Acquisition of Morphology of an Indic Language from Text Corpus

UTPAL SHARMA
Tezpur University
JUGAL K. KALITA
University of Colorado
and
RAJIB K. DAS
Calcutta University

This article describes an approach to unsupervised learning of morphology from an unannotated corpus for a highly inflectional Indo-European language called Assamese spoken by about 30 million people. Although Assamese is one of India's national languages, it utterly lacks computational linguistic resources. There exists no prior computational work on this language spoken widely in northeast India. The work presented is pioneering in this respect. In this article, we discuss salient issues in Assamese morphology where the presence of a large number of suffixal determiners, sandhi, samas, and the propensity to use suffix sequences make approximately 50% of the words used in written and spoken text inflected. We implement methods proposed by Gaussier and Goldsmith on acquisition of morphological knowledge, and obtain F-measure performance below 60%. This motivates us to present a method more suitable for handling suffix sequences, enabling us to increase the F-measure performance of morphology acquisition to almost 70%. We describe how we build a morphological dictionary for Assamese from the text corpus. Using the morphological knowledge acquired and the morphological dictionary, we are able to process small chunks of data at a time as well as a large corpus. We achieve approximately 85% precision and recall during the analysis of small chunks of coherent text.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms: Experimentation, Languages

Additional Key Words and Phrases: Morphology, machine learning, Indo-European languages, Assamese

Authors' addresses: U. Sharma, Department of Computer Science and Engineering, Tezpur University, Tezpur—784028, Assam, India; email: utpal@tezu.ernet.in; J. K. Kalita, Department of Computer Science, University of Colorado, Colorado Springs, CO 80933, USA; email: jkalita@uccs.edu; R. K. Das, Department of Computer Science and Engineering, Calcutta University, Kolkata, India; email: rajibkumardas2002@yahoo.com.

ACM Reference Format:
Sharma, U., Kalita, J. K., and Das, R. K. 2008. Acquisition of morphology of an Indic language from text corpus. ACM Trans. Asian Lang. Inform. Process. 7, 3, Article 9 (August 2008), 33 pages. DOI = 10.1145/1386869.1386871. http://doi.acm.org/10.1145/1386869.1386871.

1. INTRODUCTION

Incorporating competence in morphological analysis is crucial for natural language processing (NLP) systems. Morphology is a structural phenomenon and the effect of morphology is visible in the physical representation of words in repetitive and sequential chunks. This has led many researchers to attempt unsupervised acquisition of morphology from unannotated corpora. While most of these approaches (e.g., [Gaussier 1999; Goldsmith 2001; Creutz and Lagus 2005]) are sound as general approaches, they do not take into account many factors specific to languages, scripts, and encoding schemes used for representation. In this article, we discuss several such factors with focus on Assamese, a morphologically rich Indic language. In particular, we consider unsupervised acquisition of morphology from a text corpus.

Assamese is the easternmost Indo-Aryan language used natively by about 15 million people in the state of Assam and adjoining regions in northeast India. It is spoken by another 15 million people as the second or the third language. It is a highly inflectional language with several characteristics distinct from other Indic languages. So far little computational work has been done for this language. Ours is the first effort in this regard as far as we know and, thus, is pioneering. There are many such languages that receive very little attention from computational linguistic research in terms of both availability of funds and number of researchers. For such languages, it is very important to have suitable but inexpensive computational acquisition methods.

In Section 2, we discuss the nature of morphology in general and its significance as a structural phenomenon. In Section 3, we focus on unsupervised learning as a useful approach for acquisition of morphology and review existing methods for unsupervised acquisition of morphology. In Section 4, we consider morphological phenomena in Assamese and describe how these are represented in written form. In Section 4.2, we briefly discuss the problems faced by languages from poorer parts of the world in simply being represented in computers for reading, writing, and other computational purposes. In Sections 5, 6, and 7, we present details of experiments we perform with the goal of acquiring Assamese morphology from a text corpus using selected existing methods, and propose a new method for the task, along with experimental results. In Section 8, we conclude the article.

2. MORPHOLOGY AND SYNTAX

The structural constraints enforced by a language manifest as morphology and syntax. Loosely put, morphology and syntax are complementary to some extent, and morphologically rich languages, such as Assamese, have comparatively less rigid syntax (in particular, word order) rules. Assamese is a relatively *free word order* language.

We classify morphological transformations into two broad types from a structural perspective—one, a single word is transformed into another form, usually by changing the vowel constituents in the word. For example, *write* → *wrote*. The other kind of transformation is that in which two or more *morphemes* are concatenated to obtain a single word. For example, *cheer* + *ful* + *ly* → *cheerfully*. Sometimes as a result of concatenation of morphemes, the letter sequence in the spelling of the aggregate word is different from those in the spellings of the constituent morphemes. For example, *happy* + *ness* → *happiness.*

The acquisition of knowledge about the structural aspects of a language, including its morphology, is essential for the acquisition of the language. Work done in the 1980s regarding the lexicon led to the realization that morphology is an autonomous module on par with the phonological and syntactic modules. On the other hand, syntactic systems capable of handling word formation operations in more restricted ways were developed during that period. Such systems could avoid many of the shortcomings encountered in earlier efforts [Borer 1998]. Leiber stated that in the conceptually simplest theory, all morphology would be part of the theory of syntax [Leiber 1992]. However, most researchers have come to the conclusion that describing morphology within syntax is impossible and probably undesirable. Rewrite schema and hierarchical structures proposed for morphology are systematically incompatible with notions of phrase structures proposed for syntax [Borer 1998]. Chomsky, too, pointed out that syntax has properties completely unrelated to morphology, phonology, and semantics [Schneider 1998, p. 15].

3. UNSUPERVISED ACQUISITION OF MORPHOLOGY

Morphology is a structural phenomenon and its effect is observed in the physical representation of words, whether spoken or written. Approaches for incorporating morphological competence in NLP systems range from hand-coding of morphological "rules" provided by linguists to automatic identification of morphological rules from examples of text inputs. Perhaps the most widely cited work on hand-coding morphological rules is the Porter's method for stemming [Porter 1980]. This method deals with suffixational morphology. Many others have subsequently attempted to improve this method (e.g., [Saravanan et al. 2002]). Automatic identification of morphological rules can be performed in a supervised or unsupervised manner. The former requires a training corpus specially prepared for the purpose. For instance, Daelemans took as input a part-of-speech (POS) tagged corpus for the lexical acquisition task [Daelemans 1993]. Unsupervised approaches take raw (unannotated) corpora as input. Most unsupervised approaches are primarily probabilistic (e.g., [Goldsmith 2001; Creutz 2003; Creutz and Lagus 2004; 2005]). However, there are exceptions, for example, approaches like the one described in [Gaussier 1999] are not strictly probabilistic. We too use partial matching of words, statistical support, and set-theoretic principles for the task. We have performed experiments for Assamese using our approach. We have not come across reports of experiments in unsupervised acquisition of morphology for any other Indic language.

Unsupervised acquisition of language draws inspiration from the fact that a child learns her (or his) mother tongue by exposure to linguistic expressions. Of course, the situation for a child is not identical to that of a computer. She neither consults a dictionary nor gets explicit instructions on grammar or vocabulary, but she can perceive real-world entities and events in the environment that the linguistic expressions describe. This perception is an alternative source of information. In addition, a child possesses knowledge gathered by the human race through millions of years of evolution. For a computer provided only with an unannotated corpus of linguistic expressions, there is no alternative representation of the information. The processing in such a situation is limited to only "structure" and does not involve "meaning." Since computers are good at processing data, if there are any regularities in the structure of the expressions, a computer program should be able to discover them, at least theoretically speaking. Since concatenative morphology manifests as structural transformations of words, a computer should be able to acquire morphology by performing an unsupervised analysis of the corpus. Most morphology acquisition methods, including ours, deal with concatenative morphology.

3.1 Gaussier's Approach

In Gaussier [1999], Gaussier presented a method for acquiring suffixes used in a raw text corpus. The idea is to first find pairs of words, say w_1 and w_2, which have identical initial portions of length at least p. The portions of w_1 and w_2 after the matching portions are together referred to as a *pseudo-suffix pair*. The language-independent value of p suggested is 5. A pseudo-suffix pair (α_1, α_2) is accepted as a pair of suffixes of the language if there is at least one more pair of words, say w_3 and w_4, that also yields the same pseudo-suffix pair. That is, if

$$\begin{aligned} w_1 &= \beta_1 + \alpha_1, \\ w_2 &= \beta_1 + \alpha_2, \\ w_3 &= \beta_2 + \alpha_1, \\ w_4 &= \beta_2 + \alpha_2, \end{aligned}$$

α_1 and α_2 are two suffixes in the language.

3.2 Goldsmith's Approach

Some unsupervised morphology acquisition methods (e.g., [Snover et al. 2002]) are based on probabilistic models. A particularly interesting approach that can be seen as a special case of probabilistic modeling was presented by Goldsmith [2001]. It is based on the concept of minimum description length (MDL). The intuition is that if all the morphemes, which are the basic elements of all words, involved in an input corpus are assigned distinct numeric values in the smallest possible number space, the input can be represented as a sequence of these numbers. Identification of morphemes can be guided by the goal of minimizing the length of the representation of the input corpus, which depends on the total number of morphemes as well as the representation lengths of the individual morphemes in number of bits.

4. MORPHOLOGICAL PHENOMENA IN ASSAMESE

Our work is motivated by the fact that morphology is one of the most important structural phenomena in Assamese. Morphological transformations are more common in Assamese than in other Indic languages and in English. In a preliminary study, we find that about 48% of words in an Assamese text of around 1,600 words were inflectional or derivational whereas only about 19% of words in an English text of about 1,400 words were so. Similarly, in a sample Hindi text of about 1,000 words, 26% were inflectional and derivational. Suffixation, prefixation, and compound formation are the major morphological phenomena in Assamese [Bora 1968; Goswami 1990; Sarma 1977; Medhi 1999]. Of these, suffixation is the most common. Suffixes frequently attach to already suffixed words, giving rise to suffix sequences. For example, *l'rAkeiTAkeino* = *l′rA* + *keiTA* + *k* + *ei* + *no* (ল'ৰাকেইটাকেইনো; boy + a few + accusative + only + emphasis). There are about 200 suffixes in Assamese, but due to the use of suffix sequences, for some nouns and verbs there can be several hundred suffixal forms. The merging of words to obtain compounds is similar to those of other Indic languages to a large extent and generally follow the *sandhi* and *samas* [Vasu 1891] framework. For example,

AshA + atIt = AshAtIt (আশা + অতীত = আশাতীত)
kRhSNa + arjun = kRhSNArjun (কৃষ্ণ+ অৰ্জুণ = কৃষ্ণাৰ্জুণ).

4.1 Determiners

A class of common suffixes in Assamese is that of the *determiners*. There is a plethora of suffixes that can be used as determiners. Some examples are *To, khan, khilA, catA, pAt, khini, zan, grAkI, gac, bor, bilAk, zopA, zanI, phAl, dAl, gAl, kocA, darA,* etc. Such a large number of determiners are not seen in other Indic languages. Primarily the determiners are used as suffixes with nouns and pronouns according to certain subtle linguistic norms[1]: for example, *mAnuhTo, phulkhini, gczopA,* etc. In many situations, the corresponding noun itself is not there, and a general pronoun plays that role, for example, *eiTo, eizn, eikhan, eigrAki,* etc. (These are various forms of *this* and *that*). There are certain determiners that make the objects plural. For example, *bor, khini, brinda, bilAk,* (বোৰ, খিনি, বৃন্দ, বিলাক) etc. It is useful to compare and contrast the role of determiners of Assamese with that in other Indic languages. In Hindi there is no morpheme corresponding to the basic (for singular number) determiners in Assamese, but plurality is achieved by certain affixations. For example, *boy, the boy* and *the boys* in English are written in Hindi as *larkA, larkA* and *larke*. In Assamese, these are *l'rA, l'rATo,* and *l'rAbor* (ল'ৰা, ল'ৰাটো, ল'ৰাবোৰ). In Bengali, these are *chele*, *cheleTA*, and *chele gulo* (ছেলে, ছেলেটা, ছেলে

[1] See http://www.assam.org/assam/language/jugalpaper/node1.html for a brief discussion.

গুলো). Though in Bengali the use of determiners is similar to that in Assamese, the number of different determiners is not as large as in Assamese.

As with other Indic languages, Assamese expressions often contain nonsense words. A nonsense word simply rhymes with the preceding actual word and roughly means *and such*—for example, *kitAp citAp* (কিতাপ চিতাপ), where *kitAp* means *book* and *citAp* simply implies *other things like book*. Some nonsense words have meaning in other contexts. For example, *colA tolA* (চোলা তোলা), where *colA* means *shirt* and *tolA* can mean *pick* (imperative) in other contexts. Nonsense words can be inflected like regular words.

4.2 Representation of Morphological Phenomena in Texts

Speech is the primary form of expression for natural languages. The evolution of linguistic features, including morphology, is based mostly on the spoken form. Hence, an appropriate approach for acquisition of morphology is to consider the phonological form of utterances. For example, Gasser [1994] described a connectionist approach that takes as input phones and outputs the associated roots and inflections. In most languages, the written form actually encodes the phones in an expression using symbols from an alphabet. In some writing systems, the mapping between written symbols and phones is not very strict, and there can be some loss of information.

Unlike spoken language, the written form of a language is not naturally acquired by humans; it is usually learned through a process of formal training. The ease with which morphological phenomena can be observed in written texts depends on the orthography and the choice of word boundaries. At one extreme, there are writing systems in which the individual symbols indicate entities and concepts. Such texts may not reveal morphological features present in the spoken form. Syllabic scripts represent the phonology of expressions more realistically, but there too the mapping between written symbols and phonology is sometimes irregular.

The choice of word boundaries do not follow the same principles across languages. For example, in Hindi most case markers are written as distinct words after nouns, whereas in Assamese the equivalent case markers are written as suffixes.

4.2.1 *Irregularities in Assamese Writing Scheme.* In Assamese, sometimes an implicit vowel *a* is assumed after a consonant. For example, the word *bAr* (বাৰ) is pronounced in two different ways with two different meanings—*baar* to mean *number of times* or *day of the week*, and *baara* to mean *twelve*. However, Indic scripts do not have irregularities such as variable pronunciation for letters as is common in English written using the Roman script (e.g., the letter *u* in *but* and *put*), letters not pronounced in certain spellings (e.g., the letter *b* in *debt*), etc.

Sometimes individual sound elements undergo modifications when they occur with certain other sound elements. As a result, the spelling of a concatenated sequence of morphemes may be different from the letter sequence of the individual morphemes. For example, *garu* + *e* → *garuwe* (গৰু+ ে- → গৰুৱে, meaning *cow* in ergative case). Another noticeable characteristic of Indic scripts is the use of different vowel operators at different positions with respect to the consonant to which it is attached—left, right, above, below, and also in multiple parts at different positions, though vowel operators are invariably pronounced after the consonants to which they are attached.

A single-quote mark used in words such as *m'H* (ম’হ), meaning *mosquito*, modifies the implicit vowel associated with the preceding consonant. However, in words such as *HAiwe'r* (হাইৱে’ৰ), meaning *of highway*, the single-quote mark simply indicates that the preceding letter string is a transliteration of a foreign word (*HAiwe+r*). Assamese texts, like texts of other Indian languages, commonly contain foreign words, phrases, and abbreviations. Sometimes these are written in the original spelling (i.e., using the foreign alphabet) and sometimes transliterated into Assamese script. Often such words are also subject to inflection.

4.2.2 *Encoding of Texts in Computer.* The convenience of implementing a method for automatic analysis of texts depends on the encoding scheme employed to represent the texts inside the computer. Due to the very nature of the Roman script, straightforward encoding schemes such as ASCII are well suited for it; in fact, the ASCII script was initially designed for the Roman script as used in American English. For Indic scripts, including Assamese, special provisions are needed in encoding schemes to represent elements such as ligatures (*juktakshars*) and vowel operators. Though encoding schemes such as ISCII and Unicode have been proposed for Indic scripts as standards, many issues related to their support in current computing platforms are yet to be resolved. As a result, several nonstandard distinct encoding schemes are in use. Many such widely used encoding schemes are essentially font encodings. At present, font encodings such as Aadarsha Ratneswar, Luit, Kamakhya, Ramdhenu, etc., which do not conform to any international standards, are in use for Assamese texts even in professional software. Unlike with ASCII encoding, text processing experiments are not very convenient with these diverse encodings.

It is also possible to develop a writing system for Assamese in which each letter is denoted by a distinct Roman letter or a letter sequence so chosen that the text can be written or read unambiguously; in fact, such a system is informally used by many to write emails, SMS messages, and chat room messages using devices with Roman keyboards. This approach makes it possible to use ASCII encoding for Assamese texts and create software for automatic lossless transliteration. Such a Roman transliteration scheme has been used in our work as an encoding for Assamese texts. However, the use of the Roman alphabet for Assamese is not socially or culturally acceptable in formal settings.

5. ACQUISITION OF ASSAMESE MORPHOLOGY FROM A TEXT CORPUS

In this work, our focus is the acquisition of concatenative morphology of Assamese from an adequately large raw text corpus so that subsequently we can use this knowledge to recognize suffixes in words occurring in possibly small texts. In this section, we first specify a simple approach to produce candidate decompositions of words in the training corpus. Then, we discuss criteria to effectively select valid suffixes from these decompositions, and methods for dealing with issues due to script, irregular morphological features, and suffix sequences. In Section 6, we put together all these ideas into a sequence of steps for the morphology acquisition process.

The suffixes identified through an unsupervised method are of four types—true suffixes, composite suffixes (i.e., concatenation of more than one suffix), compound parts (i.e., words that attach with preceding words to form compounds), and invalid suffixes. To emphasize that in the analysis of a word, the portion after the base is not necessarily a true suffix, we use the term *morphological extension* to refer to that portion of the word. To quantify the performance of different methods, we compute precision, recall, and F-measure of the results. We compute precision, P, and recall, R, respectively as

$$\begin{aligned} P &= \frac{S * 100}{S + B}, \quad \text{and} \\ R &= \frac{S * 100}{T}, \end{aligned}$$

where S is the number of valid (true) suffixes identified, B is the number of invalid suffixes identified, and T is the number of suffixes actually present in the input. The denominator in the expression for precision should be the total number of morphological extensions identified, but we ignore the counts of composite suffixes and compound parts since these are valid but not our primary target. An aggregate of precision and recall is the F-measure [Chen et al. 2004] expressed as

$$f = \frac{2 * P * R}{P + R}.$$

In terms of S, B, and T it is

$$f = \frac{2 * S * 100}{S + B + T}.$$

First, we carry out experiments with the two distinct known methods mentioned earlier—Gaussier's method (see Section 3.1) and Goldsmith's method (see Section 3.2)—using a corpus of about 116,000 words (corpus A) from 231 newspaper articles. A manual analysis showed 187 true suffixes in the corpus.

Using Gaussier's method we experiment with different values for p, the base length. The results obtained are shown graphically in Figure 1. The maximum value of F-meaure is 59.85%. Precision increases from 25.68% to 78.26% as the number of letters considered in the base increases. Recall starts at 85.56% and comes down to 28.88% during the same time.

An implementation of Goldsmith's method is freely available as the software package called Linguistica. For our experiment, the input corpus needs to be preprocessed since in the Roman-script-based encoding scheme used in the

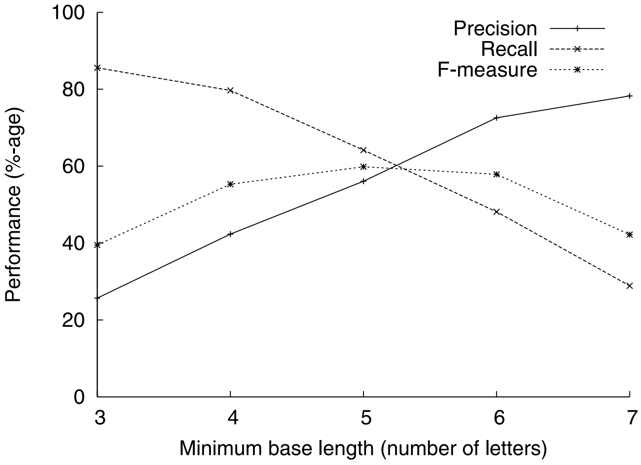

Fig. 1. Effect of base length, p, in Gaussier's method, when used with Assamese text.

Table I. Summary of Results from Linguistica (Goldsmith's Method [Goldsmith 2001])

Number of input words	: 116,096
Number of distinct input words (including hyphenated words)	: 20,685
No. of distinct morphological extensions found, n	: 167
No. of distinct valid suffixes identified, s	: 80
No. of distinct suffixes that should be further broken up, q	: 57
No. of morphological extensions that are compound parts, c	: 21
No. of invalid morphological extensions, b	: 9
Actual number of suffixes present in the input, S	: 187
Precision of single suffix identification ($s/(s+b)$)	: 89.89%
Recall of suffix identification (s/S)	: 42.78%
Proportion of noninvalid morphological extensions to total morphological extensions ($(s+q+c)/n$)	: 94.61%
F-measure	: 57.97%

corpus, certain Assamese letters are represented using more than one Roman letter, and some are represented using nonalphabetic characters. The results of this experiment are summarized in Table I.

The reported performance of Goldsmith's method for English corpora is precision = 85.9% and recall = 90.4%. (The performance figures for this task by Gaussier's method had not been provided in his article [1999].) The performance is not as high for an Assamese corpus. We notice that language characteristics such as the presence of a suffix sequence in words, presence of large number of foreign words and abbreviations, orthographic peculiarities, etc. need to be more appropriately addressed.

5.1 Experiments in Assamese Morphology

The results obtained by the aforementioned methods leave scope for improvement. The main shortcoming is that the presence of a suffix sequence in a word is not adequately addressed. We also want to perform morphological analysis of not only a large corpus, but also of smaller chunks of texts. From the experiments with a large corpus, we accumulate morphological information to use during analysis of smaller text chunks. We model the morphology acquired through an analysis of the input training corpus in the form of a collection of suffixes and suffix sequences, and by specifying the criteria for identifying their presence in different words. Simultaneously, we build a lexicon that contains the analysis of the words in a compact form. This is a morphological lexicon that provides more insight about words than a plain listing of the words encountered does. Our approach is a consolidation and extension of the work we described elsewhere [Sharma et al. 2002, 2003, 2006]. It is fully described in [Sharma 2006]. However, before we present our approach to obtaining suffixal decomposition for Assamese texts, we discuss a series of experiments we performed with an Assamese corpus and present the results of these experiments. An analysis of these results helps us in developing a step-by-step method for acquiring Assamese morphology from a corpus. Some of the experiments led to development of heuristics that are used in the suffix acquisition approach discussed in Section 6.

5.1.1 *An Initial Decomposition.* Because of the simple concatenative nature of most morphological transformations in Assamese, first we attempt to identify suffixes by considering words that can be obtained by appending some letters to some other words. The letters that are appended are possible suffixes. That is, if

$$[w_1 = w_2 + \sigma],$$

where w_1 and w_2 are two words in the corpus, and σ is a string of letters that is appended to w_2, σ is a candidate suffix. This idea is similar to the idea underlying Gaussier's method. The results of implementing this simple idea are summarized in Table II. A few sample decompositions are shown in Table III. For calculating recall, we identify and refer to the set of suffixes that are actually present in the corpus.

This attempt produces many spurious decompositions since some words can match leading portions of other unrelated words. For example, in Table III, in decomposition 7, *aH* is an invalid suffix, and *kalaH* (কলহ) is actually a root word not related to *kal* (কল). Decomposition 8 is invalid because the base *bi* and the derived word *bi/shwzy* are not semantically related. In fact, *bi* is not a true word; it has occurred in the corpus as the transliteration of an English abbreviation “B J P,” the initials of an Indian political party.

5.1.2 *Selecting Valid Suffixes.* Our initial decomposition identifies almost all the suffixes (high recall, 98.93%), but also too many nonsuffixes (only 1.35% are suffixes). Hence, we apply heuristics based on statistics as well as other language-specific and script-specific considerations, to identify valid suffixes in the initial set of candidate suffixes. Before we process the decompositions further, we remove the single-quote mark from the beginning and end of the candidate suffixes. First, we perform experiments to gauge the effect of the

Table II. Summary of Initial Decompositions from a Corpus of 231 Newspaper Articles

Number of input words	: 116,096
Number of distinct input words (original count was 20,685, but in hyphenated words only the last components are retained).	: 20,140
Number of decompositions (including multiple for same word).	: 29,054
No. of distinct morphological extensions in the decompositions, n	: 13,715
No. of distinct valid suffixes identified, s	: 185
No. of distinct suffixes that should be further broken up, q	: 654
No. of morphological extensions that are compound parts, c	: 2,218
No. of invalid morphological extensions, b	: 10,658
Actual number of suffixes present in the input, S	: 187
No. of distinct bases that occur in decompositions	: 5,186
No. of bases that occur in more than one decomposition	: 2,820
No. of bases that are, in turn, decomposed, too	: 3,638
No. of invalid decompositions	: 12,234
Precision of single-suffix identification ($s/(s + b)$)	: 1.71%
Recall of suffix identification (s/S)	: 98.93%
Proportion of noninvalid morphological extensions to total morphological extensions ($(s + q + c)/n$)	: 22.29%
F-measure	: 3.35%

Table III. Some Sample Decompositions

1.	$[kitApar = kitAp + ar]$	(কিতাপৰ)	/of book(s)/
2.	$[kitApat = kitAp + at]$	(কিতাপত)	/in book(s)/
3.	$[kitAparHe = kitAp + arHe]$	(কিতাপৰহে)	/of book(s), exactly/
4.	$[kitAparHe = kitApar + He]$		
5.	$[kitApkhanar = kitAp + khanar]$	(কিতাপখনৰ)	/of the book/
6.	$[bi/shwzy = bi/shw + zy]$	(বিশ্বজয়)	/world victory/
∗7.	$[kalaH = kal + aH]$	(কলহ)	/pot/ *kal* = banana
∗8.	$[bizy = bi + zy]$	(বিজয়)	/victory/

The decompositions marked * are invalid.

following potential criteria to use in our quest for identification of suffixes. A study of the results of these experiments is necessary to select the appropriate criteria to subsequently use actual suffix identification.

—Frequency of candidate suffixes. We count the number of distinct bases with which each candidate suffix occurs, that is, its frequency, and then retain only those candidate suffixes that have a frequency above a threshold. We experiment with different frequency thresholds, and the results are shown graphically in Figure 2. The highest value obtained for F-measure is 66.47% with suffix frequency threshold 7. At this point precision is 73.86% and recall is 60.43%.

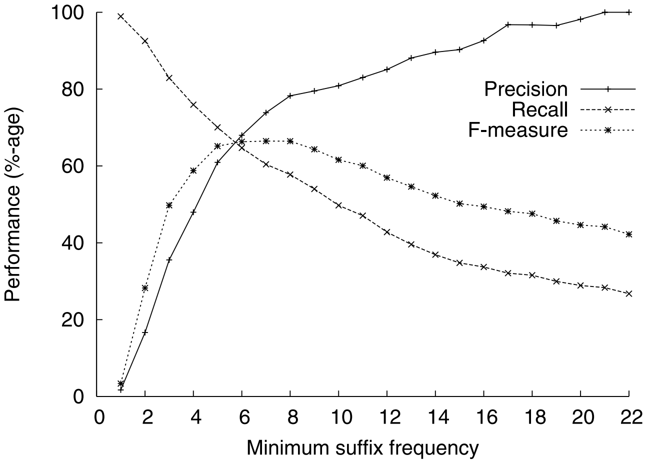

Fig. 2. Effect of frequency in suffix selection.

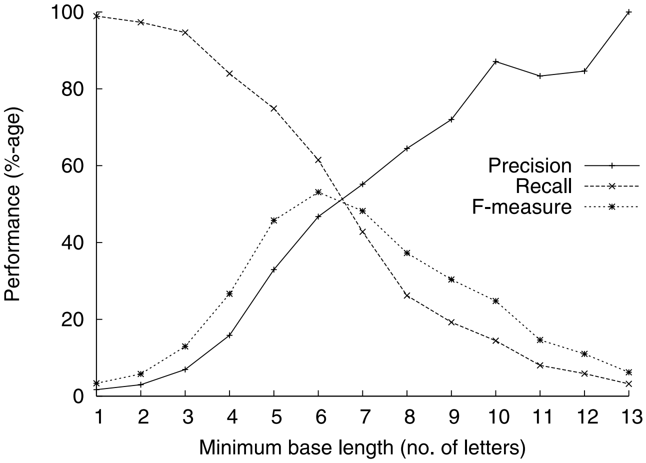

Fig. 3. Effect of base length (all letters) in suffix selection.

—Base length. Many invalid decompositions involve short bases, that is, bases with very few letters. So, we impose a lower limit on the length of the bases of the decompositions. Selecting decompositions based on the length of bases is an important criterion in the method proposed by Gaussier [1999] (see Section 3.1). Computing the length of words or portions of words must be carefully done since most of the prevalent encoding schemes for Assamese script, including the one we use, uses a non-uniform length of representation for the different letters. Figure 3 presents the results graphically. The highest value obtained for F-measure is 53.12% with base-length threshold 6. At this point precision is 46.75% and recall is 61.50%.

—Phoneme count of base. We start with a hypothesis that longer words are more stable than shorter words. That is, if a word w_2 can be obtained by

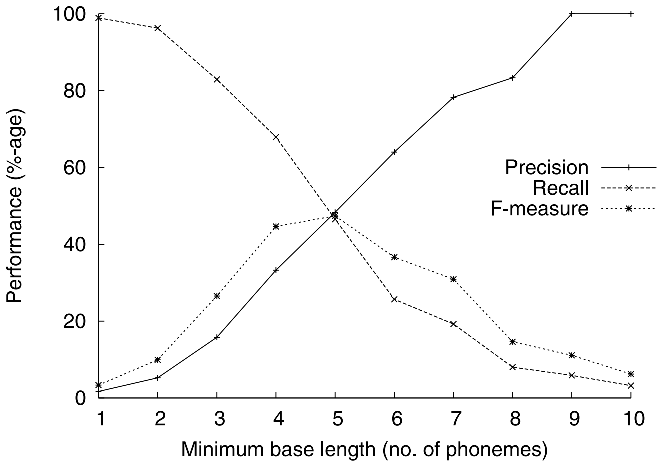

Fig. 4. Effect of base length (phonemes) in suffix selection.

concatenating one or more letters to another word w_1, the likelihood that the two words are semantically related is proportional to the phonetic length of w_1. To test the validity of this hypothesis, we count the number of phonemes in the bases of decompositions. Since character counts do not reflect the actual phonetic length of words in Assamese texts, we use the following criteria to obtain a rough approximation of the phoneme count:

(1) Each consonant is a phoneme. Each consonant in a ligature is counted independently.

(2) Each vowel that occurs at the beginning of a word or after another vowel is a phoneme.

The effects of the selection of decompositions based on the phoneme count of the bases is shown graphically in Figure 4. The highest value obtained for F-measure is 47.41% with base-length threshold 5. At this point precision is 48.33% and recall is 46.52%.

—Base frequency. In line with idea of selecting suffixes by imposing a minimum frequency threshold, we test the possibility that bases that occur with a high number of distinct candidate suffixes are more likely to be valid. Our experimental results, however, do not validate this assumption.

—Textual context. Instead of considering words from the entire training corpus together, for the purpose of decompositions, we consider one article at a time and find possible decompositions where the derived word as well as the base occur in that article. The idea is that in individual text articles, words with similar leading letter sequences are more likely to be semantically related. For the corpus A of newspaper articles mentioned earlier, the results obtained are summarized in Table IV. The results of the article-by-article decomposition exercise are along expected lines.

Table IV. Summary of Article-by-Article Decomposition of Words

Number of newspaper articles	:	231
Number of input words	:	116,096
Average number of input words per article	:	502
Actual number of suffixes present in the input	:	187
Number of distinct decompositions	:	8,585
Number of distinct morphological extensions	:	2,791
Distinct {S:154; Q:362; C:794; B:1481}		
Precision	:	9.42 %
Recall	:	82.35 %
F-measure	:	16.90 %

S: Suffix; Q: Suffix-sequence; C: Compound parts;
B: Invalid morphological extension

5.1.3 *Combination of Identification Criteria.* From the preceding discussion, it is seen that a suitable combination of multiple selection criteria for morphological extensions is likely to give better performance than any single criterion. After considering various combinations of suffix identification criteria, the following sequence of steps are seen to give the best results.

(1) Obtain the initial decompositions D from the combined corpus.
(2) Obtain the initial decompositions D_a article-by-article from the corpus.
(3) From D_a retain the decompositions in which the bases have two or more phonemes, and, the candidate suffixes have frequency f (say, three) or more in D.

The threshold occurrence count f depends on the size of the corpus. Empirically it is seen that a good estimate is

$$\begin{aligned} & f = 2, && n <= 50{,}000, \\ \text{and,} \quad & f = \lceil \tfrac{n}{50{,}000} \rceil, && n > 50{,}000, \end{aligned}$$

where n is the number of words in the input corpus. The counts of the various types of candidate suffixes obtained from the input corpus A by the steps given previously are

$$\begin{aligned} &\mathrm{B} = 92, \mathrm{C} = 141, \mathrm{Q} = 197, \mathrm{S} = 140, \\ &\text{Recall} = 74.87\%, \quad \text{Precision} = 60.34\%, \quad \text{F-measure} = 66.83\%, \end{aligned}$$

where S is the number of suffixes identified, B is the number of invalid morphological extensions, Q is the number of composite suffixes (suffix sequences), and C is the number of compound parts identified. This result is better than the other combinations we tried, because the F-measure is about the best, and recall value is good. In Figure 2 where we use suffix frequency alone, we find a slightly better F-measure for specific values of suffix frequency threshold, but the recall in those cases is below 65%.

5.1.4 *Compound Parts.* An intuitive criterion for distinguishing a compound part from other candidate suffixes is that compound parts are likely to occur as independent words or in some other derived form in a sufficiently large corpus. For example, the part *m/nt_rI* (মন্ত্রী meaning *minister*) of the word *p_rdhAnm/nt_rI* (প্রধানমন্ত্রী meaning *prime minister*) is likely to occur independently or in its other derived forms. However, in the case of Assamese, this criterion is not entirely dependable because in Assamese some suffixes are, optionally, written detached from the base, thereby making a suffix appear to be an independent word. For example, the suffixed word *chAt_rsklo* (ছাত্রসকলো meaning *the students too*) is also written as *chAt_r sklo*. In other words, the distinction between compound parts and certain suffixes in Assamese is inherently vague. On the other hand, we note that for the ultimate goal of identifying the structure of words, we do not lose anything if we continue to treat compound parts as suffixes. Decompositions of compounds can help in their recognition if the constituent parts are present in the lexicon. Hence, we consider compound parts and suffixes selected from the initial decompositions alike for the purpose of morphological analysis of words.

5.1.5 *Suffix Sequences.* Unless composite suffixes are decomposed into sequences of constituent suffixes, they would appear to be single suffixes and make the set of suffixes unduly large. The decomposition of a word is a complete decomposition if it is valid and none of the parts in the decomposition can be further decomposed. We call the number of parts in a decomposition the *degree* of the decomposition.

Suffix sequences can be identified by successively replacing the base of a decomposition by a possible decomposition of it as long as such a replacement is possible. That is, if $[w_1 = w_2 + p_1]$ and $[w_2 = \rho_1 + p_2]$ are two decompositions, a combined decomposition can be written as

$$[w_1 = \rho_1 + p_2 + p_1] \ ,$$

where we get $p_2 + p_1$ as a suffix sequence. We call this process *recursive reduction* of the bases. In the set of initial decompositions, there may be multiple decompositions for some words, each involving a different base-suffix pair. Of two alternative decompositions, the one involving the longer base is a shallower decomposition. Suppose $[w_2 = \rho_2 + p_3]$ is another decomposition of w_2. Then we have another possible combined decomposition for w_1. Hence, before we perform recursive reduction of bases, we unify the different decompositions of the same word wherever present. That is, using as input the multiple decompositions of a word, we generate a single decomposition in which the word is decomposed at each point at which it was decomposed in any of the input decompositions. For example, for the word w_2, suppose the first decomposition is shallower (ρ_1 is longer than ρ_2). Then the unified decomposition for w_2 is

$$[w_2 = \rho_2 + p_4 + p_2] \ ,$$

such that $\rho_1 = \rho_2 + p_4$ and $p_3 = p_4 + p_2$. After unifying the decompositions for all words wherever applicable, we perform recursive reduction of bases to obtain the decompositions with suffix sequences. The positions at which a word is broken up in a decomposition are partition points.

Table V. Initial Identification of Suffix Sequences

Total number of distinct words : 20,140				
Min base length	Min suff.-seq. frequency	No. of suff.-seq. identified		
		A+C	B+D	E
1	1	638	470	2468
1	2	352	240	387
1	3	260	173	156
1	4	212	140	100
1	5	172	114	70
2	1	555	411	636
2	2	325	214	170
2	3	258	161	76
2	4	207	131	43
3	1	399	293	235
3	2	239	161	44
3	3	185	123	17
4	1	276	207	129
4	2	158	108	16

We perform this exercise using the decompositions identified from the initial decompositions by further processing using the combination of criteria as stated in Section 5.1.3. The quantitative summary of the exercise is given in Table V. The column headings A, B, C, D, and E are described here:

A. correctly identified, for example, the suffix sequence $(A + b + lE)$ in the decomposition

$$[krAblE = kr + A + b + lE]$$ (কৰাবলৈ)

meaning *to get done* (*by someone else*),

B. correctly identified, but needs further decomposition, for example, the suffix sequence $(A + znk)$ in the following decomposition should actually have been $(A + zn + k)$

$$[krAznk = kr + A + znk]$$ (কৰাজনক)

meaning *one who does*

C. correct but identified in inappropriate decompositions only, for example, the suffix sequence $(A + zn + r)$ is valid, but the following decomposition from which it has been obtained is not valid

$$[mHAznr = mH + A + zn + r]$$ (মহাজনৰ)

meaning *shopkeeper's*

D. correct but needs further decomposition and identified in inappropriate decompositions only. For example, the suffix sequence $(A + zne)$ should actually be $(A + zn + e)$, and it has been obtained from the following decomposition which is not valid

$$[mHAzne = mH + A + zne]$$ (মহাজনে)

meaning *shopkeeper* (*ergative*)

E. incorrect, for example, the suffix sequence ($Ai + bor$) in the following decomposition is not valid

$$[ThAibor = Th + Ai + bor] \qquad (\text{ঠাইবোৰ})$$

meaning *the places*.

A suffix sequence may be incorrect either because one (or more) of its constituents are not a valid suffix part, or the breakup of the sequence is not correct. One important observation in the results is that most incorrect suffix sequences actually have some common defective subsequences.
Suppose we get the following decompositions by these steps:

$$\begin{aligned} & [\omega_i = \beta_i + p_1 + p_2 + p_3], \\ \text{and} \quad & [\omega_j = \beta_j + p_1 + p_2]. \end{aligned}$$

For compact representation of a set of decompositions, we record only the first decomposition since the second one can be extracted from the first.

Alternative suffix sequences. Like alternative decompositions, we also come across multiple suffix sequences that produce the same composite suffix upon concatenation. These are alternative suffix sequences. For example, the suffix sequences in the following decompositions are alternative suffix sequences:

[$kukur + bork + lE$] (কুকুৰবোৰকলৈ) meaning with the dogs
[$crAi + bor + k + lE$] (চৰাইবোৰকলৈ) meaning with the birds.

5.1.6 *Boundary Adjustment in Word Decompositions.* The suffix and suffix sequence identification method just discussed is susceptible to certain tricky morphological phenomena. For instance, if the input corpus contains the words, *mAnuH* (মানুহ, meaning *human*), *mAnuHr* (মানুহৰ, meaning *of human*) and *mAnuHrUpe* (মানুহৰূপে, meaning *as a human*), we may obtain the decompositions

$$[mAnuHr = mAnuH + r]$$
$$[mAnuHrUpe = mAnuH + r + Upe]$$

Here, the letter string *Upe*, which is not a suffix, is identified as one. The breakup of *rUpe* as $r + Upe$ is spurious. To avoid such spurious breaking up of suffixes, after the suffixes and suffix sequences are identified, for each suffix we check if all occurrences of the suffix have a common letter sequence preceding it. If so, the suffix should be extended to include that common letter sequence preceding it. We refer to this as *suffix extension*. In the aforementioned example, since the suffix *Upe* is found to be always preceded by the letter *r*, we extend *Upe* to get *rUpe*.

5.1.7 *Very Irregular Morphological Extension Parts.* There are certain suffixes that are valid but hold only in very few cases, that is, they are not regular. For example, the decomposition

$$[clothe = cloth + e]$$

is valid but the suffix *e* is not regular, in the sense that only in very few cases it adds to a base to give a valid derivative. Decompositions such as [caste = cast + e] (a valid word decomposed using *e* as suffix) are not valid. Similarly, the suffix *e* cannot be added to the word *path* though it is a noun like *cloth* and have similar structure. For Assamese, consider the following decompositions:

(1) $[thiyE = thiy + E]$ ([থিয়ৈ = থিয় + ঐ]) meaning *standing*
(2) $*[krilE = kril + E]$ ([কৰিলৈ = কৰিল + ঐ]) meaning *after doing*
(3) $*[prilE = pril + E]$ ([পৰিলৈ = পৰিল + ঐ]) meaning *after falling*
(4) $[prilE = pri + lE]$ ([পৰিলৈ = পৰি + লৈ]) meaning *after falling*
(5) $*[DAnGrkE = DAnGrk + E]$ ([ডাঙৰকৈ = ডাঙৰক + ঐ]) meaning *loudly*
(6) $[DAnGrkE = DAnGr + kE]$ ([ডাঙৰকৈ = ডাঙৰ + কৈ]) meaning *loudly*
(7) $[kE = k + E]$ ([কৈ = ক + ঐ]) meaning *saying* (participle)
(8) $[lE = l + E]$ ([লৈ = ল + ঐ]) meaning *taking* (participle)

Decompositions 1, 7, and 8 are valid, but the suffix E (ঐ) is not a regular suffix, that is, it is the valid suffix only in very few of the words where it occurs as the trailing part. Decompositions such as 2 and 3 involving this suffix are not valid, though the derivatives are valid words. In 2, the base too is invalid. A very tricky case in Assamese is decomposition 3, where the derivative and the base are both valid words and are closely related semantically. But the decomposition is not valid as the derivative *prilE* is not derived from the base *pril*. The correct decomposition is 4. Similarly, for the word $DAnGrkE$ the decomposition 6 is valid and 5 is not.

Due to the difficulty in dealing with such highly irregular suffixes, we attempt to merge them with the preceding letters in the decompositions. This requires a heuristic more complex than the one mentioned for suffix extension, since the letters preceding the irregular suffix in different decompositions are not identical. Hence, we use the criteria that an irregular suffix has a comparatively low occurrence count (say, less than three times the required threshold count to accept a part), and merging it with one or more preceding letters of the decompositions produces some known suffix that has a higher occurrence count.[2] In the just-cited example, wherever E is preceded by l or k, merging them produces lE and kE respectively, which have higher occurrence counts than E. We refer to this step of merging as *suffix consolidation*.

5.1.8 *Orthographic Peculiarities.* When two morphemes are concatenated, at the point of fusion, the pronunciation is sometimes represented in the written form by a changed spelling instead of the concatenation of the basic spelling of the fused morphemes. Such spelling modification affects the identification of suffixes. For example, the word *kakAye* (ককায়ে meaning *elder brother* in ergative case), should actually be decomposed as

$$[kkAye = kkAi + e],$$

but due to spelling modification, we fail to produce this decomposition.

[2] The occurrence count considered here is that before unification of decompositions.

We have not taken any step to deal with this type of difficulty. Some amount of supervision in the form of hand-crafted rules to deal with such irregularities can make the process of morphology acquisition as well as morphological analysis more effective.

6. A PROCEDURE FOR ACQUISITION OF ASSAMESE MORPHOLOGY

Based on the observations described in Section 5, in this section we lay out the sequence of steps we use in the acquisition of Assamese morphology and in building a morphological dictionary of Assamese. The purpose of acquiring the suffixes and suffix sequences is to subsequently use them for analysis of words of new texts. In such an exercise, we decompose the words by matching the trailing portions of the words against the suffixes or suffix sequences. Test inputs are most likely to be not-so-large chunks, such as paragraphs, essays, articles, etc. For some words, the base support for the decompositions may be poor though the decompositions are valid. Hence, we retain the word occurrence evidence from the training corpus in the form of a morphological lexicon. More specifically, we record the decompositions that we obtain for the words in the training corpus in the lexicon.

To acquire Assamese morphological knowledge, and build the lexicon, the following steps are followed:

Stage 1: Prepare initial set of suffixes, S_1

Obtain the list of identified suffixes by performing the initial decompositions and then using the combination of criteria as described in Section 5.1.3.

(1) Obtain the initial decompositions, D_{i_a}, for the words in the corpus article-by-article.
(2) Obtain the initial decompositions, D_{i_c}, for the words in the combined corpus.
(3) Perform suffix extension over the decompositions in D_{i_a} (see Section 5.1.6).
(4) Let S_1 be the set of morphological extensions that occur with bases with at least p $(=2)$ phonemes in D_{i_a}, and occur in at least f distinct decompositions in D_{i_c}.

Stage 2: Get comprehensive set of decompositions, D

Next, we use the set of suffixes S_1 to further decompose the input words in a bootstrapping way.

(1) Let $W_1 = W$, that is, the set of input words.
(2) Obtain the set, D_1, of all possible decompositions $[w = b + s]$, such that $w \in W_1$ and $s \in (S_1 \cup \{NULL\})$.
If $s = NULL$, then we term it a trivial decomposition.

(3) Obtain a set of decompositions, D, by selecting from D_1 those decompositions that are either trivial or where the base involved occurs in at least two decompositions, that is,

```
if $[w_i = b + s_i] \in D_1$ then
    $D := D \cup \{[w_i = b + s_i]\}$ iff
        $s_i = NULL$,  or,  $\exists[w_j = b + s_j] \in D_1$, where $w_i \neq w_j$.
```

(4) If there are bases involved in decompositions in D, which are hypothesized words, that is, they are not in W_1, include such bases in W_1 and *goto* step 2.

(5) Perform suffix consolidation over the decompositions in D (see Section 5.1.7).

Stage 3: Obtain higher degree decompositions, D_2

The initial set of suffixes S_1 contains composite suffixes as well. Hence decompositions in D may have scope for further decompositions. We process them further to obtain a set of higher-degree decompositions, D_2.

(1) Initialize set D_2 by unifying decompositions in D (see Section 5.1.5). Due to unification, some suffix parts that are not there in S_1 may be produced.

(2) Recursively reduce the bases of the decompositions in D_2 (see Section 5.1.5). That is,

```
if $\{[w = b_i + x_i], [b_i = b_j + x_j]\} \subset D_2$, and $x_j \neq NULL$, then
    $D_2 := (D_2 - \{[w = b_j + x_j]\}) \cup \{[w = b_j + x_j + x_i]\}$.
```

(3) Perform compaction of the decompositions set D_2 (see Section 5.1.5). That is,

```
if $\{[w = b + x_i + x], [w_i = b + x_i]\} \subset D_2$, and $x_i, x \neq NULL$, then
    $D_2 := (D_2 - \{[w_i = b + x_i]\})$.
```

Stage 4: Verify new suffix parts

Since S_1 is obtained during initial decomposition, each morphological extension in S_1 occurs as the final part of some input word. Suppose S_2 is the set of suffix parts occurring in D_2. S_2 may contain new suffix parts[3] that are not in S_1. Since D_2 is originally taken from D, some of the decompositions in D_2 may be of hypothesized *words*, that is, words not in *W*. A new suffix part might be the final part of the hypothesized word. For example, suppose D_2 contains the decomposition

$[sbhAkhnr = sbhA + khn + r]$ (সভা + খন + ৰ) meaning *of the meeting*

due to unification of the following decompositions in D:

$[sbhAkhnr = sbhA + khnr]$ (সভা + খনৰ) meaning *of the meeting*, and
$[sbhAkhnr = sbhAkhn + r]$ (সভাখন + ৰ) meaning *of the meeting*.

[3] A new suffix part would always occur as a nonfinal part in unified decomposition.

If the word *sbhAkhn* is a hypothesized word, the new suffix *khn* is the final part of the hypothesized word. If a new suffix part in S_2 occurs as the final part of hypothesized words only, we eliminate that new suffix part by merging it with the part following it in the decompositions where it occurs. For example, suppose D_2 contains the decompositions

(1) $[crAiTi = crAiT + i]$ (চৰাইট + ই) meaning *the bird*,
(2) $[crAiTo = crAiT + o]$ (চৰাইট + ও) meaning *the bird*, and
(3) $[crAiTo = crAi + To]$ (চৰাই + টো) meaning *the bird*.

where *crAiT* is a hypothesized word and is, in fact, invalid. Then by unification of decompositions 2 and 3, we obtain

$$[crAiTo = crAi + T + o].$$

Now, the new suffix part T is actually invalid, and would not occur as the final suffix part of the decomposition of any real word. Since we do not find any input word whose decomposition has T as the final part, we merge T with the suffix part following it in decompositions 1 and 2, and obtain

$$[crAiTo = crAi + To], \text{ and}$$
$$[crAiTi = crAi + Ti] .$$

We state this more specifically as:

(1) Suppose $s \in (S_2 - S_1)$, that is, s is a new suffix part, $\delta_j : [w_j = b_i + x_i + s + p_i + x_j]$ and $\delta_j \in D_2$, where x_i and x_j are, possibly NULL, parts-sequences, that is, δ_j is a decomposition involving s.
(2) If $b_i x_i s \notin W \; \forall \delta_j$, (i.e., all words with s as the final suffix part extracted from decompositions in D_2 are hypothesized words), then for each δ_j, do

$$D_2 := (D_2 - \{\delta_j\}) \cup \{[w_j = b_i + x_i + sp_i + x_j]\}.$$

That is, merge s with the part following it in the decompositions.

Stage 5: Generate more likely alternative decompositions, D_3

In building the lexicon, our primary objective is to obtain decompositions that are valid and have a high degree.[4] For validity of a morphological extension, we define a threshold value, q, for the minimum occurrence count[5] of valid morphological extensions. Empirically, we find that the suitable value of q is

[4] High degree implies that a greater number of morphemes present are identified.
[5] Occurrence is counted for distinct words formed.

3 for a corpus larger than 100,000 words. For each decomposition in D_2 we consider the alternative morphological extensions that contain all the partition points of the original morphological extension. For example, for the decomposition

$$[mAnuHznrprAHe = mAnuH + znr + prAHe]$$

(মানুহজনৰপৰাহে) meaning *from the person only*,

we get the additional decompositions

$$\begin{gathered}[mAnuHznrprAHe = mAnuH + zn + r + prAHe] \\ [mAnuHznrprAHe = mAnuH + znr + prA + He] \\ [mAnuHznrprAHe = mAnuH + zn + r + prA + He]\,.\end{gathered}$$

If such a morphological extension is not valid, we successively merge its initial parts with the base until the remaining morphological extension is valid or is $NULL$. For example, from the invalid decomposition

$$[bzArkhnrprAHe = bzA + r + khn + r + prA + He]$$

(বজাৰখনৰপৰাহে) meaning *only from the market*,

we get the valid decomposition

$$[bzArkhnrprAHe = bzAr + khn + r + prA + He].$$

From the alternative decompositions thus obtained, we select the one with the shortest base, and highest degree, in that order. If there is more than one such decomposition, we unify them.

We state this more specifically as follows:

(1) Suppose $C(X)$ denotes the occurrence count of the morphological extension X, in D_2.
Suppose $\delta : [\omega = \beta + x]$ is a decomposition in D_2.

(2) Find the decompositions

$$\delta_i : [\omega = \beta + x_i]$$

such that $x_i =_a x$ (i.e., x_i is an alternative suffix sequence of x).

(3) Suppose $x_j = (a_1 + a_2 + ... + a_n)$. If $C(x_j) < q$ (i.e., the occurrence count of x_j is lower than threshold q), modify δ_j as

$$\delta_j : [\omega = \beta_k + x_{j_k}]$$

where $x_{j_k} = (a_k + ... + a_n)$, $\beta_k = (\beta a_1 ... a_{k-1})$, and k is the smallest number such that

$$C(x_{j_k}) \geq q, \text{ and } C(a_{k-1} + ... + a_n) < q.$$

(4) From δ_i select the one that has the shortest base. If there is more than one such decomposition, from among them select the one that has the highest degree. If there is more than one such decomposition, unify them to get a single decomposition.

Actually, step 3 is required only if all the alternative decompositions obtained in the previous step have low occurrence counts, since in step 4 we prefer the decomposition that has a shorter base.

Stage 6: Final suffix and suffix-sequence sets

D_3 is the final decomposition set obtained from the input corpus. It is the lexicon that may be used for morphological analysis of any other text. The set of suffix sequences, Q in D_3, is the final set of suffix sequences obtained, and the set of suffix parts, S_2 in Q, is the set of suffixes obtained. We develop the outlined morphology acquisition process based on experiments over corpus A. Then we run the process over a larger corpus B, and build a morphological lexicon.

The results of the experiments using corpus A are summarized here.

Total number of words in the input corpus	:	116,096
Number of distinct words in the input corpus	:	20,140
Number of entries in the lexicon, D_3	:	15,707
Number of bases in the lexicon	:	10,203
Number of morphological extension parts in S_2	:	428
Actual number of suffixes present	:	187
Precision of suffix identification	:	65.71%
Recall of suffix identification	:	73.80%
F-measure	:	69.52%
Number of suffix sequences in Q	:	810

When the exercise is carried out over a corpus of 301,271 words (corpus B) from 525 news articles, in the initial suffix list S_1 we have:

Number of entries in the initial suffix list S_1	:	500
No. of valid suffixes, s	:	136
No. of compound parts, c	:	89
No. of composite suffixes, q	:	188
No. of invalid morphological extensions, b	:	87
Actual no. of suffixes present, n	:	190
Precision, $(s/(s+b))$	:	60.99%
Recall, (s/n)	:	71.58%
F-measure	:	65.86%

The final lexicon obtained from corpus B can be briefly summarized as:

Final Lexicon:		
Number of entries in the lexicon	:	26,509
No. of words that can be extracted from the lexicon	:	39,098
Number of bases in the lexicon	:	15,094
Number of entries in final suffix list	:	381
No. of valid suffixes, s	:	136
No. of compound parts	:	102
No. of composite suffixes	:	76
No. of invalid suffixes, b	:	67
Number of suffix sequences in Q	:	1741
Actual number of suffixes present, n	:	190
Precision of suffix identification $(s/(s+b))$	:	67.00%
Recall of suffix identification (s/n)	:	71.58%
F-measure	:	69.21%

The set of suffix sequences appearing in the lexicon is not exhaustive, and new texts may contain other suffix sequences too. Further, the lexicon may not provide the complete decompositions for some words. This is because the decomposition of each word depends on the presence of other related words. Words for which a sufficient number of related forms have not occurred may be left incompletely decomposed.

7. MORPHOLOGICAL ANALYSIS OF TEXTS

Once we have built a morphological lexicon from the training corpus, we use it for analysis of the words of new text chunks. To analyze each word, a careful consideration of the set of input words as well as the lexicon is necessary. We refer to the set of words in the input text as T, and the set of suffixes and composite suffixes (i.e., concatenated suffix sequences) as S. We assume that the input text is coherent in the sense that words with similar initial letter strings are derived from the same base if the differing trailing portions match known suffixes or suffix sequences. The steps followed to analyze a piece of text are discussed next.

Stage 1: Produce decompositions relating different input words

We identify decompositions

$$\delta : [w = b + x_1 + \ldots + x_n]$$

such that $x_{i=1\ldots n} \in (S \cup \{NULL\})$, b has a support greater than 1, $bx_1\ldots x_i \in T$, and $1 <= i <= n$. The steps to obtain such decompositions follow.

(1) Identify decompositions $[w = b + x]$, where $w \in T$, $x \in S$, and support of b is greater than 1.

(2) Recursively reduce the bases of the decompositions (see Section 5.1.5). If all the partition points of a decomposition of a word are present in some other decomposition of that word, drop the former decomposition and retain the latter.

(3) Perform compaction of the decompositions (see Section 5.1.5).

We refer to the set of decompositions so obtained as D_1.

Stage 2: Find lexicon entries for decompositions in D_1

For some of the decompositions in D_1, higher-degree decompositions can be actually possible. For example, if the input text contains the words, $rA/ST\lrcorner rIy$ (ৰাষ্ট্ৰীয় meaning *national*) and $rA/ST\lrcorner rIytAb\,AdIsklr$ (ৰাষ্ট্ৰীয়তাবাদীসকলৰ meaning *of the nationalists*), we have the following decomposition in D_1:

$$\delta : [rA/ST\lrcorner rIytAb\,AdIsklr = rA/ST\lrcorner rIy + tAb\,AdIsklr].$$

The actual decomposition should be

$$\delta : [rA/ST\lrcorner rIytAb\,AdIsklr = rA/ST\lrcorner r + Iy + tA + b\,Ad + I + skl + r].$$

We look up the lexicon to find relevant entries. Since our lexicon is built from the training corpus using an unsupervised method, we consider two possible cases of available lexicon entries:

Case 1. The lexicon contains the decomposition

$\delta_{l1} : [rA/ST\lrcorner rIytAb\,AdIsklrHe = rA/ST\lrcorner r + Iy + tA + b\,AdI + skl + r + He]$
(ৰাষ্ট্ৰীয়তাবাদীসকলৰহে meaning *of the nationalists rather*)

which contains all the partition points present in the decomposition δ. We take the relevant portion of the decomposition in the lexicon, that is,

$$[rA/ST\lrcorner rIytAb\,AdIsklr = rA/ST\lrcorner r + Iy + tA + b\,AdI + skl + r].$$

Case 2. The lexicon contains the decomposition

$$\delta_{l2} : [rA/ST\lrcorner rIytAb\,AdIsklr = rA/ST\lrcorner rIytA + b\,AdI + skl + r].$$

Here, δ_{l2} is shallower (i.e., it has a longer base) than δ, but its morphological extension portion contains all the partition points present in the corresponding portion of δ. Hence, we take the relevant portion of the decomposition δ_{l2} and unify it with the decomposition δ (see Section 5.1.5) to obtain the decomposition

$$[rA/ST\lrcorner rIytAb\,AdIsklr = rA/ST\lrcorner rIy + tA + b\,AdI + skl + r].$$

Let us refer to the set of decompositions we obtain by the preceding steps as D_2. Both D_1 and D_2 may contain more than one distinct decomposition for some words.

Stage 3: Words not decomposed in D_2

For input words for which no nontrivial decomposition (i.e., decomposition with non-NULL suffix) is present in D_2, we consider the decompositions in D_1. Recall that for some decompositions in D_1, higher-degree decompositions may be possible. Suppose such a decomposition in D_1 is

$$\delta : [w = b + x_1 + ... + x_n]$$

where some of $(x_{i=1...n})$ are composite suffixes. If there exists any alternative suffix sequence for $(x_1 + ... + x_n)$ that contains all its partition points, using this sequence we obtain all the alternative decompositions for w (see Section 5.1.5). Otherwise, for these decompositions we generate new alternative suffix sequences such that each two-part subsequence (such as $x_1 + x_2$, $x_2 + x_3$, etc.) in them already exist as suffix sequences. Some decompositions in D_2 may match the leading portions of these words. We refer to a pair of distinct words as *siblings* if one can be extracted from the decomposition of the other or they have identical leading portions and their decompositions have one or more common partition points. From the alternative decompositions, we select the ones with longest sibling match with some decomposition in D_2. If there is more than one such decomposition, we select the one that has a degree not higher than the others. We add the selected decompositions in the set D_2.

Stage 4: Root words and compound decompositions

For those words for which no nontrivial decomposition is found, we try to identify compound decompositions, that is, decompose into two parts both of which can be extracted from the lexicon or D_1.

7.1 Summary of the Steps

In Stage 1, we put together related words to form the longest decompositions possible. These long decompositions provide the contextual evidence that can help the program avoid invalid decompositions. The decompositions that we obtain may have scope for subsequent decompositions. In Stage 2, we seek relevant evidence from the lexicon to further analyze the input words represented in the decompositions. The reason why we seek lexicon entries only after forming the decompositions in D_1 is that for each word we want to take into account the longest context available in the input. In Stage 3, we deal with the words for which suitable decomposition evidence is not found in the lexicon. Some of the words that are left undecomposed after this may be compounds. In Stage 4, we attempt to recognize compounds that are formed from other known words.

From among the undecomposed words left, the ones for which no decomposition using the given set of suffixes is possible are actually root words. For those words for which decompositions are possible but the bases involved have very poor support (i.e., the base does not occur in any other decomposition), we consider the number of occurrences of the word. If the word has occurred several times, say more than ten times, the word is likely to be an actual root word.

7.2 Measuring Quality of Morphological Analysis

The quality of our morphological analysis is measured by the extent to which the parts identified are valid morphemes. However, it is important to clarify certain notions before the performance can be quantified. For example, the ideal analysis of the word $l'rAborklE$ (ল’ৰাবোৰকলৈ meaning *with the boys*) is

$$[l'rA + bor + k + lE].$$

If the computational method produces the analysis

$$[l'rAbor + klE],$$

neither of the two parts is actually ideal. So both precision and recall would be 0%, though it is clear that the partition point identified is valid. Hence, to quantify the performance in terms of precision and recall, we count the partition points identified in the words, and compare them with the number of partition points that should ideally be identified in the words. To account for the undecomposed words we consider the ends as partition points, and refer to them as trivial partition points. Each trivial partition point is a valid partition point. Thus, recall, which denotes the ratio of the number of valid partition points identified to the number of partition points to be ideally identified, can be computed as

$$\text{recall} = \frac{V + C}{A + R} \qquad (1)$$

where V is the number of valid nontrivial partition points identified, C is the number of undecomposed words that are actually root words (each presents a trivial partition point), A is the total number of nontrivial partition points to be ideally identified, and R is the number of root words present (each presents a trivial partition point).

Precision denotes the ratio of the number of valid cases identified to the total number of cases identified. In our exercise, we can compute this as

$$\text{precision} = \frac{V + U}{I + U} \qquad (2)$$

where U is the number of undecomposed words (each presents a valid trivial partition point), and I is the total number of nontrivial partition points identified. Alternatively, if the trivial partition points in words which should ideally have been decomposed are treated as invalid, we can compute precision as

$$\text{precision} = \frac{V + C}{I + U}. \qquad (3)$$

The numerator in Eq. 2 is greater than (or equal to) that of Eq. 3 since in the former if a word that should have been decomposed is left undecomposed, it is treated as a “missed” partition point, and not an invalid partition point. As a case of missed partition point, it is accounted for in the recall value.

Thus, for the analysis [$l'rAborklE = l'rAbor + klE$], recall is 33% and precision is 100%. For the analysis [$l'rAbor = l'rAbor$], the recall is 0%, the precision according to Eq. 2 is 100%, and according to Eq. 3 is 0%.

7.3 Results of Morphological Analysis Experiment

We tested the morphological analysis approach just outlined over text chunks from different sources. We used the lexicon and the set of suffixes and suffix sequences obtained from corpus B (Section 6) using our unsupervised morphology acquisition process. Corpus B is a collection of 525 newspaper articles that include general news, sports news, and editorial articles. For testing, we ran our process over 84 other newspaper articles totalling 32,271 words from the same newspaper source, and 66 articles from the Emille corpus for Assamese (http://www.ling.lancs.ac.uk/fass/projects/corpus/emille/) totalling 138,131 words. The Emille corpus articles used for testing are from various domains, namely, agriculture, anthropology, astrology, astronomy, biographies, business, industry, media, music, novels, stories, translated literature, and travel.

We observed that for almost all newspaper articles our method provides an analysis for over 95% of the words, that is, either the words are decomposed or are conclusively declared as root words. For only the remaining small fraction of words the method does not provide any analysis. In case of the Emille corpus, for most articles this ratio is over 92%. To evaluate the results of morphological analysis, we verified the analysis produced for each word in the input as described in Section 7.2. This required intensive manual effort. We obtained test samples from different types of newspaper articles and from the Emille corpus. We manually prepared the correct analyses of all the words in a test article, and compared the counts and appropriateness of the partition points with those produced by our morphological analysis method (see Section 7.2). More specifically, we computed the following:

—Total number of words in the input file (Col. T in Table VI)

—Total number of partition points generated (Col. A in Table VI)

—Number of spurious partition points generated (Col. B in Table VI)

—Total number of undecomposed words (Col. C in Table VI)

—Actual number of partition points required (Col. D in Table VI)

—Actual number of roots (Col. E in Table VI)

—Number of valid partition points missed (Col. F in Table VI)

—Number of undecomposed words that are roots (Col. G in Table VI)

—Precision, P_1 (Eq. 2) ((A+C-B)/(A+C) in Table VI)

—Precision, P_2 (Eq. 3) ((A+G-B)/(A+C) in Table VI)

—Recall (Eq. 1) ((A+G-B)/(D+E) in Table VI).

Note that methods such as the ones by Gaussier [1999] and Goldsmith [2001] work with large input corpora. On the other hand, methods such as Porter's [1980] use hand-coded rules. The nature of the problem we tackle is thus distinct, so we do not compare the result of those methods with ours.

Table VI. Evaluation of Morphological Analysis

Text id	T	A	B	C	D	E	F	G	Precision P_1	Precision P_2	Recall
1.	350	197	36	198	181	196	20	181	90.89	86.58	90.72
2.	615	358	38	331	366	317	46	301	94.48	90.13	90.92
3.	468	246	30	263	262	248	46	232	94.11	88.02	87.84
4.	213	109	17	122	101	123	9	114	92.64	89.18	91.96
5.	351	228	20	173	242	165	34	157	95.01	91.02	89.68
6.	419	296	47	207	296	203	47	182	90.66	85.69	86.37
7.	292	190	34	153	180	147	24	135	90.09	84.84	88.99
8.	770	438	63	395	426	403	51	360	92.44	88.24	88.66
9.	792	437	65	441	416	437	44	412	92.60	89.29	91.91
10.	514	322	36	271	356	245	70	234	93.93	87.69	86.52

(a) Evaluation for newspaper articles

Text										Precision		
Id	Type	T	A	B	C	D	E	F	G	P_1	P_2	Recall
1.	Mu	2193	1177	251	1273	1248	1187	322	1056	89.76	80.90	81.40
2.	Mu	2079	1072	198	1175	1154	1063	280	946	91.19	81.00	82.09
3.	Cm	1644	871	141	971	887	924	157	860	92.35	86.32	87.80
4.	Cm	1734	902	189	1004	907	1000	194	889	90.08	84.05	84.01
5.	As	2037	1036	205	1190	1080	1124	249	1019	90.79	83.11	83.94
6.	As	1893	948	204	1092	985	1044	241	936	90.00	82.35	82.80
7.	Bg	1961	1095	244	1119	1121	1052	270	940	88.98	80.89	82.42
8.	Bg	2049	1150	243	1182	1277	1036	370	944	89.58	79.37	80.03
9.	Nv	2395	1418	349	1290	1610	1136	541	974	87.11	75.44	74.40
10.	Nv	2423	1695	394	1184	1743	1079	442	922	86.31	77.21	78.77
11.	St	2334	1311	263	1313	1528	1096	480	999	89.98	78.01	78.01
12.	St	2391	1511	290	1229	1737	1055	516	945	89.42	79.05	77.58
13.	Tr	3041	1657	298	1660	1758	1555	399	1386	91.02	82.76	82.86
14.	Tr	2889	1550	242	1652	1711	1526	403	1409	92.44	84.85	83.94
15.	Md	1901	1122	199	1033	1204	974	281	880	90.77	83.67	82.78
16.	Md	1516	845	218	822	783	850	156	713	86.92	80.38	82.06
17.	Ot	3133	1719	314	1731	1873	1567	468	1410	90.90	81.59	81.83
18.	Ot	1320	680	115	733	730	678	165	619	91.86	83.79	84.09
19.	Ag	1271	700	151	724	729	669	180	603	89.40	80.90	82.40
20.	An	2427	1244	239	1439	1394	1280	389	1174	91.09	81.22	81.49
21.	An	2493	1261	201	1457	1444	1321	384	1231	92.60	84.29	82.86
22.	An	2273	1129	178	1272	1190	1200	239	1095	92.59	85.21	85.61
23.	An	2328	1101	234	1374	1139	1326	272	1188	90.55	83.03	83.37
24.	Bs	2369	1482	172	1263	1632	1177	322	1081	93.73	87.10	85.12
25.	TL	2086	1121	211	1199	1233	1085	323	987	90.91	81.77	81.84
26.	TL	1400	826	198	776	853	712	225	656	87.64	80.15	82.04
27.	TL	1650	954	192	906	1040	809	278	743	89.68	80.91	81.40

(b) Evaluation for Emille corpus articles

Columns:

T:	Total input words	A:	Total nontrivial partition points identified
B:	Spurious partition points identified	C:	Total undecomposed words
D:	Actual nontrivial partition points required	E:	Actual roots
F:	Valid nontrivial partition points missed	G:	Correct root recognitions
P_1:	Precision using Eq. 2	P_2:	Precision using Eq. 3
Recall:	using Eq. 1		

Text types:

Mu:	Music	Cm:	Commerce	As:	Astrology	Bg:	Biography
Nv:	Novel	St:	Story	Tr:	Travel	Md:	Media
Ot:	Other	Ag:	Agriculture	An:	Astronomy	An:	Anthropology
Bs:	Business	TL:	Translation literature				

8. CONCLUSIONS AND FUTURE WORK

In this article, we discuss the nature of morphology of a natural language in general and an Indic language, Assamese, in particular. We discuss characteristics of the Assamese language and computer representation of Assamese texts. We consider two existing approaches for unsupervised acquisition of morphology from text corpora, and present the results obtained using them for Assamese. Because of the structural nature of morphology, simple computational methods can serve as the initial steps for acquisition of morphology of a language and morphological analysis. Additional efforts are required to tackle different language-specific and script-specific issues. We propose an unsupervised approach suitable for Assamese. Using this approach, we acquire the suffixation morphology of the language from a text corpus of about 300,000 words and build a morphological lexicon. The F-measure of the suffix acquisition is about 69%. The suffixes and the lexicon can facilitate subsequent morphological analysis of small text chunks too.

The morphological knowledge acquired using our unsupervised approach is less than perfect, but using this knowledge we obtain fairly good results in morphological analysis of input texts. A direct application of this competence can be in building a spelling checker. For a highly inflectional language a spelling checker can be effective only if it has substantial morphological analysis capabilities.

Since morphology evolves according to the spoken form of a language, for its unsupervised acquisition from a written corpus it will be helpful if the script clearly and unambiguously reflects the phonological structure of the expressions. This depends on the orthography of the languages. In English, the pronunciations of words often cannot be accurately surmised from their spellings alone. In Indic language scripts it is not so. Again, with respect to the phones used in languages, the scripts have redundancy. Among Indic scripts, Hindi, which uses the Devnagri script, has less redundancy compared to Assamese, but Hindi is not as morphologically rich as is Assamese. Further, since different nonstandard encoding schemes are in use for Assamese texts in computers, suitable transliteration software needs be developed to interoperate between these schemes. It will enhance the benefits obtained from work such as ours, making them more effective.

We believe ours is the first serious effort in computational acquisition of the morphology of Assamese, and hence is pioneering. Our work is particularly significant because morphology is the dominant structural phenomenon in Assamese, and the overall structural analysis of Assamese texts can greatly benefit from the morphological analysis.

It will be relevant to see how effective our morphology acquisition approach is for other languages. Though an unsupervised approach, the heuristics incorporated in it are influenced by the issues in a particular language. Many of these issues are present in other languages too, particularly the Indic languages. Bengali, for example, uses the same script (except for a couple of characters) and has very similar morphological properties as Assamese, although we believe that the number of possible suffixes is sufficiently lower. We believe that for such languages, our approach will produce interesting results, possibly better than for Assamese.

The morphological structure of a word provides clues regarding the category of the word. The category attribute of words hypothesized from the morphological analysis can again be used as feedback to improve the quality

Table VII. The Assamese Script Transliteration Scheme

Sl.No.	Assamese letter	Roman transcription used in this document	read-as (approx.)	example
Vowels:				
1.	অ	a	*o*	the *a* in *tall*
2.	আ	A	*aa*	the *a* in *part*
3.	ই	i	*hraswa-e*	the *i* in *bit*
4.	ঈ	I	*dirgha-e*	the *ee* in *feet*
5.	উ	u	*hraswa-oo*	the *u* in *pull*
6.	ঊ	U	*dirgha-oo*	the *oo* in *school*
7.	ঋ	Rh	*ri*	the *ri* in *Krishna*
8.	এ	e	*a*	the *a* in *pack*
9.	ঐ	E	*oi*	the *ai* in *Jain*
10.	ও	o	*o*	the *oa* in *coat*
11.	ঔ	O	*ou*	the *ow* in *rowed*
Consonants:				
12.	ক	k	*ka*	the *ca* in *call*
13.	খ	kh	*kha*	the *kha* in *Jharkhand*
14.	গ	g	*ga*	the *ga* in *gall*
15.	ঘ	gh	*gha*	the *gh* in *ghost*
16.	ঙ	nG	*unga*	the *ng*^{1} in *hanger*
17.	চ	c	*pratham-sa*	the *s*^{1} in *gas*
18.	ছ	C	*dwitiya-sa*	(similar to *pratham − sa*)
19.	জ	z	*bargiya-za*	the *z*^{1} in *Amazon*
20.	ঝ	jh	*jha*	the *Jh*^{1} in *Jharkhand*
21.	ঞ	nY	*nya*	the *ian* in *fiance*
22.	ট	T	*murdhanya-ta*	the *to* in *top*
23.	ঠ	Th	*murdhanya-tha*	the *th*^{1} in *thousand*
24.	ড	D	*murdhanya-da*	the *do* in *doctor*
25.	ঢ	Dh	*murdhanya-dha*	the *Dh*^{1} in *Dhaka*
26.	ণ	N	*murdhanya-na*	the *n*^{1} in *Ganesh*
27.	ত	t	*dantya-ta*	(similar to murdhanya-ta)
28.	থ	th	*dantya-tha*	(similar to murdhanya-tha)
29.	দ	d	*dantya-da*	(similar to murdhanya-da)
30.	ধ	dh	*dantya-dha*	(similar to murdhanya-dha)
31.	ন	n	*dantya-na*	(similar to murdhanya-na)
32.	প	p	*pa*	the *po* in *point*
33.	ফ	ph	*pha*	the *ph*^{1} in *phone*
34.	ব	b	*ba*	the *ba* in *ball*
35.	ভ	bh	*bha*	the *Bh*^{1} in *Bharat*
36.	ম	m	*ma*	the *ma* in *mall*
37.	য	j	*ja*	the *jo* in *jog*
38.	ৰ	r	*ra*	the *ro* in *rock*
39.	ল	l	*la*	the *lo* in *lost*
40.	ৱ	w	*wabba*	the *wo* in *world*
41.	শ	sh	*talabya-sa*	(roughly) the *sh*^{1} in *posh*
42.	ষ	S	*murdhanya-sa*	(roughly) the *sh*^{1} in *posh*
43.	স	s	*dantya-sa*	(roughly) the *sh*^{1} in *posh*

Table VII. (Continued)

Sl.No.	Assamese letter	Roman transcription used in this document	read-as (approx.)	example
44.	হ	H	*ha*	the *ha* in *hall*
45.	ক্ষ	X	*khya*	(absent in English)
46.	ড়	R	*dare-ra*	the *r*[1] in *Orissa*
47.	ঢ়	rh	*dhare-ra*	the *rh*[1] in *Chandigarh*
48.	য়	y	*ya*	the *you* in *young*
Partial consonants:				
49.	ৎ	_t	*byanjan-ta*	the *t* in *Utpal*
50.	ং	#	*anuswar*	the *ng* in *king*
51.	ঃ	:	*bisarga*	the *h* in *eh*
52.	ঁ	*	*sandra-bindu*	the *n* in *Ranchi*
53.	্র	_r	*ra-kAr*	the *r* in *product*
54.	র্	r̂	*ref*	the *r* in *form*
55.	্য	J	*ja-kAr*	the *y* in *Myanmaar*

[1] the inherent vowel *a* is to be added.
To denote a *juktakshar* (ligature), a / (slash) is placed before the consonant sequence forming the *juktakshar*.

of analysis of the words, by ruling out candidate decompositions that are not valid for that category. Similarly, feedback from the syntax analysis stage can provide hints for word decomposition as well as word classification. Also, to make morphological analysis more effective, varying degrees of supervision can be introduced. Research on these and other issues is in progress.

A. THE TRANSLITERATION SCHEME USED

For the experiments described in this article the Assamese texts are encoded using Roman letters. The transliteration scheme is shown in Table VII.

REFERENCE

Bora, S. 1968. *bahal byaakaran*. Jnananath Bora, Guwahati.

Borer, H. 1998. Morphology and syntax. In *The Handbook of Morphology*, Spencer, A. and Zwicky, A. M. eds., 151–190, Blackwell Publishers Ltd.

Chen, T. Y., Kuo, F.-C., and Merkel, R. 2004. On the statistical properties of the F-measure. In *Proceedings of the 4th International Conference on Quality Software (QSIC'04)*. IEEE Press, Los Alamitos, CA, 146–153.

Creutz, M. 2003. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL'03)*, 280–287.

Creutz, M. and Lagus, K. 2004. Induction of a simple morphology for highly-inflecting languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON'04)*, 43–51.

Creutz, M. and Lagus, K. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, 106–113.

DAELEMANS, W. 1993. Memory-based lexical acquisition and processing. In *Proceedings of the Third International EAMT Workshop on Machine Translation and the Lexicon (EAMT'93)*, 85–98.

GAUSSIER, E. 1999. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the Unsupervised Learning in Natural Language Processing Workshop (ACL'99)*. ACL, 24–30.

GASSER, M. 1994. Acquiring receptive morphology: A connectionist approach. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL'94)*, 279–286.

GOLDSMITH, J. 2001. Unsupervised learning of the morphology of a natural language. *Comput. Linguist. 27*, 2, 153–193.

GOSWAMI, G. 1990. *asamiyaa byaakaranar moulik bisaar*. Bina Library, Guwahati, India.

LEIBER, R. 1992. *Deconstructing morphology: Word formation in syntactic theory*. University of Chicago Press, Chicago, IL.

MEDHI, K. 1999. *অসমীয়া ব্যাকৰণ আৰু ভাষাতত্ত্বAssamese Grammar and Origin of the Assamese Language*. 3rd Ed. Lawyer's Book Stall, Guwahati, India.

PORTER, M. 1980. An algorithm for suffix stripping. *Autom. Library Inform. Syst., 14*, 3, 130–137.

SARAVANAN, M., REGHV RAJ, P. C., MURTY, V. S., AND RAMAN, S. 2002. Improved porter's algorithm for root word stemming. In *Proceedings of the International Conference on Natural Language Processing (ICON'02)*, 21–30.

SARMA, D. D. 1977. *sahaj byaakaran*. Assam State Textbook Production and Publication Corporation Ltd., Guwahati-1, India.

SCHNEIDER, G. 1998. An introduction to government and binding. University of Zurich. http://www.ifi.unizh.ch/CL/gschneid/dreitaegig.ps.gz.

SHARMA, U. 2006. Unsupervised acquisition of morphology of a highly inflectional language. PhD dissertation, Department of Computer Science and Engineering, Tezpur University, Assam, India.

SHARMA, U., DAS, R., AND KALITA, J. 2006. Unsupervised acquisition of morphological features of Assamese from a text corpus. In *Proceedings of the National Workshop on Trends in Advanced Computing (NWTAC'06)*, 178–184.

SHARMA, U., KALITA, J., AND DAS, R. 2002. Unsupervised learning of morphology for building a lexicon for highly inflectional language. In *Proceedings of the Workshop on Morphological and Phonological Learning (ACL'02)*, 1–10.

SHARMA, U., KALITA, J., AND DAS, R. 2003. Root word stemming by multiple evidence from corpus. In *Proceedings of the 6th International Conference on Computational Intelligence and Natural Computing (CINC'03)*.

SNOVER, M. G., JAROSZ, G. E., AND BRENT, M. R. 2002. Unsupervised learning of morphology using a novel directed search algorithm: Taking the first step. In *Proceedings of the Workshop on Morphological and Phonological Learning (ACL'02)*, 11–20.

VASU, S. C. 1891. *The ashtadhyayi of panini* (edited and translated into English), vol. I. Motilal Banarsidass, Delhi, India.

Received March 2007; revised February 2008; accepted February 2008