **a little more** on what he has done.
RBPR	Adverb of defense	**Aoka izay** ! **Stop it** !

2.9 Second level of tags: preposition's case

In the Malagasy language, prepositions are classified according to the relationship between the word to be completed and the complement[1].

Table -8: Second level of tag for the adverb

Second level of tags	Description	Example
IN	Preposition for a possession	Trano**n'**i Rakoto The house **of** Rakoto
INA	Preposition for an attribute	Miasa **ho an'ny** vahoaka izahay We work **for** the people
INP	Preposition for a place	Eo ankavanan'ilay lehilahy no misy azy She is located **on the right of** this man.
INT	Preposition for a time	**Mandritra** ny lanonana **During** the event.
INM	Preposition for a manner	Tsy fantatra mazava tsara ny **momba** an'ilay zaza very We do not know much **about the** missing child.
INR	Preposition for a reason	Tsy niasa izy **noho** ny antony ara-pahasalamana He didn't attend work **because of** health issues.
INW	Preposition for a mean	Mipetraka **amin'ny** rainy izy He lives **with** his father.
INC	Preposition for a comparison	**Toa** ireny mpanakanto ireny ianao You **look like** one of those artists.
INE	Preposition for an exception	Nalainy daholo ny entany **ankoatra** ireto akanjo ireto He took all his stuff with him **except** those clothes.

2.10 Combination of tags: the suffix or linked personal pronouns case

Here are the rules of suffixation of these pronouns to the words of which they are complements:

- If the word does not end by **KA, TRA, NA** we add the pronouns **ko, nao, ny, ntsika, nay,nareo and ny**.
- If the word ends by **NA**, we look for this final **NA** and we add the same forms of pronouns as before.
- If the word is terminated by **KA**, **TRA** we use the forms **o, ao, ny, tsika, ay, areo, ny** so that we drop the A of **KA**, **TRA** in front of the vowels **o, eo, ay, areo** , and the whole syllables **KA**, **TRA** in front of the consonants **ny, tsika**

The table below shows the application of these rules:

Table -9: Example of rule of suffixation

	Word who doesn't end with ka, tra or na	**Word who end with na**	**Word who end with ka**	**Word who end with tra**
Words	Loha / Head Loha**ko** / **my** head Loha**nao** / **your** head Loha**ny** / **his** or **her** head Loha**ntsika** / **our** head Loha**nay** / **our** head Loha**nareo** / **your** head Loha**ny** / **their** head	Tana**na /** hand Tana**ko** Tana**nao** Tana**ny** Tana**ntsika** Tana**nay** Tana**nareo** Tana**ny**	Soroka / shoulder Soro**ko** Soro**kao** Soro**ny** Soro**ntsika** Soro**kay** Soro**kareo** Soro**ny**	Tongotra / feet Tongotr**o** Tongotr**ao** Tongo**ny** Tongo**tsika** Tongotr**ay** Tongotrar**eo** Tongo**ny**
Tags	**NNPRP**	**NNPRP**	**NNPRP**	**NNPRP**

2.11 Combination of tags: The case of the preposition with possession report

In English, a case in point is: The driver's house, in Malagasy *tranon'ny mpamily*

- *tranon'ny* is composed of the name **trano** and the preposition **n'**
- the combination therefore gives as tag: **NNIN**

2.12 Combination of tags: The case of the derived name and the determiner

In English, a case in point is: Christ's disciple, in Malagasy *Mpianatr'i Kristy*

- *Mpianatr'i* is composed of the noun derived from the verb **mianatra** and the nominal determiner **i**
- the combination gives the tag then: **NNDVBDTNOM**

2.13 Combination of tags: The case of the participle and the personal pronoun

In English, a case in point is: At last, my work is done, in Malagasy *Vitako ihany ny asako* /

- *Vitako* is composed of the verb **Vita** and the personal pronoun **ko**,
- The combination gives the tag: **VBNPRP**

3. PRESENTATION OF THE LEARNING CORPUS

The learning corpus includes 6 468 grammatical elements provided from newspaper articles, advertisements and dialogues taken from Malagasy forums.

Table -8: Learning corpus characteristics

Grammatical elements	Frequency
Noun	1636
Adverb	732
Verb	742
Determiner	592
Punctuation	655
Adjective	618
Preposition	337
Particle	369
Conjunction	380
Pronoun	191
Foreign word	150
Number	66

4. OPERATION OF THE MALAGASY PART-OF-SPEECH TAGGING

A part-of-speech tagger is a computer program that is able to recognize the grammatical nature of a word and assign a tag to that word. For our research, we used the NLTK tool.
The program operation itself can be divided into two stages:

- The first step is the part-of-speech tagging that consists of using our learning corpus and implementing our algorithm on how to assign a tag to a word according to the level of the desired set of tags.
- The second step is to tag each word of the test corpus and give a score on the percentage of correctly tagged words.

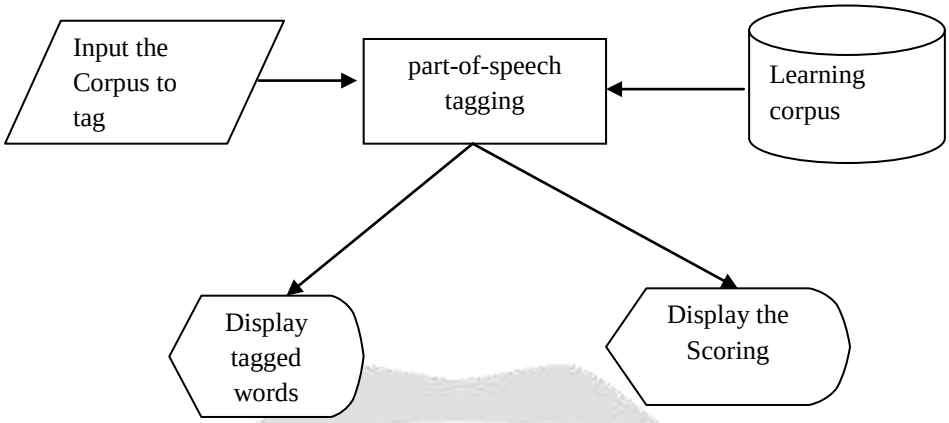

Fig-1: Our Part-of-speech tagger operation

With our learning corpus, there appear to be two major issues for our tagger:

- **An issue of precision:** our learning corpus is far from describing all the possible contexts of the Malagasy language, which means that even if the word is present in the corpus but understood in a context different from the corpus to be analyzed, it will be considered **unknown**.
- **An issue of coverage**: our learning corpus is very far from containing all the Malagasy words and the word order and tags possible, which means that a word or the word order that is not present in the learning corpus will be considered **unknown**.

Three models solve these problems but separately:

- The Unigram model, which solves most of the coverage problem but leaves a big flaw with regard to the precision problem. The reason is that this model relies on a simple statistical algorithm: for each word, it assigns the most probable tag for this word. For example, it attributes to the tag **JJ** any occurrence of the word *vaovao*, because *vaovao* is used more often as an adjective (for example, *hopitaly vaovao* / new hospital) than as a noun (for example, *Mijery vaovao* / Watch the news).
- The bigram model which resolves a part of the problem of precision and coverage. The reason is that this model is an improvement of the unigram model whose context is the current word associated with the tags of the word previously marked, so that we consider the previous word tagging in addition to the current word for tagging.
- The Hidden Markov Model (HMM) is the ideal model for maximizing accuracy in the case where our learning corpus is as rich and varied as Brown's corpus with 1,014,312 words, this model is a statistical mathematical model derived from Markov Chains, except that one cannot directly observe the sequence of states: the states are hidden. Each state issues "observations" which are observable. The purpose of the MMC is to associate with a sequence $w = w_1 ... w_i$ of words, a sequence $t = t_1 ... t_i$ of tags belonging to a set of tags.

Our tagger is based on a program that uses precision models, and then fills its coverage gaps with algorithms with more extensive coverage when needed. And for cases of unknown words, we have created a function that exploits the Malagasy grammar rules to recognize verbs or nouns.
The combination of these models works as follows:

- Use the tagger based on the NLTK Hidden Markov Model
- If the tagger based on the Hidden Markov Model cannot find a tag for the word, it refers to the tagger based on the NLTK bigram model.
- If the bigram tagger cannot find a tag for the word, it refers to the NLTK unigram tagger.
- If the unigram tagger is unable to find a tag for the word, it returns the unknown word to the function which tags the word with the first level of tag in verb or noun.

The flowchart below reflects this process:

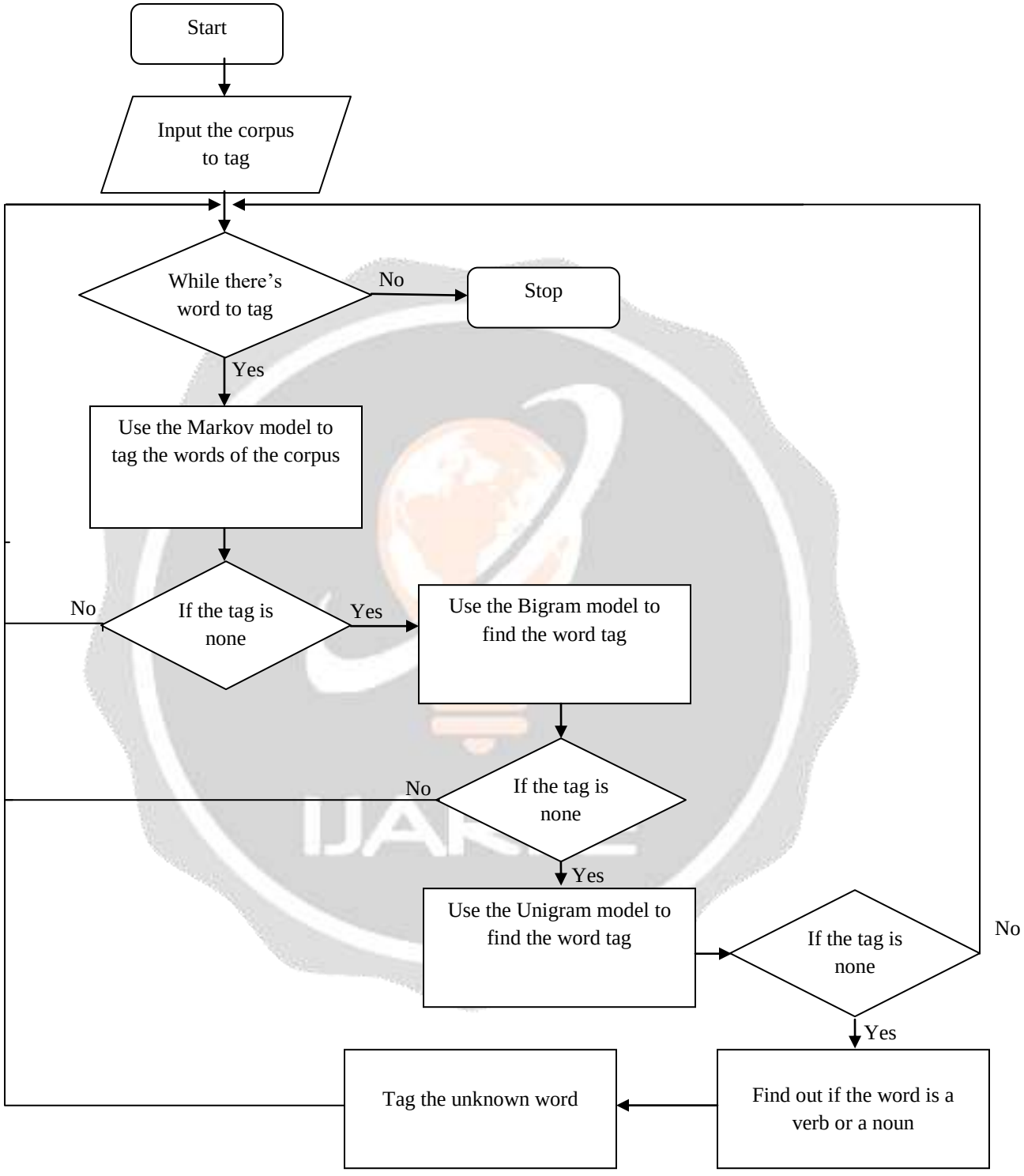

Fig-2: The tagging process flowchart

5. RESULT AND INTERPRETATION

In order to test the effectiveness of our part-of-speech tagging, we have decided to make a simulation on four types of corpus of different nature and show the results obtained with a special focus on the quantity of word contained in our learning corpus and the different mathematical models.

Table -9: Result obtained according to the words contained in our corpus

	Newspaper article 4000 words	**Advertising 2000 words**	**Blog 2700 words**	**Forum 1700 words**
Words contained in our learning corpus	3875	1889	2008	950

Table -10: Results obtained with some mathematical models

Tagger Precision	**Newspaper article 4000 words**	**Advertising 2000 words**	**Blog 2700 words**	**Forum 1700 words**
Correct tag with Unigram Model	78,68 %	73,45 %	64,25%	55,35 %
Correct tag with Bigram Model	45,56 %	40,36 %	37,02 %	28,13 %
Correct tag with Hidden Markov Model	5,04 %	6,02 %	2,56 %	0,02 %
Correct tag with our POS tagging	80,03 %	75,63 %	67,25 %	57,85 %

5.1. Interpretation according to the nature of the analyzed corpus

For the case of the newspaper articles, the analyzed corpus comes from the newspaper “Midi Madagascar” and the type of the treated information concerns the politics and various facts. The table shows that the newspaper articles present better results than the other corpus which reason might be:

- More than 3/5 of our training corpus is composed of newspaper article which makes the majority of the words that compose the corpus is already known.
- Newspaper articles are written in official Malagasy without abbreviations
- The authors of the articles use a sustained register when the newspaper deals with information on the policy, and a current register when dealing with the various facts, both are identical to the one applied in our learning corpus.

For the case of the advertisements, the analyzed corpus comes from various ads in the newspapers "Midi Madagascar". The results obtained do not differ too much from the results obtained with the articles. The reason might be:

- The use of official Malagasy
- The use of the same active verbs
- The non-respect of Malagasy grammatical structures because the purpose is to pass information quickly.
- The use of the current and sometimes familiar register
- Few use of technical term

For the case of the blogs, the corpus blog addresses to analyze young people's interest towards studies.
Those results might come from:

- Use of foreign words, usually French language but also English, to express technical terms
- The permanent use of a familiar register
- The blog deals with a particular topic, some words or vocabulary are not present in our learning corpus

For the case of the forums, the worst results are obtained with the forums in a Malagasy discussion group where the reasons might be:

- The amalgamation of Malagasy and French language
- The use of other Malagasy dialects except the official Malagasy
- The use of abbreviations and SMS languages

5.2. Interpretation according to the mathematical model applied

- The Unigram model shows high success rate in the chart which might be explained by the fact that the majority of the words that make up the different test corpus are contained in our learning corpus. The reason why the precision does not match lies in the numerous grammatical natures of a same word in Malagasy Language. Indeed, its nature depends on the sentence in which it is used.

- The Bigram model has a below average result which reason lies in the several concept (potential tag sequence) that are not present in our learning corpus.

- The Markov model presents the least satisfactory results which main cause lies in the incorrect tagging of unknown words. The Hidden Markov model is a statistical model which means that a mistagged word leads to a chain reaction and thus, all the following word / tag combination will be distorted.

- As far as our POS tagging is concerned, it was possible to obtain a result superior to the three models mentioned above. This is because our combination technique proves to be more efficient than the three individual models.

In order to obtain a more relevant interpretation through the mathematical model, we propose an example that accurately illustrates the difference between the results obtained for each model used.
The corpus to tag: *Ny Polisy, ny Bianco ary ny vahoaka*
In our example, the word "Bianco" is considered as an unknown word and is tagged with the tag **None**.

- Results obtained with the Unigram model:

[(u'Ny', u'**DTDEF**'), (u'Polisy', u'**NN**'), (u',', u'**.**'), (u'ny', u'**DTDEF**'), (u'Bianco', **None**), (u'ary', u'**CC**'), (u'ny', u'**DTDEF**'), (u'vahoaka', u'**NN**')]

As far as the Unigram model is concerned, despite the fact that this word "Bianco" is unknown, the remaining words are tagged correctly.

- Result obtained with the Bigram model:

[(u'Ny', u'**DTDEF**'), (u'Polisy', u'**NN**'), (u',', u'**.**'), (u'ny', u'**DTDEF**'), (u'Bianco', **None**), (u'ary', **None**), (u'ny', **u'DTDEF**'), (u'vahoaka', **u'NN**')]

Then for the Bigram Model, only the determiner followed by the name are tagged correctly because this sequence of tag exists in our learning corpora.

- Result obtained with the Hidden Markov model:

[(u'Ny', **u'DTDEF**'), (u'Polisy', **u'NN**'), (u',', u'**.**'), (u'ny', **u'DTDEF**'), (u'Bianco', **u'DTDEF**'), (u'ary', **u'DTDEF**'), (u'ny', **u'DTDEF**'), (u'vahoaka', **u'DTDEF**')]

Concerning the hidden Markov model, all the word that follows the unknown word “Bianco” is tagged with the tag **DTDEF** due to its recurrence in the corpus learning.

- Result obtained with our POS tagging:

[(u'Ny', u'**DTDEF**'), (u'Polisy', u'**NN**'), (u',', u'**.**'), (u'ny', u'**DTDEF**'), (u'Bianco', **NN**), (u'ary', u'**CC**'), (u'ny', u'**DTDEF**'), (u'vahoaka', u'**NN**')]

Due to our algorithm, the word “Bianco” is tagged as a noun and the rest of the word is tagged correctly.

6. CONCLUSIONS

In a nutshell, we can say that the accuracy of our POS tagging depends mainly on the right balance between the coverage of our learning corpus and the accuracy of the mathematical models used to deduce the nature of a word according to its place in the corpus.

As a perspective of future study, an improvement of the learning corpus could be set up. This improvement will carry a more assorted corpus in addition to the use of algorithms to recognize the SMS languages and the improvement of the algorithms of recognition of an extended grammatical item.

7. REFERENCES

[1]. Régis RAJEMISA-RAOLISON. GRAMMAIRE MALGACHE. Fianarantsoa 1971

[2]. Joro Ny Aina RANAIVOARISON. Modélisation de la morphosyntaxe du malgache : Construction d'un dictionnaire électronique des verbes simples. Université d'Antananarivo, 2014

[3]. Abdelhamid EL JIHAD, Abdellah YOUSFI. Etiquetage morpho-syntaxique des textes arabes par modèle de Markov caché. Université Mohamed V, Rabat, Maroc, 2005

[4]. Kristina Toutanova and Christopher D. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), 200