# paṭi (படி) in Tamil: A corpus based Investigation

S.Thennarasu, Dr. R.Prabagaran,
L.R.Premkumar, A.Vadivel and R.Amudha
LDC-IL, CIIL, Mysore

# Introduction

The word paṭi (படி) is one of the most frequently occurring words in the corpus, showing various morpho-syntactic functions.

Morphologically, the word can be analyzed as noun, verb and particle occurring after a noun, pronoun and relative participle.

Because of these categorical variations, the determining status of this lexical item ‘paṭi’ has become a challenging field for POS tagging.

In this paper, we explored the various functional properties of ‘paṭi’ and as a result, identified basic morpho-syntactic categories along with their functions.

We also explored how a machine can recognize the lexical word ‘paṭi’ by giving algorithms to resolve categorical ambiguities in language without recourse to semantic level information.

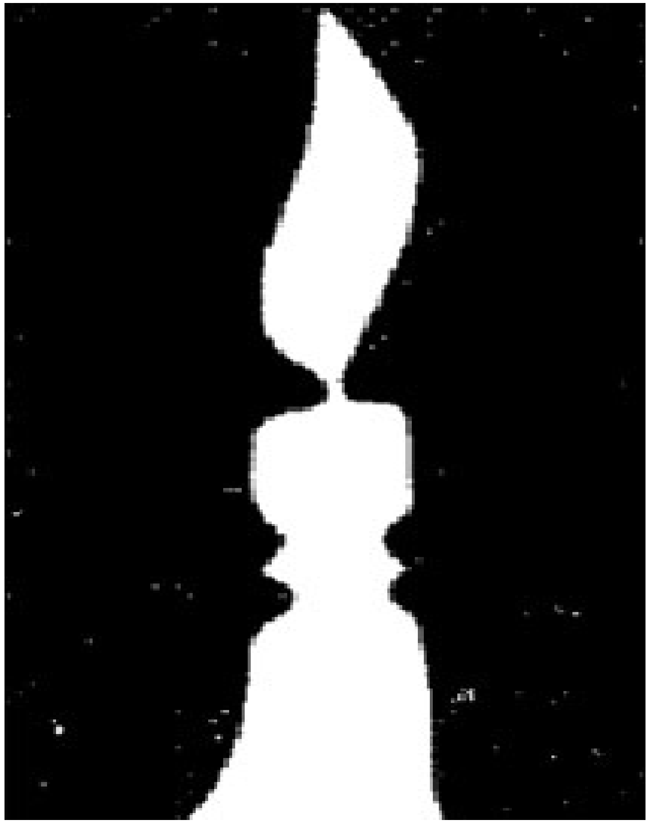
Based on the annotated corpus, the lexical item ‘paṭi’ in various occurring environments is applied and fitted to the feasible algorithm, a recursive operation, which gives command to the NLP tools.

Our Team have developed a method called LDCIL-WS which has been developed for Tamil.

This algorithm uses a systematic calculation based upon the probabilities of co-occurrence of particular tags and for which it disambiguate in linear order.

Tests of the algorithm using the 50k CIIL-Tamil annotated corpus are reported; the overall accuracy is nearly 90-96%.

It is also suggested that this algorithm can provide an accurate front end to any POS tagging system for languages like Tamil.



The word ‘paṭi’ in which sense it comes it depends upon the following verb.

For Example
paṭi ēṟiṉāṉ (ēṇippaṭi ‘step of ladder’, māṭippaṭi ‘upstairs’)

nī naṉṟāka tamizai paṭi eṉṟār. ’He said, you read Tamil properly’

Ippalkalaikkazakattil 2011ām āṇṭuk kaṇakkiṉ paṭi 2,500 māṇavarkaḷ payilkiṉṟaṉa.
‘According to the year 2011 censes, 2,500 students are studying in this university.’

paṉi paṭinta malaittoṭarkaḷaik kaṇṭēṉ
‘I saw the mounts which covered with snows’

atikārikaḷ muṉ paṭintu pēciṉār
‘He was submissive in speaking in front of officers’

avarkaḷ iruvarum vāykkuvantapaṭi oruvarukkoruvar tiṭṭikkoṇṭārkaḷ. ‘without restraint’

# Source for the study

For the study of lexical item 'paTi', we have used kriyavin tarkaalath tamizh (Cre-A: A dictionary of contemporary Tamil (Tamil-Tamil-English) as a secondary source.

The annotated corpus, on the other hand, are similar to primary data and are especially useful from the perspective of NLP.

The word ‘paṭi’ is the root word can be morphologically analyzed in the following three ways:

(1) Verb meanings as ‘read’

(2) Noun meanings such as ‘step’,

(3) Particle (after a noun or pronoun,) meaning such as providing the interpretation of ‘as per’, ‘according to’,

We automatically extracted sentences containing the occurrences of ‘paṭi’ including the suffixed form, from the CIIL Tamil corpus.

The extracted sentences were manually tagged and analyzed we tried to come up with a few basic morpho-syntactic classes under which all occurrences of ‘paṭi’ been categorized.

In order to define these basic classes in an unambiguous fashion from Tamil as described below.

Syntactic function as perceived by native speaker intuition as well as linguistic analysis;

Distribution based on the lexical categories of the preceding and following words within a sentence;

Our analysis revealed that including the noun, verb, particle and adverb as category for ‘paṭi’ been identified based on morpho-syntactic functions with different sense in Tamil corpus.

# ‘paṭi’ in the Corpus

When we checked CIIL corpus manually it revealed that ‘paṭi’ has the following:

## 1. paṭi as verbs

1a. paṉi paṭinta malaittoṭarkaḷaik kaṇṭēṉ. ‘be covered with’

1b. eṇṇey taṭavit talaiyaip paṭiya vāriyiruntāḷ. ‘be firmly pressed’

1c. tantaiyiṉ kuṇam avaṉiṭam appaṭiyē paṭintiruntatu. ‘be ingrained’

1d. ivaṉ paṭintu vēlai ceyvāṉ eṉṟa nampikkai eṉakku illai. ‘be submissive’

1e. pēram pēci vilai paṭintāl tāṉ vāṅkuvēṉ. ‘(of price, bargain) be settled’

1f. kaṭitattaip pirittu urakkap paṭittār. ‘read’

1g. itu nāṉ paṭitta kallūri. ‘study (in a college)’

1h. eṉ makaḷ oru mātamākat taiyal paṭikkiṟāḷ. ‘learn (how to do sth)’

1i. nāṉ uṅkaḷuṭaṉ vēlai ceyvatiliruntu palavaṟṟaip paṭittukkoṇṭēṉ. ‘learn (one’s experience)’

## 2. paṭi as Noun

2a. mālai nēramāṉāl paṭiyil uṭkārntu pēcikkoṇṭiruppāḷ. ‘step; staircase’

2b. vāzkkaiyiṉ ovvoru paṭiyilum kaṣṭṭappaṭṭu muṉṉēṟiyavaṉ. ’stage’

2c. paṭippilum aṟivilum uṉṉaiviṭa avar oru paṭi mēltāṉ. ‘a certain degree’

2d. ariciyai aḷakka vēṇṭum; paṭiyai koṇṭuvā. ‘a measure (of above capacity)’

2e. 52, 43, 89 eṉṟa kaṇakkil 5iṉ paṭi 2 ākum. ‘(to the) of (of)’

2f. avarukku tiṉapaṭi kiṭaikkātāl avar vēlaiyai viṭa niṉaittār. ‘allowance paid to an employee in addition to the basic pay’

2g. inta puttakattiṉ ainūṟu paṭikaḷum viṟṟuviṭṭaṉa. ‘copy (of a book, document, etc.)’

2h. anta mūṉṟu ciṟu māṇavikaḷum vācalpaṭiyil mauṉamāka uṭkārntiruntaṉar. step on the doorway.

## 3. paṭi as Particle

3a. avar mītu caṭṭap-paṭi naṭavaṭikkai eṭukkappaṭum. iṉṟaiya nilavarappaṭi taṅkam  vilai kūṭiyirukkiṟatu. ‘Particle used after a noun or pronoun, in the sense of ‘as per’, ‘according to’

3b. nāṉ coṉṉapaṭi cey. (After relative participle) in the sense of ‘as’, ‘in the manner of ’.

3c. nāṉ azuvalaka vēlaiyāka eṅku ceṉṟālum eṉakku payaṇap paṭi koṭuppārkaḷ. travelling allowance (abbreviated to T.A.).

## 4. paṭi as Adverb

4a. avarkaḷ iruvarum vāykkuvantapaṭi oruvarukkoruvar tiṭṭikkoṇṭārkaḷ. ‘without restraint’

# Pați in Trigram

| PrevTag1 | C_Tag | NextTag1 |
|---|---|---|
| QT_QTO | N_NN | JJ |
| N_NN | N_NN | V_VM_VINF |
| QT_QTO | N_NN | N_NST |
| QT_QTO | N_NN | RB |
| QT_QTO | N_NN | N_NN |
| V_VM_VNF | RP_RPD | V_VM_VNF |
| V_VM_VNF | RP_RPD | V_VM_VINF |
| N_NN | RP_RPD | N_NN |
| N_NN | RP_RPD | PR_PRP |
| N_NN | RP_RPD | QT_QTF |
| V_VM_VNF | RP_RPD | PR_PRP |
| V_VM_VNF | RP_RPD | N_NNP |
| N_NN | RP_RPD | V_VM_VF |
| V_VM_VNF | RP_RPD | V_VM_VF |
| N_NN | RP_RPD | RD_PUNC |
| N_NN | RP_RPD | N_NNP |
| QT_QTF | RP_RPD | QT_QTF |

# Treatment of ‘paṭi’ in NLP

We have seen that the lexical item ‘paṭi’ is ambiguous at every level of linguistic analyses.

It has three possible morphological analyses, 5 distinct morpho-syntactic functions that need to be resolved during POS-tagging.

And 20 different senses that should be resolved during semantic analysis.

Therefore, here we make some recommendations on how ‘paṭi’ should be handled at various levels of morpho-syntactic analysis.

# Conclusions

In this paper, we have analyzed the various morpho-syntactic functions of the lexical item ‘pati’ in Tamil, established the interconnections and evolution of these apparently divergent functionalities, and based on this analysis made some recommendations for treatment of ‘pati’ in NLP at different levels of morpho-syntactic analysis.

And also, we build a LDICL-WS for this lexical item ‘pati’ consisting of annotated examples of the different functions of the word and trained learning the algorithms for disambiguation.

# Reference

Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal. “AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages “.2006.

Arulmozhi P, Sobha L and Kumara Shanmugam B. 2004. Parts of Speech Tagger for Tamil, Symposium on Indian Morphology, Phonology & Language Engineering, March 19-21, IIT Kharagpur. :55-57.

D. Cutting, J. Kupiec, J. Pederson, and P. Nipun. “A Practical Part-of-speech Tagger”. In Proceedings of the 3rd Conference of Applied Natural Language
Gim´enez, J. and L.M`arquez. “Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited”. In Proceedings of the Fourth RANLP, 2003.

J. M. Kupiec. “Robust part-of-speech tagging using a hidden markov model”, Computer Speech and Language, 1992.

K. W. Church. “A stochastic parts program and noun phrase parser for unrestricted text”. In Proceedings of the Second Conference on Applied Natural Language Processing (ACL), 1988, pp 136-143.

Lehmann, T. 1989. A Grammar of Modern Tamil, Pondicherry: Pondicherry Institute of Linguistics and Culture.

# Reference cont…

M Anand kumar, V Dhanalakshmi, K P Soman, S Rajendran (2009),“A Novel Apporach For Tamil Morphological Analyzer”, Proceedings of Tamil Internet Conference 2009 , Cologne, Germany, Page no: 23-35, October 2009.

M Anand kumar, V Dhanalakshmi, R U Rekha, K P Soman, S Rajendran (2010), “Morphological Generator for Tamil a new data driven approach”, Proceedings of Tamil Internet Conference 2010, Coimbatore, India, 2010.

Mallikarjun B, et al. (2010) Indian Languages and Part-of-Speech Annotation published by Linguistic Data Consortium for Indian Languages, CIIL, Mysore
Processing, ANLP, 1992, pp 133-140. B. Merialdo. ”Tagging English Text with a Probabilistic Model”. Computational Linguistics, 1994, pp 20(2):155-171.

Priyanka Biswas et al. (2008) A Corpus-based Study of kare in Bangla: Theoretical and Computational Perspectives published in ICON 2008, IIIT Hyderabad, Hyderabad.

Thorsten Brants, “TnT -- A Statistical Part-of - Speech Tagger”, In Proceedings of the 6th Applied NLP Conference, ANLP-2000, April 29 – May 3, 2000.

V Dhanalakshmi, M Anand kumar, K P Soman, S Rajendran (2009), “POS Tagger and Chunker for Tamil language”, Proceedings of Tamil Internet Conference 2009, Cologne, Germany, October 2009

# Any Query ?

Thank you!