



Research Report

on

Lao Parallel Corpus

Ms Kittisack PASOMSOUK, Ms Viengkham CIDA

Information Technology Research Institute

PAN Lao Localization Project

TABLE OF CONTENTS

I. Introduction3
II. Lao and English Parallel Corpus.....6
III. Pos Tagging Implementation.....7
IV. Conclusion.....8
V. References.....8

I. Introduction

- Parallel Text is the text that can translate into one or more languages
- Parallel corpus, is the collection of large Parallel text

II. Lao and English Parallel corpus

Lao - English Parallel have been produced by translated 3110 English sentences into Lao according to main English language, the source is comes from Penn Treebank Corpus and manually translated into Lao.

Penn Treebank Text corpus

.\Penn Tree\Treebank-3\PARSED\MRG\WSJ\00\

.\Penn Tree\Treebank-3\PARSED\MRG\WSJ\01\

III. POS tagging implementation

Lao POS tagging composed 41 tags as you can see in the table bellow:

| Category | Remarks | POS Tag ID No. | POS Name | POS Tag |
|----------|---------|----------------|---|---------|
| Noun | | 1 | Noun: ດິນ ນ້ຳ ລົມ ໄຟ... | N |
| | | 2 | Proper noun: ຄົນ ສັດ ນົກ ປ່າ ທາງ ແມ່ນ້ຳ... | PRN |
| | | 3 | Cardinal number: 1 2 3 4. 100 1000... | CNM |
| | | 4 | Ordinal number: ທີ່1, ທີ່ໜຶ່ງ, ທີ່ສອງ, ທີ່2..... | ONM |
| | | 5 | Label noun: 1 2 3 , a b c ຫມຸ້. ສອງ. .. | LBL |
| | | 6 | Common noun: ໝາ ແມວ, ເຈ້ຍ, ຈອກ.... | CMN |
| | | 7 | Title noun: ທ້າວ, ນາງ, ປັກ, ອີ້, ປ້າລູງ, ແມ່,ພໍ່, ພໍ່ຕູ້, ເອື້ອຍ, ນາຍໝໍ, ອາຈານ, ດອກເຕີ້, ສາສະດາຈານ,ທ່ານ,ພະນະທ່ານ, ຄູບາ, ຍາທ່ານ, ມະຫາ,ຈົວ, ມອ່ມ, ທິດ, ຊຽງ, ພະອົງ | TTL |
| Pro noun | | 8 | Personal pronoun: ຂ້ອຍ, ເຮົາ, ກັນ, ໂຕ, ມັນ, ມີງ, ກູ, ສູ, ຕູ, ຜູ້ຂ້າ, ຕົນເອງ, ຂ້ານ້ອຍ, ຂ້າພະເຈົ້າ, ອ້າຍ, ເອື້ອຍ, ນ້ອງ, ທ່ານ... | PRS |
| | | 9 | Demonstrative pronoun: ນີ້, ນັ້ນ,ພຸ້ນ, ຫນ້, ທີ່ນັ້ນ, ທີ່ນັ້ນ, ນີ້ເດ, ນັ້ນເດ | DMN |
| | | 10 | Interrogative pronoun: ໃຜ, ໃສ,ຫຍັງ, ອັນໃດ, ແນວໃດ, ຢ່າງໃດ, ຜູ້ໃດ | NTR |
| | | 11 | Relative pronoun: ທີ່, ເຊິ່ງ, ຊຶ່ງ, ຜູ້, ອັນທີ່ | REL |
| Verb | | 12 | Verb: ກິນ ຢູ່, ນຸ່ງ ທົ່ມ, ຫົວ ໄຫ້ ເຈັບ ຕາຍ... | V |
| Pre verb | | 13 | Pre verb auxiliary: ເກືອບ, ຈົນ, ຫາກໍ່, ກຳລັງ, ພວມ, ອາດ, ຄົງ, ຄົງຈະ, ອາດ, ອາດຈະ, ຄ່ອຍ, ສູ້, ຫນ້າ, ໄດ້, ຕ້ອງ, ຄວນ, ພໍດີ,ຈົນວ່າ, ຈົນກວ່າ, | PRA |

PAN Localization Project

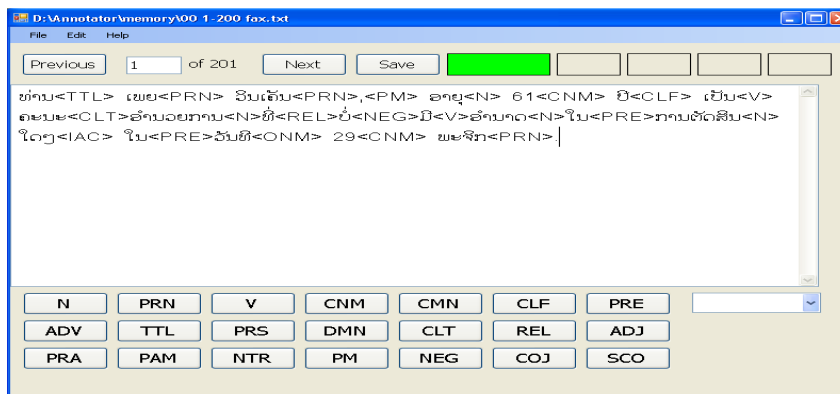
| | | | | |
|------------|--|----|--|-----|
| | | 14 | Pre verb auxiliary in imperative mood: ຢາ, ຫ້າມ, ກະລຸນາ, ຈົ່ງ, ໃຫ້ | PAM |
| | | 15 | Post verb auxiliary: ມາ, ຢູ່, ໄປ. (ເກັບມາກັບຕາ, ຢູ່ມາແລ້ວ, ຢູ່ໄປຢູ່ມາ, ຢູ່ຕາມຢູ່, ເບິ່ງໄປແລ້ວ, ເບິ່ງໄປພັງໄປ...) | PVA |
| Adjective | | 16 | Adjective: ແດງ ດຳ ຂາວ ເຫຼືອງ, ແສດ, ພ້າ, ມູ່ຍ, , ຫຼາຍ, ໜ້ອຍ, ນ້ອຍ, ໃຫຍ່, ຜອມ, ຕູ້ຍ,ພີ, ງາມ,ດີ,ຊົ່ວ, ສາມານ, ໂຫດ,ປາເຖື່ອນ, ງອກງາມ, ເທງົາຫງອຍ, ງຽບເທງົາ, ອ່ອນເພຍ, ເສຍຂັ້ວນ, ຂຶ້ນໃຈ, ຊັກຊ້າ, ຫຼ້າຫຼັງ, ທົດຖອຍ, ອີກກະທຶກ, ໜັກໜ່ວງ, ປອດໄພ, ສຸຂຸມ, ຮອບຄອບ, ເປີດເຜີຍ, ລີ້ລັບ, ສັບສົນ,ນອງເລືອດ, ເດັດຂາດ, ຫ້າວຫັນ, ທໍລະທົດ, ເຕັມໃຈ, ເສຍໃຈ,ຮອບຄອບ, ຫ່າງ, ຖີ່, ຕ່ຳ, ສູງ, ຜ່ອຍ, ຫຍາຍ,ກວ້າງ, ແຄບ, ເປົາ, ຫວານ, ເຄັມ, ຈາງ, ຈິດ, ໂກ, ມໍ່, ສົ້ມ, ຮອມ, ຝືດ, ໂລ່ງ, ໂງ່,ຄ້ານ, ໝັ່ນ, ເຖົ້າ, ແກ່, ຫຸ່ມ, ປູ້, ແຫຼມ, ມີ່ນ, ຊາ, ເຢັນ, ອຸ່ນ, ຄົດ, ຊື່, ແລບ, ຫຍຸ້ງ, ຍາກ, ລຳບາກທີ່ສຸດ, | ADJ |
| Determiner | | 17 | Definiite determiner after noun: ເຫຼົ່ານັ້ນ, ເຫຼົ່ານີ້, ທັງປວງ, ທັງມວນ, ປວງ, ທັງໝົດ, | DAN |
| | | 18 | Definite determiner preceding quantitative : $\frac{2}{4}$ ແຕ່, ພຽງ, ອີກ, ອີກຈັກ, ທັງ, ກວ່າ, ທຸກໆ, ໃຜໆ, ໃສໆ, ບໍ່ວ່າແຕ່, ຫຍັງໆ, ແມ່ນກະທັ້ງ, ຈົນເຖິງຂັ້ນ, ຈົນຮອດ, ປານນີ້, ປານນັ້ນ, ສ່ຳນັ້ນ | DBQ |
| | | 19 | Definite t determiner following quantitative:ເຕັມ, ພໍດີ, ຖ້ວນ. | DAQ |
| | | 20 | Indefinite Determiner following noun: ກໍ່, ຕ່າງໆ, ອື່ນ, ອື່ນໆ, ອື່ນໃດ, ໃດໆ, ໃຜໆ | IAC |
| | | 21 | Indefinite determiner, preceding quantitative:ເກືອບ, ປະມານ, ຮາວ, | IBQ |
| | | 22 | Indefinite determiner after quantitative: ກວ່າ, ປາຍ. | IAQ |
| | | 23 | Determiner cardinal number, before noun: ເດືອນ, ປີ, ຫນຶ່ງ, ສອງ, ສາມ... | CNM |
| Adverb | | 24 | Adverb in normal form:ດັ່ງ, ຄ່ອຍ, ແຮງ, ຍາວ, ໄວ, ຊ້າ, ອ່ອນ, ວິວ, ວາກໆ, ປົກໆ, ບັອກໆ, ຈວາກ, ທົດໆ,ທັດໆ, ຍາບໆ, ຍຸບໆ, ຈົກຈັກ, ຈິດ, ຈິດຈັດ, ເບິ່ງໆ, ແຊວໆ, ງຸບໆ, ແຈ້ບໆ, ກາດ, ຈາດ, ປົກ, ເປັກ, ຈວາດ, ໂດຍບັງເອີນ, ແບບບັງເອີນ, ລຳພັງ | ADV |
| | | 25 | Adverb iterative form: ເລື້ອຍໆ, | ADI |

PAN Localization Project

| | | | | |
|---------------|--|----|--|-----|
| | | | ຕິກງ, ເປັນປະຈຳ, ສະເໝີງ | |
| | | 26 | Adverb in prefixed form: ຢ່າງໄວ, ຢ່າງຊ້າ, ເຕັມທີ່, ຢ່າງສຸດ, ຊ້າສຸດ, ຢ່າງເກັ່ງ | ADP |
| Classifier | | 27 | Unit classifier: ອັນ, ຫົວ, ເຫຼັ້ມ, ໜ່ວຍ, ກ້ານ, ໃບ, ລູກ, ຫາບ, ແບກ, ມັດ, ມື້, ວັນ, ເດືອນ, ອາທິດ, ປີ, ລຳ, ພົ້ວ, ແຊງ, ໝໍ້, ຊາມ, ບ່ວງ, ຕົ້ວ, ກຳ, ຢ່າງ, ເຍື້ອງ, ດວງ, ມາກ, ຄັ້ງ, ບາດ, ເທື່ອ, ຮອບ, ແຫ່ງ | CLF |
| | | 28 | Collective classifier: ຝູງ, ພິງ, ໝູ່, ພວກ, ແຊງ, ພົ້ວ, ຮວດ, ຮ້ອຍ, ຫາບ... | CLT |
| | | 29 | Measurement classifier: ວາ, ຫຼາ, ສອກ, ຄືບ, ນິ້ວ, ກ້າວ, ໂຍບ, ໂຍດ, ມິລິ, ຊັງຕີ, ແມັດ, ກິໂລແມັດ... | MTR |
| | | 30 | Frequency classifier: ຖ້ຽວ, ຮອບ, ຄາບ | FQC |
| | | 31 | Verbal classifier: ກີ້, ມ້ວນ, ມັດ, ຫາບ, ແບກ | VBL |
| Conjunction | | 32 | Coordinating conjunction: ແຕ່, ແຕ່ຫາກ, ຫາກວ່າ | COJ |
| | | 33 | Comparative conjunction: | COP |
| | | 34 | Subordinating conjunction: ຍ້ອນ, ເພາະ, ເພາະວ່າ, ສະນັ້ນ, ດັ່ງນັ້ນ, ເຫດນັ້ນ, ຫາກແມ່ນ, ແມ່ນວ່າ, ຫາກວ່າ, ເຖິງຢ່າງນັ້ນ, ກໍ່ຕາມ, ຍ້ອນແນວນັ້ນ | SCO |
| Preposition | | 35 | Preposition: ເທິງ, ກ້ອງ, ລຸ່ມ, ໃຕ້, ເໜືອ, ພາຍໃຕ້, ພາຍໃນ, ໃນ, ຈາກ, ນັບແຕ່, ແຕ່ນັ້ນ, ເມື່ອນັ້ນ, ແຕ່ກ່ອນ, ເມື່ອກ່ອນ, ພາຍລຸນ, ຕາມຫຼັງ, ຂ້າງ, ຄູ່ຂ້າງ, ໃກ້, ໄກ, ມໍ່, ຕໍ່ໜ້າ, ແຈ, ຈອມ, ຍອດ, ຕີນ, ກິ້ນ, ແຄມ, ອ້ອມຮອບ, ແຖວ, ແຖບ, ບໍລິເວນ, ລະຫວ່າງ, ຕັ້ງແຕ່, ນັບແຕ່, ຕົ້ນ, ປາຍ, ຫ້າຍ | PRE |
| Interjection | | 36 | Interjection: ໂອຍ, ອຸຍ, ໂຫ, ບັດໂຫ, ໂຫ, ໂອ້ຍ, ເອີ, ເອະ | INT |
| Prefix | | 37 | Normal prefix: ການ, ຄວາມ | FIX |
| Final article | | 38 | Ending fore affirmative sentence: ເຈົ້າ, ໂດຍ, ໄດ້, ໄດ້ແທ້, | EAF |
| | | 39 | Ending fore interrogative sentence: ບໍ່, ຫວ່າ, ເຈົ້າ, ຊັ້ນຫວ່າ. | EIT |
| Negator | | 40 | Negator: ບໍ່, ບໍ່ໄດ້, | NEG |

Lao Part of speech in detail [1]

- POS tagging tool



IV. Conclusion

In tradition grammar was divided part of speech of Lao into 8 categories as: noun, pronoun (proper noun), verb, adjective, adverb, conjunction, and preposition. It was followed of the grammar for part of speech dividing of France in early 20 century, up to noun it still used to the teaching Lao grammar in high schools and for foreign language learning, In recent part of speech of Lao have been distinguished into details by using of the words function kind, position and their relation with neighbor. Now we have divided Lao Part of speech into 41 tag set and have been used in Lao-English Parallel corpus.

V. References

[1] www.PANL10n.net/wiki/PartOfSpeech