8.8 Morphological Analyzer for Great Andamanese Verbs: Implementing a Concatenative Template

Narayan K. Choudhary, Anvita Abbi and Girish Nath Jha, Jawaharlal Nehru University, New Delhi

Abstract—This paper presents an account of the verb phrase morphology of Great Andamanese, an endangered language of the Andaman Islands. The paper is based on the research work done in the Andaman Islands among the people of the endangered tribe. The verb phrase in Great Andamanese takes as constituents the morphemes carrying the content of causatives, subject and object clitic, negative, prohibitive negative, class marking consonant or thematic consonant (Abbi 2003, 2006) and TAM markings. All of these features are affixed to the verb root or lexeme -the only obligatory element in the verb phrase. An illustrative schema to the Great Andamanese verb phrase can be given like the following

CAUS-SBJ.CL-OBJ.CL-REFL-CAUS/NEG→VR/VL←NEG-CLSM-TAM

The schema works on constraints at the affixal level which include order, optionality and obligatoriness. The morphophonemic rules such as epenthesis, vowel deletion, assimilation that operate in the varying forms of verb phrase are not discussed here.Using a lexicon based approach to develop a morphological analyzer for the verb phrase in Great Andamanese, the paper presents the mechanisms used in developing a program that analyzes the verb phrase given the Great Andamanese text as input.

Index Terms—Concatenative Morphological Template, Great Andamanese, Natural Language Processing, Verb Morphology.

I. INTRODUCTION

THE Andaman Islands are a group of more than 500 islands situated in the Bay of Bengal. It is inhabited by a community that has been living there for long, in complete isolation. The earliest record of these people belonging to the Negrito stock (Hagelberg et. al., 2003, etc.) is found in, among others, Ptolemy (2nd C. AD), I-Tsing (672 AD) and Marco Polo (14th C. AD).

Among the four primitive *tribes* the Great Andamanese, the Jarwas, the Onges and the Sentenelese - of the Andaman Islands, the Great Andamanese, till a hundred years back, were the most populated and influential people.

The linguistic study of the rapidly vanishing voices of the Great Andamanese can be said to start with M.V. Portman's *Manual* in 1887 followed by other major works like that of E.H. Man's *Dictionary* (1919), Manoharan (1989) and Abbi (2001, 2006).

The Great Andamanese is a cover term assigned to a conglomerate of the ten tribes most of whom succumbed to the colonial pressure that started with the British and is still continuing in its new avatar. The present population (around 50) is dominated by the Jeru tribe with a few speakers (around 7) of the language. As the new generation is reluctant to learn the language of their forefathers, the language is under an imminent threat of extinction. Great Andamanese is an unwritten tribal language. The data presented in this paper is drawn from first-hand data elicitation in the field.

II. UNRAVELING THE VERB PARADIGM SCHEMA OF GREAT ANDAMANESE

Great Andamanese is an agglutinating language and is of the SOV type, meaning thereby it is a verb final language. The verb phrase of the language is a complex entity constituted of several grammatical morphemes. A verb root in a verb phrase is preceded by several prefixes as well as followed by two or more suffixes. These prefixes and suffixes encode several grammatical functions such as subject and object information, various modalities such as negation and mood. In addition, tense marking is suffixed to the verb stem. In all, the possibility of various types of affixation to the verb root or lexeme can be illustrated using the following schema (see Fig. 1).

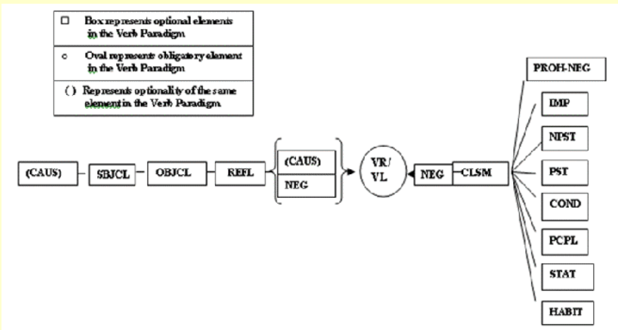

Figure1. Verb Schema of Great Andamanese

Narayan Kumar Choudhary and Anvita Abbi are with the Centre for Linguistics, Jawaharlal Nehru University, New Delhi (e-mail: choudhary_narayan@rediffmail.com, anvitaabbi@gmail.com)

Girish Nath Jha is with the Special Centre for Sanskrit Studies, Jawaharlal Nehru University, New Delhi (e-mail: girishj@gmail.com).

For example :-

i	ʈʰutconnepʰobe
	ʈʰut-connc-pʰo-b-c
	1SG.CL-go-NEG-CLS-IND
	I do not go.

ii	uʈʰuncikamo
	u-ʈʰu-n-ci-k-amo
	3SG.CL-1SG.CL-REFL-comes-CLS-COND
	If he comes to me

There are at most five morphemes that can possibly be prefixed to the Verb Root (VR) or verb lexeme (VR) while at most three morphemes that can be suffixed to it. The only obligatory element in the verb phrase (VP) is the VR or VL. Thus a verb phrase with maximum number of affixes will have the structure as the following-

CAUS-SBJ.CL-OBJ.CL-REFL-NEG→VR←CLSM-TAM
Or,
CAUS-SBJ.CL-OBJ.CL-REFL→VR←NEG-CLSM-TAM
Or,
SBJ.CL-OBJ.CL-REFL-CAU→SVR←NEG- CLSM-TAM

For example we have verb phrases like /pʰutɛʃamo/ and /ʈʰuŋolobom/ as in example sentence number iii below, /uʈʰuncikom/ as in iv and /ŋutuncɛkʰo/ as in v.

iii ŋuʈʰi miʈʰaibi tɛʃe pʰutɛʃamo ʈʰoŋolobom
ŋu-ʈʰi miʈʰai-bi tɛʃ-e pʰu-tɛʃ-amo ʈʰo-ŋol-o-b-om
2SG-1SG.OBJ sweet-ACC give-IMP NEG-give-COND
1SG.CL-cry-EPV-CLS-NPST
If you do not give me the sweets I will cry.

iv cya:k ocikom kɔil tɔ uʈʰuncikom
cya-k o-ci-k-om kɔil tɔ u-ʈʰu-inci-k-om
what-DIREC 3SG.SBJ.CL-come-CLS-NPST later
EMPH 3SG.SBJ.CL-1SG.OBJ.CL-come-CLS-NPST
Where will he go, later he will come only to me.

v ca:y kʰudi ŋutuncɛkʰo
ca:y kʰudi ŋu-tun-cɛkʰ-o
what for 2SG.SBJ.CL-REFL-angry-PST
Why did you get angry?

III. A FRAMEWORK FOR THE ANALYZER

The Great Andamanese Verb Analyzer (GAVA) is a five module program that takes Great Andamanese text as input, in IPA (*using Lucida Sans Unicode or Arial Unicode MS fonts*) and analyzes the verb phrases in it. The five modules are in fact the five processes that the input text undergoes. This has been illustrated in the following diagram (see Fig. 2).

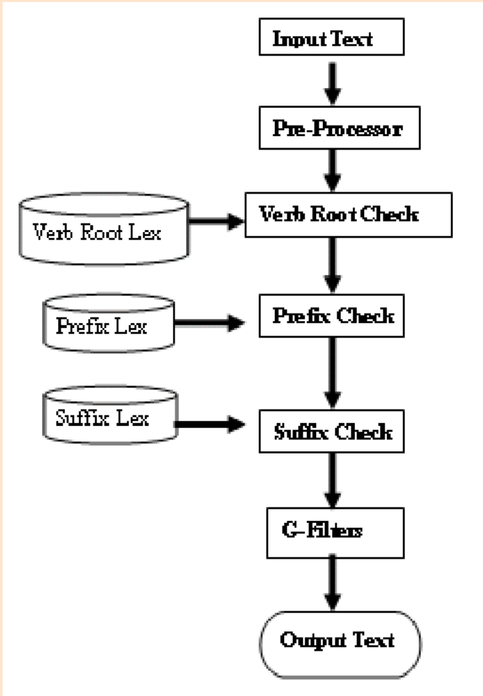

Figure2. A Model Diagram of GAVA

The pre-processor module first filters the input and checks whether any unwanted elements are there in the input text or not. If this is the case, it either corrects the input or leaves it as it is for the consideration of user.

The verb root module searches for the verb roots or lexeme in the input text and segments them from the string. The remaining part of the input string is sent for further analysis in the next modules.

The prefix module takes the elements that are to the left of the verb lexeme and analyzes them by matching each of the possible strings with the prefixes in the prefix lexicon and stores the results for display.

The suffix module takes the elements that are to the right of the verb root or lexeme and analyzes them matching each of the possible strings with the suffixes in the suffix lexicon and stores the results for display.

The G-Filters module is the last module that implements the grammatical rules. If the system has not found the right analysis of the input text or if there is some ambiguity or violation of some rules, these are checked through rules here.

The final result is displayed as Unicode HTML on a JSP web front.

A. *POS tags for Great Andamanese verbs*

Each of the verbs have been tagged with their meaning and an additional identifier of VR in the lexicon. No classification of the verbs as per transitive/intransitive or on any other criteria has been made. The linguistic resources used have all been prepared on the basis of first- hand data collected by Abbi (2001, 2005) and Choudhary (2005-06) and compared with available other printed forms. The program uses lexicon that is basically text files of small sizes.

B. *Tagged Lexicon*

There are three types of grammatical categories that are used for the Great Andamanese Verb Analyzer.

Verb Roots

Prefixes

Suffixes

All these categories are tagged properly. The prefixes and suffixes have an additional tag of PREF and SUFF respectively. This is for specific use of the program and is also displayed in the output text.

The lexicon of verb lexemes[1] contains about 120 verb roots and non-verb roots. All are verbal lexemes. The longest verb lexeme in Great Andamanese is of three syllables containing eight characters. The frequency of monosyllabic roots is higher than disyllabic roots and that of latter is much higher than tri-syllabic roots. All these roots have been arranged in the lexicon in an ascending order of the number of characters present in the lexeme to facilitate better search by the program.

There are a total of 52 prefixes and 20 suffixes at present. The number of affixes has grown up because there are allomorphic variations. Thus a morpheme with a gloss of 1SG.SBJ.CL has 6 variations, 1SG.OBJ.CL has 4 variations etc. The following table gives a list of variations in the clitics attached to verbs. A clitic is a morpheme that has the syntactic characteristics of a word but shows evidence of being phonologically bound to another word. The clitics given in table-1[2]

TABLE I
A list of pronominal verbal clitics in Great Andamanese

Name of the Prefix	No. of Variants	Variant Forms
1SG.SBJ.CL	6	ʈʰu, ʈʰa, ʈʰo, ʈʰe, ʈʰ, ʈʰut
1PL.SBJ.CL	2	me, mut
1SG.OBJ.CL	4	ʈʰu, ʈʰa, ʈʰi, ʈʰɛ,
2SG.SBJ.CL	7	ŋu, ŋa, ŋe, ŋɛ, ŋi, ŋ, ŋut
2SG.OBJ.CL	4	ɲu, ɲa, ɲe, ɲɛ, ɲi, ɲut
3SG.SBJ.CL	7	u, o, a, e, aka, uku, dut
3SG.OBJ.CL	8	aka, ek, ɛk, ek, ik, it, ut, et, i
3PL.SBJ.CL	2	nu, n
3PL.OBJ.CL	NA	Not Available

(The first person and second person clitics are homophonous in subject and object position while third person clitics are not)

C. *Rules*

There is only one rule implemented. This rule takes care of the ordering problem of the prefixes that emerges due to identical forms of the clitics. For example, /ʈʰu/ can be used as both subject clitic and as object clitic. Similarly there are other clitics that have homophonous entries. This problem can be solved by the constraint of ordering. In a verb phrase there can be no more than two clitics. As the language is of SOV nature, the subject clitic precedes the object clitic no matter what the phonetic shape is.

If a single clitic in a verb phrase is found, it is assigned the tag of subject clitic. However, it is not always necessary that the single clitic in the VP is subject clitic. If the subject is omitted or is not a pronominal category in the verb phrase, it can be an object clitic in case of transitive verb root. In this case, the solution lies only in the context of the whole clause.

There are also morphophonemic changes involved in the verb morphology of Great Andamanese, which will constitute the subject matter of the next paper.

D. *Implementation Strategies*

The program has been prepared on a Windows platform with tools and techniques as described below. This program however is platform independent and can run on any platform.

1 Verbal lexeme in Great Andamanese consists of minimum of a verb root in case of verb intransitive and maximum of two argument markers prefixed to the verb root in case of Verb transitive. This implies that transitive verb root is always prefixed by an optional subject and obligatory object clitic to gain a lexemic status to take part in the verb analyzer.

2 The list is not final as it is based on a limited source of data. There may be more or less variants, their names and forms. More specific study on this topic is warranted.

IV. AN OVERVIEW OF THE TOOLS AND TECHNIQUES USED

Table II gives an overview of the tools and techniques used in developing the program.

TABLE II
An Overview of the Tools and Techniques Used

- I. Front end
  - A. JSP, HTML, CSS, Java Script
- II. Java Objects
  - A. Pre-processor
  - B. Analyzer
    - 1) Search Parts()
      - a) Verb root
      - b) Prefixes
      - c) Suffixes
    - 2) gFilter()
    - 3) reorder()
- III. Back-end
  - A. Data files stored in UTF-8
- IV. Webserver
  - A. Apache-Tomcat

A. *Front End*

At the front end of the program the technologies like the JSP, HTML, CSS, Java Script have been used. The following is a brief introduction to these technologies and how they have been implemented in developing the program.

The front end opens in a web browser that is based locally on the user's computer.

1) Java Server Pages

The java server page used here utilizes all of the four items discussed above. It uses first, the html coding convention and initializes the style sheet, the java objects from AVT*agger.class* as servlets.

Using small Java programs (called "Applets"), web pages can include functions such as animation's, calculators, and other fancy tricks. Java programs are of three kinds

Stand-alone executable programs
Applets
Servlets

a) Cascading Style Sheets

Here the CSS has been used to bring text in a particular font namely the Lucida Sans Unicode. Another font named Arial Unicode MS can also be used for the purpose of entering the input text in Great Andamanese.

b) HTML

HTML or Hyper Text Mark-up Language is the base of the front end of the interface on which other objects namely that of Java Objects and CSS has been embedded.

B. *Java Objects*

The JSP file called the *andverbs.jsp* uses a java object called AVTagger which uses the services of Pre-processor. The *Pre-processor* object filters the input text and checks whether the input text is a potential Great Andamanese text or not. The *AVTagger* object is in fact the analyzer program that processes the input text as rendered by the pre-processor.

As described briefly above, there are five modules of the GAVA program. Below given is the description of each of the modules.

1) Pre-processor

The input first goes to the pre-processor module and checks whether any undesired elements such as punctuation marks or other control characters, numbers etc. are not given in input. If this is the case, either it corrects the input text itself or removes them from going into further analysis.

2) Analyzer

AVTagger is the file that is the most important to the program. Two Java APIs from the Java library have been imported to be used in this object.

The analyzer uses several functions and methods to analyze the GA verb.

a) parseVerbs

This is the main calling function which gets all the work done by using services of other functions/methods. This function first gets the pre-processing done on the entire text. Then it tokenizes the output of the pre-processor based on space character. Then by calling the search_Parts() function, it processes each word for verb, and affixes (to a maximum of 5 prefixes and 3 suffixes).

b) Search Parts

The search is then for the parts starting from the whole of the input to the last available string in the input text until the search is complete or there are no characters left to be searched and matched with the lexicon (or first five prefixes and first three suffixes have been searched). The search is processed in three modules. The search_Parts module assumes the role of searching verbs, prefixes or suffixes when an appropriate call is made for each kind of search.

c) G-Filter

It is here that the grammatical rules not covered in the previous modules are taken care of. The rules that are applied can be classified broadly into three categories, namely, reordering, constraints and recursivities.

(1) Reordering

The same key may have more than one value. There are pronominal clitics that have identical shapes as subject and object clitics. In this case, simple search results in a random choice may be wrong. To bring surety of the results, some rules have been drawn.

(a) Ordering of the Segmented Items

Meta Rule: Follow the ordering rule as prescribed in the verb paradigm. Take the order as given in the input string.

(b) Clitics Reordering

For the clitics having homophonous forms (e.g. 1^{st} and 2^{nd} person clitics), the following rules apply:

Rule A. If there is only one clitic preceding the verb root, take it to be Sbj.cl by default

Rule B. If there are two clitics preceding the verb root, take the first one as Sbj.cl and the second one as Obj.cl

(2) Constraints

The input verb phrase in Great Andamanese has a limited number of prefixes and suffixes. These numbers work as constraints and the system would not recognize the input if it has more than the required number of affixes.

(3) Recursivity

As there may be systemic ambiguities regarding the verb roots or the prefixes after the first round of processing of the input text, to handle this, the options/multiple values are again sent back for better results.

As there may be more than one affix in the input word, the system must analyze all of this, one by one. For this, the system must be recursive to search for different affixes in the same lexicon.

C. Back-End: Data files stored in UTF-8

The GAVA uses data files of three types of lexicon as described above. These are annotated lexicon of verb roots, prefixes and suffixes.

D. Webserver: Apache Tomcat 4.0

We have used Apache Tomcat technology for the web server.

V. EVALUATING THE PROGRAM

After successful testing of the verb phrases of Great Andamanese, more than 90% results were found correct. The verb types may be divided on the following basis: 1. Number of prefixes and suffixes 2. Types of Verb Roots based on the number of characters or syllables.

So far, I have tested a list of verb phrases extracted from a set of model sentences containing a total of 129 verb phrases (Choudhary, 2006), with a satisfying correct result of 94%.

VI. CONCLUSION

As the ambition of this project is to develop a computational framework for the verb morphology of the language, the GAVA program does not aspire to account for an exhaustive list of the verb roots and lexemes in the language under discussion. It uses a list of about 130 verb lexemes. It is basically a morphological analyzer. It is highly scalable and portable system. As an NLP program, it can be used in several ways. It can serve as a template for further work on computing of this language or other languages having morphological systems. As the system developed is highly scalable, it can be easily adapted and extended to suit the needs of other languages as well.

GAVA can also serve as a subsystem for major NLP systems on this language or other languages with like structures. The major programs may be a general purpose parser, machine translation systems, speech recognition systems, corpus anralyzers etc.

ABBREVIATIONS USED

1=First Person 2=Second Person 3=Third Person Arg=Argument Marker AUX=Auxiliary CAUS=Causative CL=Clitic CLS=Class Marker Consonant or Thematic Consonant COND=Conditional EPV=Epenthetic Vowel EXCL=Exclusive EXIST=Existential Gen=genitive HABIT=Habitual IMP=Imperative INCL=Inclusive IND=Indicative NEG=Negative NPST=Non-Past OBJ=Object PCPL=Participle PL=Plural PREF=Prefix PST=Past PROH.NEG=Prohibitive Negative REFL=Reflexive SG=Singular STAT=Stative SBJ=Subject SUFF=Suffix VL=Verb Lexeme VR=Verb Root

APPENDIX

Lexicon A: The Verb Roots and lexemes <verbroots.txt>

emp^{h}orol = turn_VR

kaɲyɔrɔ = come_frequently_VR

kaɲɔrɔ = come_frequently_VR

ereŋk^{h}ol = play_VR

ravu fro = winnow_VR

ektɛrʈɔ = throw_VR

untɛle = call_with_happiness_VR

emp^{h}il = die_VR

bɔk^{h}um = know_(neg)_VR

tabiŋo = think_VR

aratta = convince_VR

ekak^{h}u = open_VR

embele = overflow_VR

akaile = return_VR

tert^{h}u = take_out_VR
raliʃo = finish_VR
bɔrɔt^{h} = fall_VR
ɛrence = fight_VR
conne = go_VR
cɔnne = go_VR
rɛp^{h}o = climb_tree_VR
erŋol = write_VR
itp^{h}u = cut_VR
tɛrta = tell_VR
utlub = open_VR
mek^{h}u = bloom_VR
birəŋ = redden_VR
tɛbol = run_away_VR
erteɖ = see_VR
raʃui = cook_VR
beliŋ = cut_VR
eluk^{h} = pick_(caus)_VR
ʈ^{h}ibi = live_VR
bereŋ = pour_VR
ʃerep = cut_VR
rap^{h}o = cut_VR
ʈ^{h}ulu = kick_VR
k^{h}ole = laugh_VR
ektɛr = push_VR
ip^{h}il = throw_VR
ɛʃilo = shake_VR
ka:ra = rise_VR
tɛrto = shoot_arrow_VR
bat^{h}e = slap_VR
rok^{h}o = ready,_to_get_VR
bilup = remember_VR
boʃuʈ = hit_VR
olam = tire_VR
t^{h}ud = pierce_VR
belɔ = aux-clsm-pst_VR
boʃo = beat_VR
eban = make_VR
biŋo = hear_VR
ɖuoc = hear_VR
eule = see_VR
meli = return_VR

bit^{h} = sink_VR
jiyo = stay/ebb/AUX_EXIST/VR
koin = wake_up_VR
cɛk^{h} = to_be_angry_VR
ʈɔp^{h} = bathe_VR
ʃuɲe = blow,_of_nose_VR
ʈɔl = break(intr.)_VR
unɖu = break_VR
buli = defecate_VR
juvu = fly_VR
emfe = jump_VR
inci = go_VR
tɔle = mix_VR
ralɛ = moonset_VR
bele = overflow_VR
ʈɛnɔ = pull_VR
cok^{h} = row_VR
koʈɛ = serve_food_VR
ʃimu = soak_VR
buli = take_away_VR
cɔp^{h} = to_be_enough_VR
beno = sleep_VR
jira = speak_VR
ʈoya = stand_up_VR
kɛle = stay_VR
lele = swing_VR
emaʈ = run_VR
coŋ = get/find_VR
cɔŋ = get/find_VR
ʃɔr = sing_VR
noe = knit_VR
boi = ask_VR
bɔi = ask_VR
eɲo = come_VR
t^{h}u = come_out_VR
ŋol = cry_VR
ŋɔl = cry_V
caʈ = do_VR
bɔl = peel_VR
tɔl = roam_around_VR
eul = see_VR
iye = catch_VR

tok^h = close_VR

ʃui = cook/burn_VR

kaɲ = touch_VR

buʈ^h = fall_VR

iji = eat_VR

tɛʃ = give_VR

ʃol = walk/hang_VR

mok = leave_VR

muk = leave_VR

ɲyo = live_(home)_VR

rɔʃ = love_VR

oɖu = paste_VR

k^{h}i = pour_VR

k^{h}u = drink_VR

cɛr = rain_VR

bor = scratch_VR

lɛb = sweep_VR

cɔk = do_well_VR

ʃiʈ = hunt_VR

lub = pluck_VR

uno = sit_down_VR

ʈob = steal_VR

ŋɔʈ = swim_VR

ʃir = wash_VR

ɲa = bark_VR

ku = burn_VR

ɲa = eat_VR

cu = have_VR

ɖe = shut_up_VR

eb = take_VR

co = tie_VR

ie = give_VR

REFERENCES

[1] Abbi, Anvita 2003. *Vanishing Voices of the Languages of the Andaman Islands*. Paper presented at the Max Planck Institute, Leipzig.

[2] Abbi, Anvita. 2005. *Is Andamanese Typologically Divergent from Standard Average Andamanese*. In the 6th Biennial Meeting of Association for Linguistic Typology. Padang, West Sumatra, Indonesia. 21-25 July.

[3] Abbi, A. 2006. *Endangered Languages of the Andaman Islands*. Lincom-Europa: Munich.

[4] Choudhary, Narayan K. 2006. *Developing a Computational Framework for the Verb Morphology of Great Andamanese*. Unpublished Dissertation, Jawaharlal Nehru University, New Delhi.

[5] Endicott, Phillip, M. Thomas, P. Gilbert, Ch. Stringer, C. Lalueza-Fox, E. Willerslev, A.J. Hansen, A. Cooper. 'The Genetic Origins of the Andaman Islanders' The American Journal of Human Genetics. No. 72 (1), January 2003. Report no. 178.

[6] Hagelberg, Erika, Lalji Singh, K. Thangaraj, A.G. Reddy, V.R. Rao, S.C. Sehgal, P.A. Underhill, M. Pierson, I.G. Frame. 'Genetic Affinities of the Andaman Islanders. A Vanishing Human Population'. *Current Biology*, January 21, 2003:13, pp: 86-93

[7] Man, E.H. 1919. *A Dictionary of the South Andaman Language*, Indian Antiquary.

[8] Manoharan, S. 1989. *A Descriptive and Comparative Study of the Andamanese language*. Anthropological Survey of India: Calcutta.

[9] Portman, M.V [1898] 1992 (reprint). *Manual of the Andamanese Languages*. Manas Publications: Delhi.

[10] Radcliffe-Brown, A.R. 1948. *The Andaman Islanders*. Free Press: Illinois.

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.