What we want

- Something to automatically do the following kinds of mappings:

Cats	cat +N +PL
Cat	cat +N +SG
Cities	city +N +PL
Merging	merge +V +Present-participle
Caught	catch +V +past-participle

Lexeme : roughly corresponds to a set of words that are different forms of "the same word".

Ex:*run*, *runs*, *ran* and *running* are forms of the same lexeme.

Morpheme: smallest lingual unit that carries a semantic interpretation.

Ex: "unpredictable" has three morphemes.

Types of morphemes:

Free morphemes

Allomorphs: the plural marker in English is sometimes realized as /-z/, /-s/ or /-ɪz/.

Bound morphemes

Inflectional morphemes

Derivational morphemes

Cranberry morpheme is a type of bound morpheme that cannot be assigned a meaning nor a grammatical function, but nonetheless serves to distinguish one word from the other. Ex:

mit in *permit*, *commit*, and *submit*

ceive in *receive*, *perceive*, and *conceive*

- **Morphology** : sub-discipline of linguistics that studies word structure. While words are generally accepted as being the smallest units of syntax. Ex:
- *dog*, *dogs* and *dog-catcher* are closely related.
- *dog* → *dogs* or *encyclopædia* → *encyclopædias*;
- *dog* → *dog-catcher* or *dish* → *dishwasher*.

Search for singular and Plurals:

1) Spelling rules: y→ i and adding an es (pluralized)

2) Morphological rules: fish – null plural

goose – geese (changing the vowel)

Parsing: produce some sort of structure for the input

Morphological Parsing: oxen→ ox and en (two morphemes)

Morphological Parsing:ending in –ing (input form talking→ talk (verb) and ing (gerund)

Inflectional Morphology

- Two kinds of inflection:

An affix that marks Plural

An affix that marks possessive

Regular nouns:

Cat→ cats thrush→ thrushes

Irregular nouns:

Mouse→ mice ox→ oxen

Regular plural: -s, -z, -sh, -ch, x

Possessive suffix: ‘s

Cont..

- Verbal inflection:

i) Main verbs, ii) modal verbs and iii) primary verbs

4 morphological forms for regular verbs:

Stem

-s form

-ing participle

Past form or –ed participle

Modal verbs

Examples:

I **can** ride a horse. *ability*

We **can** stay with my brother when we are in Paris. *opportunity*

She **cannot** stay out after 10 PM. *permission*

Can you hand me the stapler? *request*

Any child **can** grow up to be president. *possibility*

Morphological Analysis

- Inflectional
  - duck + s = [N duck] + [plural s]
  - duck + s = [V duck] + [3rd person s]
- Derivational
  - kind, kindness
- Spelling changes
  - drop, dropping
  - hide, hiding

Cont.

- Irregular verbs

Often have 5 morphological forms for irregular verbs:

Stem

-s form

-ing participle

Past form

–ed participle

3 forms : cut or hit

? forms: be

Derivational Morphology

- Nominalization: formation of new nouns from verbs or adjectives.

Suffix	base Verb/Adjective	Derived Noun
-ation	combine (V)	combination
--ee	pay	payee
-er	kill	killer
-ness	happy	happiness

Cont..

- Adjectives from Noun/Verb

Suffix	Base Noun/Verb	Derived Adjective
-al	computation (N)	computational
-able	predict	predictable
-less	clue	clueless

What do we need to build a morphological parser?

- Lexicon

- Morphotactics of the language: model of how and which morphemes can be affixed to a stem

- Orthographic rules: spelling modifications that may occur when affixation occurs
  - City → cities instead of citys
  - Most morphological phenomena can be described with regular expressions – so finite state techniques often used to represent morphological processes

FINITE – STATE MORPHOLOGICAL PARSING

- Productive Nominal Plural (-s):

Input	Morphological Parsed output
Dogs	Dog + N + PL

The verbal progressive (-ing)

Input	Morphological Parsed output
going	go + V + PRES-PART

To build morphological parser, one needs:

1. Lexicon
2. Morphotactics
3. Orthographic rules

Compounding

- Compounding: Two base forms join to form a new word
  - Bedtime
  - Careful? Compound or derivation?

Morphotactics: What are the ‘rules’ for constructing a word in a given language?

- Pseudo-intellectual vs. *intellectual-pseudo

- Semantics: In English, un- cannot attach to adjectives that already have a negative connotation:
  - Unhappy vs. *unsad
  - Unhealthy vs. *unsick
  - Unclean vs. *undirty
- Phonology: In English, -er cannot attach to words with more syllables
  - great, greater
  - Happy, happier
  - Competent, *competenter
  - Elegant, *eleganter

Morphological Parsing

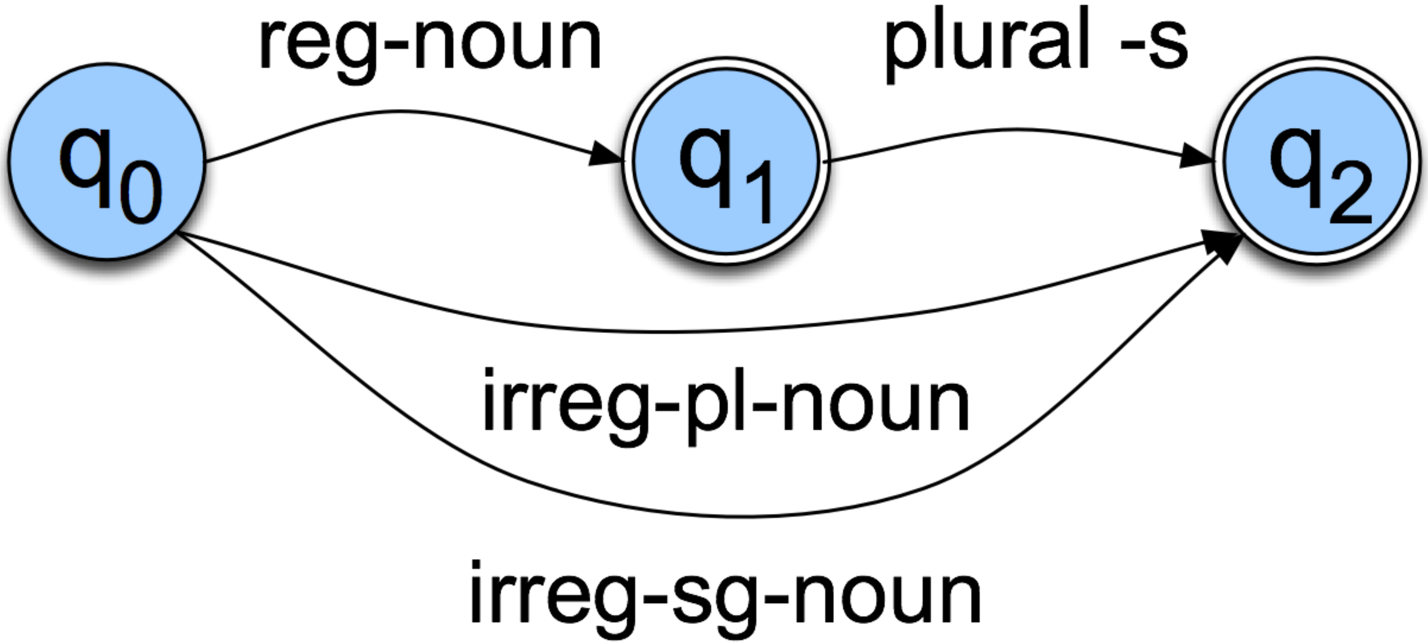
- These regularities enable us to create software to parse words into their component parts

Start Simple

- Regular singular nouns are ok
- Regular plural nouns have an -s on the end
- Irregulars are ok as *is*

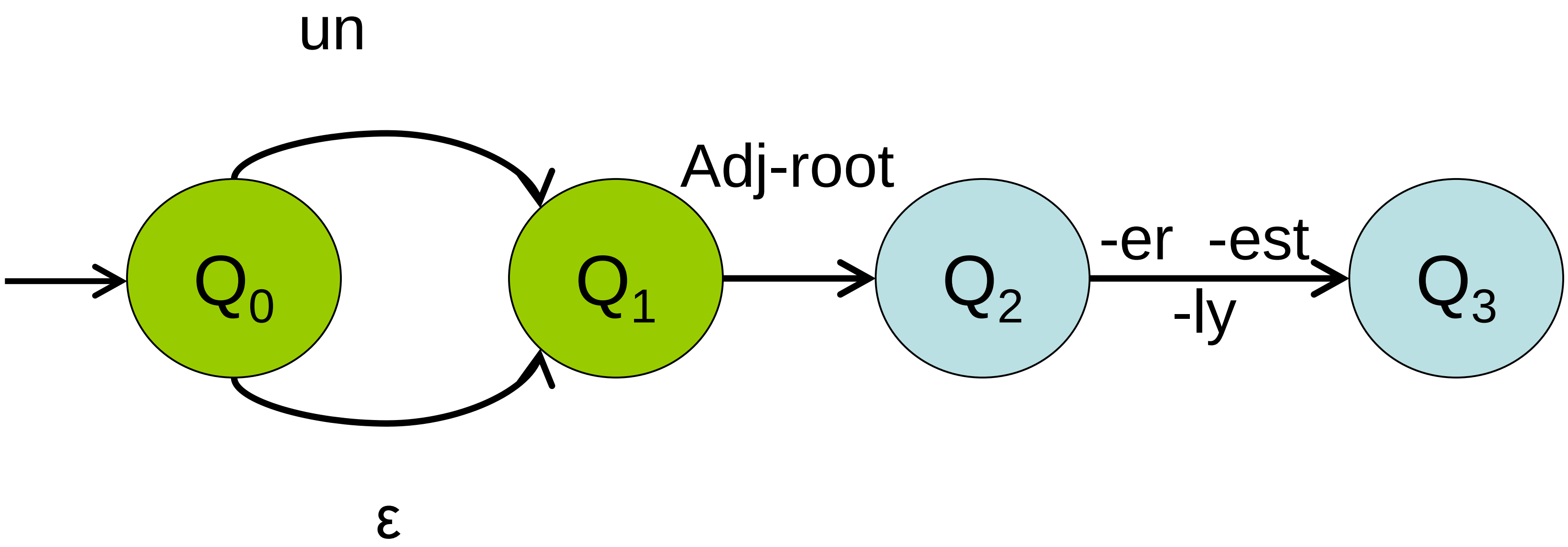
Simple Rules

reg noun → cat, irrg-pl-noun → mice, irreg-sg-noun → goose



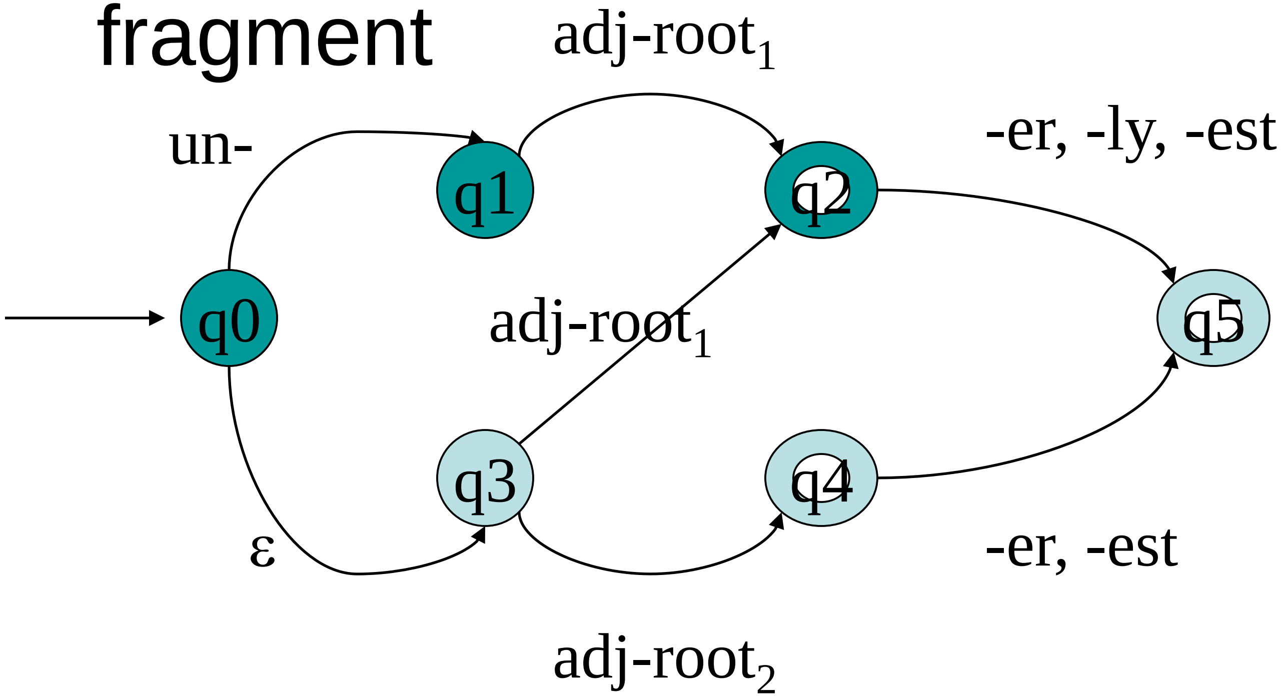
big, bigger, biggest

clear, clearer, clearest, unclear, unclearly



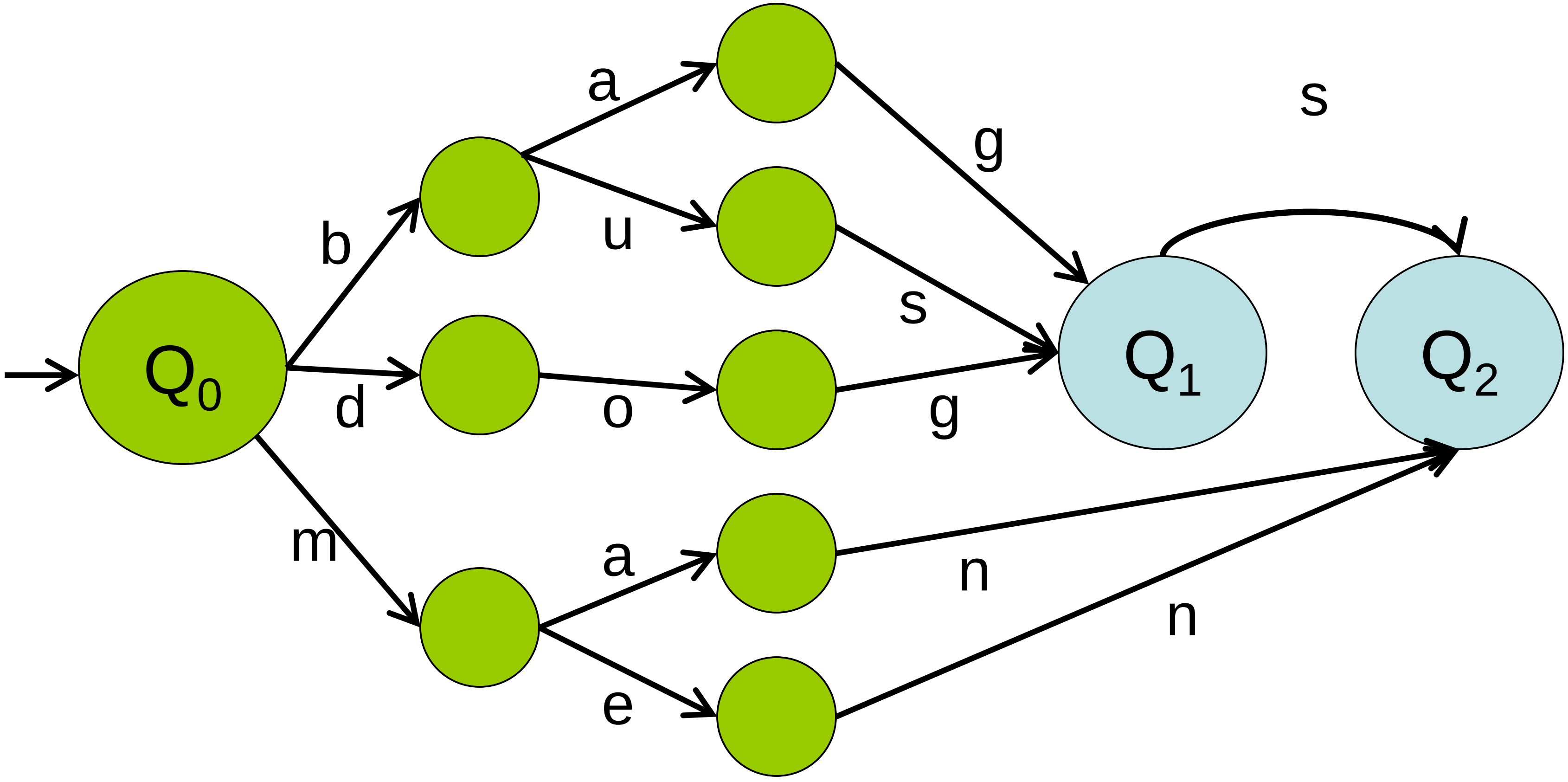
un?ADJ-ROOT{*er* | *est* | *ly*}?

- Derivational morphology: adjective fragment

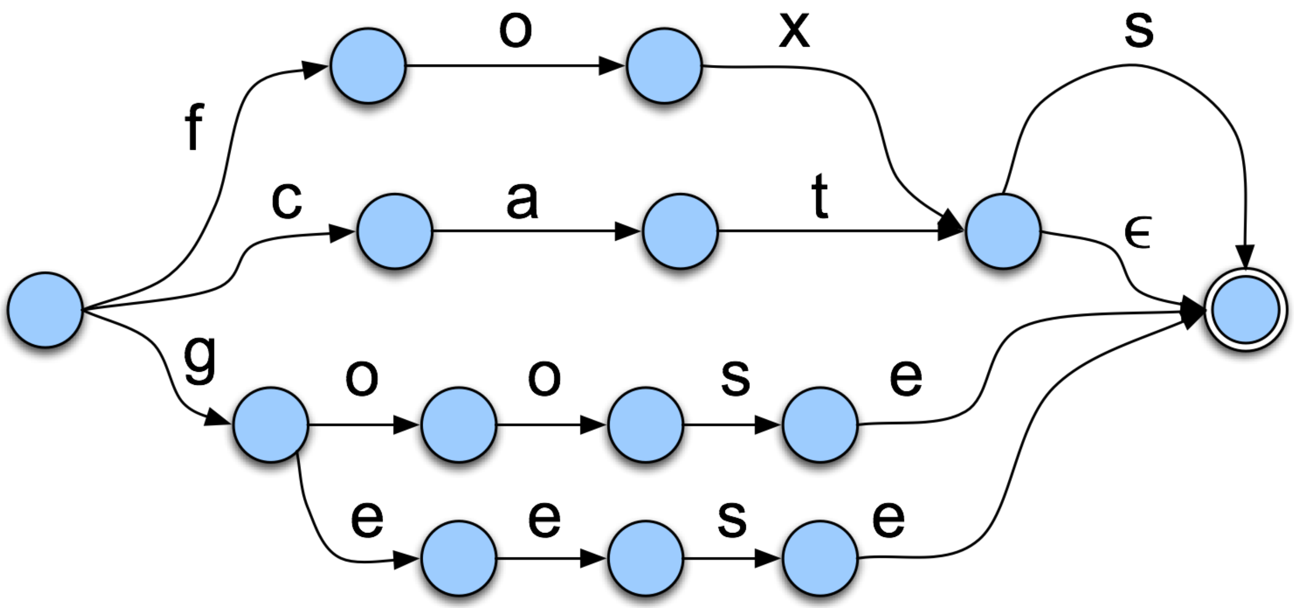


- Adj-root$_1$: clear, happy, real (clearly)
- Adj-root$_2$: big, red (*bigly)

After adding a mini-lexicon



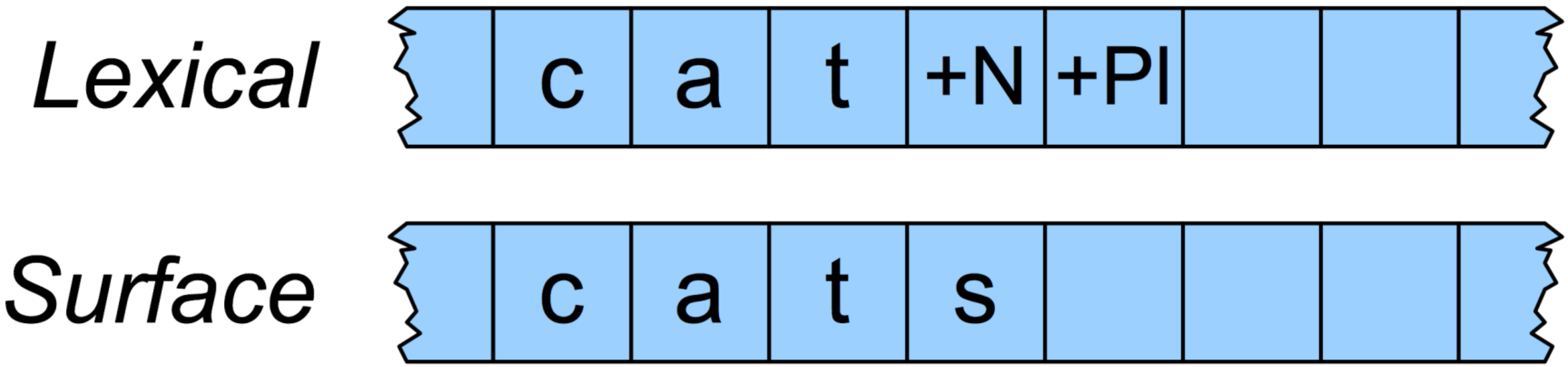
Contd…



FSTs

Kimmo Koskenniemi's two-level morphology

Idea: word is a relationship between **lexical** level (its morphemes) and **surface** level (its orthography)

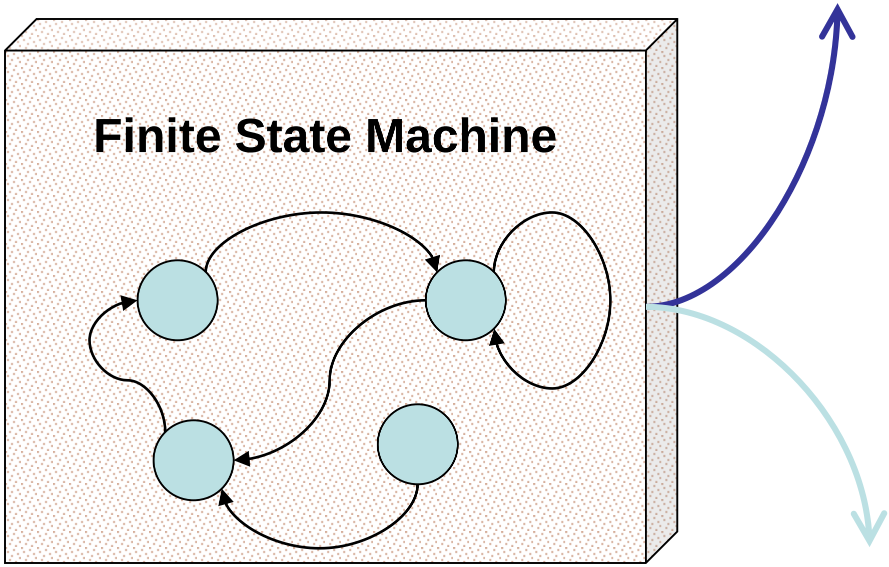


Lexical		c	a	t	+N	+Pl			
Surface		c	a	t	s				

Finite State Transducers

s	*i*	*n*	*g*	*s*	

Surface form



Finite State Machine

Lexical form

s	*i*	*n*	*g*	*#*	*v*	*+sg*	

Ambiguity

- What’s the right parse (segmentation) for
  - Unionizable
  - Union-ize-able
  - Un-ion-ize-able
- Each represents a valid path through the derivational morphology machine.

Ambiguity

- There are a number of ways to deal with this problem
    - Simply take the first output found
    - Find all the possible outputs (all paths) and return them all (without choosing)
    - Bias the search so that only one or a few likely paths are explored

Algorithm to add the plural s

```
FUNCTION add_s (word)
    FOR each rule IN the set of rules LOOP
        IF the last letters of the word match the lhs of the rule THEN
            RETURN the first part of the word + the rhs of the rule
        END IF
    END LOOP
    RETURN word
END add_s
```

Problem (man will return as mans) ‘dictionary lookup’

```
FUNCTION plural (word)
    IF word is in the dictionary with an irregular plural THEN
          RETURN the irregular plural
    ELSE
          RETURN add_s (word)
    END IF
END plural
```

Generic morphological engine

To develop Generic morphological engine as a part of Machine Translation System, which would:

i. Intake the POS tagged data as the input,

ii. Extract the root word of each word in the input corpus

iii. Analyses these root forms on the basis of eight parameters defined in the language,

iv. Subsequently performs a dictionary look up of the analyzed word and generates the word

v. Presents the output in the target language.

Contd…

2. This engine would stem the words into there root form and would analyze the words in accordance to eight parameters which will be passed as an output to the other module.
3. This engine will also contain a rule editor which would be linked to the rule base made earlier and also contain a feature of validation through which new rules could be added in future. This feature will be independent of the language.

Contd…

4. This engine would be generic in the sense that we could link different rule bases to it and this engine would extract the root words from the input corpus, analyze these words in accordance to eight parameters and would pass these root words and its information to the dictionary from where the word in target language is generated.
5. This rule generation system will involve consultation with language experts and experimenting with different representation of rules.

Features explanation of a word

- **ROOT**
- **CATEGORY**
- **GENDER**
- **NUMBER**
- **PERSON**
- **CASE**
- **TAM** Case marker for noun or Tense Aspect Mood(TAM) for verb of the word
- suff : Suffix of the word