A statistical model for morphology inspired by the Amis language

Isabelle Bril*
Lacito-CNRS

Achraf Lassoued
University Paris II

Michel de Rougemont
University of Paris II and IRIF-CNRS

Abstract

We introduce a statistical model for analysing the morphology of natural languages based on their affixes. The model was inspired by the analysis of *Amis*, an Austronesian language with a rich morphology. As words contain a root and potential affixes, we associate three vectors with each word: one for the root, one for the prefixes, and one for the suffixes. The morphology captures semantic notions and we show how to approximately predict some of them, for example the type of simple sentences using prefixes and suffixes only. We then define a *Sentence vector s* associated with each sentence, built from the prefixes and suffixes of the sentence and show how to approximately predict a derivation tree in a grammar.

1 Introduction

The representation of words as vectors of small dimension, introduced by the Word2vec system [1], is based on the correlation of occurrences of two words in the same sentence, or the second moment of the distribution of words[1]. It is classically applied to predict a missing word in a sentence, to detect an odd word in a list of words and for many other predicting tasks. Computational linguists also studied how to extend the vector representation of words to a vector representation of sentences, using the *Recurrent Neural Networks* [2, 3] and the attention mechanisms [4].

Words have an internal structure, also called morphology. The word *preexisting*, for example, has a prefix *pre-*, a root *exist* and a suffix *-ing*. In this case, we write *pre-exist-ing* to distinguish these three components. Given some texts, we can analyse the frequency distribution of affixes. Given a root, the same analysis can be conducted. We call these statistical distributions the *Morphology Statistics* of the language. There are similar approaches such as [5] and [6] to enrich the word embedding with character and subword information. In this paper we focus only on some prefixes and suffixes.

We consider the first and second moment of the *Morphology Statistics*. We can then determine which prefix is the most likely in a missing word of a sentence, which suffix is unlikely given a prefix and a sentence, and the syntactic structure of simple sentences. We argue that these statistics are very useful to approximately capture some key semantic and syntactic parameters. *Amis*, a natural language with a profuse morphology is well suited for this analysis. This approach can be applied to any other language in order to capture some key semantic notions such as Tense, Voice, Mood, Illocutionary force and other discourse notions.

*This research is financed by the "Typology and dynamics of linguistic systems" strand of the Labex EFL (Empirical Foundations of Linguistics) (ANR-10-LABX-0083/CGI). A preliminary version appeared in the *Language,Ontology, Terminology, and Knowledge Structures Workshop (LOTKS), 2017*.

[1]The second moment is the distribution of pairs of words, the third moment is the distribution of triples of words and the k-th moment is the distribution of k words in the same sentence.

DOI: 10.5121/ijwest.2022.13201

Amis is one of the fifteen surviving Austronesian languages spoken in Taiwan. *Amis* is spoken along the eastern coast of Taiwan and has four main dialects with significant differences in their phonology, lexicon and morphosyntactic properties. The analysis bears on Northern Amis; the data were collected during fieldwork. Some simple sentences are presented in appendix A, where affixes, infixes and clitics are denoted by special symbols.

We built a tool to represent the statistical morphology of *Amis*, given a set of texts where each word has been decomposed into components (i.e. prefix, infix, root and suffix). The tool is similar to the OLAP (Online Analytical Processing) Analysis used for Data Analysis.

- We analyse the global distribution of roots and affixes, i.e. the most frequent occurrences.
- Given a root (or a prefix, or a suffix), we obtain the distribution of the pairs (Prefixes;Suffixes) of that root, and the distribution of the prefixes, or the distribution of the suffixes by projection. Similarly for a given prefix, or a given suffix.

We then study the second moment of the *Morphology Statistics* and are able to predict the most likely prefix, root or suffix given a sequence of words. In *Amis*, as affixes carry some semantic and syntactic information, we construct a *Sentence vector* which we use to choose the preferred parsing of a sentence. We propose:

- A statistical representation of words, made of prefixes, roots and suffixes, as structured vectors.
- A vector representation for a sentence, the *Sentence vector* which captures general properties of the sentence. We show its use to approximately predict the syntactic class of a sentence and the most likely derivation tree.

In the next section, we introduce the basic concepts. In the third section, we present our statistical model to capture the morphology of a natural language and apply it to *Amis*. In the fourth section, we give a syntactic outline of Amis. In the fifth section, we study how to approximately predict the syntactic class of a simple sentence. In the sixth section, we describe which derivation tree is preferable for a given grammar.

2 Preliminaries

We review some basic statistics in the context of natural languages in section 2.1 and the *Amis* language in section 2.2.

2.1 Basic Statistics

Let $s = w_1.w_2...w_n$ be a text with the words w_i on some alphabet Σ. Let $\mathsf{ustat}(s)$ be the *uniform statistics*, also called the 1-gram vector of the sentence s. It is a vector whose dimension is the size of the dictionary, the number of distinct words. The value $\mathsf{ustat}(s)[w]$ is $\#w$ the number of occurrences of w divided by n, the total number of occurrences.

$$\mathsf{ustat}(s) = \frac{1}{n}. \begin{pmatrix} \#w_1 \\ \#w_2 \\ ... \\ \#w_m \end{pmatrix}$$

We can also interpret $\mathsf{ustat}(s)$ as the distribution over the words w_i observed on a random position in a text. When the context is clear, we may also display the absolute values as opposed to the relative

values of the distribution. Variations of these distributions are used in Computational Linguistics [7, 8].

Suppose we take a random sentence, then two random positions i, j in that sentence and define the $\mathsf{ustat}^2(s)$ vector as the density of the pairs (w_i, w_j). It would be the second moment of the distribution of the words. For simplicity, we consider the symmetric covariance matrix $M(w_i, w_j)$ which gives the number of occurrences of the pair (w_i, w_j), i.e. without order. One can view the covariance matrix as the probability to observe a pair of words in a sentence and the diagonal values of the matrix give the first moment.

Given a (n, n) covariance matrix, one can associate a vector of v_i dimension n to each w_i such that the dot product $v_i.v_j$ is equal to $M(w_i, w_j)$. If we only select the large eigenvalues of M, we can obtain vectors of smaller dimension such that $w_i.w_j \simeq M(w_i, w_j)$. This PCA (Principal Component Analysis) method goes back to the 1960s, and may use the SVD (Singular Value) Decomposition of the (n, n) matrix with a $O(n^3)$ time complexity. In [1], a more efficient learning technique is used to obtain vectors of dimension 200 when the dictionary has $n = 10^4$ words, i.e. reducing the dimension. In this paper, we apply the classical technique to the covariance matrices of prefixes and suffixes. As we observe 30 distinct prefixes and 10 distinct suffixes, a direct SVD decomposition is efficient as n is small.

2.2 The Amis language

A fundamental property of Amis is that roots[2] are most generally underspecified and categorially neutral [9]; they are fully categorised (as nouns, verbs, modifiers, etc.) after being derived and inflected as morphosyntactic word forms and projected in a clause.

Primary derivation operates on roots and is basically category attributing; it derives noun stems and verb stems. Noun stems are flagged by the noun marker *u* or by demonstratives. Verb stems display voice affixes, the basic ones are the Actor Voice *mi-* (AV), the Undergoer Voice *ma-* (UV), the passive voice *-en*, the Locative voice *-an*. [10].

Secondary derivation occurs on primarily derived verb stems: (i) operating category-changing derivation (i.e. deverbal nouns, modifiers, etc.). (ii) deriving applicative voices[3] (Instrumental *sa-*, and Conveyance *si-*). For instance, *mi-* verb stems are derived as instrumental *sa-pi-* forms, *ma-* verb stems are derived as instrumental *sa-ka-* forms.

A syntactic outline is given in section 4.

3 A statistical model for morphology

Given numerous transcribed oral texts, we first built a tool which constructs the distribution of prefixes, suffixes and roots, i.e. the number of occurrences. Given a root, we can display the distribution of its affixes. Similarly, we can give a prefix (respectively a suffix) and represent the distribution of roots and suffixes (resp. prefixes). We then consider the second moment distributions of prefixes, suffixes and roots and build their vector representations. Our statistical model combines these three representations for each word composed of prefixes, a root and a suffix. It also introduces a vector for sentences.

[2] A *root* is an atomic word without affixes. Affixes are either inflectional (i.e. express a semantic or syntactic function), or derivational (i.e. create different categories).

[3] With applicative voices, the promoted non-core term (i.e. locative, instrumental, conveyed entity) becomes the nominative subject of the derived verb form, with the same syntactic alignment as Undergoer Voice.

3.1 Basic Statistics for the Amis language

The distribution of all the prefixes and suffixes, given 70 Amis texts with more than 4000 words, is given in Figure 1. All the charts use absolute values. The *Morphix* tool provides an interface where

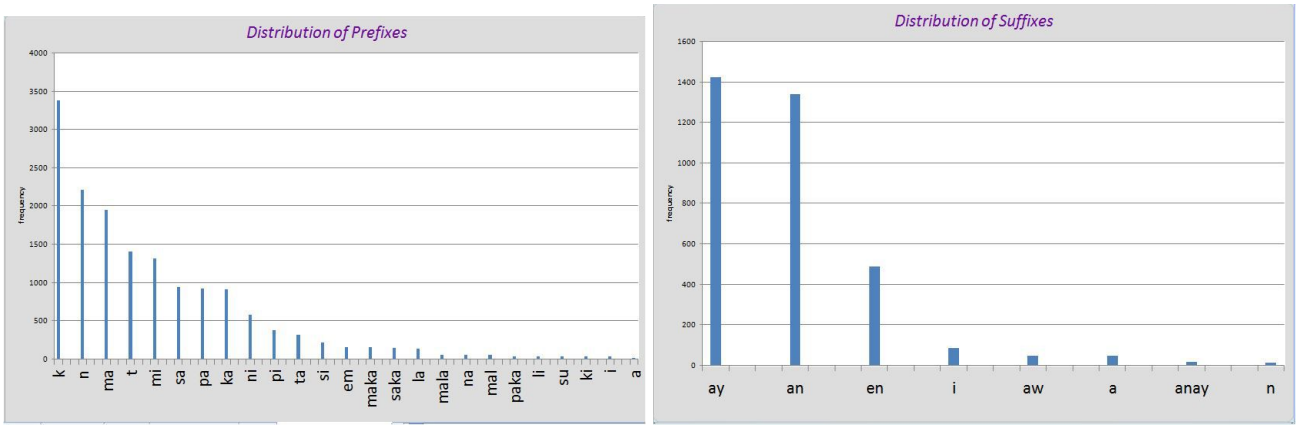

Figure 1: Most frequent prefixes and suffixes.

a root (resp. a prefix or a suffix) can be selected; the distribution of prefixes and suffixes for a given root are graphically displayed, as in Figure 2.

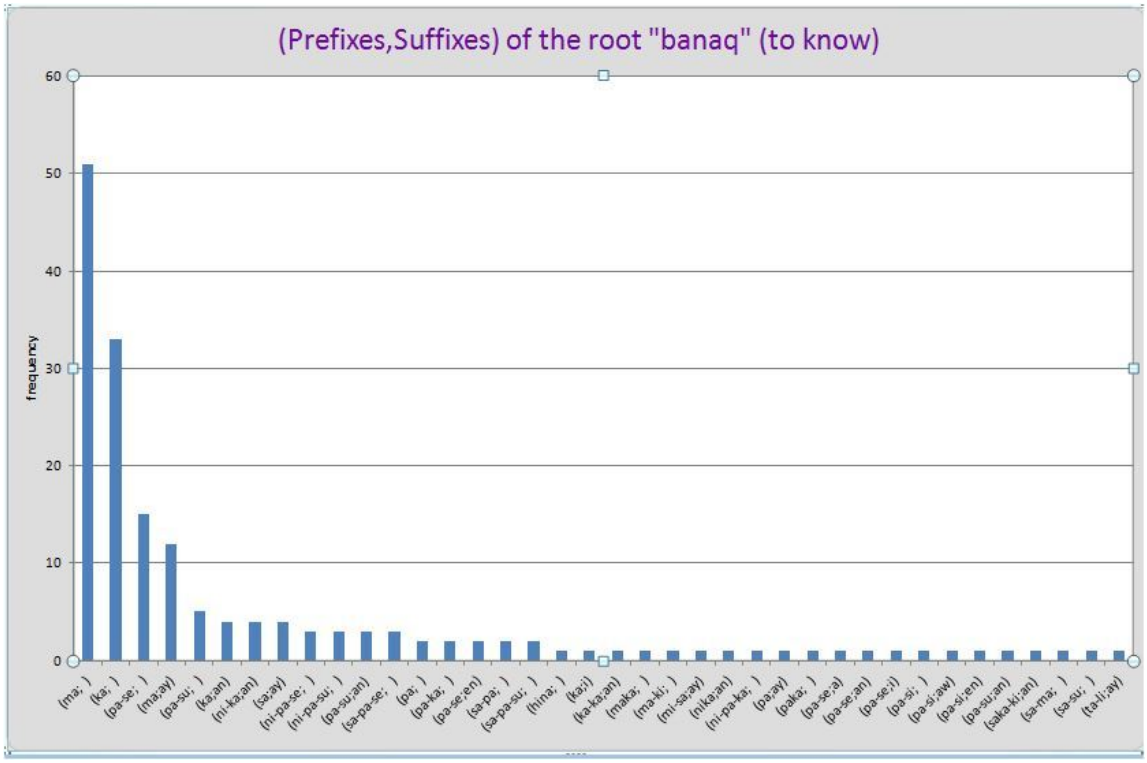

Figure 2: Most frequent (prefixes;suffixes) of the root *banaq* (' know').

Given the distribution of (prefixes;suffixes)[4] of Figure 2, which at this stage are unordered according

[4]A word can have several prefixes and suffixes. In Figure 2, the most frequent pairs (prefixes;suffixes) are *(ma-;)*, i.e. the prefix *ma-* with no suffix, *(ka-;)*, i.e. the prefix *ka-* with no suffix, *(pa-se-;)*, i.e. the two prefixes *pa-* and *se-* with no suffix and *(ma;ay)*, i.e. the prefix *ma-* with the suffix *-ay*.

to root forms and functions, we obtain by projection the distribution of prefixes and suffixes in Figure 3 for this specific root.

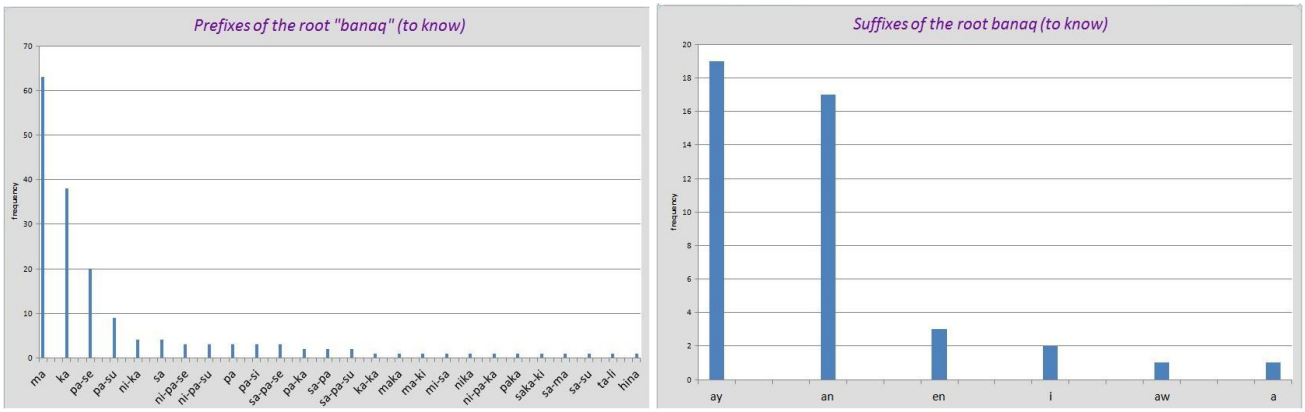

Figure 3: Most frequent prefixes and suffixes of the root *banaq*.

3.2 Vector representation of prefixes, roots and suffixes

Given a list of n prefixes, the (n, n) correlation matrix M measures the number of co-occurrences of the prefixes in the same sentence: $M(i, j)$ is the number of occurrences of the prefixes i and j in the same sentence and $M(i, i)$ is the number of occurrences of the prefix i. As M is a positive semi-definite matrix, there exists a matrix an (n, n) U such that $M = U.U^t$. We can interpret each prefix i as a vector v_i of dimension n such that $v_i.v_j = M(v_i, v_j)$. The classical *PCA* (Principal Component Analysis) allows to reduce the dimension of these vectors, along their principal components defined by the large eigenvalues such that $v_i.v_j \simeq M(v_i, v_j)$. The SVD (Singular Value decomposition) produces the eigenvectors (matrix S) and the eigenvalues λ_i (diagonal matrix V). If we project the eigenvectors on the dimensions defined by the large eigenvalues, we reduce the dimension.

Consider the following 4 structured Amis sentences with the root *padang* 'help, support'[5]:

1. *Mi-padang k-u tumuk t-u suwal n-ira tatakulaq.*
   AV-help NOM-ART chief OBL-ART word GEN-that frog[6]
   The tumuk supported the words of the frog.

2. *Isu Kungcu, yu ira k-u pa-padang-an,...*
   You Princess when exist NOM-ART RED-help-NMZ
   You Princess, when (you) had some help,...

3. *Sulinay mi-padang k-u taw,...*
   Indeed AV-help NOM-ART people
   Indeed when people help,...

4. *Aka-a ka-pawan t-u ni-padang-an n-u taw.*
   PROH-IMP NFIN-forget OBL-ART PFV.NMZ-help-NMZ GEN-ART people

[5]The first line is the original text where prefixes and suffixes are identified. The second line is the morphological analysis with labels such as AV, OBL,.... described below. The third line is the translation.

[6]Abbreviations: AV Actor Voice; ART article; CV conveyance voice; GEN genitive; IMP imperative; INST.V instrumental voice; LOC locative; LV locative voice; NFIN non-finite; NOM nominative; NMZ nominaliser; OBL oblique; PFV perfect; PROH prohibitive; RED reduplication; UV undergoer voice.

Then, you mustn't forget people's help.

In these four sentences, there are seven prefixes in the order: *k,ka,n,ni,mi,pa,t* and two suffixes *-an, -a*. The matrix M_p for these prefixes is:

$$M_p = \begin{array}{c} \\ k \\ ka \\ n \\ ni \\ mi \\ pa \\ t \end{array} \begin{array}{c} \begin{array}{ccccccc} k & ka & n & ni & mi & pa & t \end{array} \\ \begin{pmatrix} 2 & 0 & 1 & 0 & 2 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 2 & 1 & 1 & 0 & 2 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 2 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 2 & 1 & 1 & 0 & 2 \end{pmatrix} \end{array}$$

The first line of the Matrix indicates 2 occurrences of *k-*, 1 occurrence of the pair *(k-, n-)* (first sentence), 2 occurrence of the pair *(k-, mi-)* (first and third sentence), 1 occurrence of the pair *(k-, pa-)* (second sentence) and 1 occurrence of the pair *(k-, t-)* (first sentence). The large eigenvalues of M_p are 2.5 and 1.7. Two other eigenvalues are close to 1 and the three others are close to 0. If we decompose the vectors[7] on the two largest eigenvectors, we obtain 7 vectors of dimension 2, one for each prefix in the matrix below and in Figure 4.

$$B = \begin{array}{c} \\ k \\ ka \\ n \\ ni \\ mi \\ pa \\ t \end{array} \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{pmatrix} -1.07140232 & 1.05796864] \\ -0.61507482 & -0.69205651 \\ -1.35624875 & -0.37685711 \\ -0.61507482 & -0.69205651 \\ -0.89917084 & 0.67091885 \\ -0.19923754 & 0.51352862 \\ -1.34624875 & -0.36685711 \end{pmatrix} \end{array}$$

$B * B^t$ is approximately M_p. The first vector for *k-* has coordinates $-1.07, 1.06$. We represent graphically the 7 prefixes in Figure 4. A similar approach can be followed for suffixes and for the roots. Figure 4 can be used to predict, given a prefix v, the most likely next prefix v_{next}. It is the vector v' which maximizes the dot product $|v.v'|$. For example, given the vector for the prefix *mi-*, the most likely next prefix is *k-*, in these 4 sentences. For the 95 simple sentences of the corpus, the vectors would be slightly different.

3.3 Distributions and representative vectors

Let δ be the distribution of the most frequent words, δ_P the distribution of the most frequent prefixes (resp. δ_R the distribution of the roots) and let π_p be the mapping which associates the prefix of a word. For example, π_p(*mi-padang*)=*mi-*. Similarly π_r associates the root of a word, π_r(*mi-padang*)=*padang*. These distributions are related, mostly by projections. Let $\pi_p(\delta)$ be the distribution such that

$$\pi_p(\delta)(p) = \sum_{p \text{ is a prefix of } w} \delta(w)$$

Then $\delta_P = \pi_p(\delta)$ and $\delta_R = \pi_r(\delta)$. Similarly for the other distributions. The correlation matrix M_p of the prefixes is also the projection of the correlation matrix M of the words, i.e. $M_p = \pi_p(M)$.

For each correlation matrix M_p, M_r, M_s for the prefixes, roots and suffixes, we apply the dimension reduction and obtain vectors v_p of dimension n_p for the prefixes, v_r of dimension n_r for the roots and v_s of dimension n_s for the suffixes. The important reduction is for M_r, whose dimension is as large as the size of the vocabulary. For M_p, M_s, the dimension is small and the reduction is less important. In the previous example, the dimension n_p of the prefixes is 7 without reduction and 2 with the reduction. Consider a word *w=pre-root-suf* with one prefix *pre*. Let v_p(*pre*) be the vector for the prefix *pre*, v_r(*root*) be the vector for the root *root* and v_s(*suf*) be the vector for the suffix *suf*. The vector $v(w)$ is the union of the three vectors:

[7] We a package for linear algebra in Python to obtain the SVD decomposition and the projections.

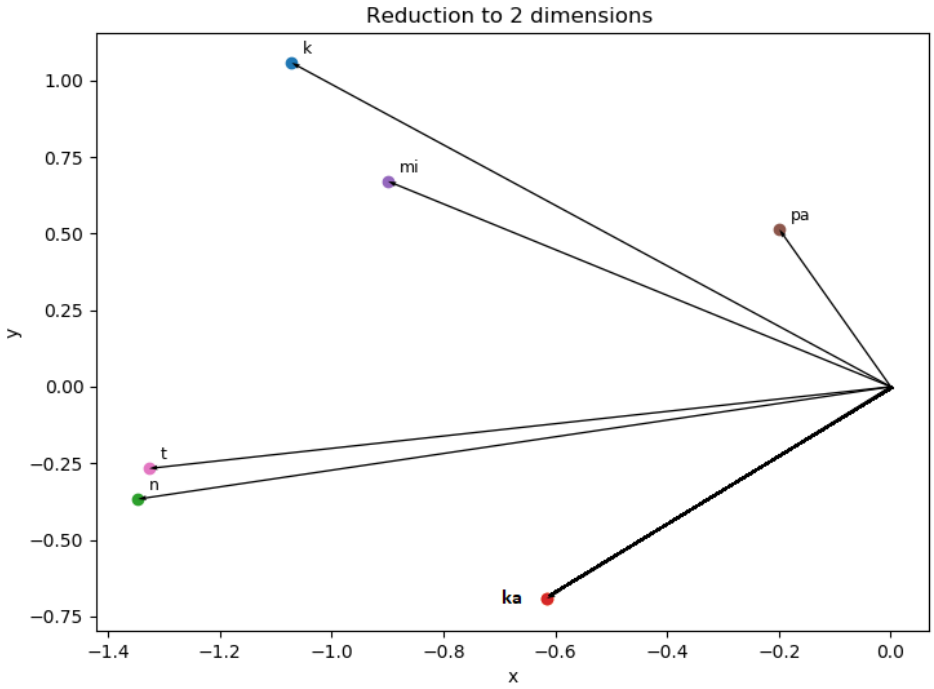

Figure 4: The vectors for the 7 most frequent prefixes *k-,ka-,n-,ni-,mi-,pa-,t-* in two dimensions.

$$v(w) = \begin{pmatrix} v_p(pre) \\ v_r(root) \\ v_s(suf) \end{pmatrix}$$

If the word has several prefixes, we enumerate them in the vector $v(w)$ and increase the dimension. For two words w_i, w_j with one prefix, let $\widetilde{M}(w_i, w_j) = M_p(pre_i, pre_j) + M_r(root_i, root_j) + M_p(suf_i, suf_j)$ be the sum of the correlations of the prefixes, roots and suffixes. The fundamental fact is that for any two words w_i, w_j, the dot product $v(w_i).v(w_j) \simeq \widetilde{M}(w_i, w_j)$. Indeed, $v(w_i).v(w_j) = v_p(pre_i).v_p(pre_j) + v_{r,root_i}.v_{r,root_j} + v_{s,suf_i}.v_{s,suf_j}$. The dot product $v_p(pre_i).v_p(pre_j)$ approximates $M_p(pre_i, pre_j)$ and similarly for the roots and suffixes. Hence $v(w_i).v(w_j) \simeq \widetilde{M}(w_i, w_j)$. For words with several prefixes, this approximation can be generalized.

Notice that $\widetilde{M}(w_i, w_j)$ can be very different from $M(w_i, w_j)$. It is possible that $M(w_i, w_j) = 0$, but that its prefixes, suffixes and roots have strong correlations, hence $\widetilde{M}(w_i, w_j)$ can be large. A rich theory of these structured vectors can be developed using cross-correlations, which we do not use at this point.

3.4 Vectors for Sentences

An important topic is to extend the vectors from words to sentences and more generally to paragraphs and texts. *Recurrent Neural Networks* [2, 11] or other learning based approaches [3, 12] construct such a vector of latent variables. We follow a different approach as we want to capture more general properties of sentences which can be captured by some prefix and some suffix. Let us define the probabilistic *Sentence* vector s of a sentence as a vector of dimension 5 whose components are distributions over some finite domains:

- Valence: $\{0, 1, 2, 3\}$,
- Voice: {AV, UV, LV, INST.V},
- Tense: {Present, Past, Future},

- Mood: {Indicative, Imperative, Hortative, Subjunctive},
- Illocutionary Force: {Declarative, Negative, Exclamative},

The first component s^1, the *Valence*, is the number of arguments of the verb. We view it as a distribution over the domain $\{0, 1, 2, 3\}$. It has 4 values, hence a vector of size 4 such that the sum of the values is 1. Similarly for the other components, and the vector s of dimension 5 has a global size of $18 = 4 + 4 + 3 + 4 + 3$. More dimensions could be used, but we keep the vector short for this example. For example, if the third component s^3 over {Present, Past, Future} is $[0, 1, 0]$, it indicates a PAST (with probability 1). If it were $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$, it would indicate a uniform distribution. We read the sentence $w_1, w_2, ...w_n$, and the vector $v_i = v(w_i)$ associated with each word w_i, as defined inthe previous section. Let:

$$s_i = F(s_{i-1}, v_i)$$

with s_0 is an initial state (zero vector) and F is a function. We construct F by cases following the prefixes and suffixes. As an example, consider the following sentence:

5. *Tengil-i* *isu* *k-aku !*
   hear-IMP.UV GEN.2sg NOM-1sg
   Listen to me ! (lit. let me be listened to by you)

The suffix *-i* expresses the imperative mood in Undergoer Voice. The suffix thus carries specific syntactic and semantic instructions, such as mood and UV voice, which itself has a specific type of alignment (a nominative patient and a genitive agent). In this case s^2_i, the second component of F is defined as:

$$s^2_i(c_{i-1}, v_i) = \begin{cases} [0, 1, 0, 0] & \text{if } [v_i]_s = \text{"–i"} \\ s^2_{i-1} & \text{otherwise} \end{cases}$$

The vector s^2_i becomes a Dirac distribution after the suffix *–i*. In general, each component of F is built as a decision tree, with rules and possible learnt components. At the end of a sentence, we have the *Sentence vector* s_n. We describe more advanced syntactic rules of *Amis* in section 4.

4 A syntactic outline of Amis

The basic word order of Amis is predicate initial. Arguments are case-marked: nominative is marked by *k-*, the agent is marked as genitive by *n-*, themes and oblique arguments are marked by *t-* [10]. The voice affixes (AV) *mi-*, (NAV) *ma-*, (UV) *ma-*, also identify verb classes ; class membership can be exclusive or allow voice alternations with different semantics. Some stative property verb stems are prefix-less.

AV *mi-* verb stems denote activities or accomplishments. UV *ma-* verbs denote achievements carried out by an agent on specific or definite, and fully affected patients ; NAV *ma-* verbs include states and psych states, properties, verbs of cognition (*ma-banaq* 'know'), bodily functions, position and motion[8] (*ma-nanuwang* 'move'). See [9] for details.

The root's ontology and semantic features pair up with the semantic and syntactic properties of voice affixes. The voice system is thus based on the co-selection of a nominative argument (the subject), and a voice affix whose semantics matches the semantics of the nominative subject. The AV *mi-*, NAV *ma-* and UV *ma-* voices are restricted to declarative sentences. In non-declarative sentences (such as negative, imperative, hortative), they have different forms. Compare *ma-butiq cira* '(s)he is asleep/sleeping' and *ka-butiq!* 'go to sleep !'.

Transitivity and alignment. Alignment[9] is split. AV *mi-* verbs and NAV *ma-* verbs (Non-Actor Voice) have an "accusative-like" alignment with an argument marked by *t-* as in (6) and (8). The subject of *mi-* verbs

[8]Motion verbs are not activities despite their dynamic feature, but intradirective verbs.

[9]Alignment refers to the kind of case-marking of the verb's argument(s). In accusative languages, the nominative subject is case-marked differently from the accusative object. In ergative languages, the agent of transitive verbs is case-marked as ergative, while the subject of intransitive verbs and the patient of transitive verbs are case-marked similarly as nominative/absolutive.

is an Actor, while the subject of NAV *ma-* verbs is a Non-Actor (i.e. a theme or experiencer, the seat of some property or state). On the other hand, transitive UV *ma-* verbs have ergative alignment with a patient subject and a genitive agent as in (7). The word order (V-S-0 or V-S-A) is given for each example:

6. *Mi-melaw k-u wawa t-u tilibi.*
   AV-look NOM-ART child OBL-ART TV V-S-O
   ' The child is watching TV. '

7. *Ma-melaw n-uhni k-u teker.*
   UV-look GEN-3pl NOM-ART trap V-A-S
   'They saw the trap.'(lit. the trap was seen by him)

8. *Ma-hemek k-aku t-u babainay. (*mi-)*
   NAV-admire NOM-1sg OBL-ART boy V-S-O
   'I admire the boy.'

All other voices align ergatively like UV *ma-*, so UV *-en*, LV *-an*, INST.V *sa-*, BV (Beneficiary Voice) *si-*, i.e. they have a nominative subject matching the voice affix's semantics (i.e. a patient, location, instrument, beneficiary subject), and a genitive Agent (if expressed).

5 Sentence classification from prefixes and suffixes

The prefixes and suffixes occurring in a sentence provide some syntactic information. Here, we restrict the analysis to 95 simple sentences, mainly in AV and UV verb forms, given in appendix A. We encode each sentence as in Figure 5 below, with its prefixes, infixes, roots and suffixes. We then ask if it is possible to recover the syntax of a sentence by observing only the prefixes and suffixes. The lexical roots of words are replaced by a "—" as in Figure 5.

Sentences	**Sentences Encoding**
Ma-pa-tangasa n-umisu k-aku iri.	Ma-pa— n— k— —.
Adihay k-u ni-urung t-u kidudung.	— k— ni— t— —.

Figure 5: Sentence and their encodings

The *simple clauses* considered belong to one of the following 10 syntactic templates:

. *V-A-S-O.* Verb-Agent-Subject-Object
. *V-A-S.* Verb-Agent-Subject
. *V-A-O.* Verb-Agent-Object
. *V-A.* ...
. *V-S-O.*
. *V-S-A.*
. *V-O-S.*
. *V-S.*
. *V-O.*
. *V.*

All the 95 simple sentences have been labeled as belonging to one of the 10 classes. We now construct a simple decision tree which reads the *encoding of a sentence* and approximately predicts its class. The decision tree presented in Figure 7 first decides if the sentence contains a *ma–* or a *mi–* prefix or a *k–* prefix without *ma–* or *mi–*. It then observes the presence or absence of the case prefixes *n–*, *k–* and *t–*. A branch with the label *n–* assumes the prefix *n–* is present and with the label ¬*n–* assumes the prefix *n–* is absent. Each leaf of the tree is labeled with one of the syntactic classes.

The statistics which represents the frequency of these 10 different classes among the corpus of 95 sentences, are given in Figure 6.

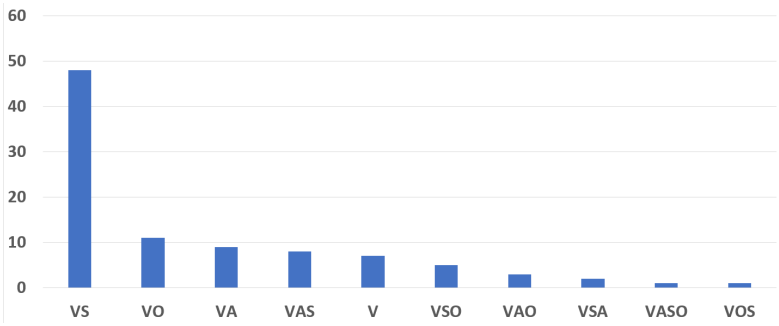

Figure 6: Frequencies of the different existing classes in the corpus

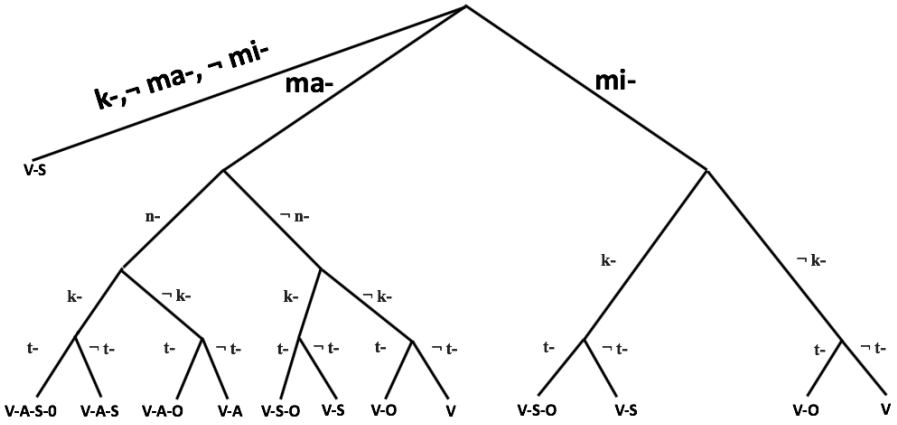

Figure 7: Decision Tree to predict the syntactic class

In the sentences of appendix A, there are 95 simple affirmative sentences and 4 negatives sentences. From the 95 affirmative sentences, 22 sentences do not display the prefixes *ma–* or *mi–* or *k–*, hence 73 sentences were classified by the decision tree. The prediction is considered *average* if the predicted class is at distance 1 for the Edit distance with the exact class, *bad* if the predicted class is at distance greater than 1 from the exact class. For example, *V-S-O* is an *average* prediction of the class *V-A-S-O* as it is at distance 1 and *V-S* is a *bad* prediction as it is at distance 2. We reached a good prediction in 25 cases, an average prediction in 42 cases, and a bad prediction in 6 cases. The prediction statistics over the 95 sentences are given in Figure 8.

This simple decision tree classifies 77% of the 95 sentences presented in the appendix A. Most of the time, the classification is of average quality: it only gives an approximate prediction.

5.1 Sentence vector and attention mechanisms

The Valence, the first component of the *Sentence vector* is encoded by the verb stem (see section 4). The other 4 components are partially defined by the prefixes and suffixes and are therefore probabilistic distributions. As the total number of prefixes and suffixes is small, the description of the function F is essentially a definition by cases.

Attention mechanisms [13, 4] provide for a given word w_i of a sentence, the distribution of the most correlated other words of a sentence. The correlation of a word w_j with w_i is approximately the value $v(w_i).v(w_j)$. We can then compute $\{v(w_i).v(w_j) : j \neq i\}$ for a fixed w_i and normalize the values (or with a Softmax function)

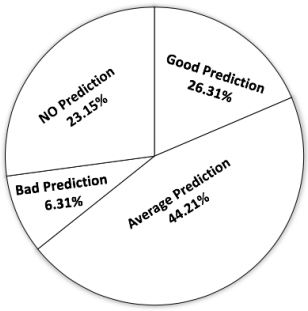

Figure 8: Statistics of the predictions

to obtain a distribution. We can also specialize on the prefixes and the suffixes and similarly define the prefix and suffix attention. In the example of figure 9, consider the sentence: *Mi-melaw k-u wawa t-u tilibi,* with the prefixes *mi–*, *k–*, *t–*. Let us compute the attention for the prefix *mi–*. We compute $v_p(mi).v_p(k) = 1,149$ and $v_p(mi).v_p(t) = 1,446$, using the vectors given in the matrix B of section 3.2. With a normalization, we obtain the distribution in Figure 9. We could extend the *Sentence vector* with this information.

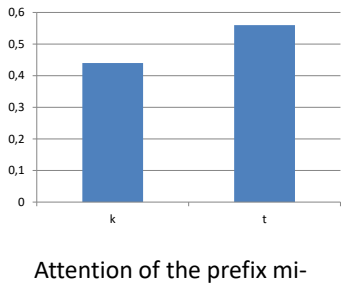

Figure 9: Attention distribution for *mi–* in the sentence *Mi-melaw k-u wawa t-u tilibi.*

6 Grammars and statistics

We now investigate the impact of the prefixes and suffixes of a sentence on the possible derivation trees defined by parsers in classical grammars. Given a grammar and a sentence, which derivation trees are the most likely? It is a central issue for any parser, including dependency parsers [14] built by learning algorithms from labeled data. In this case, the distribution of the examples defines a probabilistic space.

This is a fundamental problem, from RNA folding in biology to computational language understanding. An RNA folding is a possible derivation tree with a potential energy and the most likely folding has a minimum energy level. In our context, we have various statistics on prefixes and suffixes, and the Entropy of each distribution, obtained from the linguistic data. We just lack energy levels for each tree and turn to the linguistic interpretation, first given in section 4. Verbs with *mi-* prefix denote activities and accomplishments and we will therefore prefer some derivation tree. A grammar G for words represented as *prefix-root-suffix* can be represented by rules of the type:

$$\text{Non} - \text{terminal} \ \rightarrow \ \text{regular expression}$$

A typical example using Non-terminals[10] which include words, prefixes, roots and suffixes is:

$S \rightarrow VP.KP + VP.KP^*$
$VP \rightarrow Voice.V.KP^*$
$KP \rightarrow K.DP$
$DP \rightarrow D.N + D.N.PossP$
$PossP \rightarrow K.DP$

Another set of rules enumerates the possible words and affixes.

$K \rightarrow t\text{–} \ + ...$
$V \rightarrow padang \ + ...$
$Voice \rightarrow mi\text{–} \ + ...$
$N \rightarrow suwal + ...$
$D \rightarrow u + ...$
$PossP \rightarrow$*n-ira* $+ .....$

Each non-terminal symbol $S, VP, KP, ...$ generates a simple regular expression for the first part. Some non-terminal $K, V, ...$ enumerate all possible words for the second part. We can generate a sentence applying several rules from the root S, and conversely a derivation tree is the inverse construction. There are several possible derivation trees associated with a sentence such as *mi-padang t-u suwal n-ira tatakulaq*, repeated from (1) and described in Figure 10, and we argue about the best representation.

6.1 Best derivation trees

Given a *Sentence vector* s, we can then decide that the (a) derivation tree of Figure 10 is better suited than the (b) for the sentence *mi-padang t-u suwal n-ira tatakulaq* ('(he) supports the words of the frog'), in which the subject is dropped, since it is known and referential. We partly repeat the explanation for the *mi-* verbs given in section 4.

A verb stem selects a voice, one to three arguments with a given thematic role and case-marking. The voice affixed verb stem assigns a certain alignment and case-marking to its arguments, and we know that an AV *mi-* verb assigns nominative to the Actor and oblique to the theme. Therefore the VP structure of the derivation tree (a) encodes all the arguments whereas they are separated in the derivation tree (b). Consequently the derivation tree (a) is a better representation.

6.2 Other stochastic models

In a stochastic grammar [7], derivations with the same non-terminal symbol have a probability p such that the sum of the probabilities for each Non-terminal symbol is 1. The probabilistic space associates with each sentence s and derivation tree t, the product of the probabilities of the rules used, noted $p(s, t)$. Given a sentence, a classical task is to predict the most likely derivation tree, and it can be achieved in $O(n^3)$ for a sentence of n words.

In our context, the probabilistic space is entirely different. The structured vectors allow us to predict the most likely word, prefix or suffix, given a context of previous words. They determine the probabilistic distribution of *Sentence vector* defined in section 3.4. The positions of the prefixes and suffixes determines the probabilistic distributions of the vector s. We look at the preferable derivation tree, given this distribution of semantic components.

[10]KP stands for Case Phrase, DP stands for Determiner Phrase, K stands for Case, PossP stands for Possessive Phrase. K and Voice represent Case prefixes and Voice prefixes respectively.

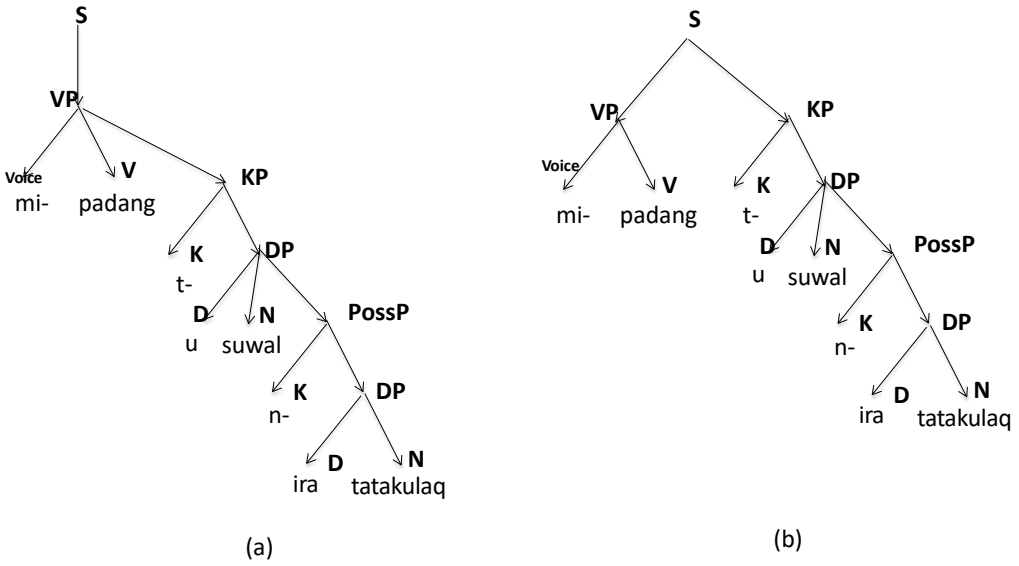

Figure 10: Tree derivations of the sentence *mi-padang t-u suwal n-ira tatakulaq* for the grammar G.

7 Conclusion

We introduced a statistical model for the morphology of natural languages, inspired by *Amis*, a language with a rich morphology. We built the classical distributions of the most frequent prefixes, roots and suffixes, and the corresponding distributions given a possible root, prefix or suffix. From the second moments of the distributions, we built vectors for prefixes, roots and suffixes which capture their correlations. The morphology of a word w=*pre-root-suf* is the concatenation of the vectors for $v_p(pre)$ for the prefix *pre*, $v_r(root)$ for the *root* and $v_s(suf)$ for the suffix *suf*. There are about 30 most common suffixes, and 15 of them carry 90% of the mass. Among the 10 most common suffixes, 4 of them carry 90% of the mass as Figure 3 shows. Hence, the dimensions of the corresponding vectors are small.

We defined a probabilistic *Sentence vector* as a simplified model for the semantic analysis of a sentence. The online analysis of the prefixes and suffixes, realised by the function F, determines most of the components of the *Sentence vector* S. We showed how to predict the syntactic class of a simple sentence with a decision tree. Given a grammar G and a sentence $w_1, w_2, ...w_n$, we then looked at the preferable tree decomposition for S. All the predictions are approximate and the voice and affixal morphology provides valuable information.

We argue that in Amis, most syntactic and semantic features can be approximated from the voice morphology.

References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[3] R. Socher, C. D. Manning, and A. Y. Ng, "Learning continuous phrase representations and syntactic parsing with recursive neural networks," in *In Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, 2010.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[5] K. Cao and M. Rei, "A joint model for word embedding and word morphology," *CoRR*, vol. abs/1606.02601, 2016.

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *CoRR*, vol. abs/1607.04606, 2016.

[7] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[8] R. Baayen, *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, 2008.

[9] I. Bril, "Roots and stems: Lexical and functional flexibility in amis and nêlêmwa," in *Studies in Language. Special issue on lexical flexibility in Oceanic languages (In Press)* (E. van Lier, ed.), pp. 358–407, 2017.

[10] T. Chen, "Verbal constructions and verbal classifications in nataoran-amis," in *Series C. Canberra: Pacific linguistics*, 1987.

[11] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *CoRR*, vol. abs/1708.02709, 2017.

[12] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," in *In Proceedings of the ACL conference*, 2013.

[13] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *CoRR*, vol. abs/1703.03130, 2017.

[14] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 740–750, Association for Computational Linguistics, Oct. 2014.

A Amis Sentences

As in the text, affixes are marked by "-", infixes by "<...>", clitics by "=".

1. Itiya satu a remiad, mi-kasui i lakelal ci Sadaban hananay a tamdaw.
2. Sulinay, adihay=tu k-u ni-pi-kasui.
3. Ta-lumaq=tu cira.
4. Si-pa-qeses cira i liyal.
5. A tubu=tu k-ita.
6. Na sa k-u balucuq n-ira.
7. Pa-tangsul han=tu i liyal.
8. Sulinay, ma-patay k-u balucuq n-ira.
9. Pa-keda han=tu n-ira k-u tireng.
10. U nuka n-u Kawas, si-pa-cahcah cira i niyaruq n-u Balaisan.
11. I cuwa k-ita anini hakira?
12. Saasaan, ira=tu k-u Balaisan.

13. Lepel han=tu cira.
14. U maan a tamdaw-an k-u ma-tini-ay?
15. U pulut han-an-ay iri, u panga k-u han.
16. Pa-habay-ay n-umita.
17. Tuduh-en=ita.
18. Tuwa mi-sangaq t-u pa-pulul-an n-i Sadaban.
19. Sa-kapah-en k-u rangat.
20. A pa-habay-an cira n-u Balaisan a niyaruq.
21. Tuwa si-balucuq k-iya pa-habay-ay i ci Sadaban-an.
22. Araw si-balucuq-ay cira.
23. Tayra-an n-ira t-u labi.
24. Sasaan ma-puyapuy cira.
25. Tuwa ka-tangasa-an=tu k-u ka-subuc-an.
26. Sulinay ma-subuc n-ira.
27. U niyaruq hantu n-uhni, ma-banaq t-u ni-ka-subuc n-ira.
28. ka-ta-tengteng han amin k-u panga n-iya wawa.
29. Ma-susu=tu.
30. Ma-dengay=tu.
31. A patay-en k-ina babuy numita.
32. Uruma k-u talaw n-ira t-u a=patay-en.
33. A maan-en=ita ?
34. Tuwa, tumitumian pa-ka-labi sa=tu cira-an.
35. Melaw han n-i Sadaban, u puut.
36. Tanu ulah sa=tu cira.
37. Ira=tu k-iya Balaisan.
38. Mi-kilim i cira-an.
39. Tanu dungudungus sa=tu k-iya Balaisan.
40. U mi-maan-ay k-isu ?
41. Kiya tireng itini?
42. Tuwa meduk sa=tu tayra i ni-ka-sadak-an n-ira.
43. Pa-sa-tip sa cira.
44. Ira=tu k-iya Balaisan.
45. A ma-lepel=tu k-ita.
46. Tangasa sa cira i ungcung iri.
47. Ci Sadaban hantu iri ma-tini k-u kakawaw numaku.
48. U mi-laliw-ay k-aku t-u Balaisan.
49. U mai-dudu-ay itakuwan k-ira Balaisan.
50. Na ma-lepel k-aku.

51. Urasisa tayni k-aku.
52. Urasisa tangic k-aku.
53. A maan-en=ita.
54. Pa-ta-lumaq-aw-aku k-isu.
55. Kalat-i k-u tangila numaku haw.
56. Han=tu n-ira ci Sadaban.
57. Hay-i sa=tu.
58. Tangasa sa=tu k-uhni.
59. Palita-an n-i Sadaban k-iya isu.
60. Cima k-isu haw aru?
61. Han=tu n-ira ci Sadaban.
62. Ma-pa-tangasa n-umisu k-aku iri.
63. A caay=tu k-aku ka-pawan t-u ni-pa-urip numisu.
64. ka-si-ayam t-u buhcal-ay.
65. Han=tu n-ira.
66. Sulinay dutuc han=tu nu niyaruq a paysin k-u na-u-suwal n-ira.
67. Ala-en k-u besi numaku.
68. Han=tu n-ira.
69. Sulinay ma-patay sa=tu ci Sadaban.
70. Sa cira.
71. Sa=tu cira.
72. Sa k-aku.
73. Sa cira.
74. Sa=tu k-uheni.
75. Ma-wacay-ay k-uhni.
76. Awaay k-u buduy
77. Ma-'pud=tu k-uhni namaka lutuk haw.
78. Ma-ruqruq itini i 'enar.
79. Mi-awit tu kidudung.
80. Adihay k-u ni-urung t-u kidudung saan iri.
81. Mi-kapet tu sa-sait.
82. Mi-kapet=tu t-u kidudung sa.
83. Mi-sangaq tu talip.
84. Araw ma-banaq k-iya niyaruq itira saan.
85. Ma-banaq tu ngudu.
86. Na caay hen ka-banaq tu ngudu.
87. Dihku=tu k-uhni t-u ka-siqnaw-an iri.
88. Ma-hemek k-uhni.

89. Pa-beli-en k-aku t-u belac iri.
90. Pa-beli-en-aku k-amu t-unian u kidudung.
91. Pa-beli han 'amin n-uhni t-iya taw-an pa-cakay-ay t-u ciyapu.
92. Manay si-buduy sa ku Pangcah.
93. Pa-se-banaq n-iya pa-cakay-ay tu ciyapo sananay.
94. Ma-rimurak ku niyaroq.
95. Mi-cakay t-iya kidudung-an.

Negative sentences.

1. Caay ka-demec k-u nanum.
2. Caay ka-tala-mana k-u hadui n-ira.
3. Caay=tu ka-wacay iri.