A statistical model for morphology inspired by the Amis language

Isabelle Bril*
Lacito-CNRS
Isabelle.Bril@cnrs.fr

Achraf Lassoued
University Paris II
achraflassoued985@gmail.com

Michel de Rougemont
University of Paris II, IRIF-CNRS
mdr@irif.fr

Abstract

We introduce a statistical model for the morphology of natural languages. As words contain a root and potentially a prefix and a suffix, we associate three vector components, one for the root, one for the prefix, and one for the suffix. As the morphology captures important semantic notions and syntactic instructions, a new *Content vector* c can be associated with the sentences. It can be computed online and used to find the most likely derivation tree in a grammar. The model was inspired by the analysis of *Amis*, an Austronesian language with a rich morphology.

1 Introduction

The representation of words as vectors of small dimension, introduced by the Word2vec system Mikolov et al. (2013), 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 or to detect an odd word in a list of words. Computational linguists Socher et al. (2013) also studied how to extend the vector representation of the words to a vector representation of the sentences, capturing some key semantic parameters such as Tense, Voice, Mood, Illocutionary force and Information structure.

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 then analyse the most frequent prefixes, the distribution of prefix occurrences, the distribution of suffixes given a root, and so on. We call these statistical distributions the *Morphology Statistics* of the language.

In this paper, we consider the second moment of the *Morphology Statistics* and can 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 many other predictions. We argue that these informations are very useful to associate a vector representation to sentences and therefore to capture some key semantic and syntactic parameters. As an example, we selected *Amis*, a natural language with profuse morphology which is well suited for this analysis. *Amis* is one of the twenty-four Austronesian languages originally spoken in Taiwan, only fifteen of which are still spoken nowadays. This approach can be applied to any other language.

Amis belongs to the putative Eastern Formosan subgroup of the great Austronesian family Blust (1999); Sagart (2004); Ross (2009). *Amis* is spoken along the eastern coast of Taiwan and has four main dialects which display significant differences in their phonology, lexicon and morphosyntactic properties. The analysis bears

*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).

[1]The third moment is the distribution of triples of words and the k-th moment is the distribution of k words.

on Northern Amis; the data were collected during fieldwork. A prior study of the northern dialect Chen (1987) dealt mostly with verbal classification and the voice system.

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 can analyse the global distribution of prefixes, roots, suffixes, 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. As some prefixes or suffixes carry some semantic and syntactic information, as it is the case in *Amis*, we build a *Content* vector for a sentence, and then predict the parsing of a sentence. Our results are:

- A statistical representation of prefixes, roots and suffixes, as structured vectors,
- A vector representation for a sentence, the *Content vector*. We show its use to predict 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 describe its use for a syntactic and semantic analysis.

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 sentence with the words w_i on some alphabet Σ. Let $\mathtt{ustat}(s)$ 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 $\mathtt{ustat}(s)[w]$ is $\#w$ the number of occurrences of w divided by n, the total number of occurrences.

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

We can also interpret $\mathtt{ustat}(s)$ 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 Manning and Schütze (1999); Baayen (2008).

Suppose we take two random positions i, j and define the $\mathtt{ustat}^2(s)$ 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, uses the SVD (Singular Value) Decomposition of the (n, n) matrix and has an $O(n^3)$ time complexity. In Mikolov et al. (2013), a learning technique is used to obtain vectors of dimension 200 when the dictionary has $n = 10^4$ words. In this paper, we refine this approach by separating the covariance matrices of prefixes, roots and suffixes. As we observe 30 distinct prefixes and 10 distinct suffixes, a direct SVD decomposition is efficient.

2.2 The Amis language

A fundamental property of Amis is that roots[2] are most generally underspecified and categorially neutral Bril (2017); 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, Actor Voice *mi-* (AV), Undergoer Voice *ma-* (UV), passive voice *-en*, Locative *-an*.

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-* stems are derived as instrumental *sa-pi-* forms, *ma-* stems are derived as instrumental *sa-ka-* forms.

Some other brief indications (see section 4.4 for further details), nouns are case-marked; voice-affixed verbs select a nominative pivot/subject with the same semantic role.

3 A statistical model for morphology

We first built a tool *Morphix* which, given several texts, constructs the distribution of prefixes, suffixes and roots. Given a root, we can display the distribution of its affixes. Similarly, we can give a prefix (resp. a suffix) and represent the distribution of roots and suffixes (resp. prefixes). We then consider the second moment distributions of prefixes, suffixes and roots. We build their vector representations. If we combine them, we obtain a structured decomposition of the original words.

3.1 Basic Statistics for the Amis language

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

[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 pivot of the derived verb form, with the same syntactic alignment as Undergoer Voice.

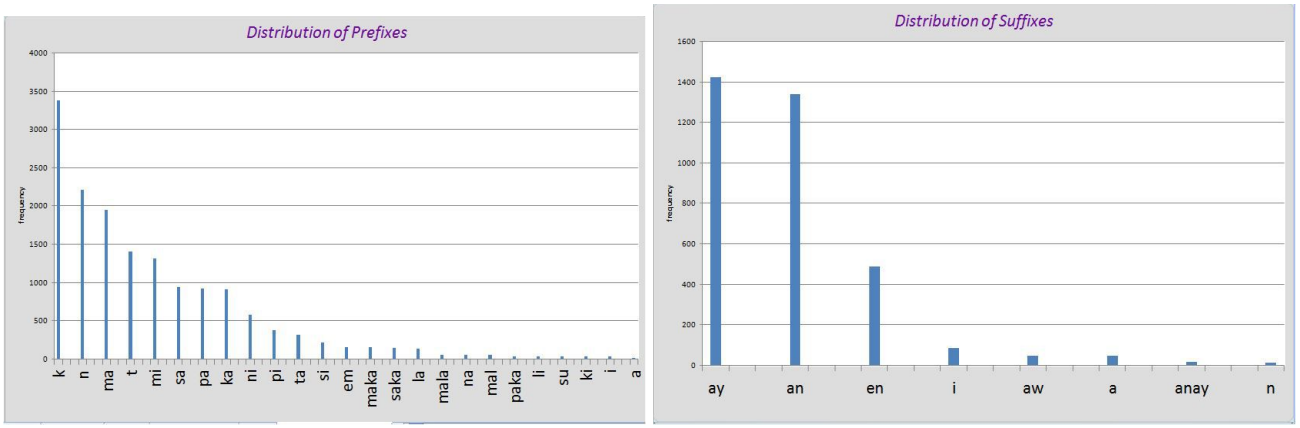

Figure 1: Most frequent prefixes and suffixes.

Given the distribution of (prefixes;suffixes)[4] of Figure 2, we obtain by projection the distribution of prefixes and suffixes in Figure 3 for this specific root.

3.2 Vector representation of prefixes, roots and suffixes

Given a (n, n) correlation matrix M, the SVD (Singular Value decomposition), produces n vectors v_i of dimension n such that $v_i.v_j = M(v_i, v_j)$. If we project v_i on the large eigenvalues of M, we reduce the dimension and obtain vectors such that $v_i.v_j \simeq M(v_i, v_j)$.

Consider the following 4 structured Amis sentences[5]:

Nika ina Hungti, mi-padang t-u suwal n-ira tatakulaq;
but that King AV-help OBL-ART word GEN-that frog[6]
But as for the king, he supported the words of the frog;

"Isu Kungcu, yu ira k-u pa-padang-an;
you Princess when exist NOM-ART RED-help-LOC
"You Princess, when (you) had some help;

Sulinay mi-padang k-u taw;
indeed AV-help NOM-ART people
indeed when people help;

aka-a ka-pawan t-u ni-padang-an n-u taw."
PROH-IMP NFIN-forget OBL-ART PFV.NMZ-help-LOC GEN-ART people
then, you mustn't forget people's help."

[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*.

[5] The first line is the original text where words are structured as prefix-root-suffix. The second line is the morphological analysis with labels such as AV, OBL,....The third line is the translation.

[6] Abbreviations: AV Actor Voice; ART article; CV conveyanve 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.

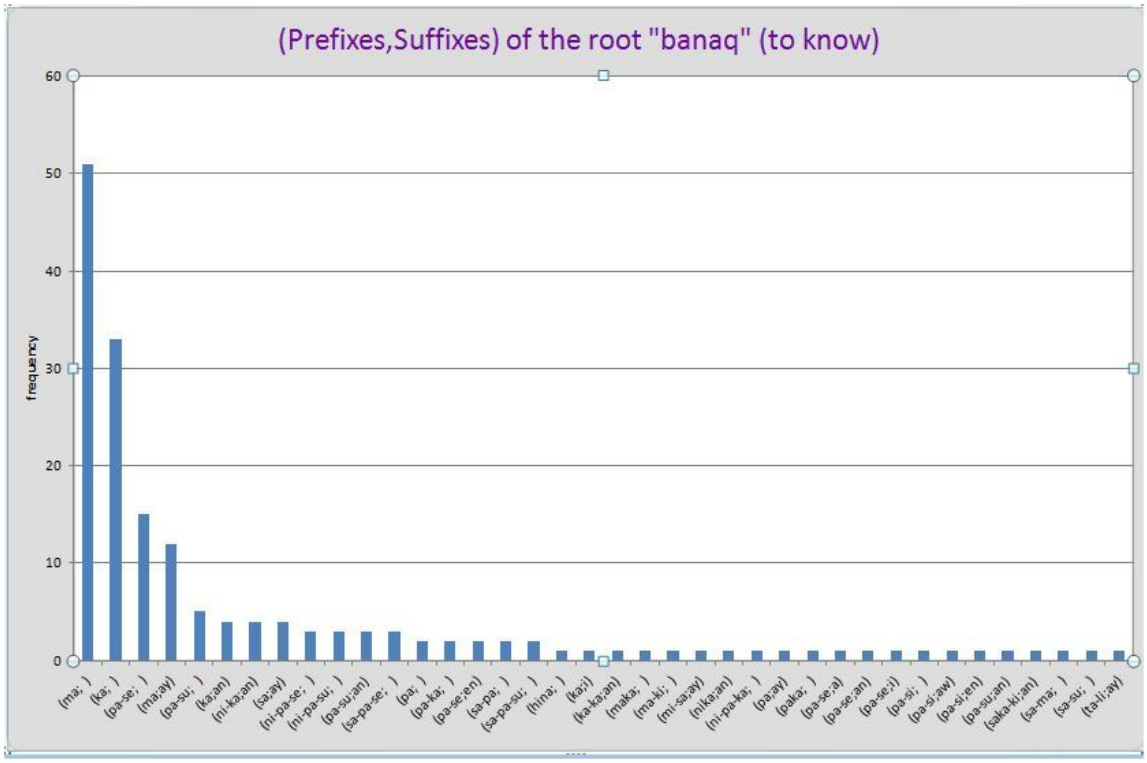

Figure 2: Most frequent (prefixes;suffixes) of the root *banaq* (' know').

In these sentences, there are seven prefixes: *k,ka,n,ni,mi,pa,t*. The matrix M_p for these prefixes is:

$$M_p = \begin{bmatrix} 4 & 0 & 0 & 2 & 0 & 2 & 0 \\ 0 & 2 & 2 & 2 & 0 & 0 & 2 \\ 0 & 2 & 4 & 2 & 2 & 0 & 4 \\ 2 & 2 & 2 & 4 & 0 & 0 & 2 \\ 0 & 0 & 2 & 0 & 2 & 0 & 2 \\ 2 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 2 & 4 & 2 & 2 & 0 & 4 \end{bmatrix}$$

The actual values in M_p are doubled to be consistent with the probability measure. The first line indicates 2 occurrences of *k-*, 1 occurrence of *k-, pa-* (second sentence) and 1 occurrence of *k-, mi-* (third sentence). The large eigenvalues of M_p are 6 and 3.2. Two other eigenvalues are close to 1 and the three others are close to 0. If we decompose the vectors[7] on the large eigenvectors, we obtain 7 vectors of dimension 2, one for each prefix.

$$B = \begin{bmatrix} 1.8860e+00 & -4.7065e-01 \\ -9.9611e-17 & 6.5699e-01 \\ -4.7150e-01 & -2.8430e-01 \\ 9.4301e-01 & 9.6547e-01 \\ -4.7150e-01 & -9.4129e-01 \\ 9.4301e-01 & -7.7913e-01 \\ -4.7150e-01 & -2.8430e-01 \end{bmatrix}$$

and $B * B^t$ is approximately M_p. In this example the absolute L_2 error is 11.5. The first vector for *k-* has coordinates $1.88, -0.47$. We can therefore represent graphically the 7 prefixes as in Figure 4. A similar approach can be followed for suffixes and for 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 $|v.v'|$. Given the vector for the prefix *k-*, the most likely next prefix is *pa-*.

[7]We used Octave, a tool for linear algebra to obtain the SVD decomposition and the projection.

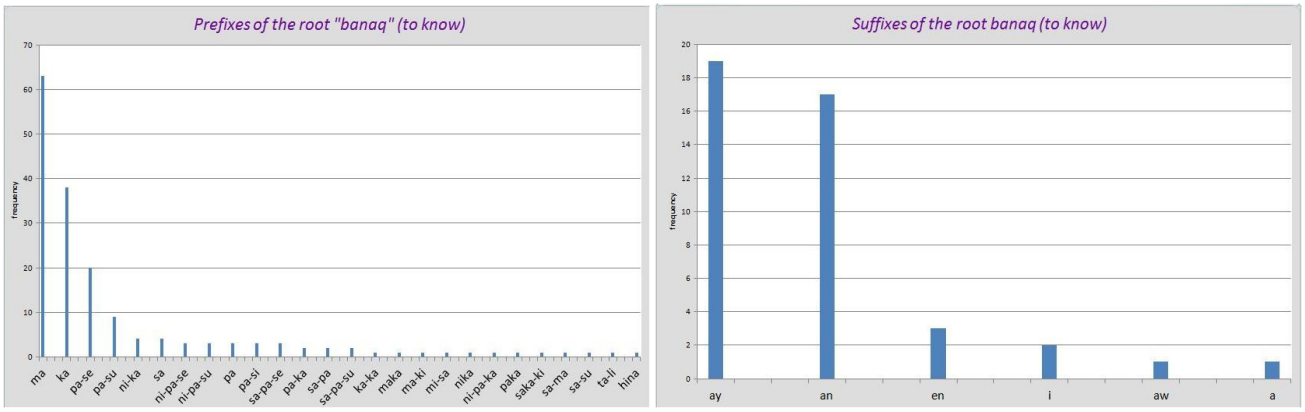

Figure 3: Most frequent prefixes and suffixes of the root *banaq*.

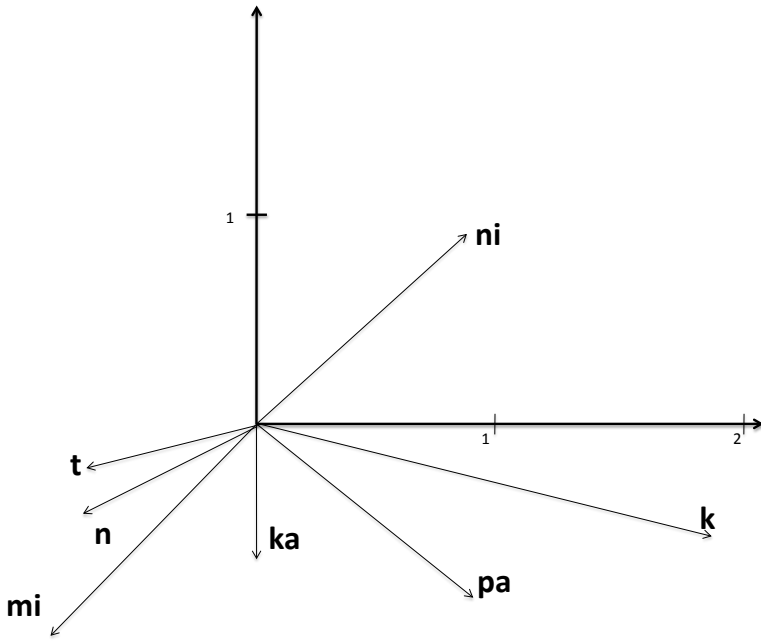

Figure 4: The vectors for the 7 most frequent prefixes *k-,ka-,n-,ni-,mi-,pa-,t-* in two dimensions.

3.3 Distributions and representative vectors

All the distributions are related, mostly by projections. Let δ be the distribution of the words, δ_P the distribution of the 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(*mi-padang*)=*padang*. 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, we apply the dimension reduction and obtain vectors $v_{p,i}$ of dimension n_p for the prefixes, $v_{r,i}$ of dimension n_r for the roots and $v_{s,i}$ of dimension n_s for the suffixes. We associate the union of the three vectors to a word w=*pre-root-suf* :

$$\mathsf{ustat}(w) = \begin{pmatrix} v_{p,pre} \\ v_{r,root} \\ v_{s,\text{suf}} \end{pmatrix}$$

For two words w_i, w_j, let $\widetilde{M}(w_i, w_j) = M_p(pre_i, pre_j) + M_r(root_i, root_j) + M_p(\text{suf}_i, \text{suf}_j)$ be the sum of the correlations of the prefixes, roots and suffixes. The fundamental fact of the approach is that for any two words w_i, w_j, $\mathsf{ustat}(w_i).\mathsf{ustat}(w_j) \simeq \widetilde{M}(w_i, w_j)$. Indeed, $\mathsf{ustat}(w_i).\mathsf{ustat}(w_j) = v_{p,pre_i}.v_{p,pre_j} + v_{r,root_i}.v_{r,root_j} + v_{s,suf_i}.v_{s,suf_j}$. 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 $\mathsf{ustat}(w_i).\mathsf{ustat}(w_j) \simeq \widetilde{M}(w_i, w_j)$.

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 developped using cross-correlations, which we do not use at this point.

4 Grammars and statistics

We now study how to extend the vectors from words to sentences, as in Socher et al. (2010, 2013). We follow a different strategy as we fix a probabilistic *Content Vector* with specific dimensions which depend directly on the prefixes, roots and suffixes. We then show its use for a syntactic decomposition. A grammar G is classically represented by rules of the type[8]:

$S \to VP.KP + VP.KP^*$
$VP \to Voice.V.KP^*$
$KP \to K.DP$
$DP \to D.N + D.N.ModP$
$ModP \to K.DP$
$K \to t + ...$
$V \to padang + ...$
$Voice \to mi + ...$
$N \to suwal + ...$
$D \to u + ...$

Our goal is to compare the possible derivation trees of the sentence *mi-padang t-u suwal n-ira tatakulaq* and to use the *Content Vector* to infer the "most likely" tree in the grammar G.

4.1 Stochastic grammars

In a stochastic grammar Manning and Schütze (1999), derivations with the same non terminal symbol have a probability p such that the sum of the probabilities for each non terminal 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 also determine the distribution of *Content Vector* defined in section 4.2 which predicts some key semantic components. Hence we look at the most likely derivation tree, given this distribution of semantic components.

4.2 Semantic representation

Let us define the *Content* vector of a sentence as a vector of dimension 6 whose components are:

- 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},
- Information Structure: {Topicalisation, Cleft Focus},

[8] KP stands for Case Phrase, DP stands for Determiner Phrase, ModP stands for Modifier Phrase.

This is a just an example and more dimensions could be used. Let c be such vector of dimension 6 where values are distributions over each finite domain. For example, the third component c^3 over {Present, Past, Future} is $[0, 1, 0]$ to indicate a PAST or $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ to indicate a uniform distribution. We read the sentence $w_1, w_2, ...w_n$, and a vector $v_i = \texttt{ustat}(w_i)$ is associated with each word w_i. Let us define:

$$c_i = F(c_{i-1}, v_i)$$

with c_0 an initial state and F a function, we construct by cases or by learning techniques. As an example, consider the following sentence:

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

In this case, 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 encodes a type of alignment (a nominative patient pivot and a genitive agent). In this case c_i^2, the second component of F is defined as:

$$c_i^2(c_{i-1}, v_i) = \begin{cases} [0, 1, 0, 0] & \text{if } [v_i]_p = \text{"mi-"} \\ c_{i-1}^2 & \text{otherwise} \end{cases}$$

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 Content vector c_n. We describe more advanced rules of *Amis* in section 4.4.

4.3 Rules and Correlations

The previous rule for the imperative mood is simple. It is also possible to learn this rule from positive and negative examples, i.e. sentences in imperative mood and sentences not in imperative mood, as suggested in Socher et al. (2013). In that case, we would get a correlation and a neural network could approximate the imperative mood given enough examples.

This is a general paradigm, often called *Causality versus Correlation*. It is however far more difficult to learn the structure of the Content vector, i.e. the decomposition in 6 independent components. Notice that 5 of the components are set by the prefixes and suffixes. The Valence is set by the roots. As the number of prefixes and suffixes is small, the description of the function F is much simplified.

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-*, oblique themes and oblique arguments are marked by *t-* Chen (1987). The voice affixes (AV) *mi-*, (UV) *ma-*, also identify verb classes, (i) verbs which only accept *mi-* voice, (ii) verbs which only accept *ma-*, (iii) verbs which accept both *mi-* and *ma-* with different semantics, and (iv) stative, property verb stems which accept none of these prefixes.

AV *mi-* verb stems denote activities or accomplishments. *Ma-* verbs denote non-actor or undergoer oriented events (depending on their semantics and valency); *ma-* verbs include states and psych states, properties, verbs of cognition (*ma-banaq* 'know'), bodily functions, position and motion[9] (*ma-nanuwang* 'move for object').

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 pivot), and a voice affix whose semantics matches the semantics of the nominative pivot. AV *mi-* and UV *ma-* voices are restricted to declarative sentences. In non-declarative sentences (such as negative, imperative, hortative), *mi-* occurs as *pi-* and *ma-* as *ka-*. Compare *ma-butiq cira* '(s)he is asleep/sleeping' and *ka-butiq!* 'go to sleep !'.

[9]Motion verbs are not activities despite their dynamic feature; their nominative pivot is not an Actor but a theme.

4.4.1 Transitivity and alignment

Alignment[10] varies with transitivity. *Mi-* verbs and extended intransitive *ma-* verbs (labelled Non-Actor Voice,NAV) have an oblique argument marked by *t-* as in (1a-2). The nominative pivot of *mi-* verbs is an Actor, while that 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 a nominative (generally fully affected) patient pivot and a genitive agent as in (1b).

1a. *Mi-melaw k-u wawa t-u tilibi.*
AV-look NOM-ART child OBL-ART TV
' The child is watching TV. '

1b. *Ma-melaw n-uhni k-u teker.*
UV-look GEN-3pl NOM-ART trap
'They saw the trap.'(lit. the trap was seen by him)

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

Ma- verbs are thus generally oriented towards a non-actor, or an undergoer nominative pivot; the case assignment of the non-pivot argument varies with transitivity: with extended intransitive NAV *ma-* constructions (2), the theme is oblique; with transitive UV *ma-* constructions, the agent is genitive (1b). All other voices, UV *-en*, INST *sa-*, LOC *-an*, CV *si-*, have a nominative pivot which is the corresponding semantic argument (i.e. patient, instrument, location, transported theme), and a genitive Agent (if it is expressed).

4.5 Best derivation tree

Given c_n, we can then decide that the (a) derivation tree of Figure 5 is better suited than the (b) for the sentence *mi-padang t-u suwal n-ira tatakulaq* ('he supports the words of the frog'). We follow the explanation of the *mi-* verbs given in section 4.4.

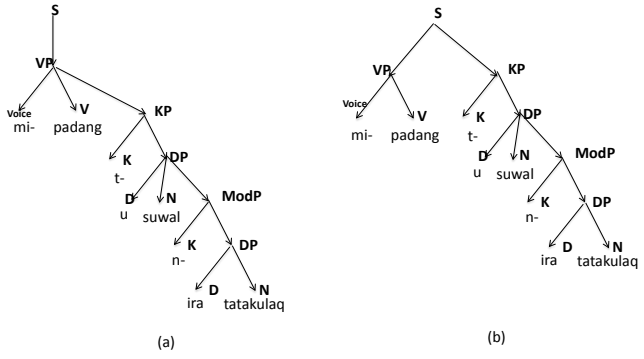

Figure 5: Tree derivations of the sentence *mi-padang t-u suwal n-ira tatakulaq* for the grammar G.

The conceptual structure of a verb stem selects the voice, the number and type of arguments. Case-assignment takes place in the domain of the VP and correlates with Voice which assigns theta-roles to its arguments (for ex. an AV mi-verb

[10] Alignment refers to the morphosyntactic encoding of the grammatical relationship between the two arguments of transitive verbs, and the single argument of intransitive verbs. In accusative languages, the subjects are marked in the same way independently of transitivity, and differently from the object. In ergative languages, the single argument of intransitive verbs and the patient of transitive verbs are similarly marked as nominative/absolutive, but differently from the agent of transitive verbs.

assigns nominative to the Actor and oblique to the theme; an UV ma-verb assigns nominative to the Patient and genitive to the agent). Consequently the derivation tree (a) is a better representation.

5 Conclusion

We introduced a statistical model for the morphology of natural languages and applied it to *Amis*. The *Morphix* tool builds the classical distributions of prefixes, roots and suffixes, given a possible root, prefix or suffix. From the second moments of the distributions, we build vectors for prefixes, roots and suffixes which capture their correlations. 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. Hence, the dimensions of the corresponding vectors are small.

We defined a probabilistic *Content vector* as a simplified model for the semantic and syntactic analysis of a sentence. The online analysis of the prefixes and suffixes, realised by the function F, determines most of the components of the *Content* vector c. Given a grammar G and a sentence $w_1, w_2, ...w_n$, we then looked at the most likely tree decomposition for c.

Other languages have different types of morphology or no morphology, but we argue that the most likely tree decomposition is dependent on semantic features in a probabilistic way.

References

Baayen, R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Blust, R. (1999). Subgrouping, circularity and extinction: Some issues in austronesian comparative linguistics. In E. Zeitoun and P. Li (Eds.), *Selected Papers from the Eighth International Conference on Austronesian Linguistics*, pp. 31–94. Taipei: Institute of Linguistics, Academia Sinica.

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

Chen, T. (1987). Verbal constructions and verbal classifications in nataoran-amis. In *Series C. Canberra: Pacific linguistics*.

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

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*.

Ross, M. (2009). Proto austronesian verbal morphology: a reappraisal. In A. Adelaar and A. Pawley (Eds.), *Austronesian historical linguistics and culture history. A festschrift for Robert Blust*, pp. 285–31. Canberra: Pacific Linguistics.

Sagart, L. (2004). The higher phylogeny of austronesian and the position of tai-kadai. *Oceanic Linguistics 43*, 411–444.

Socher, R., J. Bauer, C. D. Manning, and A. Y. Ng (2013). Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*.

Socher, R., C. D. Manning, and A. Y. Ng (2010). 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*.