Computing Morphology

G. Uma Maheshwar Rao

University of Hyderabad

- Language is perceived as sequences

of one or more words.

- Understanding Language begins with

Understanding of words

✓ Hence, words are analyzable.

Constituency

- Nature of Words:
  - Atomic
  - Non-Atomic
  - Continuous
  - Discontinuous

Distribution

Words in a text CAN be

A certain number of tokens and

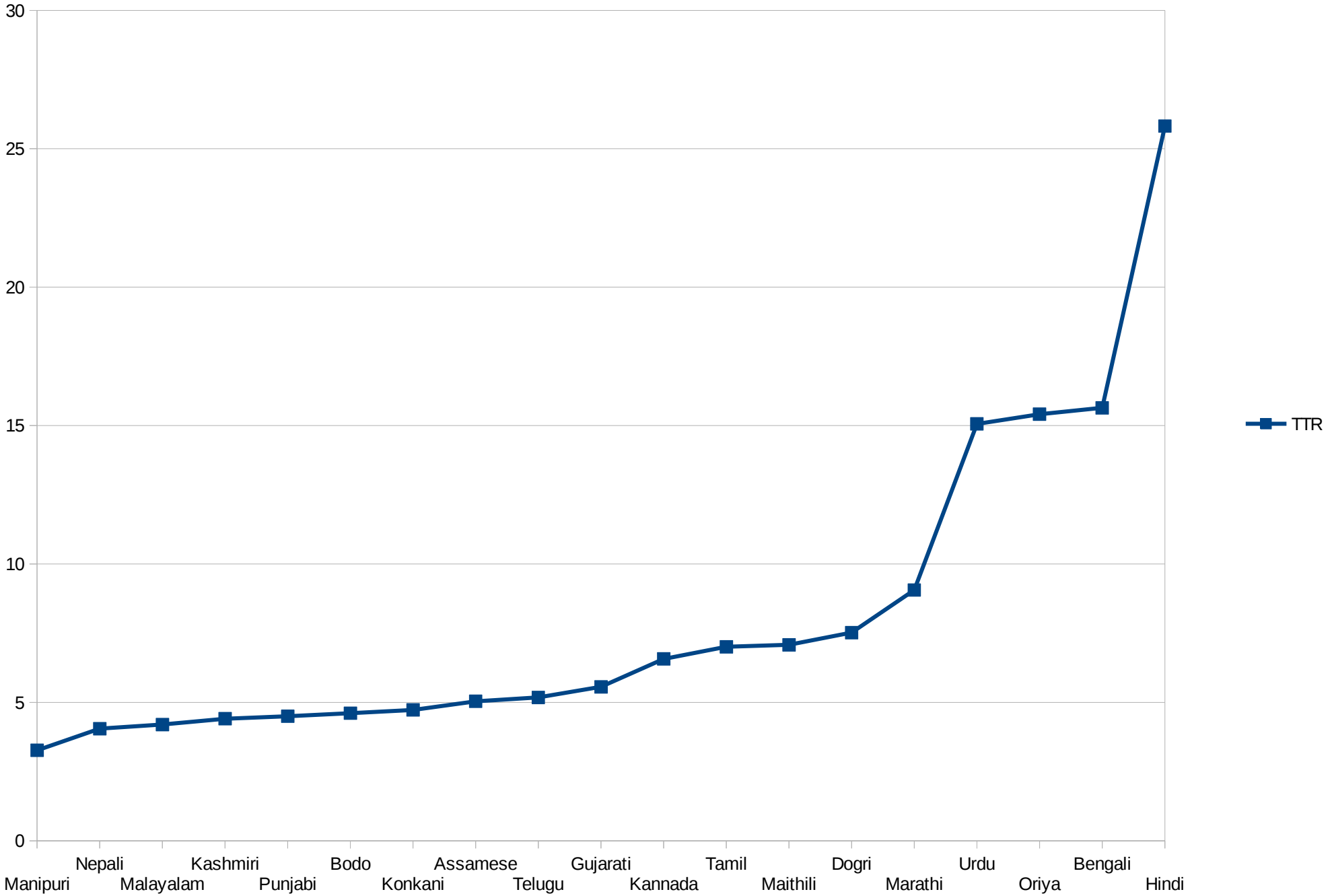
A certain number of types

Corpus Analysis

S.No	Lang	Tokens	Types	token-type ratio(parsity)	Type-Token-Ratio (density)
1	Assamese	201854	39974	5.04	19.8
2	Bengali	2531295	162,454	15.64	7.9
3	Bodo	204382	44311	4.61	21.6
4	Dogri	204473	27160	7.52	13.2
5	Gujarati	198837	35752	5.56	17.9
6	Hindi	3,104,668	120,227	25.82	3.8
7	Kannada	3,118,987	474,066	6.57	15.1
8	Kashmiri	198194	44934	4.41	22.6
9	Konkani	199626	42150	4.73	21.1
10	Maithili	209409	29553	7.08	14.1
11	Manipuri	201694	61563	3.27	30.5
12	Malayalam	2313855	542,657	4.2	15.58
13	Marathi	1784198	196916	9.06	11.0
14	Nepali	201322	49652	4.05	24.6
15	Oriya	2966417	192464	15.41	6.4
16	Punjabi	2308030	104368	4.5	22.11
17	Telugu	2,769,797	534,628	5.18	19.30
18	Tamil	3,124,447	445,361	7.01	14.2
19	Urdu	117240	7783	15.06	6.6

Token Type Ratio (Sparsity)

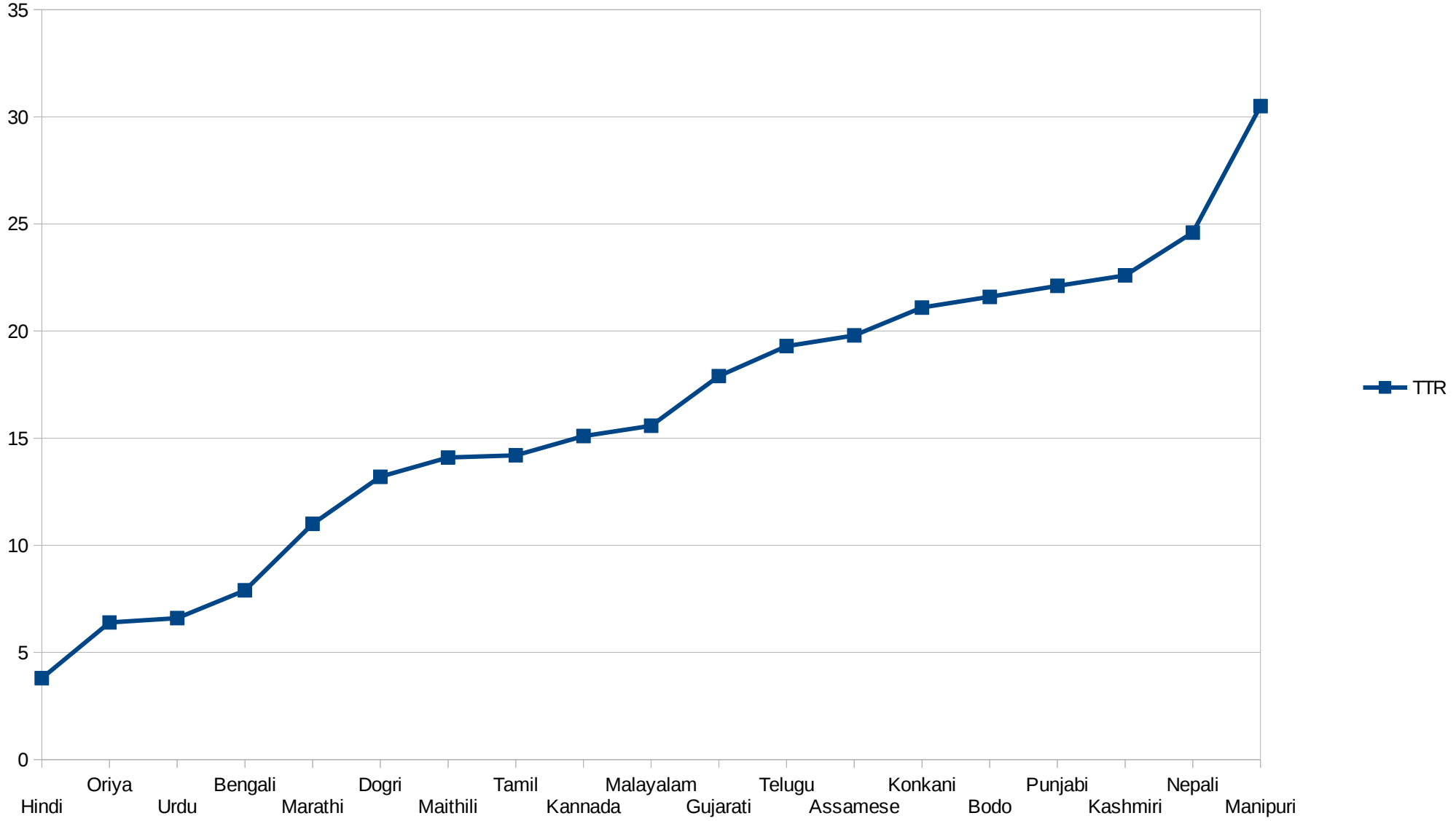
Lang	TTR
Manipuri	3.27
Nepali	4.05
Malayalam	4.2
Kashmiri	4.41
Punjabi	4.5
Bodo	4.61
Konkani	4.73
Assamese	5.04
Telugu	5.18
Gujarati	5.56
Kannada	6.57
Tamil	7.01
Maithili	7.08
Dogri	7.52
Marathi	9.06
Urdu	15.06
Oriya	15.41
Bengali	15.64
Hindi	25.82


30
25
20
15
10
5
0
TTR
Manipuri
Nepali
Malayalam
Kashmiri
Punjabi
Bodo
Konkani
Assamese
Telugu
Gujarati
Kannada
Tamil
Maithili
Dogri
Marathi
Urdu
Oriya
Bengali
Hindi

Type –Token Ration (density)

Lang	Type-Token-Ratio (density)
Hindi	3.8
Oriya	6.4
Urdu	6.6
Bengali	7.9
Marathi	11
Dogri	13.2
Maithili	14.1
Tamil	14.2
Kannada	15.1
Malayalam	15.58
Gujarati	17.9
Telugu	19.3
Assamese	19.8
Konkani	21.1
Bodo	21.6
Punjabi	22.11
Kashmiri	22.6
Nepali	24.6
Manipuri	30.5

Density


35
30
25
20
15
10
5
0
TTR
Hindi
Oriya
Urdu
Bengali
Marathi
Dogri
Maithili
Tamil
Kannada
Malayalam
Gujarati
Telugu
Assamese
Konkani
Bodo
Punjabi
Kashmiri
Nepali
Manipuri

What is Morphology?

There are two dominant views:

1. … Study of word Structure

2. …Study of formal relationships between words

The Null Hypothesis

Morphological processing can be undesirable since every word in a language may be stored and accessed as and when required.

However, in any human language -

- ✓ possible words are infinite in number!

Contd....

Actual and attested words are unmanageably large in number.

Hence, it is necessary to model morphology in terms of

<Morphological rules or Word Formation Strategies>

to permit us to recognize or produce new words.

The basic concepts of Morphology:

Native speakers create new words from the existing ones;
Borrow from other languages as and when necessary.

Discovery of these mechanisms and the intuitive knowledge underlying this creativity is *morphology.*

Speakers possess intuitive knowledge about:

that words are related to each other

By form/shape and semantics/meaning

Contd....

Knowledge of the existence of

patterns, rules and the other details of the processes involved is what is all about morphology.

Ability to *Form* or *Recognize* that a group of words are related and they are derived from common base is due to morphology at work.

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

Contd....

Speaker's ability to derive or relate words like,

act, *active*, *activity*, *activate*, *activator* and *activation*, in terms of their shape and meaning.

Alternatively, ability to reject-

*kætən, *kætz, *kætəz for cats,

walk – *walken; drive – *drived; read – *readed; active – *activement, *activance, and *activant

as ilformed is due to the knowledge of morphology.

Morphological typology:

... basis for the classification of Languages of the world into four major Morphological types:

Isolating/Analytic	Ex. Chinese
Agglutinating/Synthetic	Ex. Altaic, Dravidian
Inflectional/Fusional	Ex. Indo-European
Incorporating/Polysynthetic	Ex. Icelandic/Aleutian

Semitic languages exhibit a very peculiar type of morphology, often called *root-template morphology.*

Eg. Arabic root ``ktb'' produces the following wordforms:

Template	aa (active)	ui (passive)	gloss
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'

Contd....

A Correspondence

between a word and it's parts

i.e.

morphemes per word ratio

in terms of their nature and function;

A range from one-to-one to one-to-many characterizes Analytic to Polysynthetic types.

Morphological Modelling:

- Modelling speaker's knowledge about words

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

1. Item and Arrangement (IA) :

a. Conceived as object oriented concatenation.

b. No notion of basic allomorph

contd....

2. Item and process (IP):

a. Conceived as processing of abstract units of Lexicon.

b. Basic allomorph is at the centre of the concept.

3. **Word and Paradigm (WP):**

a. Assumes morpho-syntactic Property (P)

associated with the root X.

b. Words (XP) are viewed as exponents of P.

Which Morphology?

1. Concatenative Morphology

(dubbed as Neo-Paninian)

-is the main stream morphology

-is the most popular

and the dominant approach till date;

-numerous representational variants exist;

Contd....

-sub-word units (root/stem, affix) are building blocks

-distinguishes between inflection and derivation

-easy to manage in pedagogy and computation

-exceptions are too numerous

-directionality assumed

cont..

2. Non-concatenative Morphology

(Non-Paninian),

also known as Relational Morphology

-most promising and convincing in terms of psychological reality

-multi-directional

-reject multiple morphologies- not many variants

contd....

-***morphologically complex languages***

may need **n*n-1/2 WFSs**

-not an easy task for computational implementation

-claims to capture native speaker's

morphological

knowledge

-no exceptions

The basic building blocks of Morphology

words are composed of one or more of *small indivisible or minimal but meaningful units often called as morphemes*.

walk (one morpheme), walk-s (two morphemes), walk-ed (two morphemes), walk-ing (two morphemes), establish-ment-ary (three morphemes), establish-ment-ari-an (four morphemes), establish-ment-ari-an-ism (five morphemes), anti-establish-ment-ari-an-ism (six morphemes), anti-dis-establish-ment-ari-an-ism (seven morphemes) and so on so forth.

contd....

Morphemes do not always come

in the same shape in all their occurrences.

Ex. /laĭf/ *life* : /laĭv/ *live*-s,

/vaĭf/ *wife* : vaĭv/ *wive*-s;

-s, -z, -əz, rən, -ən in the case of plural marker

The variants: /laĭf/ and /laĭv/, /vaĭf/ and /vaĭv/,

-s, -z, and -əz, are often technically called allomorphs.

Contd....

words are often spoken together as continuous stream of sounds without any silence or punctuation.

native speakers are well equipped to deal with this situation .

native speakers have knowledge of-

- word beginnigs and and endings.

Word (internal) structure is the source of this knowledge.

Inflection Vs. Derivation

words are either inflectional or derivational.

Inflectional:

words used in syntax,

and carry

- exponents of morpho-syntactic formatives
- explicate morpho-syntactic functions

contd....

Derivational:

derives new words;

- used as a reservoir of words to be used in inflection.
- often hidden in inflection
- tradition recognizes two kinds of derivation;
- proper derivation or affixal derivation
- compounding.

involves two or more words rather than affixes.

Contd....

Word: is the most commonly used term in morphology

-ambiguous in common usage.

Ex: walk, walks, walked, walking,

- share *sense* and *shape* among them
- But they are different in that they can't generally be used in the same syntactic structures.

Words vs. Lexemes

Similarities and differences between these "words/ wordforms" have the most significant theoretical import in morphology.

Distinct 'words' with essentially the same 'sense' but each occurring in a distinct syntactic context with distinct morphological realization are subsumed under the concept called 'lexeme'.

contd....

These words are to be considered as different forms of the same lexeme (usually represented in CAPITALS).

words like WALK, WALKER, *WALKOUT*, *WALKATHON* etc. are different lexemes,

because they refer to different kinds of semantic entities viz. 'an act of motion involving locomotory organs', `a person or device that walks or helps in walking', 'walk away in protest from meeting', and 'a marathon walking'.

Contd....

Inflection and Derivation :

- word-forms are organized into paradigms,
- derivational forms are not
  - word-forms are syntactically motivated
  - lexemes are conceptually motivated
    - wordforms enter syntax
    - lexemes enter lexicon

Contd....

A Word-form is an exponence of a morpho-syntactic projection of the functions

overtly marked by the corresponding formative (bound morphemes or affixes).

Inflectional morphology involves the formation of wordforms from the bases (roots/stems) of words/lexemes by the addition of certain affixes to express certain grammatical relationships and functions.

Inflection and Paradigm

The term paradigm refers to an exhaustive set of *morpho-syntactically* related word-forms associated with a given lexeme.

Members of a paradigm are all those word-forms that are obtained through the conjugation of verbs, and the declensions of nouns, pronouns etc.

contd....

Language: English

Lexeme: Book

Category: N

Case	word
Nom.	sg. book pl. books
Gen.	sg. book's pl. books'

Lexeme: DRINK

Category: V

non-3rd sg. pr.t.	Drink
3rd sg. pr.	Drinks
pt. any	Drank
Ppl any	Drunk
gerund	Drinking

contd....

Lexeme: GREAT

Category: A

Normative	Great
Comparative	Greater
Superlative	Greatest

Cont

Language: Hindi

Lexeme: LADAKA

Category: N

Function	Sg.	Pl.
Direct	ladakA (0/0)	ladake (e/A)
Indirect	ladake (e/A)	ladakoM (oM/A)
Vocative	ladake (e/A)	ladako (o/A)

Lexeme: CAL

Category: V

Imp. m/f 2p. sg.	cal
Imp.	calo
Imp.	calie
Opt.m/f 1p.sg.	calUM
Pr.m.sg.	calwA
Pr.f.sg.	calwI
Pr.m.pl	calwe
Pr.f.pl.	calwI
Ft.m.1p.sg.	calUMgA
Ft.f.1p.sg.	calUMg
Ft.m.2p.sg.	caloge
ft.f.2p.sg	calogI
Ft.m./f.2hon.sg./pl.	caliyegA
Ft.m.3p.sg.	calegA
Ft.f.3p.sg.	calegI
Ft.f.3p.pl.	caleMgI
Ft.m.3p.pl	celeMge
Pt.m.any p.sg	calA
Pt.f.any p. sg.	calI
Pt.m.any p. pl.	caleM
Pt.f.any p.pl.	caleM
Gerund	calnA
Adverbial	calkar

contd....

Variation in form realization

in inflectional categories like

gender, number, person and case in Nouns

tense, aspect, modals etc. in Verbs

is the source for paradigms.

Inflectional categories are determined

by the relevant

morpho-syntactic functions of the language.

English lacks gender as an inflectional category as its subject-verb agreement does not require gender information of the subject noun but only the person and number.

Hindi and Telugu use gender along with person and number to mark the verbform showing agreement with the subject.

Allomorphy

Sources of complexity in morphology

Failure of one-to-one correspondence

between ‘meaning’ and ‘form increases allomorphy demanding complex rule system

Languages differ from each other with regard

to the degree of complexity of allomorphy.

contd....

Often allomorphs can be related to the basic underlying form by a set of phonological or morphological rules

However, there are several cases that cannot be related.

Suppletives, allomorphs by non-phonological basis are sources of morphological irregularities.

Ex. Eng. *go:* *we*-nt;

Hi. *jA :* *ga*-yA;

Te. *vaccu* : *rA (imp.)*

Analyzing word-forms as if they were made of morphemes attached to each other like beads on a string, is called *Item-and-Arrangement* model of morphology. In this approach words are viewed as pure concatenation of various sorts of morphemes:

thus, morphology as basically involves cut and paste method.

However, the relationship between the allomorph

like -s, -z, -əz, rən, -ən and 0, is missed out.

The *Item-and-Process* model underlie the Lexeme-based approach to Morphology. Analyze a word-form as a set of morphemes arranged in sequence;

A word is said to be the result of applying rules that alter a given lexeme.

An inflectional rule takes a lexeme, changes it as required by the rule, and outputs a word-form.

It bypasses the difficulties inherent in the Item-and-Arrangement model. The problematic cases like *men* can start with *man* and apply the rules of plural formation

The *Word-and-Paradigm* model of morphology is the basis for Word-based morphology

the notion of paradigm is at its core.

No rules to combine morphemes into word-forms, or to generate word-forms from roots!

Word-based morphology makes generalizations that hold between various forms of inflectional paradigm.

Words are maximal projections of Lexemes in morpho-syntactic contexts mediated by formatives.

IA and IP models assume discreteness of morphemes and one-to-one correspondence between the units of form and the units of functional categories.

However,

geese, *men*, *feet*, *deer*, *sung*, *sang* etc.

are not composed of linear sequences of discrete units like morphemes.

Levels in Morphology:

Underlying Representation/Lexical Representation vs.

Surface Representation/Wordform representation

$$\text{R+R*+} \sum_{i=o}^{i} Aff\ i <=> \text{WORD}$$

- Words are Analyzed and generated in isolation
- Words are Analyzed or generated as orphans

Morphological Analysis
for
possible morphology

Complexity

An estimation of the number of forms derived from each verb in Telugu:

Simple finite verbforms=650

Simple non-finite verbforms=330

total verbforms==1000

Total number of verbforms in compound verbs involving 2to 7:

1. x]itr.(v2)8=8000
2. x]itr.(v2)6+(v3)5=30,000
3. x]itr.(v2)6+(v3)5+(v4)4=120,000
4. x]itr.(v2)6+(v3)5+(v4)4+(v5)2=240,000
5. x]itr.(v2)6+(v3)5+(v4)4+(v5)2+(v6)2=480,000
6. x]itr.(v2)6+(v3)5+(v4)4+(v5)2+(v6)2+(v7)2+= 960,000

i. maximum verbcombinations-Total verbforms(excluding 6)=878,000

7. x]itr.(v2)6+(v4)4=24,000

8. x]itr.(v2)6+(v5)2=12,000

9. x]itr.(v2)6+(v6)2=12,000

10. x]itr.(v2)6+(v7)2=12,000

11. x]tr.(v3)5+(v4)4=20,000

12. x]tr.(v3)5+(v5)2=10,000

13. x]tr.(v3)5+(v6)2=10,000

14. x]tr.(v3)5+(v7)2=10,000

15. x]tr.(v4)4+(v5)2=8,000

16. x]tr.(v4)4+(v6)2=8,000

17. x]tr.(v4)4+(v7)2=8,000

18. x]tr.(v5)2+(v6)2=4,000

19. x]tr.(v5)2+(v7)2=4,000

20. x]tr.(v5)2+(v8)2=4,000

21. x]tr.(v6)2+(v7)2=4,000

22. x]tr.(v6)2+(v8)2=4,000

i.total verbforms from three verb combinations = 164,000

Total verbforms: from all combinations = 1,042,000

Computation

Lexeme:	WALK	d	a	EAT	d	a
v,pr,any,1,2,3pl	walk	0	0	eat	0	0
v,pr,any,3sg	walks	1	0	eats	1	0
v,pt,any,any	walked	2	0	ate	3	eat
v,ppl	walked	2	0	eaten	2	0
v, ger	walking	3	0	eating	3	0

Input:	walked	ate
delete	-ed	-ate
add	+0	+eat
	---	---
Output:	walk	eat

Relationship between MA & MG

Ma and MG have reverse relationship

MA	MG
Root+Af0~n	well-formed word-forms
Multiple Analyses	Unique Output

Computational model: Finite State Technology

- Finite State Automata
- Finite State Transducers
- Finite State Automata (FSA) is an abstract mathematical device which describes
- processes and processing.

FSA may have several states and switches between them.

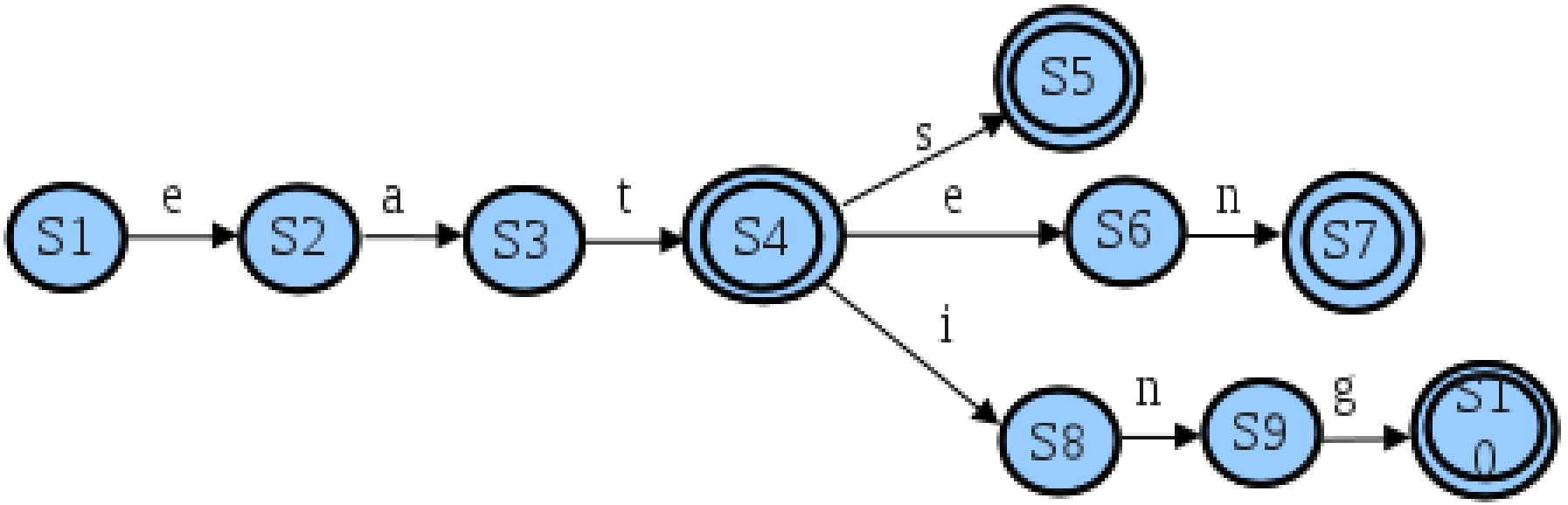
Each state is crossed depending on the input symbol and

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.

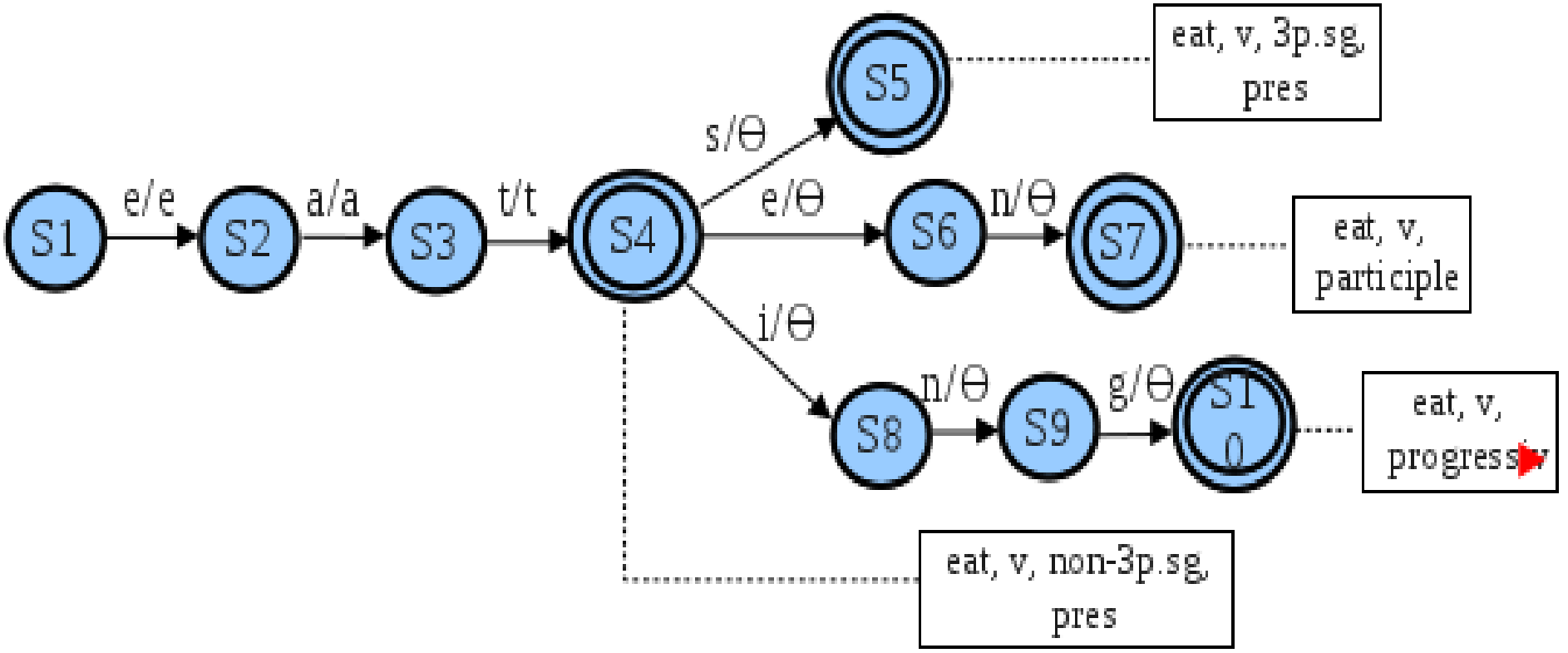


Finite State Transducer:

- 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.

Cont…

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



end

THANK YOU