A Way to Break Them All: A Compound Word Analyzer for Marathi

Raj Dabre
CFILT
IIT Bombay
prajdabre
@gmail.com

Archana Amberkar
CFILT
IIT Bombay
amberkararchanaa
@gmail.com

Pushpak Bhattacharyya
CFILT
IIT Bombay
pushpakbh
@gmail.com

Abstract

In this paper we describe and evaluate a compound word analyzer for Marathi, a highly inflectional language with agglutinative suffixes. Compound words are one of the most frequently used constructs in the spoken language and as such have been a thorn in most Natural Language Processing activities. We explain the morphological phenomena occurring in a variety of compound words and then go into the methods used to handle them. A thorough analysis of an objective evaluation of the analyses of such words will establish the efficacy of the analyzer. We believe that this will give the readers a good idea on how to develop compound word analyzers for their own languages.

1 Introduction

Morphological Analysis is the task of taking as input individual words and providing the linguistic features of those words. Marathi is a language that uses agglutinative, inflectional and analytic forms. The number of Marathi speakers all over the world is close to 72 million[1]. Marathi displays an impressive amount of derivational and inflectional morphology. About 15% of the word forms are in the participial form known as *Krudantas*, which come from the influence of the Dravidian languages. These form a part of the derivational morphology of Marathi wherein attachment of suffixes to a word form will change its grammatical category. Traditional grammars of Marathi classify the derived forms in Marathi into two categories- *Krudanta*s and *Taddhita*s. *Krudanta*s are the adjectives, adverbs and nouns derived from verbs, while *Taddhita*s are the nouns, adjectives and adverbs derived from words of any category other than verb. This is also accompanied by inflectional processes which help lend the words linguistic features of gender, number, person, case, tense, aspect and modality (the latter 3 for verbs only).

Having said this one will have realized that Marathi has a number of suffixes, which denote a number of features when attached to words. Consider this example: "शहाणपणा दाखवणारा मुलगा" {*shahanpanna dakhavnara mulgaa*} {*smart-acting boy*} {*the boy acting smart*}. Here "शहाणपणा" is a derived noun which arises when the morpheme "पणा" {*panaa*} {*-ness*} is attached to the noun "शहाणा" {*shahanaa*} {*smart*} and the *Krudanta* "दाखवणारा" is an adjective derived by attaching the morpheme "णारा" {*naara*} {*-ing*} to the verb root "दाखव" {*dakhav*} {*show*}, both having a single root. Handling these phenomena is a difficult task for Marathi and similar languages. The existence of compound words, which are built on top of the above constructs, further complicates the situation.

1.1. Compound Words

Consider the word मामामामीबरोबरचा {*mama-mamibarobarcha*} {*the one with the uncle-aunt (maternal)*}, a noun and a compound word. There are two root words (both of which can be of different grammatical category) namely मामा {*mama*} {*maternal uncle*} and मामी {*mami*} {*maternal aunt*} (both nouns). Moreover this

[1] http://en.wikipedia.org/wiki/List_of_Indian_languages_by_total_speakers

word has two suffixes namely बरोबर {*barobar*} {*with*} and चा {*cha*} {*a suffix which causes the word to behave as a pronoun*}. In general, more than two root words can exist in a compound word, but in Marathi these are very rare. For the sake of simplicity of understanding, in rest of the paper, we will address compound words with only 2 components. It will be seen that the methods and descriptions easily extend to compound words with more than 2 components. At this point we would like to make a brief digression into the difference between Multi-Word Expressions and Compound Words.

1.1. Multi-Word Expressions and Compound Words

A Multi-Word Expression (MWE) is a lexeme made up of a sequence of two or more lexemes, separated by spaces, that has properties (meaning being one) that are not usually predictable from the properties of the individual lexemes or their normal mode of combination[2]. In linguistics, a Compound Word is a lexeme that consists of more than one stem[3]. They are a kind of MWE's. "सरड्याची धाव कुंपणापर्यंत" {*sardyachi dhav kumpanyaparynt*} {*A person will only be able to do things of what he is capable of*} is a MWE where the individual words are non-compositional. But compound words, in Marathi, are single words.

1.2. Motivation for handling Compound Words

Compound words occur very frequently not only in Marathi but in all languages of the world. In order to translate from Marathi to any other language it becomes necessary to obtain Morphological Analyses of these words. Even obtaining the individual components is quite valuable and thus the morphological analysis of compound words is quite crucial for high quality translation. They also exhibit interesting word formation phenomenon which needs to be studied and uncovered. Multi Word Expression handling in any language, due to the very fact of "multiple words", is quite complex because of non-compositionality. Marathi compound words, however, being single words can be detected. Moreover, भाऊबहीण {*bhaubahin*} {*brother-sister*} has a Hindi equivalent भाई-बहन {bhai-bahan}. Here the individual components are directly translated. There are many more cases and thus even getting the breakdown of words of this type helps in translating from one language to another, especially when it comes to close languages like Marathi and Hindi.

1.3. Identification of the problem

Given a word containing two word components (and hence roots) **a** and **b**, which have been inflected and appended with suffixes, identify each one and provide linguistic information in the following format:

1. Field 1 : <input word>
2. Field 2 : <'root-word-a, CGNPTAM-a'&'root-word-b, CGNPTAM-b, suffix-b'&fincat=lexical category>...

CGNPTAM means 'grammatical category, gender, number, person, tense, aspect, modality and form'. For verbs we have additional features namely "*Krudanta* Type" and "*Krudanta* Case Marker/Suffix". Krudantas are non-verbal class words derived from verbs. Fincat is a feature which tells the grammatical category of the resultant word. In case the features of the root words are not identifiable then only the root words are reported with a short description.

1.4. Related Work

Dongli et al. (2002) used corpus and statistics based methods for handling structural analysis of compound words, allowing them to eliminate word combinations not likely to form compounds. Dasgupta et al. (2005) used a combination of finite state morphological analysis and morphological parse trees for detection and analysis of compound words. Their method involved the creation of a number of rules for analysis which can be quite tedious. Damle and M.K. (1970) and Dhongde and Wali (2009) have had provided a classification of compound words in Marathi. Dabre et al. (2012) and Bapat et al. (2010) had developed a two level morphological analyzer (Koskenniemi et al., 1983, Antworth et al., 1990 and Kim et al., 1994) for Marathi using a finite state-based approach and morphological parsing but did not handle compound words since incorporating the rules for their formation in a finite state transducer is difficult. We extend their work to ensure efficient detection and analysis of compound words. In the next section we briefly describe the Marathi Morphological analyzer.

[2] http://en.wikipedia.org/wiki/Multiword_expression
[3] http://en.wikipedia.org/wiki/Compound_words

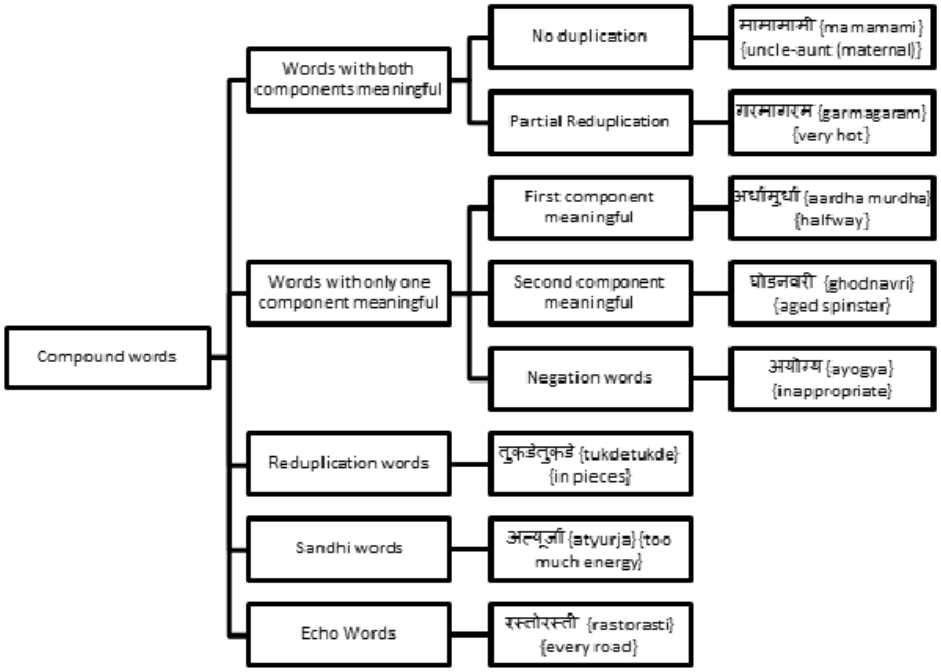

Figure 1 : Taxonomy of Compound Words

2 Marathi Morphological Analyzer

The Marathi Morphological Analyzer (MA) is fully rule-based and thus relies on string manipulation and file lookup. The quality is greatly affected by the size of the resources that are used for morphological analysis. The important components of this analyzer are given below:

1. **Lexicon**: This is a list of all Marathi root word forms associated with their inflectional category.
2. **Suffix Replacement Rules**: This is a list of all inflectional categories and the rules of suffix replacements which give rise to inflected forms of root words and the grammatical features that the word acquires.
3. **Inflector and REPO**: This module applies the suffix replacement rules to the lexicon to give a repository of all possible inflected forms of the words in the lexicon which is called the REPO file. This file is crucial for our compound word analysis. Words have 2 forms: Direct Form (DF) where they take no suffixes and Oblique Form (OF) where they take suffixes. All forms are present in the REPO file. A lot of lookup is done in this file for analysis.
4. **Finite State Transducer**: The rules of Marathi morpheme combinations are specified to the finite state transducer compiler which creates a finite state machine that can segment of Marathi word into its individual morphemes.
5. **Morphological Parser**: The segmented components are looked up in the REPO file and the morphological analyses are retrieved.

We use the REPO file and the Marathi morphological analyzer as a black box for the purpose of analyzing compound words. We now describe the different kinds of Marathi compound words and the methods used to handle them. To the best of our knowledge, this is the first of its kind work on Computational Compound Word analysis for Marathi.

3 Compound Words and their Handling in Marathi

To solve a problem it is best to understand what it is. This chapter attempts to classify compound words into subcategories with as much linguistic information as possible. The linguistic knowledge comes from the books by Damle and M.K. (1970) and Dhongde and Wali (2009). For each type the method and algorithm used to handle it is given. Figure 1 gives the taxonomy of compound words. We explain each type of compound words handled and the method used for analysis. We would like the reader to note that no resources in terms of morphology rules or words for a lexicon were added during these experiments. If a word falls under more than one cate-

gory then our methods try to identify all possible ones. We had a relatively small list of 465 compound words to work with and it is possible that some types of words, not in the list, might not be handled by the methods we developed.

3.1 Words with both components meaningful

Compound words of this category have meanings associated their individual components. By this, we mean that both make sense. There are two subcategories of this type.

3.1.1 No duplication words

These are words wherein the first component is a different word from the other. These words can be viewed as full words (including those with suffixes) prefixed with a morpheme. This morpheme is either an uninflected word or an inflected form with no suffixes attached to it. To put it formally:

1. Type 1: (DF of Word1) (Word2 with suffixes if any).
2. Type 2: (OF of Word1) (Word2 with suffixes if any).

In general all compound words have this form. An example of a type 1 word is मुलबाळ {*mulbala*} {*children*} and its type 2 equivalent is मुलांबाळांसाठी {*mulabalansathi*} {*for children*}. Here मुलां {*mulan*} {*children*} (plural noun) is an inflected form of मुल {*mul*} {*child*} where as बाळांसाठी {*balansathi*} {*for children/babies*} is a full word resulting from बाळ {*bal*} {*child/baby*} undergoing inflection and affixation. Note that the two words may belong to different grammatical categories and thus will have a resultant category. We have a set of rules of the form: <Category1> + <Category2> = <Resultant Category> : <Example> called as the Compound Category Rule Set (CCRS), in which a lookup is done which we describe below:

1. N +N = N : भाऊबहिण {*bhaubahin*} {*brother- sister*}
2. V + V = N: नेआण {*neaan*} {*take away-bring*}
3. ADJ + N = N: वीरपुरुष {*veerpurush*} {*brave man*}
4. ADJ + ADJ = ADJ : गोराघारा {*gooraghara*} {*pale*}
5. N + ADJ = ADJ : रोगमुक्त {*roagmukta*} {*disease free*}
6. N + V = ADJ : ईश्वरनिर्मित {*ishwrnirmt*} {*god-created*}
7. CAR + N = N : सातरस्ता {*saatrasta*} {*seven roads*}
8. CAR + CAR = CAR : दोनतीन {*doantean*} {*two-three*}
9. CAR + ORD = ORD : साठसत्तरावा {*sath-sattarava*} {*60-70th*}
10. ADV + ADJ = ADJ : सदासुखी {*sadasukhi*} {*eternally happy*}

To handle this, the following algorithm is used:

1. Start at the first position of the word. (If दगड {*dagad*} {*stone*} is a word then first position is between द and गड)
2. Split the word into 2 parts.
3. If the first part is in the REPO file hash table and the second part can be analyzed by the MA as a valid full word then
   a. Report the word as a compound word.
   b. If c1 is the grammatical category of the 1st part and c2 that of the second then
      i. Do lookup in the CCRS and report the final category along with the features of the individual words.
      ii. In case the CCRS doesn't have the entry, report only the features and final category as unknown.
4. Else go to the next position and go to step 2.

If the word cannot be analyzed report that "No analysis for word".

3.1.2 Partial Reduplication words

These are words wherein the first and second components are identical except that the first component is in the inflected form. These are mostly nouns and adjectives. Below we give the type of words handled. The format is: <Inflecting matraa> : <Component> : <Resultant Form>.

1. ा : जुळव {*julav*} {*mix*} : जुळवाजुळव {*julva-julav*} {total mixture}
2. े : लाल {*lal*}{*red*} : लालेलाल {*lalelal*} {*red everywhere*}
3. ो : घर {*ghar*} {*house*} : घरोघर {*gharo-ghar*} {*every house*}
4. ें : दिवस {*divas*} {*day*} : दिवसेंदिवस {*divsendivas*} {*day after day*}

To handle these, the following algorithm is used:

1. For each inflecting character given in the above table identify its position in the word.
2. See if the part of the word before this position is the same as the part of the word after this.
3. If so then report it as a partial reduplication word along with the features of the root/component identified.

3.2 Words with only one component meaningful

Compound words of this type have one of their components meaningful with the other one being senseless. By senseless we mean that these components are typically not used independently in Marathi but in combination with other Marathi words. This stage is performed only after the analyzer is unable to detect two meaningful root words. These come under 3 subcategories:

3.2.1 First component meaningful

Here the first component has a meaning with the second component, mostly, being some senseless word. These constructions are mostly used in dialogue for dramatic or poetic effect. Following are some of the word types currently handled. Format is <Word> : <Sensible part> : <Insensible part>:

1. अर्धामुर्धा {*aardhamurdha*} {*halfway*} : अर्धा {*aardha*} {*half*} : मुर्धा {*murdha*}
2. बागबगिचा {*baagbagicha*} {*garden*} : बाग {*baag*}{*garden*} : बगिचा {*bagicha*}
3. जुनापुराणा {*junapurana*} {*old*} : जुना {*juna*} {*old*} : पुराणा {*purana*}

To analyze these the same method under 3.1.1 is used, only here the second component won't be analyzed by the morphological analyzer and will be reported as such. Some words of this type are also echo words. Here बगिचा {*bagicha*} means *garden* but since it is not a native Marathi word, it is not present in the lexicon of Marathi MA. Similar is the case of पुराणा {*purana*} which means *old*.

3.2.2 Second component meaningful

Here the second component has a meaning with the first component being some senseless word. Following are some of the word types currently handled. Format is <Word> : <Insensible part> : <Sensible part>:

1. घोडनवरी {*ghodnavri*} {*aged spinster*} : घोड {*ghod*} : नवरी {*navri*}{*bride*}
2. महामूर्ख {*mahamurkha*} {*very stupid*} : महा {*maha*} : मूर्ख {*murkha*}{*stupid*}

Sometimes the insensible parts like महा {*maha*} may seem sensible which gives the sense of *great*. Similarly घोड {*ghod*} giving the sense of *old*. But note that they cannot be used independently as words in their own right. The method used to handle this is the same as in 3.1.1, only here it will be reported that the first component cannot be analyzed.

3.2.3 Negation words

These are words in which adding a negatory prefix to a root word which gives its negation. These prefixes are "अ" {*a*} and "ना" {*na*} and "गैर" {*gair*}. An example is अन्याय {*anyaya*} {*injustice*}. The algorithm used is:

1. If negatory prefix exists:
   a. If rest of the word can be analysed by the MA then report the words features with a marker indicating that it is a negative word.
   b. Else it is not a negation word.

The prefix "अन" {*ana*} is also a negatory prefix but it undergoes Sandhi formation which is explained later.

3.3 Reduplication words

This category of words is those wherein both parts are identical. An example would be कावकाव {*kaavkaav*} {*sound of crows crowing*}. Detecting them requires identifying that the first component matches the second one. A slight variation of this category is one where although the first and second components are identical and the second component has a suffix. There are three suffixes we encountered; namely, ाट {*aat*}, णे {*ne*} and ीत {*eet*}. 3 types handled are given

below in the format - <Word> : <Suffix> : <Resultant word>:

1. भर {*bhar*} : ाट {*at*} : भरभराट {*bharbharat*} {*a word indicating that something is full*}
2. सण {*sana*} : ीत {*it*} : सणसणीत {*sunsunit*} {*a word indicating sharpness of sound or action*}
3. दुम {*duma*} : णे {*ne*} : दुमदुमणे {*dumdumne*} {*to resonate*}

The algorithm used to handle this is:

1. If the ending of the word is either ाट {*aat*}or णे {*ne*}or ीत {*eet*} then strip it and
   a. Divide the remaining word into 2 exact halves
   b. If the first half resembles the other report the word as a reduplication word with the appropriate suffix.

The णे {*ne*} suffix (a Krudanta suffix) words are nouns that are derived from verbs. Note that for these compound words either both parts are sensible or insensible. They lend onomatopoeic effect when used in a sentence and thus it needs only to be reported whether the word is a Reduplication type or not.

3.4 Sandhi words

This category of compound words is a very peculiar one because it does not follow regular inflection. Normally a word, say रामाबरोबरच्याने {*ramabarobarchyane*} {*the one with Ram did*} split as रामा-बरोबर-च्या-ने, is formed by inflecting a word (राम {*ram*} to रामा {*rama*}), appending a suffix to it, then inflecting that suffix if needed (बरोबर) and appending more suffixes in the same way (चा to च्या followed by ने). Notice that at the word boundary only the previous morpheme undergoes orthographic change (inflection). Sandhi words are those in which orthographic changes occurs at the boundary between both words or at the beginning of the next word to give a new word.

Thus there are 2 cases. One where a matraa (example: ो) at the end of a word and a vowel (example: आ) at the beginning of another, are replaced by either a matraa or a consonant preceded by a halant (example: ् + य = ् य) or a consonant followed by a matraa and preceded by a halant (example: ् + य + ा = ् या). The second case is where the vowel at the beginning of the next word is replaced by a matraa. The replacement is called as a Sandhi. These two cases together have 26 possible instances (which we identified in our data), which are given in below. The list below, with format <Sandhi Characters> : <Replacement/Expansion Characters> : <Example word>, tells what Sandhi is formed by what combination of characters. The expansion characters are separated by a hyphen indicating whether they belong to the 1st or 2nd component.

1. ा : -अ : न्याय+अन्याय=न्यायान्याय {*nayayanyaya*} {*justice-injustice*}
2. ा : -आ : देव+आलय=देवालय {*devalaya*} {*temple*}
3. ा : ा-आ : विद्या+आरंभ=विद्यारंभ {*vidyaarambha*} {*beginning of education*}
4. ी : ी-इ : देवी+इच्छा=देवीच्छा {*deviccha*} {*will of god*}
5. ी : ी-ई : पार्वती+ईश=पार्वतीश {*parvatish*} {*Goddess Parvati*}
6. ू : ू-उ : भू+उद्धार=भूद्धार {*bhuuddhar*} {*great gratitude*}
7. ू : ू-ऊ : भू+ऊर्जित=भूर्जित {*bhuurjit*} {*fully energetic*}
8. े : -इ : ईश्वर+इच्छा=ईश्वरेच्छा {*ishwariccha*} {*gods will*}
9. े : -ई : गण+ईश=गणेश {*ganaisha*} {*The lord of the Ganas-Ganesh*}
10. ो : -उ : सूर्य+उदय=सूर्योदय {*suryodaya*} {*sunrise*}
11. ो : -ऊ : समुद्र+ऊर्मी=समुद्रोर्मी {*samudrormi*} {*desire for the sea*}
12. े : ा-इ : रमा+इच्छा=रमेच्छा {*ramechha*} {*the will of Rama*}
13. े : ा-ई : रमा+ईश=रमेश {*ramesh*}{*a name*}

14. ो : ा-उ : गंगा+उत्साह=गंगोत्साह {*gangotsaha*} {*the glee of seeing Ganga*}
15. ो : ा-ऊ : गंगा+ऊर्मी=गंगोर्मी {*gangormi*}
16. ्य : ि-अ : अति+अल्प=अत्यल्प {*atyalpa*} {meager}
17. ्या : ि-आ : अति+आनंद=अत्यानंद {*atyananda*} {*extreme happiness*}
18. ्यु : ि-उ : अति+उत्तम=अत्युत्तम {*atyuttam*} {*most apt*}
19. ्यू : ि-ऊ : अति+ऊर्जा=अत्यूर्जा {*atyurja*} {*extreme energy*}
20. ्ये : ि-ए : प्रति+एक=प्रत्येक {*pratyek*} {*every-one*}
21. ्यौ : ि-ओ : मति+ओघ=मत्यौघ {*matyogh*}
22. ्य : ी-अ : नदी+अर्पण=नद्यर्पण {*nadyarpan*} {*offering to the river*}
23. ्या : ी-आ : गौरी+आनंद=गौर्यानंद {*gauryananda*} {*happiness of the goddess Parwati/Gauri*}
24. ि : -इ : अन+इष्ट=अनिष्ट : {*anishtha*} {*evil*}
25. ो : -ओ : अन+ओळखी=अनोळखी {*anolkhi*} {*unknown/ not known*}
26. ा : ा-अ : विद्या+अभ्यास=विद्याभ्यास {*vidyabhayasa*} {*acquisition of knowledge*}

The following algorithm helps handling Sandhi words.

1. For each entry in the Sandhi replacement rule list, identify the Sandhi in the word
   a. Replace the Sandhi by its replacement characters
   b. If the first component is present in REPO and the next word can be analyzed by the MA then report the word as a Sandhi word along with the features of both words.

Sometimes the first word might be meaningless and can be a negatory marker (अन) or an intensifier (अति). The negative compound word is reported.

3.5 Echo words

Echo words are of two types. One type resembles partially reduplicated words with the main difference that both the first and second components (both identical) are inflected. The following are some examples we have encountered. Representation is: <Inflecting matraa for 1st part> : <Inflecting matraa for 2nd part> : <Example word/root> : <Resultant form>

1. ो : ी : गाव {gaav} {village} : गावोगावी {gaavogaavi} {every village}
2. ो : ी : रस्ता {*rasta*} {*road*} : रस्तोतास्ती {*rastorasti*} {*every road*}
3. ो : ी : घर {*ghar*} {*home*} : घरोघरी {*gharoghari*} {*every home*}

Semantically speaking, this kind of a formation associates an "every" meaning as also a locative meaning with the word. The next type is of three subtypes as below. The categorization is very fine grained in order to understand the specific underlying phenomena.

1. Both components are in DF where the first character of the 1st and 2nd components is different.
2. The first component is in DF or OF and the second is in OF which takes suffixes where the first character of the 1st and 2nd components is different.
3. Both components are in DF where the first syllable of the 1st and 2nd components is different.

3.5.1 Subtype 1

This type of echo words has a difference only in the first character of both components. Both components are in DF. Not all character combinations are possible. Using the format: <1st Component Character> : <2nd Component Character> : <Common Remainder> : <Resultant Word>, we give the 20 types uncovered:

1. क : म : डूक : किडूकमिडूक {*kidukmiduk*} {*senseless*}
2. ग : घ : ोड : गोडधोड {*goadghod*} {*sweets*}
3. ल : प : ेचा : लेचापेचा {*lechapecha*} {*weak*}

4. स : फ : टर : सटरफटर {*sutterfutter*}{*miscellany*}
5. अ : प : ंगत : अंगतपंगत {*angatpangat*} {*gathering of people for casual talk*}
6. प : झ : ड : पडझड {*padzhad*}{*fall*}
7. च : म : ंगळ : चंगळमंगळ {*changalmangal*} {*chaos*}
8. अ : ट : ळं : अळंटळं {*alantalan*} {*procratination*}
9. ल : ब : ग : लगबग {*lagbag*} {*approximately*}
10. अ : ग : लबत्या : अलबत्यागलबत्या {*albatyagalbatya*} {*a fanciful formation*}
11. ल : ट : िंबू : लिंबूटिंबू {*limbutimbu*} {*small and weak*}
12. ग : ब : ड : गडबड {*gadbad*} {*noise*}
13. न : ग : रम : नरमगरम {*naramgaram*} {*soft and warm*}
14. ज : प : ात : जातपात {*jaatpaat*} {*castecreed*}
15. म : प : ान : मानपान {*maanpaan*} {*respect*}
16. भ : स : लता : भलतासलता {*bhaltasalta*} {*respectable*}
17. अ : म : धून : अधूनमधून {*adhunmadhun*} {*intermittently*}
18. त : फ : ुटका : तुटकाफुटका {*tutkafutka*} {*broken*}
19. ख : प : ान : खानपान {*khanpan*} {*food consumption*}
20. व : स : ाईट : वाईटसाईट {*vaaitsait*} {*badness*}

The rules above were not pre-specified but discovered automatically. The main observation is that both components have same length. Moreover the first characters of both components take the same matraas.

The algorithm used to handle this is:

1. Divide the full word into two exact halves.
2. Remove the first character of both halves.
3. If the remainder of both halves are the same then
   a. Report that echo word has been detected and indicate the replacement character pair.

3.5.2 Subtype 2

This type of echo words has 2 characteristics. The first is the same as the type 1 and the second is that the second component takes suffixes. Again, the replacement of characters is not arbitrary and the rules were discovered automatically. The only requirement is that the suffixes have to be pre-specified. This type bears strong resemblance to the Reduplication words.

The suffixes are ीत {*eet*} and णे {*ne*}. The णे suffix is a noun deriving Krudanta suffix. Using the format: <1st Component Character> : <2nd Component Character> : <Common Remainder> : <Suffix> : <Resultant Word>, we give the only 3 examples uncovered:

1. ड : म : ळ : ीत : डळमळीत {*dalmaleet*}
2. ख : ब : ड : ीत : खडबडीत {*khadbadit*}
3. ध : फ : ुस : णे : धुसफुसणे {*ghusfusne*}{*to backanswer*}

To detect these words we strip the suffixes and test whether the remainder is a subtype 1 echo word.

3.5.3 Subtype 3

Instead of the single character differences between the 2 components, now there is a single syllable difference. For this purpose a syllable splitter was developed. Note that गौ {*gauu*} is a syllable and ग {*ga*} is a character. After this, the method followed to extract the syllable replacement rules is the same as that for the Type 1 words. The syllable replacement rules discovered are given below in a table. Either the first or the second component is meaningful. The words of this type have a “this and that” connotation. For example पुस्तकबिस्तक {pustakbistak} stands for “books and paraphernalia”. There is some vagueness that the word expresses and these words are mostly used in informal conversation. Note that for most echo words only one of the individual components/words is meaningful. However the combination of components creates a different meaning altogether. Three examples of this type were discovered. Representation be-

low is as : <Syllable 1 >:<Syllable 2> : <Common Part> : <Resultant Word>

1. ओ : पा : ळख : ओळखपाळख {*olakhpalakh*} {*acquaintance*}
2. आ : ति : डवा : आडवातिडवा {*aadvatidva*} {*crooked*}
3. इ : पि : डा : इडापिडा {*eedapeeda*} {*pain*}

This concludes the taxonomy of the world of compound words. As can be seen a number of interesting phenomenon take place. Sometimes the boundaries between some categories become very fuzzy. The next chapter deals with the experiments carried out and the results.

4 Evaluation and Observations

A total of 465 compound words were provided as the input to the compound word analyzer. These words consisted of a mixture of all the types of compound words discussed in the previous chapter. They were studied for the various types of compound words and algorithms were implemented in Java. The experiment is a controlled one to test if all rules and algorithms can work well. The MA for Marathi was used as a part of the compound word analyzer. The power of the compound word analyzer is depicted by its ability to split the words into their individual components and after identifying they type of compounding, give at least one useful analysis. In case analyses cannot be generated at least the individual components must be obtained by splitting. The formula for accuracy is:

$$Accuracy = \frac{\#Correctly\ Analyzed\ Words}{Total\ Number\ of\ Words}$$

The overall accuracy is 94.83% which is high mainly due to the controlled nature of the experiment. Table 1 below describes the per category results of analyses.

4.1 Error Analysis

Reduction in accuracies of analyses was due to the following: errors in splitting, missing/wrong analyses, and missing rules. In almost all cases the word was split into its components.

4.1.1 Incorrect splitting

There are cases of triple compound words such as ताटवाटीभांडे {*taatvaatibhande*}{*plate-bowl-vessels*} having 3 components for which mechanisms have not been put in place and are split as ताटवा {*taatwa*} +टीभांडे {*tibhande*} since ताटवा is a word which *means odour of the flowering plants and trees*. Since these are very rare, their analyses can be explicitly stored.

4.1.2 Missing/wrong analyses

The analysis is erroneous in some cases since some of the compound words may belong to multiple categories and mistakes in the string manipulations lead to wrong outputs. Most of them are in the cases of the single meaningful component words. A better string matching mechanism is needed. Missing analyses is attributed to the absence of the entries in the monolingual lexicon and hence in the repository of inflected forms (REPO).

4.1.3 Missing rules

Words where only one component is meaningful and Sandhi words are mostly incorrectly analyzed due to missing rules in the algorithm used to analyze them. Even after adding rules due to incorrect priorities given to some rules a particular split is preferred over others leading to wrong analyses.

5 Conclusions and Future Work

No.	Type	Word count	Split count	Analyzed count	% correctly analyzed
1	Both components distinct and meaningful	234	231	225	96%
2	Partial Reduplication	25	25	25	100%
3	Only first component meaningful	35	30	30	85.7%
4	Only second component meaningful	11	8	8	72.7%
5	Negation words	12	12	12	100%
6	Reduplication words	59	59	59	100%
7	Echo words	54	54	54	100%

Table 1 : Categorized Results of Experiments

The morphological analysis of compound words is important from the standpoint of improving the accuracy and coverage of the Morphological Analyzers for all languages. Since compound words occur quite frequently in Marathi vocabulary, their detection and analysis is important from the point of view of translation. The analyzer developed has a sufficiently high accuracy to fulfill this purpose. We believe that the methodologies used are also generic enough to be applicable to other languages sharing the properties of Marathi. The design and implementation of mechanisms to use the analyses of compound words during translation from Marathi to Hindi will be quite beneficial. The accuracy depends on the quality of rules embedded and the coverage of the lexicon of the basic words and thus can be improved by adding some more complex rules which will take care of corner cases. It also eliminates the need to store the all possible combinations of words, which form compound words, explicitly in the lexicon and Wordnet. A good combination of algorithms and pattern observation can help in the identification and analysis of compound words. A deeper insight into the formation of compound words, a future task for us, will be quite useful. The experiments mentioned before were quite controlled and for true quality testing we plan to obtain an exhaustive list of compound words by running the analyzer on a large monolingual corpus followed by automatically obtaining translations of detected compound words, using a target language corpus, which will help in populating bilingual dictionaries.

References

Raj Dabre, Archana Amberkar, Pushpak Bhattacharyya. 2012. *Morphological Analyzer for Affix Stacking Languages: A Case Study of Marathi*, Conference on Computational Linguistics (COLING).

Dongli Han, Takeshi Ito, Teiji Furugori. 2002. *Structural Analysis of Compound Words in Japanese Using Semantic Dependency Relations*. Journal of Quantitative Linguistics 9(1): 1-17

Sajib Dasgupta, Naira Khan, Asif Iqbal Sarkar, Dewan Shahriar, Hossain Pavel, Mumit Khan. 2005. *Morphological Analysis of Inflecting Compound Words in Bangla*. International Conference on Computer, and Communication Engineering (ICCIT), Dhaka, 2005, pp. 110-117

Mugdha Bapat, Harshada Gune, Pushpak Bhattacharyya. 2010. *A Paradigm-Based Finite State Morphological Analyzer for Marathi.* Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pages 26–34, the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010.

Dhongde and Wali. 2009. *Marathi*. John Benjamins Publishing Company, Amsterdam, Netherlands.

Damale, M. K. 1970. *Shastriya Marathii Vyaakarana*. Deshmukh and Company, Pune, India.

Koskenniemi, Kimmo. 1983. *Two-level Morphology: a general computational model for word-form recognition and production*. University of Helsinki, Helsinki.

Antworth, E. L. 1990. *PC-KIMMO: A Two level Processor for Morphological Analysis.* Occasional Publications in Academic Computing. Summer Institute of Linguistics, Dallas, Texas.

Kim, Deok-Bong., Sung-Jin Lee, Key-Sun Choi, and Gil-Chang Kim. 1994. *A two level Morphological Analysis of Korean*. In Conference on Computational Linguistics (COLING), pages 535–539.