Volume 4, Issue 4, April 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Morphological Generator for Kannada Nouns and Verbs

Bhuvaneshwari C Melinamath
Department of Computer Science
University of Hyderabad, India

Abstract— The morphological generator generate morphological forms for nouns and verbs when the root word is given. This paper describes a method of generating different word forms. Nouns in Kannada gets inflected for gender, number and vibhakti (cases). There are 2 numbers singular and plural and 3 genders feminine, masculine and neuter. 6 cases and another set of 6 extended case like suffixes, 4 clitics. A single noun root has around 250 morphological forms in Kannada where as a single noun root has only 2 forms in English say boy, boys. This shows the complexity of Kannada Language. Similarly a single verb root in Kannada has around 30000 different morphological forms where as verb root in English has 5 verb forms. Morphological generator are useful components of MT (Machine Translation application. The input to the generator is root word followed by category and required inflection types or all inflections. The morphological generator refers the morph dictionary to check for any morph related information present in the dictionary. We have stored morph related information in dictionary like information regarding 'real u', past participle form of verb or kinship terms, countable and uncountable features etc. These instances avoid generator to be over general which otherwise will generate in valid word forms. Our results regarding nouns and pronoun is more than 95% and around 85% for verbs. We have handled derivational morphology also.

Keywords— Natural language Processing (NLP), Part of Speech (POS), Expert Advisory Group on Language Engineering Standards) EAGLES, Machine Translation (MT).

I. INTRODUCTION

Morphology is the study of the internal structure and transformational processes of words. Words are formed from a combination of one or more free morphemes and zero or more bound morphemes. Free morpheme are units of meaning which can stand on their own as words. Bound morphemes are also units of meaning; however, cannot occur as words on their own: they can only occur in combination with free morphemes. The English word jumped is comprised of two morphemes, **jump+ed**. Since jump is an individual unit of meaning which cannot be broken down further into smaller units of meaning, it is a morpheme. And, since jump can occur on its own as a word in the language, it is a free morpheme. The unit +ed can be added to a large number of English verbs to create the past tense. Since **+ed** has meaning, and since it cannot be segmented into smaller units, it is a morpheme. However, +ed can only occur as a part of another word, not as a word on its own; therefore, it is a bound morpheme. The process by which bound morphemes are added to free morphemes can often be described using a word formation rule.

Both analysis and generation rely on two sources of information a dictionary of valid lemmas of the language and a set of inflections paradigms. The basic principle of morphological generation is to get forms from a root and a set of features (lexical category and morphological properties). Generally, there are two categories of approaches to developing a morphological generator. Approaches that use finite-state transducers (FSTs), such as Xerox Arabic analyzer (Beesley, 2003) and approaches that use rule based transformations, by (Cavalli- Sforza, 2000)

We have developed morphological generator kannada for nouns and verbs. Nouns in Kannada gets inflected for gender, number and vibhakti (cases). There are 2 numbers singular and plural and 3 genders feminine, masculine and neuter. 6 cases and another set of 6 extended case like suffixes, 4 clitics. A single noun root has around 250 morphological forms in Kannada Similarly a single verb root in Kannada has around 30000 different morphological forms. Morphological generator are useful components of MT (Machine Translation) application. The input to the generator is root word followed by category and required inflection types or all inflections. The morphological generator refers the morph dictionary to check for any morph related information present in the dictionary. We have developed 30000 plus words dictionary for this tag. The dictionary is tagged with hierarchical tag set. We have handled derivational morphology also.

The remaining part of paper comprises of 5 sections. Section 2 describes the work in this area, section 3 describes Kannada morphology, section 4 describes the proposed method, and section 5 is about performance and section 6 is about conclusion.

II. LITERATURE SURVEY

An Lots of work is carried out on development of morphological analyzers and generators for foreign languages and also for few Indian languages. Following is the gist of works cited in the literature. One of the most efficient approaches to morphological analysis and generation uses finite state transducers (FST) addressed by(Mohri, 1997). There are number of tools for the construction of FST based morphological analyzers. The best known being those developed at

Xerox by(Karttunen, 1993) and by (Chanod ,1994). (Kataja and Koskenniemi, 1988) presents a system for handling Akkadian root and pattern morphology by adding an additional lexicon component to (Koskenniemi, 1983) two-level morphology. The first large scale implementation of Arabic morphology within the constraints of finite-state methods is that of (Beesley, 1996). The approach of (McCarthy,1981) describing root-and-pattern morphology in the framework of auto segmental phonology has given rise to a number of computational proposals.(Kay,1987) proposes a framework with which each of the autosegmental tiers is assigned a tape in a multi-tape finite state machine, with an additional tape for the surface form. (Kiraz, 1994) extends Kay's approach and implements a small work kind multitape system. (Habash, 2004). has proposed an large scale lexeme approach for Arabic morphological generation. Regarding Indian Scenario we have attempts made by Akshar Bharati group using paradigm approach. An attempt for Kannada is made by K.N. Murthy, in his machine added translation but the performance of the system is 60 %. From the literature survey it is clear that the computational aspects of Kannada still need to explored. Kannada is lagging from the point of computational perspective. Clitics aspects are need to be handled. Aspect auxiliaries and modal auxiliary kind of derivational are to be touched.

.

III. ABOUT KANNADA LANGUAGE

Kannada is a Dravidian language. This is spoken in southern India. Kannada has complex morphology. Words are built up from roots by following fixed patterns that add prefixes, suffixes and infixes to the word. This system that studies how words are constructed from roots, and describes the patterns they follow which in English could be called derivational morphology. Kannada language uses 49 phonemic letters, divided into 3 groups: 13 swaragaLu (called vowels in English), 34 vyaNjangaLu (called consonants in English) and 2 yogavaahakagaLu (neither consonants nor vowels), anusvara, namely, "aM " and visarga, namely, "ah ". Verb formation in Kannada is complicated. For the above verb exists (iru) we have around 3000 verb forms. We can add any number of affixes to form complex verb root. A single root word in Kannada can give rise to a very large number of its surface forms. The richness of Kannada lies in the fact that significant part of grammar is handled by word morphology. Consider the Kannada verb form.

Kannada Word ತಪ್ಪಿಸಿಕೊಳ್ಳಬಲ್ಲವವರನ್ನೂ

Transliterated Form tappisikoLLaballavavarannoo

English Including those capable of escaping

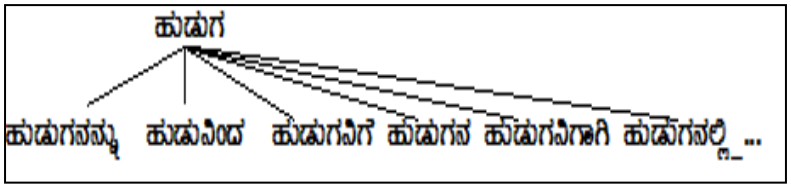
KHPOS Tag : V-IN-CAU-CJP-REF+INF+CAP+PRO.avanu+ACC+CLIT.oo IIIT-H VGNN acc

IL-POST : NV.0.0.mas.pl.acc.0

BOX 1. COMPLEX KANNADA GERUND

This Kannada word form is derived by adding 9 suffixes to the root verb tappu (wrong). The formation of this verb form is Verb-intransitive-causative-Verbal participle-reflexive + infinitive +modal (CAP) +PRO.avaru+accusative+clitic indefinite.

A noun boy in English has at the maximum 2 inflections boy and boys. But the same word in Kannada has around 250 forms. All these morphosyntactic aspects of the complex morphology has to adopted while generation. Hence designing morphological generator for Kannada language is quite challenging.



Box II. Different noun Forms of Kannada noun

A. Noun Morphology

Nouns in Kannada may be distinguished for gender, number, case and clitcs.

- **Gender:**

Nouns referring to biologically female beings are feminine in gender, that are biologically male are masculine in gender and nouns that are not thought to be "rational'(capable of thought) are referred as neuter. Sometimes young children, small kids are treated as non-rational.

- **Number:**

Kannada nouns are distinguished by two numbers, singular and plural. The singular has no particular distinguishing marker added. The plural marker 'gaLu' is used for neuter nouns, 'ru' is used for others. There is an exception in some case whether gender is not specific in human noun 'gaLu ' is also used as plural marker for in examples like prajegaLu 'citizen. Usually all masculine and feminine nouns ending in 'a','i', 'e', and consonant followed by enunciate

u have plural marker as 'ru'. Some masculine and feminine are not specially marked for gender like vyaapaari (merchants), adhikaari(officers), these words take plural marker 'gaLu' which is generally used as neuter plural marker, we are marking such words as with tag MF, to indicate dual sense. Since such word are used to represent even the feminine gender. Some nouns have irregular plurals makkaLu 'children' Some nouns have no singular forms, they have only plural usage. Like jana 'people' here janaru (people), janagaLu (people) are two plurals. aneekaru 'many people'. Some noun have no usage in plural form, say gurugaLu 'teacher' has no corresponding plural.

- **Case system:**

The case system is useful to indicate different relationships between the noun and other constituents of the sentence. For Ex to indicate whether the noun is the "object" of a verb (in which case it is marked for accusative case), or "goal" of a verb of motion (dative case), posser of something (genitive case), or the means by which something takes place (ablative case), etc.

TABLE 1. NOUN CASES AND THEIR CHARACTERISTICS

Case/vibhakti	Suffix/pratyaya	Characteristics	
		Singular	Plural
Nominative (prathama)	u	U	ru/gaLu/vu/aMdiru
Accusative (dvitiya)	annu	nannu/Lannu/vannu/annu/yannu	rannu/aMdirannu/gaLannu/yaranau
Ablative/Instrumental (tritiya)	iMda	niMda/diMda/yiMda/viniMda/LiMda/	gaLiMda/riMda/yariMda/
Dative	ge/ige/kke	nige/Lige/vige/yige	gaLige/rige/ge
Genetive (shashTi)	a	La/na/da/vina	ra/ gaLa/aMdira
Locative (saptami)	alli	nalli/yalli/Lalli/dalli/vinalli	ralli/yaralli/gaLalli

TABLE 2. EXTENDED NOUN CASE SYSTEM

Case	Suffix
Purposive	Gooskara/gaagi
Comparative	giMta/kkiMta
Locative_Dative	oLage/allege
Similarative	aMte/aMtaha/aMtha
Sociative	oDane/oMdige

2.2 Verbs Morphology

Verbs are one of the most interesting and distinguishing aspects of Kannada. The verb stems in Kannada can be divided into finite and nonfinite a finite verb ends the sentence with the exception of clitics. But the nonfinite verbs can not stand alone and requires finite verb for ending the sentences. includes infinitives modal auxiliaries, aspect auxiliaries. Finite verbs includes imperatives, it is in this form the Kannada verb roots are listed in the dictionary.

DERIVATIVE STEMS: These are formed by adding various kinds of derivative suffixes to the verbal or nonverbal stem. Consider the following examples.

Derived Stem (1) tinnu 'eat'+/alu/→ tinnalu, (to eat) tinnu is basic stem

COMPLEX VERBS: These are formed by adding various kinds of models to the primitive and derived stems. These can be further subdivided into compound verbs, conjunct verbs, modal verbs and aspectual verbs. Kannada has at least four nonfinite forms of verbs. Like the participial constructions, verbal participles, relative participles, they are aspectually distinguished.

TABLE 3 VERB FEATURES AND CHARACTERISTICS SUFFIXES

Features	**Characrtistic Suffixes**	**Example**
Infinitive	alu	maaDu+alu→maaDalu (to do)
Imperative	i,iri,oo,ee	maaDu+iri→maaDiri (you do)
Negative Imperative	beeDa,baaradu,kuuDadu	maaDu+alu+beeDa→maaDabeeDa

		(Don't do)
Optative	Li	maaDu+ali→maaDali (let him do)
Hortative	ooNa	maaDu+ooNa→maaDooNa (let us do)
Participle	a/i/ade	maaDu+uv+a→maaDuva (Which will be done) maaDa+i→maaDi (after eating) maaDu+ade→maaDade (without doing)
Verbal aspect Markers	koLLu,haaku (table 4)	maaDu+i+haaku→maaDihaaku (do non finite and put finite)
Causative Suffix	Isu	maaDu +isu →maaDisu (get it done)
Conditional Suffix	are, doDe	maaDu+id+are→maaDidare (did they do)
Concessive	Aruu	maaDu +id +aruu → maaDidaruu (even though I did)
Imprecate	A	maaDu +a→ maaDa
Emphatic clitic	Ee	maaDu+utt+aane+ee→maaDuttaaneyee (will he do)
Interrogative clitic	Oa	maaDu+uva+anu+aa→ maaDuvanaa (will he do)
Indefinite clitic	Oo	maaDu +id +anu+oo→ maaDinoo (whether he has done it)

TABLE 4: TABLE SHOWING MODAL AUXILIARIES

Modal auxiliary	**Meaning**	**Example**
Beeku	MUST (Want)	Huuvu beeku(want Flower)
kuuDadu	PROH(Should not)	Niivu hoogakuuDadu(you should not go)
beeDa	NEG(IMP)	Nanage pustaka beeDa(I donot want book)
Bahudu	PERM(May)	niivu barabahudu(you may come)
Aara	NCAP(might not)	avanu tinnalaranu(he cannot eat)
Balla	CAP(capable)	avanu maaDaballanu(he can do)
paDu	PASS(Passive voice)	haNNu tinnalapaDuttade(Fruit will be eaten)

TABLE 5: INVENTORY OF ASPECT AUXILIARIES

Aspect Marker	**Aspect Meaning**	**Lexical Meaning**	**Example**
biDu	Completion	Leave	avanu biddubiTTanu (he fell down)
Hoogu	Completion	Go	anna beMditu(rice got over cooked)
aaDu	Continuity	Play	avaru ooDaaDidaru(They ran around)
koDu	Benafactive	Give	avanu kathe baredukoTTa(He wrote story and gave)
nooDu	Attemptive	See	avanu kaaphi kuDidunooDida(he drank coffee and see
Haaku	Exhaustive	Put	avanu dose tiMduhaakidanu
koLLu	Reflexive	Purchase	avanu tale baachikoMDanu(he Combed)
Aagu	Finality	Become	avanu baMdaayitu(he finally came)
Iru	Perfective	Be	avanu hoogiddanu(he had gone and Stayed)

TABLE 6. KANNADA PNG MARKERS

Person	**Number**	**Gender**	**Tenses**			
			Present	**Futuret**	**Past**	**Contigent**
First	Singular	Masculine /Feminie	Eene	enu	Enu	(y)eenu
	Plural	Masculine /Feminie	Eeve	evu	Evu	(y)eevu
Second	Singular	Masculine /Feminie	ii(ye)	e	E	Iiye
	Plural	Masculine /Feminie	Iiri	Iiri	Iiri	Iiri
Third	Singular	Masculine	Aane	anu	Anu	(y)aanu

	Singular	Feminie	aaLe	aLu	aLu	(y)aaLu
	Plural	Masculine /Feminie	Aare	Aru	aru	(y)aaru
	Singular	Neuter	Ade	Adu	itu	Iitu
	Plural	Neuter	Ave	Avu	avu	(y)aavu

IV. PROPOSED METHOD OF DEVELOPING MORPHOLOGICAL GENERATOR

To develop a morphological generator we must consider the following factors in particular. The morphological generator consists of following sub modules as explained in the figure 1. The process of combining morphemes involves a number of orthographic rules that modify the form of created word so it is not a simple interleaving or concatenation of its morphemic components.

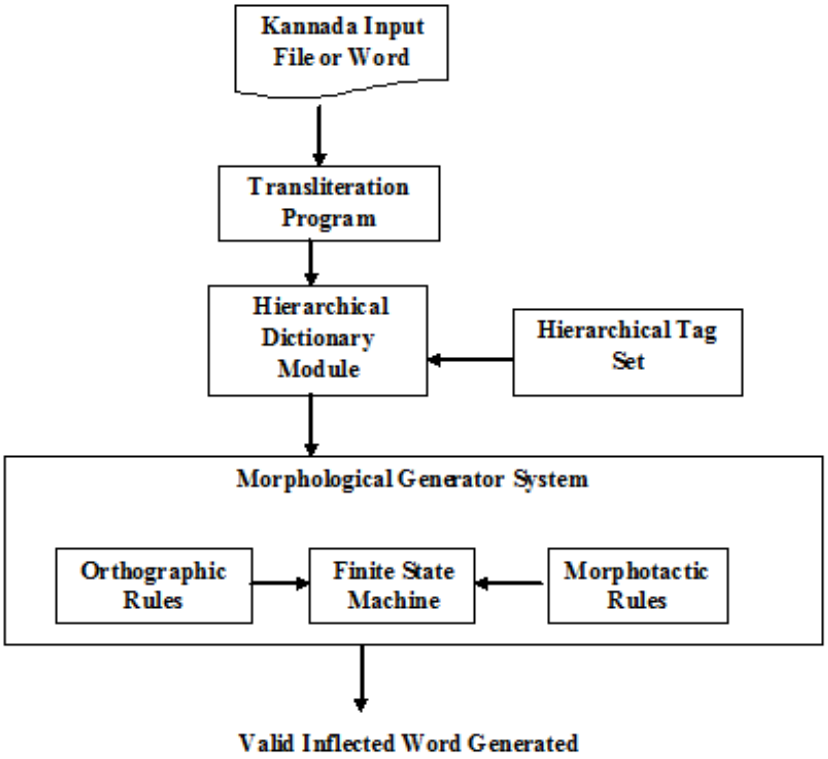


Fig 1. Proposed Method for Morphological Generator

***A. Transliteration Module*:** This module converts the raw text in Kannada in to roman form by using a convertor program. If raw text is in ISCCI or Unicode then the convertor program accepts both kind of input and transliterate the text as per our lerc notation. The input file or word should be a uninflected root.

B. RESOURCES : To develop a generator we must have some resources. These resources are

- **Dictionary:** A dictionary contains roots/stems, categories and exceptions, morph relevant information etc.
- **Orthographic Rules Set** : Morphological generators use the concatenation process while they generate a word-form. So, we must create a set of concatenation rules for the generator according to the language. The rules changes with word endings as shown in table 1. The alternation forms (spelling changes) of morphemes according to the context in which they appear.

TABLE 1: SAMDHI RULES

	Stem Ending						
	a	**E**	**I**	**o**	**U**		**Consonant**
					Real	**Enunciative**	
Glide insertion for rational	**n**	**Y**	**Y**	**n**	**v**	**u dropped**	**N**
Glide for non rational	**v**	**Y**	**Y**	**v**	**v**	**u dropped**	**-**

C. Morphotactic Rules: These rules govern the order of suffixation. Generally the additions of suffixes to noun and verb follow the pattern like below.

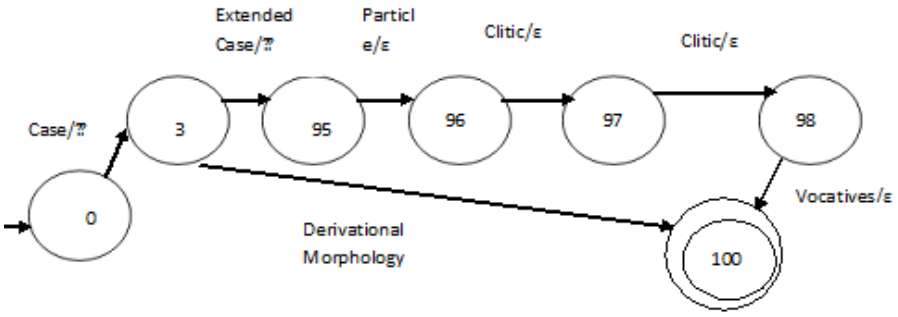


Fig 2. Morphotactic Rule for Kannada Nouns

D. Our Morphological Dictionary

The development of dictionary goes on with a specific requirement. The purpose of our development of dictionary is, we require a dictionary of the type, lemma followed by its tag, The structure of our morphological dictionary entry is shown in figure 2.

Word	Seperator1	Tag1	Tag2	Seperator2	Morph related Information

Box 3. Structure of Dictionary Content

One entry per line. <word>||<TAG>||<TAG>:: Relevant Morph information.
Grammatical information useful for our morphology system is also stored in dictionary using the separator :: (double colon) to help morphological generator. The word guru (teacher) is a word ending with real 'u'. But the word niiru (water) even though ends with u but it is not real 'u', it is enunciative. It is necessary to store such useful information for morphological generator. We have stored such morph related information in the dictionary as shown here guru|| N-COM-COU-M.SL-NOM-NULL: REAL u.

TABLE II. Sample Dictionary

Kannada Word	Transliteration	Dictionary Tag	English Meaning
ನಾನು	naanu	PRO-PER-P1.MFN.SL-NOM	I
ನಿನಗೆ	ninage	PRO-PER-P2.MFN.SL-DAT	To you
ತಮ್ಮ	tamma	PRO-REF-P23.MFN.PL-GEN‖N-COM-COU-M.SL-NOM: TYPE Kinship	Brother, Another meaning is themselves
ಅದಕ್ಕೆ	adakke	PRO PER P3.N.SL DIST DAT‖ADV-CONJ	to that
ಹೊತ್ತು	hottu	N-LOC-TIM-UNC-ABS-N.SL-NOM‖V-IN :: TYPE-Cjp-of-horu	Indicate Time instance and also Past participle form of a verb carry i.e "being carried"
ಗುರು	guru	N-COM-UNC-N.SL-NOM: LV- real -u	Teacher
ನೀರು	niiru	N-COM-UNC-N.SL-NOM	Water
ಹುಡುಗ	huDuga	N-COM-COU-M.SL-NOM	Boy
ಹುಡುಗತನ	huDugatana	N-COM-UNC-N.SL-NOM	Boyhood

D Finite State transducers

The proposed tool is developed using. Finite state transducers. Finite State Automaton (FSA) is a model of computation consisting of a finite set of states, a start state, an input alphabet, and a transition function that maps input symbol and current state to next state. The transducer is defined as T = (Q, L,δ, qI, F,) where Q is a finite set of states, L a set of transition labels, qI Є Q the initial state, F ⊆ Q the set of final states, and $\delta : Q \times L \rightarrow 2^Q$ the transition function (where 2^Q represents the set of all finite sets of states). The set of transition labels is $L = (\Sigma \cup \{ \varepsilon \} \times (\Gamma \cup \{ \varepsilon \})$ where Σ is the alphabet of input symbols, Γ the alphabet of output symbols, and ε represents the empty symbol.

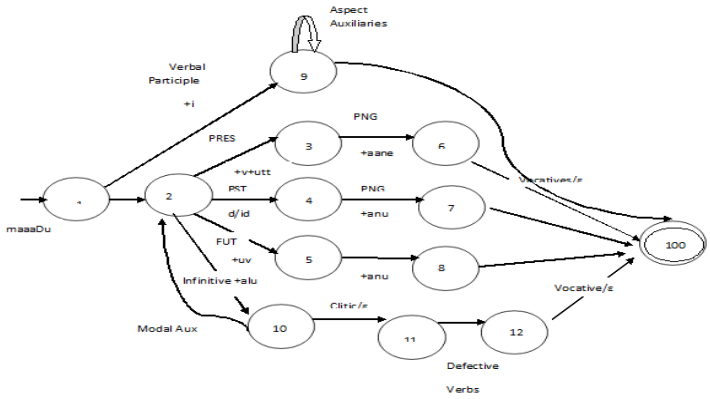


Fig 3. FSM Transition for Verb Limited Version

V. EXPERIMENTS AND RESULTS

We manually generated gold standard data for nouns by choosing 2 words from a/e/i/o/u endings and 3*6*250 combinations, covering all inflections are chosen and matched that with inflected forms generated by morphological generator. We observed that more than 95 % words are correctly matched with that of gold standard data. Similarly we prepared gold standard data of 3000 inflections for verb. And observed that the matching performance is rate is around 85 %

VI. CONCLUSION

It is shown here that each noun in Kannada has about 250 word forms for single noun. One can observe the complexity in morphology as compared to English, which has only three forms like say boy, boys, boy's. Similarly in English maximum inflections for a verb 'go' are 5 in number like, go, gone, went, going, goes. But in Kannada it is proved that more than 3000 different forms exists. One can understand the complexity of Kannada. Developing tools like morphological analyzer or generator is quite challenging. Morphological generator for Kannada nouns and verbs is developed here using finite state transducers. Morphological genearator is useful tool in Machine translation applications. A dictionary of 30000 words is developed with hierarchical tag set for this task.

REFERENCES

[1] Martin Kay. 1987. Nonconcatenative finite-state mor phology. In Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics, pp. 2–10.

[2] L. Kataja and K. Koskenniemi. 1988. Finite state description of Semitic morphology. In COLING-88: Papers Presented to the 12th International Conference on Computational Linguistics, volume 1, pp 313–15.

[3] Koskenniemi, Kimmo. 1983. : Two level Morphology: A general Computational Model, for Word Form Recognition and Production. Ph.D. Thesis, Department of General Linguistics, University of Helsinki, Helsinki, Finland.

[4] Chanod, Jean-Pierre. 1994. Finite state composition of French verbmorphology. Technical report MLTT-005, Xerox Research Centre Europe, Meylan, France.

[5] Karttunen, Lauri. 1993. Finite state lexicon compiler. Technical Report ISTL-NLTT-1993-04-02, Xerox Palo Alto Research Center, Palo Alto, California.

[6] K. Beesley 1996. Arabic finite-state morphological analysis and generation. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), volume 1, pp. 89–94,Copenhagen, Denmark.

[7] [Buckwalter, 2002] T. Buckwalter. Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Catalog No.: LDC2002L49, 2002

[8] Hopcroft, J. E., & J. D. Ullman. : Introduction to automata theory, lan guages. and computation. Reading, MA : Addition -Wesley .

[9] Mohri, Mehryar. : Finite-state transducers in language and speech processing. Computational Linguistics 23(2):269–311. (1997)

[10] Oncina, Jose, Pedro Garc`ia, & Enrique Vidal. 1993. Learning subsequential transducers for pattern recognition interpretation tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence 15:448-458

[11] John McCarthy.1981. A prosodic theory of nonconcatenative morphology. Linguistic Inquiry, 12(3):373–418.

[12] Kiraz, G.: Multi-tape Two-level Morphology: A Case study in Semitic Non-Linear Morphology. In: Proceedings of COLING'94, Vol. 1 (1994) 180-186.

[13] Cavalli-Sforza, V. Soudi A., Mitamura T.: Arabic Morphology Generation Using a Concatenative Strategy. In:Proceedings of NAACL 2000. Seattle, WA (2000).

[14] Nizar Habash. 2004. Large scale lexeme based arabic morphological generation. In Proceedings of Traitement Automatique du Langage Naturel (TALN-04). Fez, Morocco

