Computational Linguistics

Morphology

Parameswari Krishnamurthy

Language Technologies Research Centre
IIIT-Hyderabad

param.krishna@iiit.ac.in

Morphological Analysis

- Analyzing the structure and form of words.

Morphological Analysis

- Analyzing the structure and form of words.
- Breaking down words into morphemes.

Morphological Analysis

- Analyzing the structure and form of words.
- Breaking down words into morphemes.
- Understanding prefixes, suffixes, and inflections.

Morphological Analysis

- Analyzing the structure and form of words.
- Breaking down words into morphemes.
- Understanding prefixes, suffixes, and inflections.
- Important in languages with rich morphology.

Morphological Analysis

- Analyzing the structure and form of words.
- Breaking down words into morphemes.
- Understanding prefixes, suffixes, and inflections.
- Important in languages with rich morphology.
- Utilized in natural language processing tasks.

Morphological Analysis

- Analyzing the structure and form of words.
- Breaking down words into morphemes.
- Understanding prefixes, suffixes, and inflections.
- Important in languages with rich morphology.
- Utilized in natural language processing tasks.
- Crucial for understanding complex word forms.

Complex Morpheme Segmentation in Some Languages

- Some languages require complex morpheme segmentation.

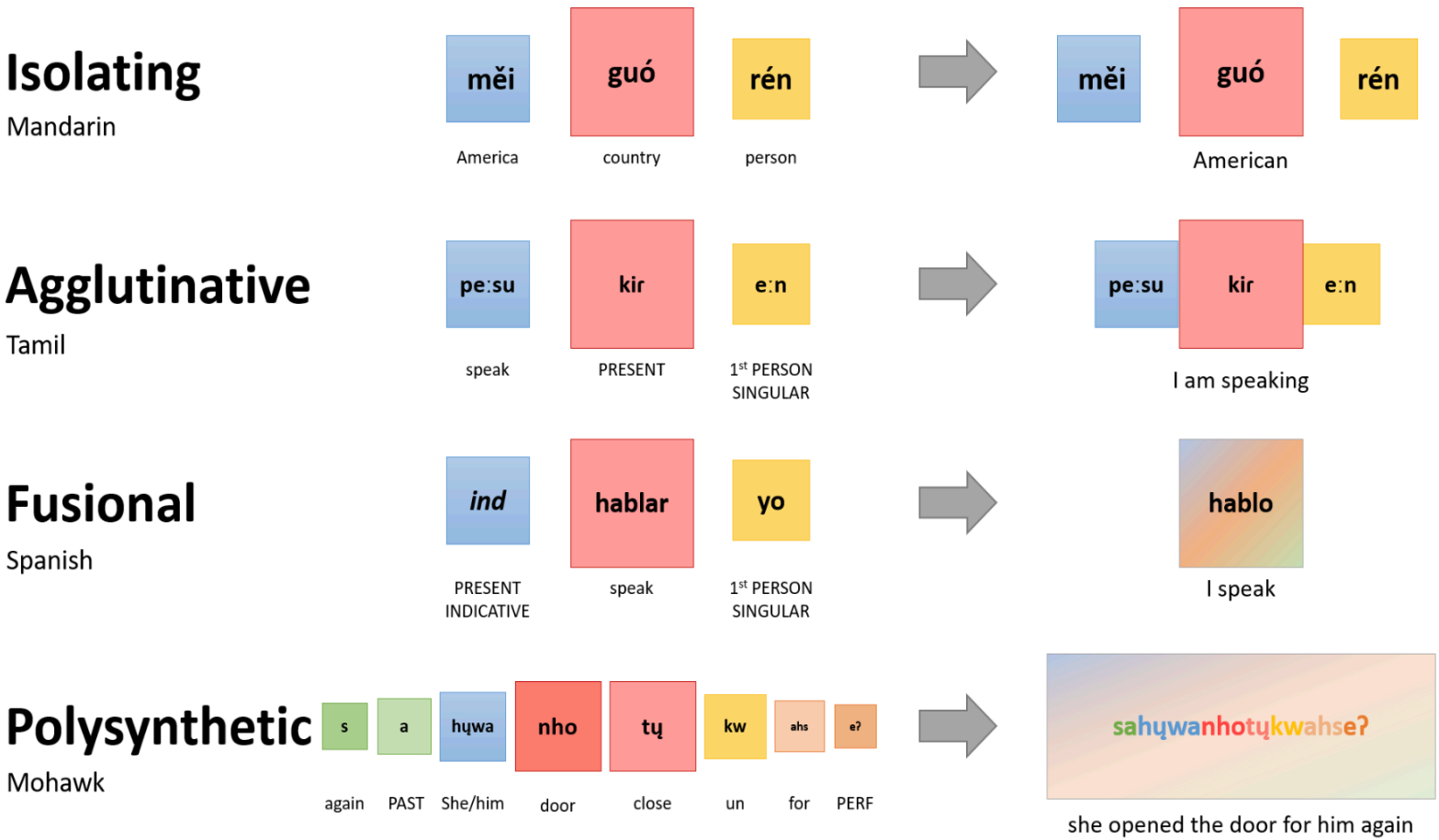
Turkish:

- Uygarlastiramadiklarimizdanmissinizcasina
  '(behaving) as if you are among those whom we could not civilize'
- Uygar 'civilized' + las 'become' + tir 'cause' + ama 'not able' + dik 'past' + lar 'plural' + imiz '1pl' + dan 'ablative' + mis 'past' + siniz '2pl' + casina 'as if'

Example for derivation from Telugu:

pagalagottiMcipettamananivvadalacukooleekapootunnaanu.
pagulu+a-kottu+iMcu+i-pettu+a-manu+a-ivvu+a-daluvu+i-konu+a-leeka-poo+tunn+1,sg,any
break+inf-strike+cause+cpm-benefactive+inf-tell+inf-permit+inf-think+cpm-reflexive+inf-neg+go+prog+1, sg
'I could not think to permit someone to tell for my sake to break something'
(pc, G. Uma Maheshwar Rao)

Morphological Typology



Introduction

Introduction

What is Computational Morphology?

- Computational morphology deals with developing techniques and theories for computational analysis and synthesis of word forms.

What do you need to understand?

- Theoretical knowledge of morphology of languages
- Computational techniques for implementation

Where is the application?

- Hyphenation, Spell Checking, Stemmers etc.
- Machine Translation, QA system, Content Analysis, Speech Synthesis etc.

What is Morphology?

What is Morphology?

Two dominant views of Morphology:-

Morphology is the study of

- the mental system involved in word formation
- words, their internal structure, and their formal relationships.

Its etymology is Greek:
morph- means 'shape, form'
morphology is the study of form or forms.
The word 'morphologie' was first used by August Schleicher in 1859.

The earliest and the first morphological analysis of a human language: *Ashtadhyayi* by Panini.

Concepts of Morphology

- Null Hypothesis: Morphological processing can be undesirable since every word in a language may be stored and accessed as and when required.
- Continuously new words are integrated while others are drifting out of use.
- However, in any human language
  - possible words are infinite in number!
  - actual and attested words are also unmanageably large in number.
- Hence, it is necessary to formulate *Morphological rules or Word Formation Strategies* to permit us to recognize or produce new words.

- Native speakers create new words from the existing ones or borrow from other languages as and when necessary.

- The discovery of these mechanisms and the intuitive knowledge underlying this creativity is what is usually known as morphology.

- Speakers possess intuitive knowledge about:
  - words are related to each other partially in the form and meaning

  eg. walk, walks, walked, walking, walker, walkathon etc.

- Native speaker's ability to derive or relate the words in terms of their form and meaning.
  active, activity, activate, activator and activation

- Alternatively native speaker' s ability to reject *cat-en, *cat-z, *cat-iz, (for cats), walk –*walken; drive –*drived; read –*readed, *readen; active –*activement, *activance, and *activant as illformed is because of the knowledge of morphology.

There are two basic divisions in morphology :

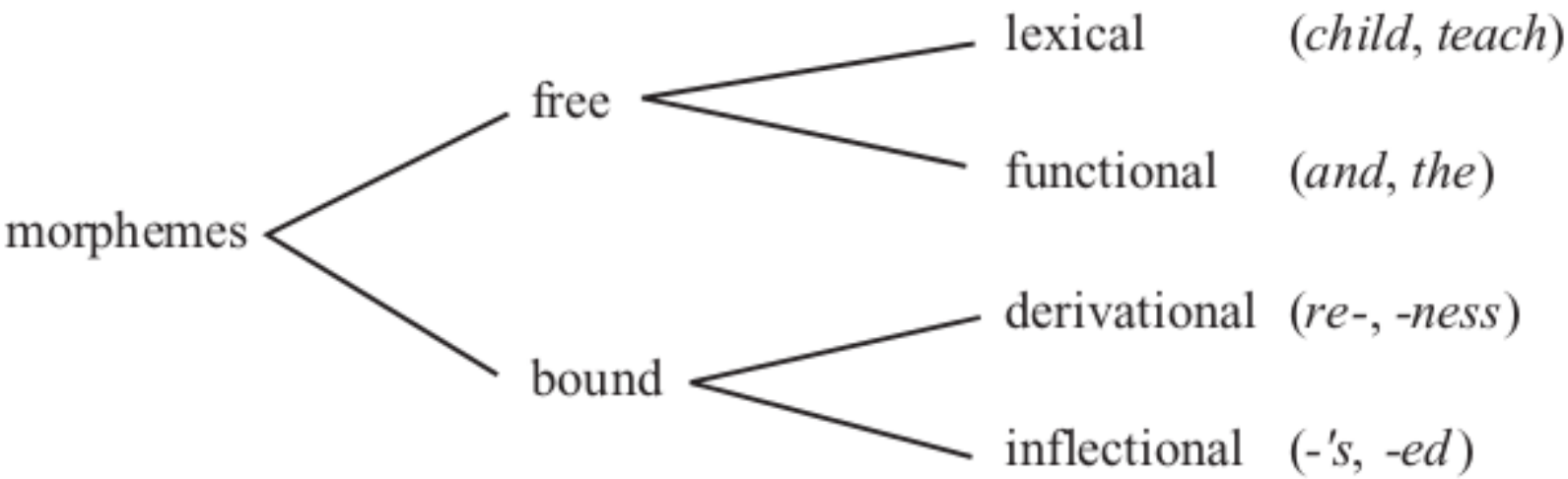
(1) inflectional morphology (conjugation/declension)
(2) lexical morphology (word formation)

Building Blocks of Morphology

Building Blocks of Morphology

Morpheme: the smallest meaningful linguistic unit. Some morphemes are identical with words, but many morphemes are smaller than words.

Morpheme $\leq$ Word



Free morpheme: a morpheme that can stand alone, that is a complete word; an independent morpheme.
eg. walk, book, but, of and etc.,

Lexical morpheme: a morpheme that denotes the content words. They receive inflection.

- Open class categories such as Nouns, Verbs and Adjectives.
- Closed class categories such as Pronouns, Number words and Nouns of space and time (NST).

Functional morpheme: a morpheme that denotes the functional words. They do NOT receive inflection and are indeclinables or *avyayas*.

- Categories such as Prepositions/Postpositions, Conjunctions, Interjections, Adverbs, Demonstratives, Intensifiers, Quotatives etc,.

Bound morpheme: a morpheme that cannot stand alone, but must be attached to something else.
eg. -ed 'past tense marker',
-s 'plural marker',
-er 'comparative maker'

Inflectional Morpheme: creates new forms of the same word with the addition of grammatical properties; the basic meaning (and the category) of the word is the same.

Inflection in English:
a. with nouns: book, books, book's, books'
b. with verbs: ride, rides, rode , ridden, riding
c. with adjectives: old, older, oldest
d. with numbers: seven, seventh

Derivational morpheme: creates a new word with a different meaning that may belong to a different or to the same grammatical category.
RE + WRITE = rewrite "write again", verb
WRITE + ER = writer "one who writes", noun

Derivation in English:
a. Verb to Noun: kill ⇒ killer
b. Noun to Verb: glory ⇒ glorify
c. Adjective to Noun: dark ⇒ darkness
d. Noun to Adjective: person ⇒ personal
e. Adjective to Verb: modern ⇒ modernize
f. Verb to Adjective: walk ⇒ walkable
g. Adjective to adverb: great ⇒ greatly

Affix

Affixes:
An affix is a bound morpheme that is attached to stem to form a word.

Affixes may be derivational
eg. -ish boy = boy-ish
-less care = care-less

Affixes may be inflectional
eg. -s book = book-s
-ed work = work-ed

Concatenative phenomena

Prefix:
a bound morpheme added before a root/stem, or at the beginning of a word.
eg. un- as in undo

Schema: prefix-root/stem

Suffix:
a bound morpheme added after a stem, or at the end of a word.
eg. -ing as in look-ing

Schema: root/stem-suffix

Non-concatenative phenomena

Infix:
an affix added within a single morpheme, i.e appears within a stem.
eg. Philippines
bili 'buy' is a root
bumili 'bought' is an example for infix and -um- is a past tense marker.

Circumfix:
A circumfix is an affix made up of two separate parts which surround and attach to a root or stem.

In Dutch, ge___te is a plural suffix
berg 'mountain' gebergte 'mountains'
vogel 'bird' gevogelte 'birds'
raam 'frame' geraamte 'frames'

Non-concatenative phenomena

Semitic languages exhibit a very peculiar type of morphology, often called *root-template morphology.*
Eg. Arabic root "ktb" produces the following wordforms:

Template	a (active)	ui (passive)	
CVCVC	katab	kutib	'write'
CVCCVC	kattab	kuttib	'cause to write'
CVVCVC	ka:tab	ku:tib	'correspond'
tVCVVCVC	taka:tab	tuku:tib	'write each other'
nCVVCVC	nka:tab	nku:tib	'subscribe'
CtVCVC	ktatab	ktutib	'write'
stVCCVC	staktab	stukib	'dictate'

Exercise:

buyers = buy-er-s (3)
walk = walk (1)
winter = winter (1)
establish = establish (1)
establishment = establish-ment (2)
establishmentary = establish-ment-ary (3)
establishmentarian = establish-ment-ari-an (4)
establishmentarianism = establish-ment-ari-an-ism (5)
antiestablishmentarianism = anti-establish-ment-ari-an-ism (6)
antidisestablishmentarianism = anti-dis-establish-ment-ari-an-ism (7)

Allomorph

- The physical variation of a morpheme is called as allomorphs.
- A morpheme may display allomorphy, i.e. have more than one form.
- The allomophs are physically different forms, but they indicate same meaning of a morpheme.

Types of allomorph:

- Phonologically Conditioned Allomorph
- Lexically Conditioned Allomorph
- Suppletive

Allomorph

Phonologically Conditioned Allomorph

The variants in the pronunciation of the plural are phonologically conditioned allomorphs, because the choice depends only on the phonological characteristics of the element to which it attaches.

The English plural morpheme has three allomorphs:
/s/ bits, tips, tacks,
/z/ dogs, slabs
/iz/ ladies, bodies
Here, [s],[z],[iz] are the allomorphs the morpheme let's say /z/.

Allomorph

Lexically Conditioned Allomorph

When the choice of allomorphs are unpredictable from the knowledge of language's Morphology and Phonology, they are lexically conditioned allomorphs. Since there is no pattern, there is no rule: the information is just in the lexicon.

English:

two oxen *two oxes *two ox

two deer *two deers *two deeren

man men

child children

sheep sheep

Allomorph

Suppletion is an extreme form of allomorph in which two completely different roots realize the same morpheme.

eg.
go | went
be | is| was | were | am
good | better | best
bad | worse | worst
one | first
two | second

Stem allomorphy: Allomorphy can also exist in stems.

Vāk (voice)

	Singular	Plural
Nominative	/vaːk/	/vā͡ːʧ-as/
Genitive	/vā͡ːʧ-as/	/vā͡ːʧ-aːm/
Instrumental	/vā͡ːʧ-aː/	/vaːg-bʰis/
Locative	/vā͡ːʧ-i/	/vaːk-ʂi/

The three allomorphs of the Sanskrit word, **vāk**

/vaːk/, /vā͡ːʧ/ and /vaːg/

Morphological Typology

Morphological Typology

- Languages can be classified into groups based on a number of different linguistic criteria.
- One such way to categorize languages is by the type and extent of **morphology** that they use.
- Some languages string many morphemes together to form words. They are called *synthetic* languages.
- While some other languages tend to realize most words as independent or mono-morphemic segments. They are called as *analytic* languages.
- This typology should be seen not as a strict dichotomy between analytic and synthetic, but rather as a scale on which languages can be placed depending on the degree to which they exhibit that type of morphology.

Morphological Typology

Morphological typology is the basis for the broad classification of Languages of the world into four major Morphological types:

- **Isolating Languages (analytic):**
  - each word tends to consist of a single, independent morpheme
  - there are no bound forms i.e affixes
  - grammatical markers, for features like tense and case, are generally realized as unattached (free) morphemes.
  - morpheme = word
  - grammatical changes are indicated by word order

Analytical languages are most common in Southeast Asia (Chinese, Vietnamese), but some such languages are also found among the Austronesian languages (Fijian, Tongan) and some Niger-Congo languages (Gbe, Yoruba).

Example from Vietnamese:
no se khong doc sach
he FUT NEG read book
'he will not read book'

Example from Chinese:
Ta ba shu mai le
He NOM book buy Asp
'He bought the book.'

Morphological Typology

Agglutinative Languages (synthetic):

- all bound forms are affixes
- they are added to a stem like beads on a string
- every affix represents different morphological feature
- each morpheme represents only one grammatical meaning
- morpheme $<$ word
- word order is slightly less important than it was in analytic languages.
- eg. Dravidian, Turkish, Finnish, Hungarian etc.

Example from Telugu (Dravidian)

illu ‘house’
iMti- ni ‘house (object)’
iMti- ki ‘to the house’
iMti- lō ‘in the house’
iMti- tō ‘with the house’
....

rA ‘to come’
vacc- A- nu ‘I came’
vacc- A- mu ‘we came’
vacc- A- vu ‘you(sg.) came’
vacc- A- ru ‘you(pl.) came’
vacc- A- ḍu ‘he came’

Example for derivation from Telugu:
pagalagoVttiMcipeVttamananivvaxalacukolekapowunnAnu.
pagulu+a-koVttu+iMcu+i-peVttu+a-manu+a-ivvu+a-xaluvu+i-koVnu+a-leka-po+ wunn+1,sg,any
break+inf-strike+cause+cpm-benefactive+inf-tell+inf-permit+inf-think+cpm-reflexive+inf-neg+go+prog+1, sg
'I could not think to permit someone to tell for my sake to break something'
(pc, G. Uma Maheshwar Rao)

Morphological Typology

- **Inflectional Languages (fusional):**
  - distinct features are merged into a single bound form (portmanteau morph)
  - morpheme boundaries are difficult to identify
  - every suffix has several grammatical functions

The classical Indo-European languages like Sanskrit, Greek, Latin etc. are examples of flexional languages where in the inflectional morphemes are said to be "fused" together.
Example from Ancient Greek:

lu-ō	1S:PRES:ACT:IND (I am releasing)
lu-ōmai	1S:PRES:ACT:SBJV (I should release)
lu-omai	1S:PRES:PASS:IND (I am being released)
lu-oimi	1S:PRES:ACT:OPT (I might release)
lu-etai	3S:PRES:PASS:IND (He is being released)

Russian:

Case	**singular**	**plural**
Nominative	knig-a	knig-i
Genitive	knig-i	knig-ø
Dative	knig-e	knig-am
Accusative	knig-u	knig-i
Instrumental	knig-oj	knig-ami

Incorporating Languages (polysynthetic):

- all bound forms are affixes
- Inflections are incorporated into the word.
- ability to form words that are equivalent to whole sentences in other languages
- morphologically extremely complex
- Generally, morphology is more important than context and syntax.
- eg. Icelandic/Aleutian

Inuktitut, for instance the word-phrase:
tavvakiqutiqarpiit
roughly translates to "Do you have any tobacco for sale?"

Yup'ik (Alaska):
angya-li-ciq- sugnar- quq-llu
boat- make-FUT- PROB- 3sg.NOM-also
'Also, he probably will make a boat'

Morphological Model

Morphological Modeling

Modelling speaker' s knowledge about words.

Morphologists propose three models (Hockett, 1954) describing morphological formations:

1. **Item and Arrangement (IA)**
2. **Item and Process (IP)**
3. **Word and Paradigm (WP)**

Item and Arrangement

1. Item and Arrangement (IA)

- Morpheme Based Morphology
- Conceived as object oriented concatenation.
- No notion of basic allomorphs
- Word-forms are analyzed as sequences of concatenated morphemes
- Cut and paste method
- *anti-dis-establish-ment-ar-ian-ism* is analyzed as *establish*, a root morpheme and the rest as bound derivational morphemes.

Example:
(i) -s plural is pronounced [s] after voiceless obstruent
e.g.
book $\Rightarrow$ books [s]
cat $\Rightarrow$ cats [s]

(ii) -s plural is pronounced [z] after voiced sounds
e.g.
bag $\Rightarrow$ bags [z]
lamb $\Rightarrow$ lambs [z]

(iii) -s plural is pronounced [iz] after sibilant consonants such as [s], [z], [3], [s] , [ts], [d3]
e.g.
bus $\Rightarrow$ buses [iz]
watch $\Rightarrow$ watches [iz]

Item and Arrangement

- In this approach, the relationship between allomorphs like [s], [z], [iz], [ren], [en] and [0] are missed out.
- Furthermore it assumes that words are always composed of discrete sequences of morphemes.
- However, in a number of languages linear sequencing of morphemes is not the favoured method of deriving words.
- Therefore, a morpheme-based model quickly leads to complications when one tries to analyze many forms of allomorphy.
- Analyzing words as sequences of morphemes simply ignores the intuition that words are related to each other in more than one aspect i.e. formally and semantically.

Item and Arrangement

For example, the word *cats* can be easily sliced into *cat* and the plural morpheme *-s*.
But a similar analysis of the words *geese*, *men*, *feet* etc. into their corresponding roots and plural morphemes runs into difficulty.

Item and Process

2. Item and Process

- Lexeme based Morphology
- Applying rules to form new words
- Notion of allomorphs
- An inflectional rule takes a lexeme, changes it as is required by the rule, and outputs a word-form.
- Bypasses the difficulties inherent in the Item-and-Arrangement approach.

Item and Process

- There is a concept of allomorph in IP.
- Let's take again English plural formation.
- If /z/ is taken as basic morpheme, it has allomorphs like [s], [z], [iz], [ren], [en] and [0].
  - Rule 1: /z/ $\longrightarrow$ [s] / [-voiced]___#
  - Rule 2: /z/ $\longrightarrow$ [z] / [+voiced]___#
  - Rule 3: /z/ $\longrightarrow$ [iz] / [+sibilant]___#
  - The problematic cases like *men* can start with *man* and apply the rules of plural formation which automatically massage the form into a well-formed word-form.
  - Rule 4: man + /z/ $\longrightarrow$ men
  - Rule 5: child + /z/ $\longrightarrow$ children

Word and Paradigm

3. Word and Paradigm

- Word based Morphology
- Paradigm as the central notion
- Good to tackle exceptions in a language
- Instead of stating rules to combine morphemes into word-forms, or to generate word-forms from stems, word-based morphology makes generalizations that hold between various forms of inflectional paradigms.
- Words are treated as whole words that are related to each other by analogical rules.
- The assumption is, a morpho-syntactic Property (P) is associated with the root/stem (X). Words (XP) are viewed as exponents of P.
- Each paradigm is different in their morpho-phonemic/ add-delete processes.

Word and Paradigm

For example,
PLAY]verb is a Lexeme which has the following paradigm.

WORDFORMS	FORMATIVES
play	present, 1-SG, 1-PL, 2-SG, 2-PL, 3-PL
plays	present, 3-SG
played	past
played	participle
playing	progressive

Word and Paradigm

For example,
GO]verb is a Lexeme that has the following paradigm.

WORDFORMS		FORMATIVES
play	go	present, 1-SG, 1-PL, 2-SG, 2-PL, 3-PL
plays	goes	present, 3-SG
played	went	past
played	gone	participle
playing	going	progressive

Word and Paradigm

For example,
CUT]verb is a Lexeme that has the following paradigm.

WORDFORMS			FORMATIVES
play	go	cut	present, 1-SG, 1-PL, 2-SG, 2-PL, 3-PL
plays	goes	cuts	present, 3-SG
played	went	cut	past
played	gone	cut	participle
playing	going	cutting	progressive

Word and Paradigm

Paradigms PLAY, GO, CUT share similar morpho-syntactic properties, but differ in their add-del rules.

WORDFORMS			FORMATIVES
play (ø,ø)	go (ø,ø)	cut (ø,ø)	present, 1-SG, 1-PL, 2-SG, 2-PL,
plays (ø,s)	goes (ø,es)	cuts (ø,s)	present, 3-SG
played (ø,ed)	went (go, went)	cut (ø,ø)	past
played (ø,ed)	gone (ø, ne)	cut (ø,ø)	participle
playing (ø,ing)	going (ø, ing)	cutting (ø, ting)	progressive

Typology Vs. Model

Morphological Typology Vs. Model

The three models of morphology (IA, IP and WP) more or less match languages with different morphological types (agglutination (synthetic), inflectional (fusional) and incorporation (polysynthetic)).

- The Item-and-Arrangement approach fits very naturally with agglutinative languages;
- while the Item-and-Process and Word-and-Paradigm approaches usually address flexional languages;
- Word-and Paradigm approach fits very well to incorporating languages.

Computational Model

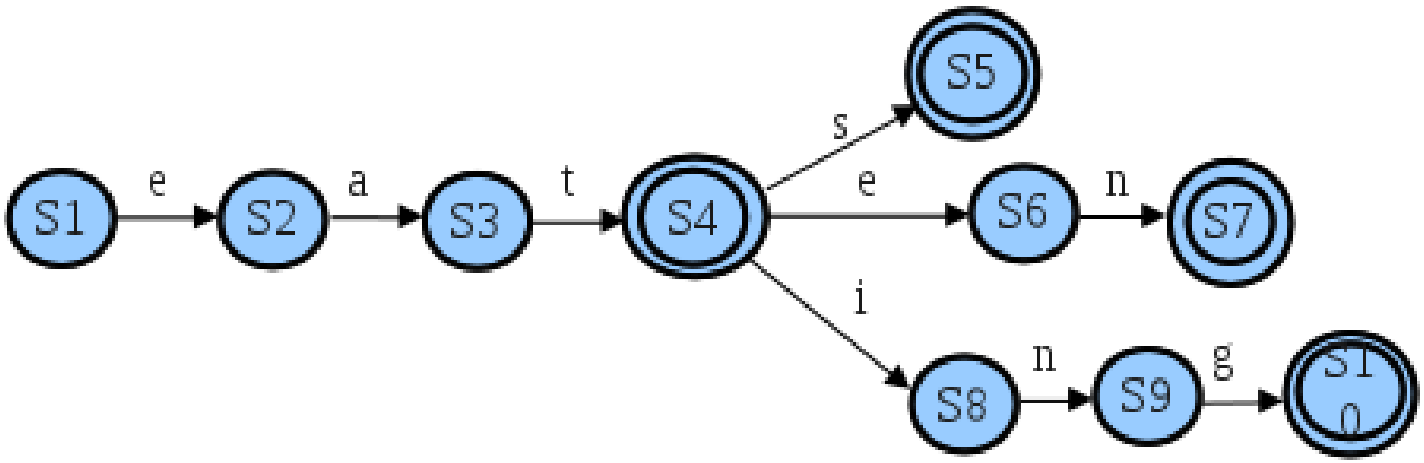
Computational model: Finite State Technology

- Finite State Automata
- Finite State Transducers

Finite State Automata (FSA) is a abstract mathematical device which describes processes involving inputs and processing it.
FSA may have several states and switches between them.
Each state is crossed depending on the input symbol and performs the computational tasks associated with the input.
A Finite State Automaton is a machine composed of

- An input tape
- A finite number of states, with one initial and one or more accepting states
- Actions in terms of transitions from one state to the other, depending on the current state and the input

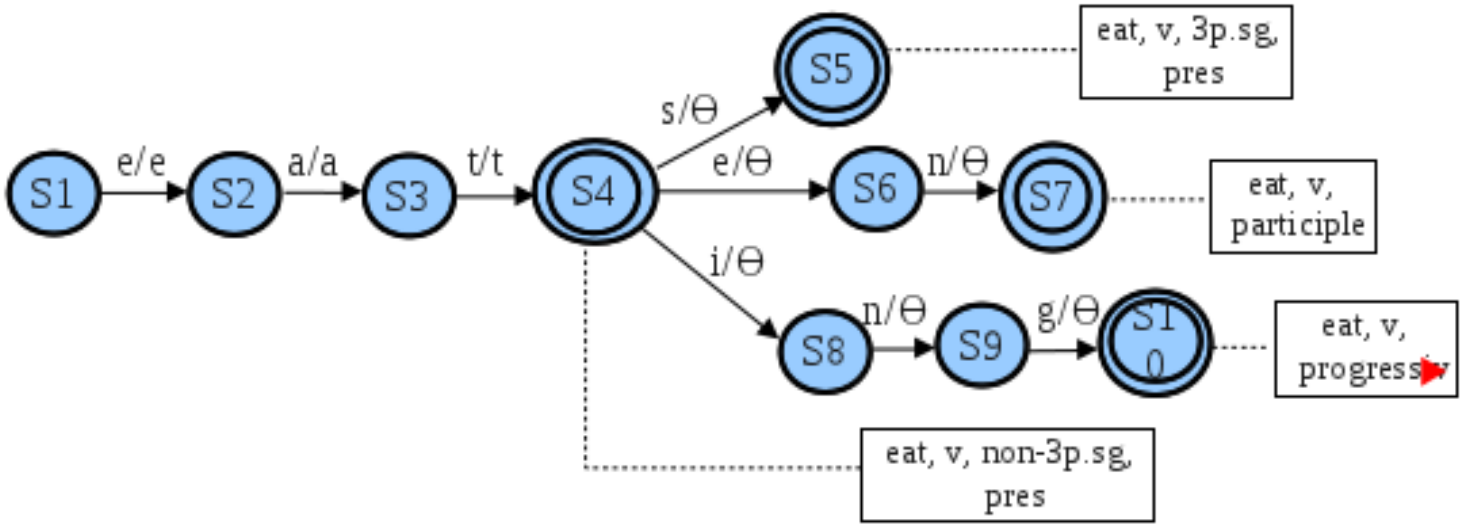
A simple FSA that recognises various verb forms of 'EAT' viz. eat, eats, eaten and eating is shown below.



Finite State Transducer:

- **FST** unlike FSA works on two tapes; input and output tape.
- FSAs can recognize a string but do not give the internal structures.
- But FSTs can recognize and able to provide the internal structure of any input.
- They read from one tape and write on another tape.
- So it is possiple to turn FST to analyse and generate the forms.

A simple FST that recognises various verb forms of 'EAT' viz. eat, eats, eaten and eating is shown below.



Morphological Analyzers & Generators

- **Morphological Analyzer:**
  Analyses given wordforms as root, lcat, gend, num, pers, cm/tam and other grammatical information.
  Basic module for any NLP tasks.
  Input: wordform Output: root +lcat + feature values

Language	Input	Output
Tamil	ArYu	rt=ArYu, lcat=n, g=n, n=sg, p=3, c=dir, cm=0
		rt=ArYu, lcat=v, g=any, n=sg, p=2, tam=IMP
		rt=ArYu, lcat=num, g=n, n=sg, p=3, c=dir, cm=0
	avarE	rt=avar, lcat=pn, g=fm, n=sg, p=3,c=obl, cm=E
		rt=avarE, lcat=n, g=n, n=sg, p=3, c=dir, cm=0

- **Morphological Generators:**
  Generates wordforms from given root, lcat, gend, num, pers, cm/tam and other grammatical information.
  Used in Machine Translation, Speech synthesis, TTS and etc.,

  Input: wordform Output: root +lcat + feature values

Language	Input	Output
Tamil	rt=ArYu, lcat=n, g=n, n=sg, p=3, c=dir, cm=0	ArYu
	rt=ArYu, lcat=v, g=any, n=sg, p=2, tam=IMP	ArYu
	rt=ArYu, lcat=num, g=n, n=sg, p=3, c=dir, cm=0	ArYu
	rt=avar, lcat=pn, g=fm, n=sg, p=3,c=obl, cm=E	avarE
	rt=avarE, lcat=n, g=n, n=sg, p=3, c=dir, cm=0	avarE

- Morphological analysis and generation: Inverse processes.
- Analysis may involve non-determinism, since more than one analysis is possible.
- Generation is a deterministic process.
- In case a language allows spelling variation, to that extent, generation also involves non-determinism

Models for Indian Languages

Best suitable Linguistic models for Indian languages:-

Word and Paradigm Model

- Not much Linguistic background required
- Anybody with adequate language background can implement
- Several fast tools are available

Resources Required:

- Paradigm Class and Table
- Morphological Lexicon
- Category, Feature Definition

Models for Indian Languages

Computational Model:-

Finite state Model

Several off-the-shelf tools available for FST which support Word and Paradigm model

1. Apertium (Lttoolbox)
2. Helsinki Finite-State Transducer Technology (HFST)
3. XFST (Xerox Finite State Tool)
4. SFST (Stuttgart Finite State Transducer Tools) etc.

Conclusion

Conclusion

- MA & MG are indispensable modules for any NLP applications in Indian languages
- Understanding morpho-phonemics and morpho-syntax are required for WP model
- Not morphology:
  - Tokenization (before morphology)
  - Stemming and Lemmatization (instead of morphology)
  - POS tagging (after morphology)
- Selecting the appropriate morphological analysis is a challenging task which are done by other specific modules.