LANGUAGE IN INDIA

Strength for Today and Bright Hope for Tomorrow

Volume 7 : 1 January 2007

COMPLEXITY OF TAMIL IN POS TAGGING

S. Rajendran, Ph.D.

COMPLEXITY OF TAMIL IN POS TAGGING

S.Rajendran, Ph.D.

The paper aims to focuses on the Morphological complexity in Tamil language from the point of view of POS tagging. Nouns get inflected for number and cases. Verbs get inflected for various inflections which include tense, finite and non-finite suffixes. Verbs are adjectivalized and adverbialized. Also verbs and adjectives are nominalized by means of certain nominalizers. Adjectives and adverbs do not inflect. Many post-positions in Tamil are from nominal and verbal sources. So, many times we need to depend on syntactic function or context to decide upon whether one is a noun or adjective or adverb or post position. This leads to the complexity of Tamil in POS tagging.

PARTS OF SPEECH IN TAMIL

The following parts of speech or word classes are identified for Tamil languages by modern grammarians:1) Noun, 2) Verb, 3) Adjective, 4) Adverb, 5) Postposition, 6) Numeral, 7) Quantifier, 8) Words of conjunction, 9) Exclamatory words, 10) Words expressing feeling, 11) Word of calling, and 13) Words accepting calling.

NOMINAL COMPLEXITY

Nominal forms show the following structure:

Noun (+Number) (+Case)

marang-kaL-ai ‘trees_PL_ACC’

Though at the underlying structure there are only two grammatical morphemes, on the phonological level, however, four types of morphs (or suffixes) can occur with the noun stem (Lehmann, 1989:12):

Plural suffix
Oblique suffix (increment)
Euphonic suffix (increment)
Case suffix

So we have the following structure (Lehmann, 1989:13):

Noun stem/Oblique stem (+euphonic increment) + case suffix)

caavi.y-(in)-aal 'key_euph_inst'
mara-tt(-in)-aal 'tree_obl_euph_inst'

Noun stem + plural suffix (+euphonic increment) + case suffix)

viiTu-kaL(-in)-ai 'house_pl_euph_acc

Nouns need to be annotated into pronoun, proper noun and common noun. Pronouns need to be further annotated for person (1^{st}, 2^{nd} and 3^{rd}), number (singular and plural), gender (masculine, feminine, neuter), status (honorific and non-honorific). Nouns need to be annotated into rational and irrational. Also nouns need to be annotated for nominative, accusative, dative, instrumental, sociative, locative, ablative, genitive, vocative cases. Nouns and Pronouns need to be annotated as oblique or non-oblique form.

In the following examples, *aaTu* is in nominative (i.e. non-oblique) form, where as *aaTTu* is oblique form; the formal difference makes difference in sense.

aaTu (nom) vs aaTTu (obl) ‘goat’
aaTu mandaiyil irukkiRatu ‘Goat is in herd’
aaTTu mandtaiyil avan oLindtaan ‘He hid himself in goat herd’
ndaan (nom) vs en (obl) ‘I’
ndaan en viiTTukkup pooneen
‘I my house_DAT go_PAST_FP

Furthermore, nouns need to be annotated for number and gender (masculine, feminine, and neuter) as the subject nouns show agreement with PNG marker at the finite verbal form. Nominaliztion makes the nominalized verbal form more complex. Nomininalized verbal forms need to be distinguished into two or three types. For example, Tamil requires the productive forms formed by the suffixation of *tal/kai/aamai* which are sentential in nature are to be differentiated from non-productive forms formed by the suffixation of *ppu* etc. which are lexical in nature. In the following examples, *paTittal* is sentential form and *paTippu* is lexical form.

kaalaiyil ezhundtu paaTangkaLaip paTittal ndallatu
morning_IOC wake up_ADV study_NOM good
‘It is good to wake up and study the lessons in the morning’

avan meel paTippu paTikka ameerikkaa cenRaan
he higher studies study_INF America go_PAS_he
‘He went to America for learning higher study’

al-suffixed nominalized forms need to distinguished into two types as one type is lexical another is sentential. In the following example *aaTal* is a nominalized form of the verb *aaTu* ‘dance’.

avaL aaTal avanaik kavarndtau
‘her dance he_ACC attract_PAS_he
‘He dance attracted him’

avaL aaTal-aam
she dance_NOM_be
'She may dance'

kai suffixed nominalized form need to be distinguished into two: lexical form and sentential form. In the following examples *vaazhkkai* is lexical and *vaazhkai* is sentential.

vaazhkkai(il) 'in life' vs vaazhkai(il) 'while living'
avan vaazhkkai tunpamaantu
'His life is sorrowful'
avan makizhcciyaaka vaazhkaiyil ndaan avanaic candtitteen
he happily live_NOM_LOC I he_ACC meet_PAS_I
'I met him while he was living happily'

But *paTukkai* is ambiguous, as both the lexical and sentential forms are homophonous.

paTukkai(il) 'in the bed' / 'while lying down'
avan paTukkaiyil irundtaan
he bed_LOC sit_PAS_he
'He sit on the bed'

avan paTukkaiyil tolaipeeci maNi aTittatu
he lay_NOM_LOC telephone ring rang
'While he was about to lie down, the telephone rang'

atu suffixed ambiguous forms in Tamil need to be distinguished into three types.

vandtatu 'that which came'/ 'that somebody came' /'it came'

vandtat(u)-illai ‘did not come’

ndaan uuTTiyilirundtu vandtatai paartteen
I Ooty_LOC_ABL come_PAS_NOM see_PAS_I
‘I saw that which came from Ooty’
avan vandtatu enakkut teriyaatu
he come_PAS_NOM I_DAT know_NEG
‘I didn’t know that he came’
atu vandtatu
it come_PAS_it
‘It came’
ndaan uuTTikku vandtatillai
I Ooty_DAT come_NOM_NEG
‘I did not come to Ooty’

Many pronominalized forms are also ambiguous in Tamil and need to be distinguished into two types: lexical and sentential (productive).

paTittavan ‘educated male person’ / ‘one who read x’
andta puttakattaip paTittavanaip paaraaTTa veeNTum
that book_ACC read_PAS_ADJ_he appreciate necessary
‘We should appreciate the person who read that book’

avan mikavum paTittavan
he very educated_person
‘He is an educated person’

VERBAL COMPLEXITY

The verbal forms are complex in Tamil. A finite verb shows the following morphological structure:

V+Tense+PNG

A number of non-finite forms are possible: adverbial forms, Adjectival forms, infinite forms and nominalized forms

vandtu ‘having come’
varaamal/varaatu ‘without coming’
vara ‘to come’
vandtatu (illai)

Distinction needs to be made between main verb followed by main verb and main Verb followed by an auxiliary verb. The main verb followed by an auxiliary need to be interpreted together, whereas the main verb followed by a main verb need to be interpreted separately. This lead to functional ambiguity as given below:

FUNCTIONAL AMBIGUITY IN ADVERBIAL FORM

vandtu caappiTTu viTTu poo ‘having come and having eaten went’
ndondtu poo ‘become vexed’

FUNCTIONAL AMBIGUITY IN INFINITIVAL FORM

vara.v-iru ‘going to come’
vara-k.kuuTaatu ‘should not come’
vara-c.col ‘ask x to come’

The adjectival forms differ by tense markings: V+Tense+Adjectivalizer

vandta ‘x who came’
varukiRa ‘x who comes’

varum ‘x who come’

Adjectival form allows several interpretations as given in the following examples.

cappiTTa ilai ‘the leaf which is eaten by x’
‘the leaf on which x had his food and ate’
vaangkiya x ‘x which is bought’
‘x who bought’
‘x (price) by which something is bought‘
‘x (money) received’
‘x (container) in which something is received’

um-suffixed adjectival form clashes with other homophonous forms which leads ambiguity.

varum paiyan ‘the boy who will come’
varum ‘it will come’
varum pootu ‘while coming’

The adjectival forms when followed by nouns such as *ceyti* ‘news’, and *uNmai* ‘fact’ etc. are ambiguous as they allow relative interpretation and non-relative interpretation.

avan paTitta ceyti
‘the news which he has read’ (when interpreted as a relativized form of the sentence avan ceyti paTittaan ‘he read the news’)
‘the information that he has read’(similar to avan vandta ceyti ‘the information that he has come’

avan paTitta uNmai (when interpreted as the relativized from the sentence avan uNmaiyaip paTittaan ‘He read the truth’) ‘the truth which he read’

avan paTitta uNmai ‘the fact that he has read’

Some adjectivialized verbal forms of verbs are lexicalized as adjectives (as against sentential ones). So there is ambiguity in the interpretation of them purely as an adjective modifying only the noun which it follow and sentential adjective modifying the noun which stands as a relative clause modifying the nominalizer (i.e. noun which moved to position after the relativized verb).

iruNTa ‘dark’ (lexicalized) as in iruNTa kaalam ‘dark period’
iruNTa ‘dim’ (relativized) as in (kaNNkaL iruNTana ‘eyes became dim’ >)
iruNTa kaNkaL ‘the eyes which became dim’
ndiiNTa kai ‘long’ vs. ndiiNTa kai ‘extended hand’

Nominals can function as adjectives modifying a noun as given in the following examples.

mara-ppeTTi (T) ‘wooden box’
akala paatai ‘wide path

mara ‘wooden’ is a reduced form of *maram* ‘tree’ and *akala* ‘wide’ is a reduced form of the noun *akalam* ‘width’

Verbal roots functions can function as adjectives as given in the following examples.

cuTu cooRu (T) ‘hot rice’
aazh kiNaRu (T) ‘deep well’

cuTu ‘be hot’ and *aazh* ‘dig’ are verbal roots. Use of verbal roots as adjective is a productive process.

A number of adverbial forms of verbs functions as postpositions. They are discussed under 'complexity in postpositions'.

COMPLEXITY IN ADVERBS

We have seen that a number of adjectival and adverbial forms of verbs are lexicalized as adjectives and adverbs respectively and clash with their respective sentential adjectival and adverbial forms semantically creating ambiguity in POS tagging.

Adverbs too need to be distinguished based on their source category. Many adverbs are derived by suffixing *aaka* with nouns in Tamil. But not all *aaka* suffixed forms are adverbial.

veekam-aaka (T) 'fast' vs. TaakTar-aaka 'as doctor'

Functional clash can be seen between adjective and adverb in *aaka* suffixed forms. This type of clash is seen among other Dravidian languages too.

avaL azhakaaka irukkiRaaL
'she beauty_ADV be_PRE_she
'she is beautiful'

COMPLEXITY IN POSTPOSITIONS

Postpositions are from various categories such as verbal, nominal and adverbial in Tamil. Many a time, the demarking line between verb/noun/adverb and postposition is slim leading to ambiguity. Some postpositions are simple and some are compound. Postpositions are conditioned by the nouns inflected for case they follow. Simply tagging one form as postposition will be misleading

There are postpositions which come after noun and also after verbs which makes the postposition ambiguous (spatial vs. temporal).

pinnaal ‘behind’ as in viiTTukkkup pinnaal ‘behind the house’
pinnaal ‘after’ avanukkup pinnaal vandtaan ‘he came after him’

Use of adverbial forms of verbs leads to ambiguity in the annotation of postpositions,

katti koNTu
knife have_ADV
‘by means of knife/having the knife’

avaLaik koNTu
she_ACC have_ADV
‘by means of her/having her’

Similarly the following adverbial forms leads to problems in POS tagging.

viiTT-il irundtu
house_LOC be_ADV
‘from the house/being in the house

The complex postpositions still makes things more complex.

viiTT-il-irundtu-koNTu
house_LOC be_ADV_have_ADV
‘being at home’

Similarly the following postpositions from verbal source (adverbial forms of certain verbs) may lead to ambiguous annotation.

oTTi 'regarding' the adverbial from of the verb *oTTu* 'stick
kuRittu 'about' the adverbial form of the verb *kuRi* 'aim, mark'
cuRRi 'around' the adverbial form of the verb *cuRRu* 'circulate'
tavirttu 'except' the adverbial form of the verb *tavir* 'avoid'
paRRi 'about' the adverbial form of the verb *paRRu* 'seize'
viTTu 'from' the adverbial form of the verb *viTu* 'leave'
vaittu 'with' the adverbial form of the verb *vai* 'put'
ndokki 'towards' the adverbial form of the verb *ndookku* 'see'
pindti 'after' the adverbial form of the verb *pindtu* 'be behind'
mundti 'before' the adverbial form of the verb *mundtu* 'precede'
tavira 'except' the infinitive form of the verb *tavir* 'avoid'
viTa 'than' the infinitive form of the verb *viTu* 'leave'

CONCLUSION

Tamil is no doubt a morphologically rich language. The relation between verb and its nominal arguments is decided by case suffixes rather than position. It is possible to have a few numbers of tagset at shallow level. But one needs to address other unique features at the deep level. Hierarchical tagset is a welcome thing.

REFERENCES

Arulmozhi, P and Sobha, L. 2006. A Hybrid POS tagger for a Relatively Free Word Order Language. In Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages, pages 79-85.

Brochures on 'Language Technology Products' of the Resource Center for Indian Language Technology Solutions – Tamil, Chennai

Kumara Shanmugam, B. 2004. Parse representation of Tamil syntax. MS Thesis, submitted to Anna University, Chennai.

Lehmann, Thomas. 1992 (second edition). A Grammar of Modern Tamil. Pondicherry Institute of Linguistics and Culture, Pondicherry.

Language Analysis and Understading. In Survey of the State of Art in Human Language Technology. (A downloaded script)

Rajendran S, Arulmozi S, Ramesh Kumar S, & Viswanathan S. 2003. Computational Morpohology of Verbal Complex In B. Ramakrishna Reddy (ed.) Word Structure in Dravidian, Kuppam: Dravidian University, 376-398.

Ranganathan, V. 1997. "A Lexical Phonology Approach to Processing Tamil Word by Computer", International Journal of Dravidian Linguistics 26.1:

Shanmugam, C. 2001. "Computer Analysis of Simple Sentence in Tamil", Paper read in UGC-SAP National Seminar on Computational Linguistics and Dravidian Languages, 22-24 February, 2001, CAS in Linguistics, Annamalai University, Annamalainagar.

---------2002. "Grammar and Parser: A Program for Syntactic Parsing in Tamil", International Seminar on Tamil Computing, 27-28 February and March 1, 2002, University of Madras, Chennai.

----------"Minimalist Program for Tamil Parsing".

Sobha, L and Vijay Sundar Ram. 2006. "Noun Phrase Chunker for Tamil Language", in Proceedings of the First National Symposium on Modeling and Shallow Paring of Indian Languages, pages 194-198.

S. Rajendran, Ph.D.
Department of Linguistics
Tamil University
Thanjavur 613 005
Tamilnadu, India
raj_ushush@yahoo.com