Burmese Grammar Model

Than Lwin Aung

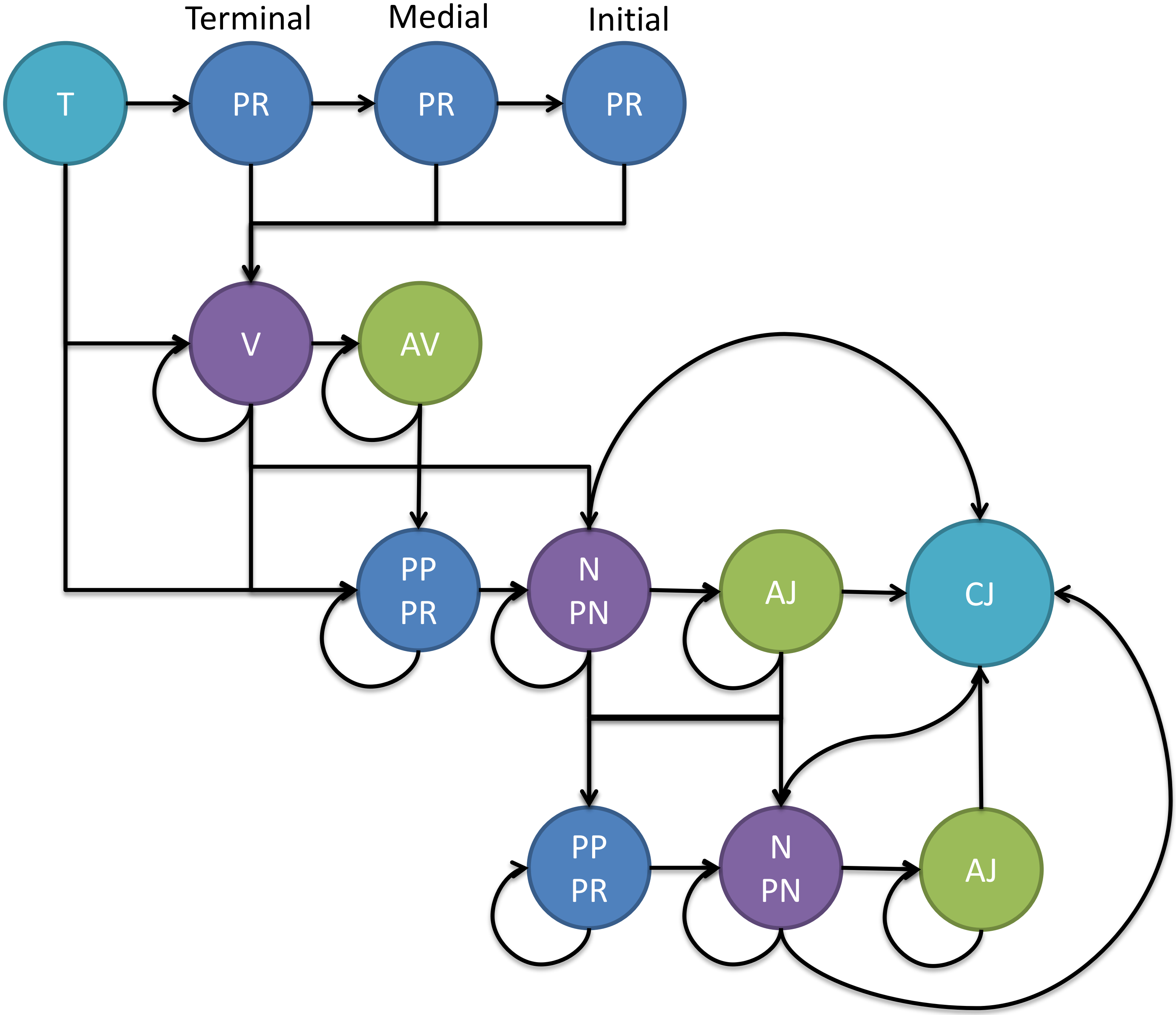
Word Classes

- N - Noun
- PN - Pronoun
- AJ - Adjective
- AV - Adverb
- V - Verb
- PP - Preposition
- PR - Particle
- CJ – Conjunction
- SJ – Sentence Conjunction
- IJ - Interjection
- T – Terminal

Basic Sentence Model (**SOV**)

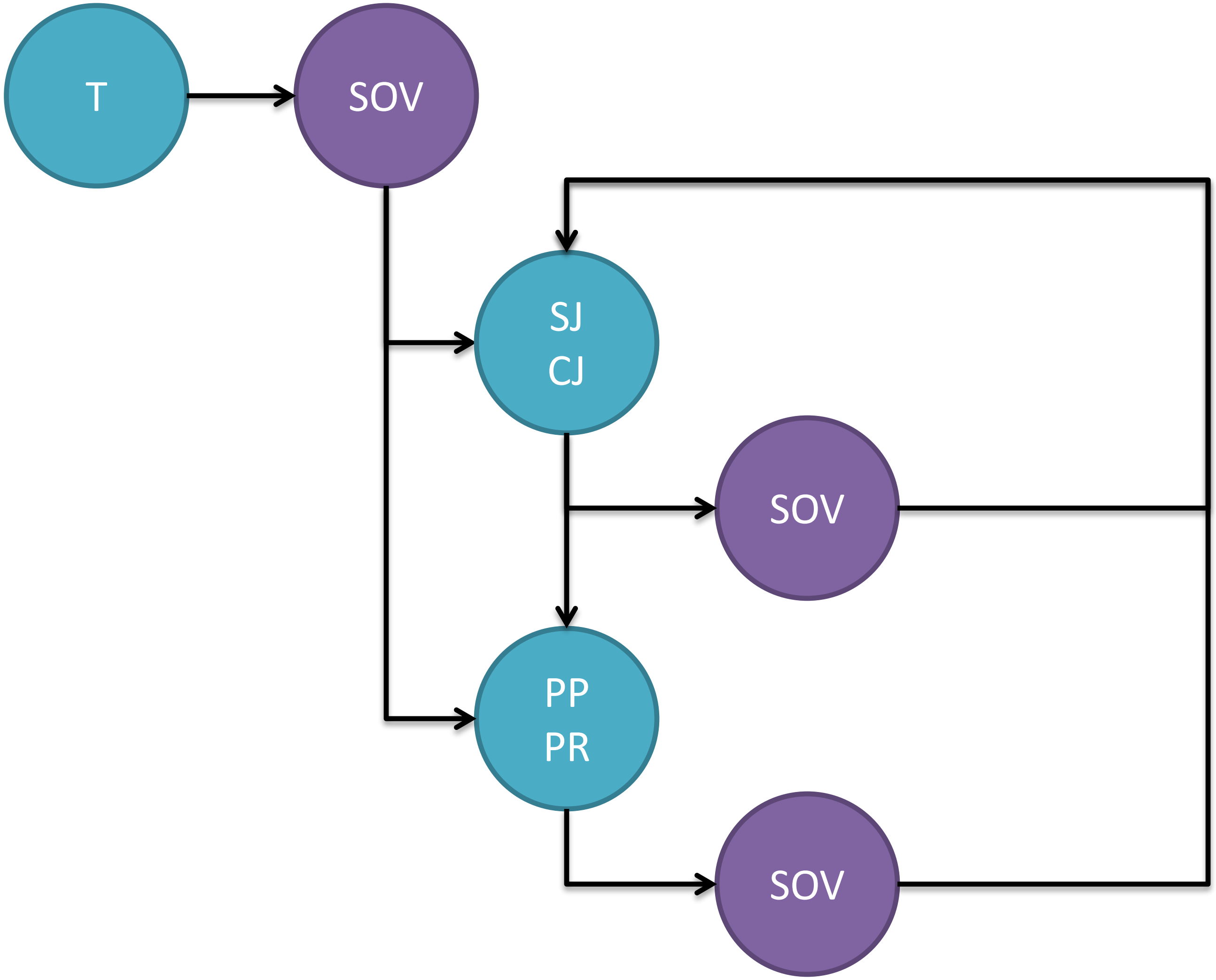
SOV Model is based on the assumption that there is only one main verb in a simple sentence.

SOV Model is the sequence model, which will accept and process lexicons in reverse.


Terminal
Medial
Initial
T
PR
PR
PR
V
AV
PP
PR
N
PN
AJ
CJ
PP
PR
N
PN
AJ

Recursive Complex Sentence (**RCS**)

Complex Recursive Sentence is formed by recursively connecting two or more SOVs with Particle, Preposition, Sentence Conjunction or Conjunction.


T
SOV
SJ
CJ
SOV
PP
PR
SOV

Production Rules

- T + [V]^n
- T + [V]^n + AV
- T + [PR]^n + [V]^n
- T + [PR]^n + [V]^n + AV
- T + PR + N
- T + PP + N
- T + PR + N + AJ
- T + PP + N + AJ
- T + PR + PN
- T + PP + PN
- T + PR + PN + AJ
- T + PP + PN + AJ

- T + [V]^n + N
- T + [V]^n + AV + N
- T + [PR]^n + [V]^n + N
- T + [PR]^n + [V]^n + AV +N
- T + [V]^n + PP + N
- T + [V]^n + AV + PP + N
- T + [PR]^n + [V]^n + PP +N
- T + [PR]^n + [V]^n + AV + PP +N
- T + [V]^n + PR + N
- T + [V]^n + AV + PR + N
- T + [PR]^n + [V]^n + PR +N
- T + [PR]^n + [V]^n + AV + PR +N

ရန်ကုန်မြို့ပြင်ပတ်လမ်းစီမံကိန်းကို လှည်းကူး၊ ဒဂုံမြို့သစ်(အရှေ့)၊ ဒဂုံမြို့သစ် (ဆိပ်ကမ်း)၊
သန်လျင်နှင့် ကျောက်တန်း စသည့်မြို့နယ်ငါးခုကို ဖြတ်သန်းတည်ဆောက်မည်ဖြစ်ပြီး အများ
ပြည်သူသုံးသပ်အကြံပြုနိုင်ရန် စီမံကိန်းဆိုင်ရာ ပတ်ဝန်းကျင်ထိခိုက်မှု ဆန်းစစ်ခြင်း
အစီရင်ခံစာကို ဆောက်လုပ်ရေးဝန်ကြီးဌာနက ထုတ်ပြန်လိုက်သည်။

။ T သည် PR လိုက် PR ထုတ်ပြန် V က PP ဝန်ကြီးဌာန N ဆောက်လုပ်ရေး N ကို PP

အစီရင်ခံစာ N ခြင်း ဆန်းစစ် N မှု ထိခိုက် N ပတ်ဝန်းကျင် N ဆိုင်ရာ N စီမံကိန်း N

ရန် SJ နိုင် PR အကြံပြု V သုံးသပ် V အများပြည်သူ N ပြီး SJ ဖြစ် V မည် PR

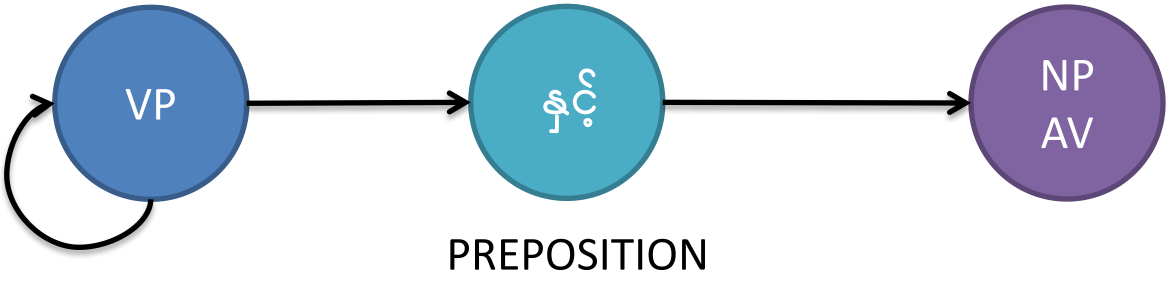
တည်ဆောက် V ဖြတ်သန်း V ကို PP ခု ငါး မြို့နယ် N စသည့် PP ကျောက်တန်း

နှင့် CJ သန်လျင် N ၊) ဆိပ်ကမ်း N (မြို့သစ် N ဒဂုံ N ၊)

အရှေ့ N (မြို့သစ် N ဂုံ ဒ ၊ လှည်းကူး N ကို PP စီမံကိန်း N

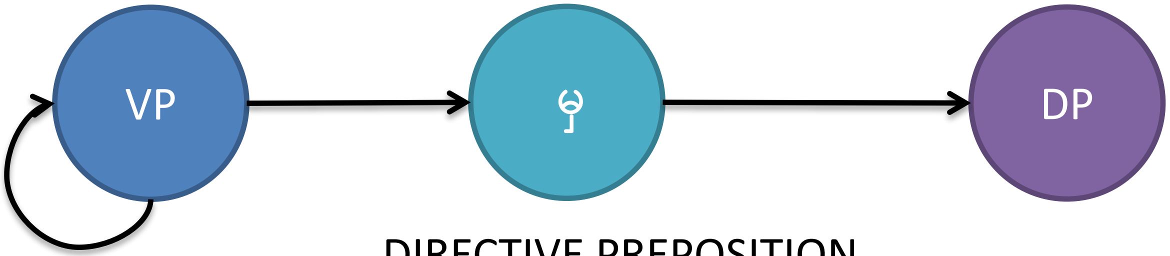
ပတ်လမ်း N မြို့ပြင် N ရန်ကုန် N

နှင့်- Ambiguity



နှင့် can be either preposition or conjunction, depending on its location.

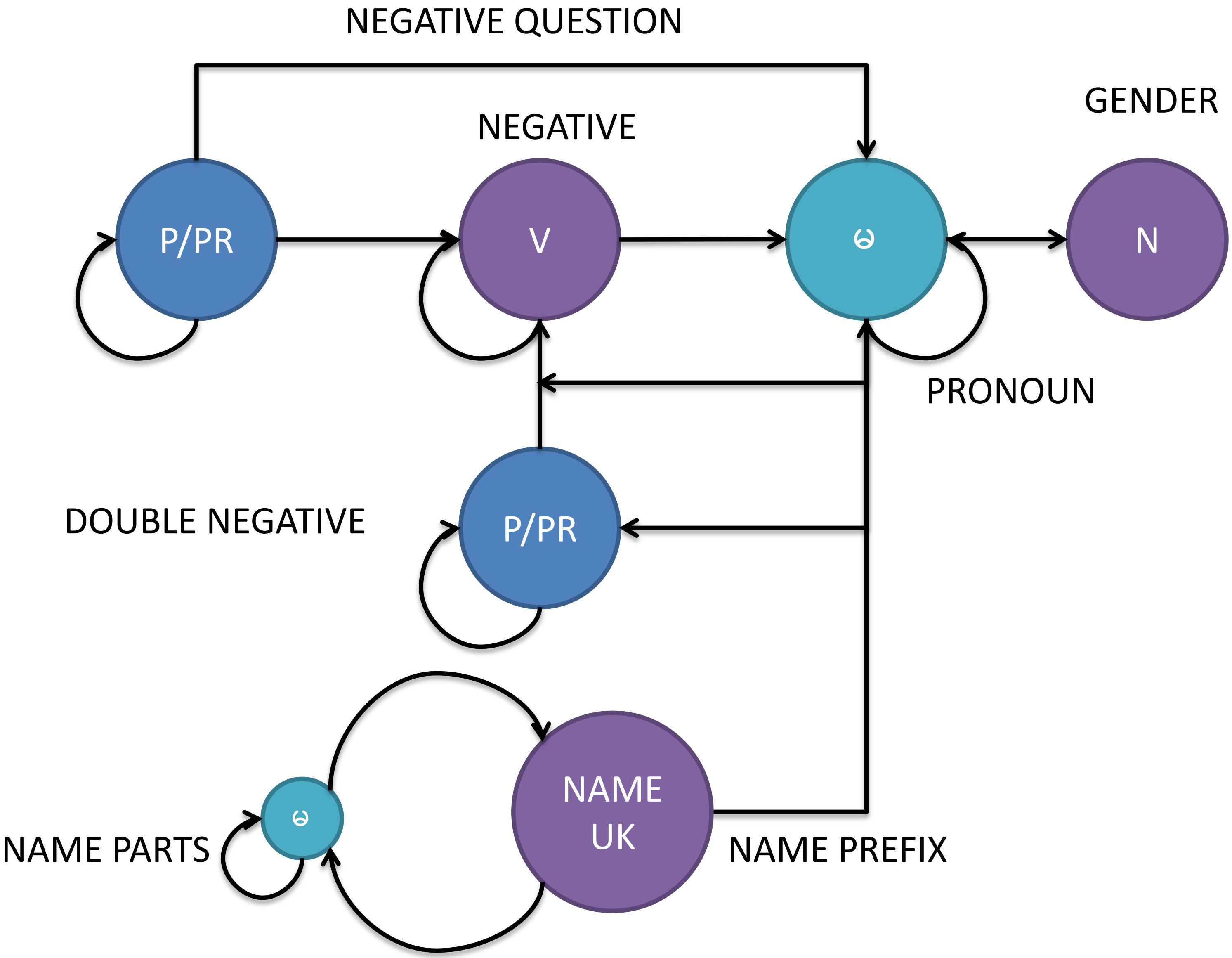
ꯗ- Ambiguity



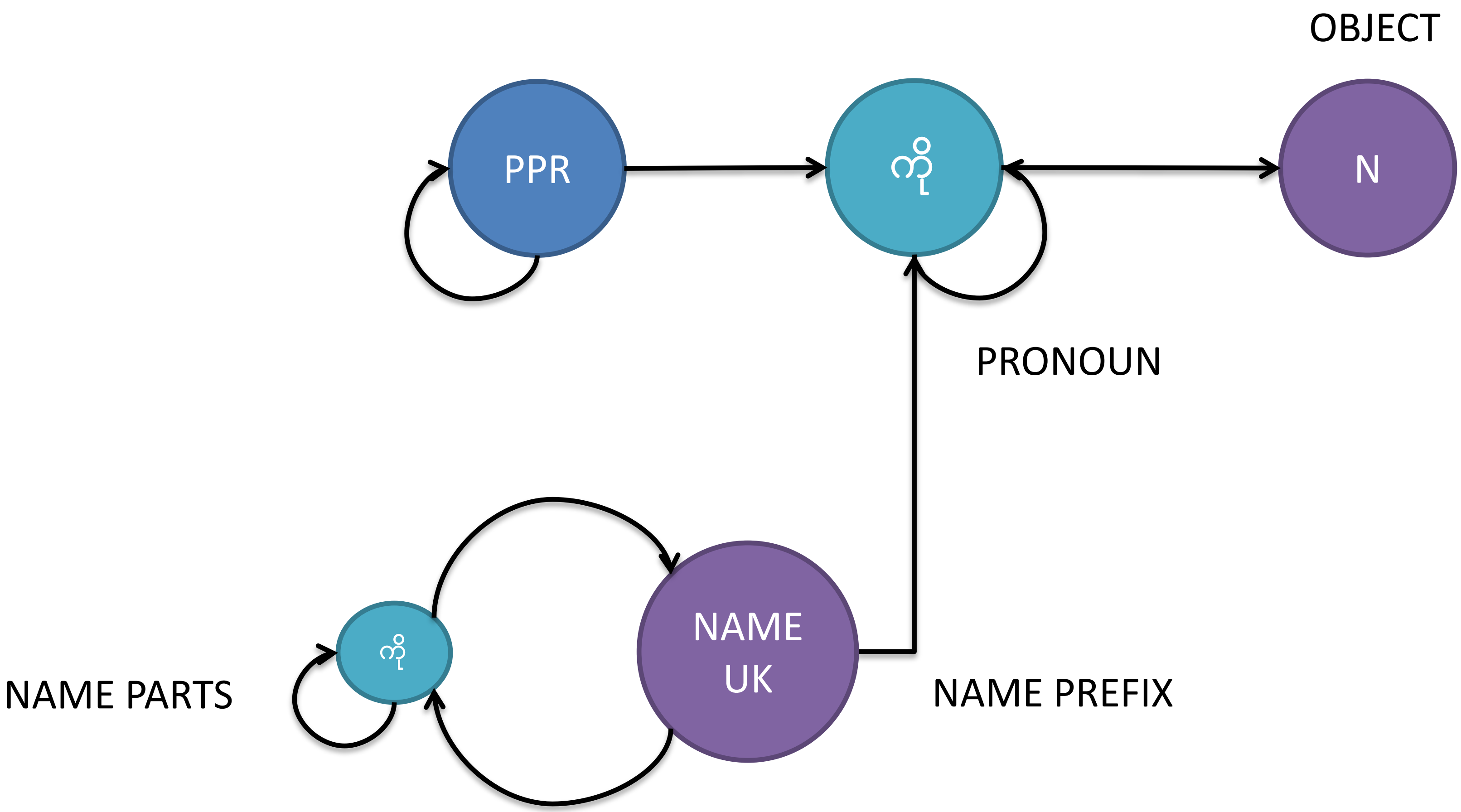
DIRECTIVE PREPOSITION

ꯗ can be either directive preposition or locative preposition, depending on its location.

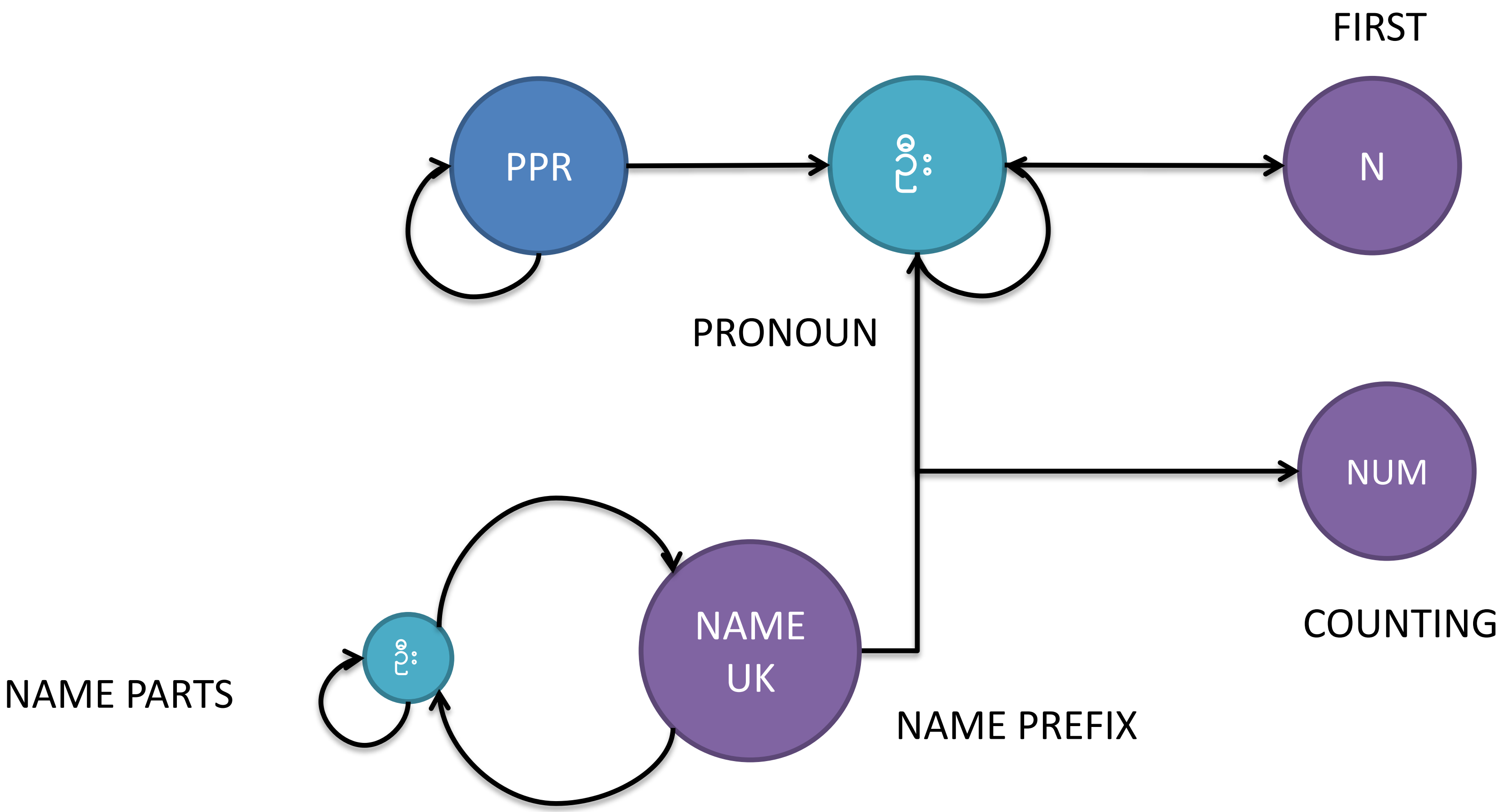
မ- Ambiguity


NEGATIVE QUESTION
NEGATIVE
GENDER
P/PR
V
မ
N
PRONOUN
DOUBLE NEGATIVE
P/PR
NAME PARTS
မ
NAME
UK
NAME PREFIX

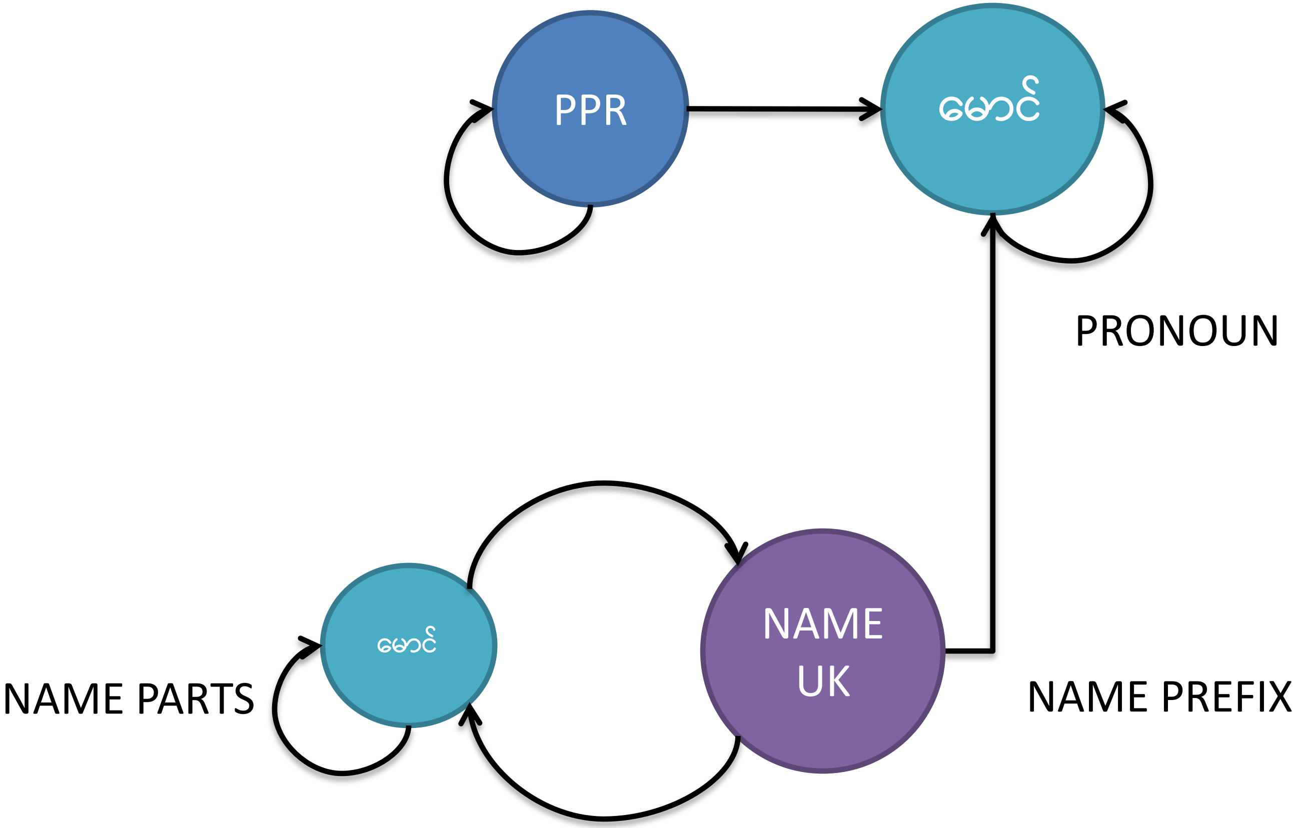
ကို- Ambiguity



ဦး- Ambiguity


FIRST
PPR
ဦး
N
PRONOUN
NUM
COUNTING
NAME PARTS
ဦး
NAME
UK
NAME PREFIX

မောင် - Ambiguity



PPR

မောင်

PRONOUN

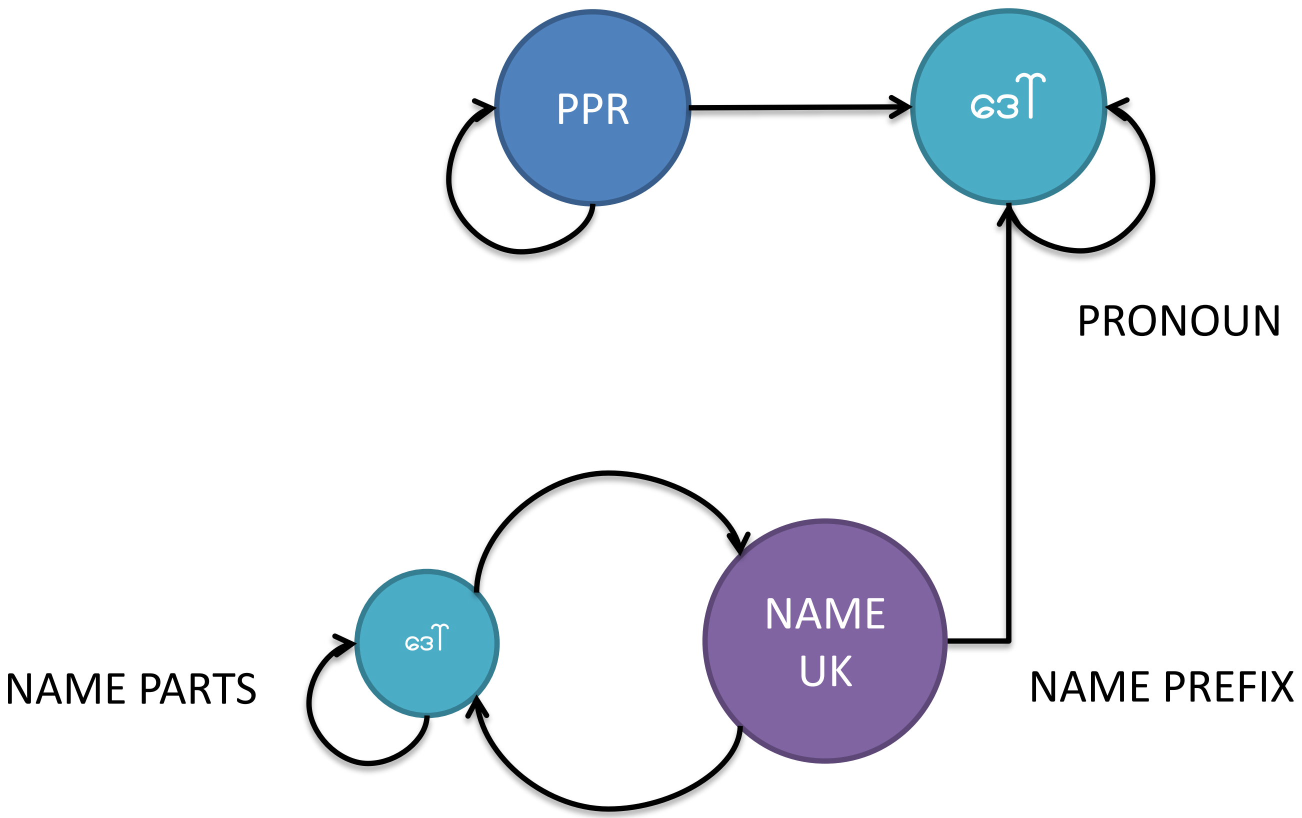
မောင်

NAME UK

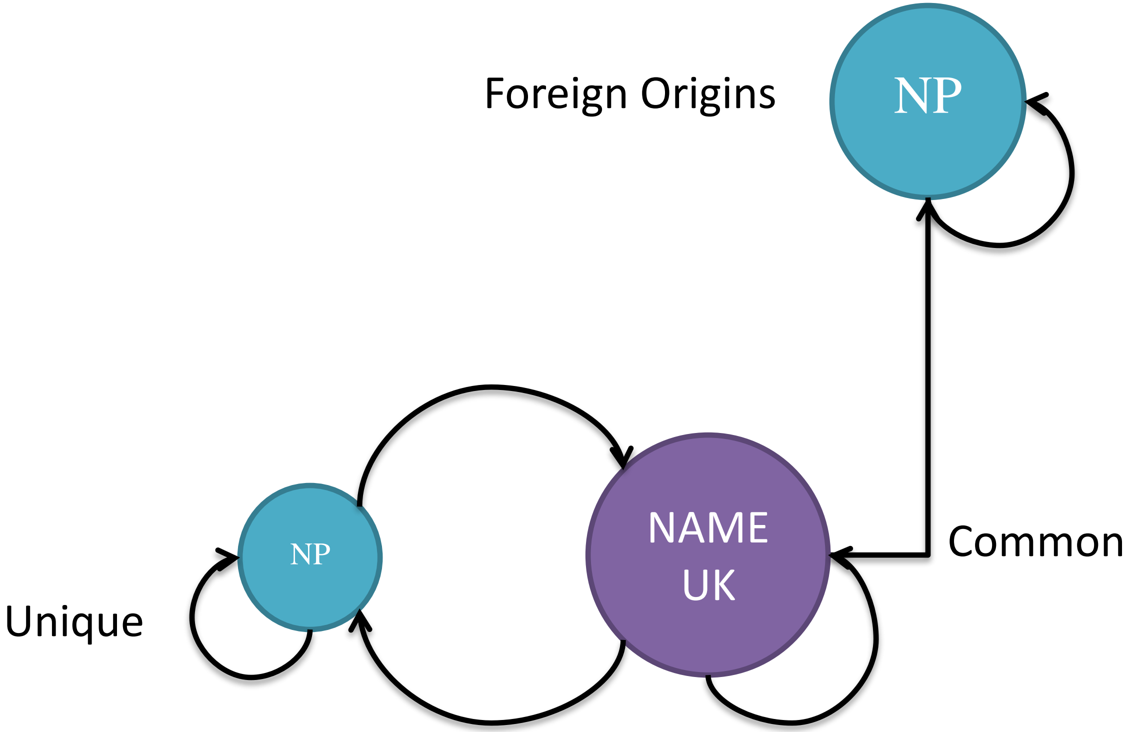
NAME PARTS

NAME PREFIX

ഡോ- Ambiguity



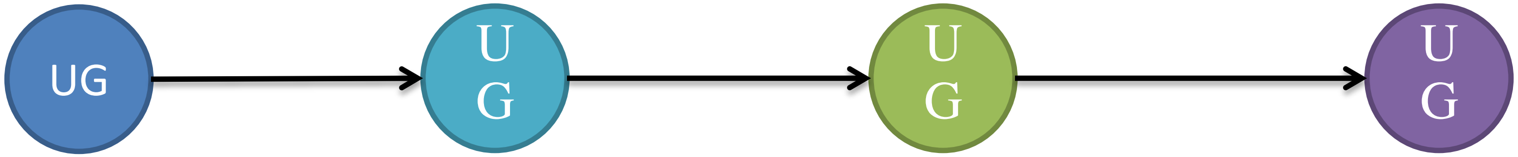
Burmese Nomenclature



Burmese Naming System (Nomenclature) has common name parts, and as a rule of thumb, most Burmese Names are linear combination of common name parts.

However, there are some unique names which are name after something special.

Foreign Words



ဘွန်+ ဒစ် + လီ + ဂါ

All Unigrams are non-repeating, and they are all unknown unigrams.

လေ + ဗာ + ကူ + ဆင်

All Unigrams are non-repeating, and some of them are unknown unigrams.

ဟာ + သာ + ဘာ + လင်

All Unigrams are non-repeating, and all of them are known unigrams.

N-Gram Rules

1. Unigrams cannot repeat more than 2 times
2. Bigrams cannot repeat more than 2 times
3. Bigrams, which are made up of 2 repeating unigrams, cannot repeat more than 1 time.
4. 2 different bigrams can juxtapose next to each other.
5. Trigrams, which are made up of different 3 non-repeating syllables, are generally the upper limit in Myanmar Native Words.
6. N-Grams, which have more than 3 syllables, are either Compound Words or Foreign Adoptions which have foreign origins.

Tokenizing

- Tokenizing is a process, which identifies individual words from a sentence or statement.

- In a language, such as Burmese, where there is no distinct word boundary, it is quite a challenging task.

- In Burmese, the atomic units of language are syllables – not words.

- Therefore, Burmese Tokenizing starts with 2 fundamental processes: Syllable Breaking and Word Segmentation.

Syllable Breaking

- Syllable Breaking is LTR (Left To Right) process, which reads each alphabet from the left, and combines them into basic syllable.

- Syllable Breaking is rule-based approach.

- We used **Ko Zin Maung Maung & Yoshiki Mikami** method and further enhanced it to deal with Foreign Words, Numbers and different Symbols and Punctuation.

Word Segmentation

- After carefully researching on different approaches on Word Segmentation, we have developed our Non-Ambiguous Shortest Match Algorithm.

- Many have suggested to approach with Longest Match Algorithm with Dictionary-Based Method.

- However, LM (Longest Match) has given rise to some ambiguity, and it is harder to resolve the intersection of 2 ambiguous words.

သဘာဝဟာသဘာဝ

မမကောင်းဘူး

- Therefore, NASM (Non-Ambiguous Shortest Match) is implemented to avoid ambiguity issues.

Word Class Identifications

- NASM results in Raw Words, which will be further processed by tagging with Word Classes (POS).

- In addition, some ambiguity will be resolved with Number, Nomenclature rules.

- Although many major problems are solved after Word Class Identifications, some ambiguity will still exist. The primary reason is many words can have many different meanings and classes according to the context.

- In order to solve this, Chunking is necessary to make it more clear.

ဆရာကဆရာလုပ်တာမမှားပါဘူး။

ရချင်တာကိုရဖို့ရသလိုလုပ်ရလိမ့်မယ်။

မမမနိုင်လို့ကိုကိုကိုမခိုင်းတာ။

Chunking

- Chunking is a process which takes a sequence of classified words and put them together to form a phrase.
- In Language Processing, phrases are more useful than individual words.
- If an unknown word is encountered, it is possible to guess the meaning of the word by its surroundings.
- To put it simply, a phrase can give the same meaning even if there are some unknown words in a phrase.
- Our Chunking Algorithm is a RTL (Right To Left) processing based on SOV Model.

ဆရာက(Noun Phrase) ဆရာလုပ်တာ (Verb Phrase) မမှားပါဘူး(Verb Phrase) ။

Foreign and Native Nomenclature

- One of the major issues in SOV Chunking is identification of Foreign and Native Nomenclature.

- Nomenclature, basically, is a systematic approach to give “Names” to Proper Nouns.

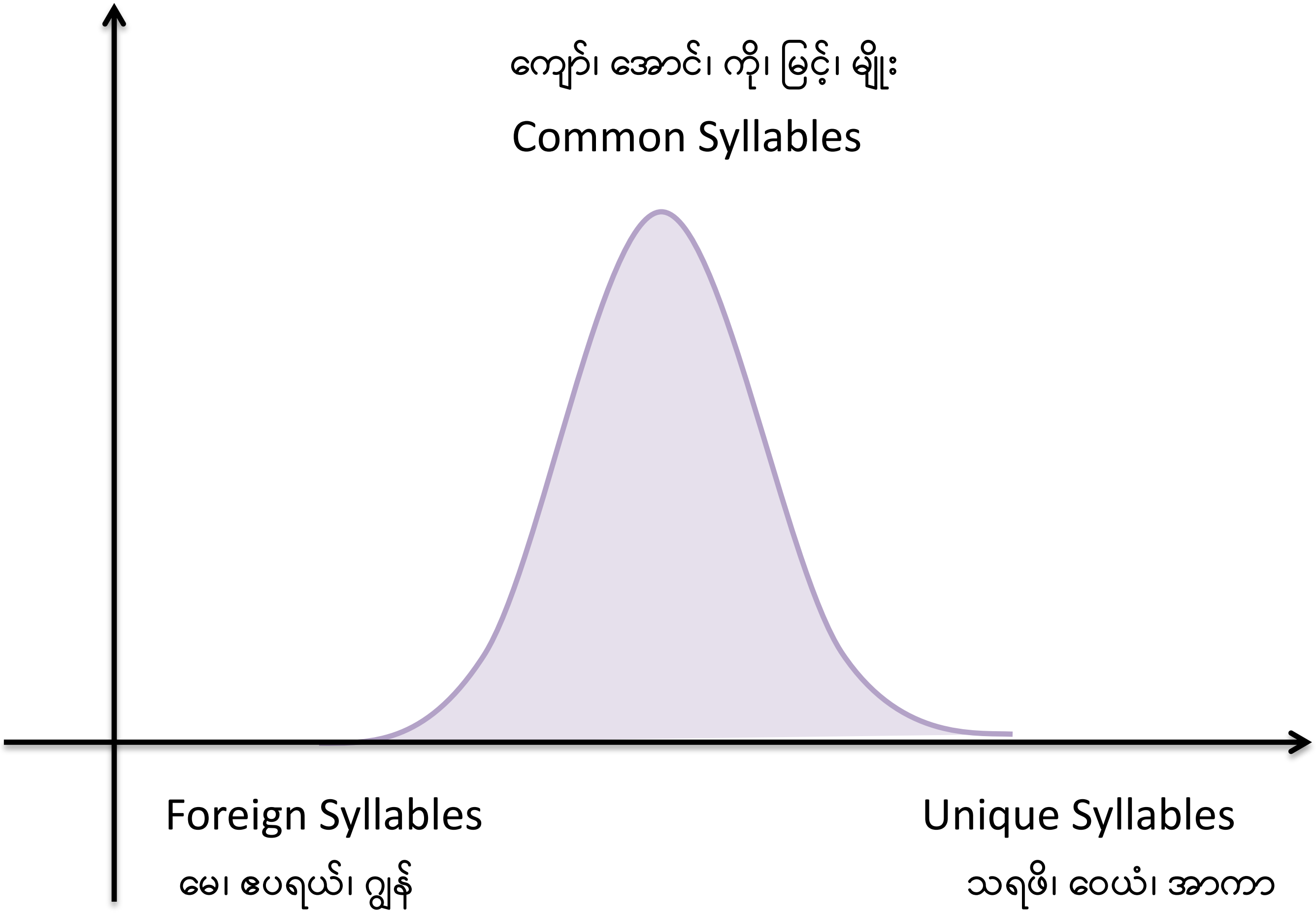
- In English, all Proper Nouns are capitalized, which makes them easier to identify.

- However, in Burmese, there is no such thing as Capitalization.

- To make things worse, Burmese Adaptation of Proper Nouns are literally no difference from Gibberish – words which has no proper meanings.

- Fortunately, while Proper Nouns from Foreign Origins are harder to identify, Proper Nouns from Native Origins at least follow as some standard rules.

Native Nomenclature Distributions



ကျော်၊ အောင်၊ ကို၊ မြင့်၊ မျိုး

Common Syllables

Foreign Syllables

မေ၊ ဧပရယ်၊ ဂျွန်

Unique Syllables

သရဖိ၊ ဝေယံ၊ အာကာ

Foreign Nomenclature

- To identify Foreign Nomenclature is one of the hardest problems to solve in Chunking.
- First of all, there is no specific rules to identify Foreign Origin Words, which will come from different languages: English, Korean, Japanese, French etc.
- In addition, it will be unwise to include all the Foreign Words into a dictionary, which will result in millions of Words.
- Currently, Identifying Foreign Nomenclature is a roughly a Post-Processing Method.
- We are currently looking for a better method to solve this issue.
- Although Machine Learning can be used, we would like to avoid the storage of millions of words, which will not only increase the search space, but also increase the processing time.