CS4025: Morphology and the Lexicon

Words

- The Lexicon
- Morphology

Reading: J&M (chapter 3 in both editions)

Types of Words

Open-class, content words:

- Nouns (shoe, tablecloth, cause, E-insertion…)
- Verbs (see, walk, cause, forget, promise, …)
- Adjectives (big, small, reliable, three-legged, …)
- Adverbs (quickly, well, reliably, …)

Function (closed-class) words

  - » prepositions, determiners etc

See J&M, sect 8.1

Lexicons

- Lexicons are databases of word information.
- “Dictionary” of NLP system
- A good lexicon is critical to performance
  - » “the system with the bigger lexicon always wins”

Dictionary: Child

The following is, of course, written for humans:

- child \'chi-(*)ld\ *n pl* children [ME, fr. OE cild] 1a: an unborn or recently born person 1b: *dialect:* a female infant 2a: a young person of either sex esp. between infancy and youth 2b: a childlike or childish person 2c: a person not yet of age...
  - » [From Webster's on-line dictionary]

Word Information

An NLP system needs to know

- Spelling
- Category and subcategory
- Inflections (plurals, past, etc)
- What word corresponds to in DB or KB
- Statistical information
- maybe pronunciation
- probably not derivation

Example: *Person*

- *Person*
  - » Category: noun
  - » Subcategory: count noun
  - » Inflections: plural = *people*
  - » Database correspondence: **person** class.
  - » Frequency: .03%

Word Meaning

- What do these words mean
  - » He told a lie
  - » The temperature fell in the afternoon
- Many subtleties, difficult to represent
- Most NLP systems ignore, use very simple models of meaning

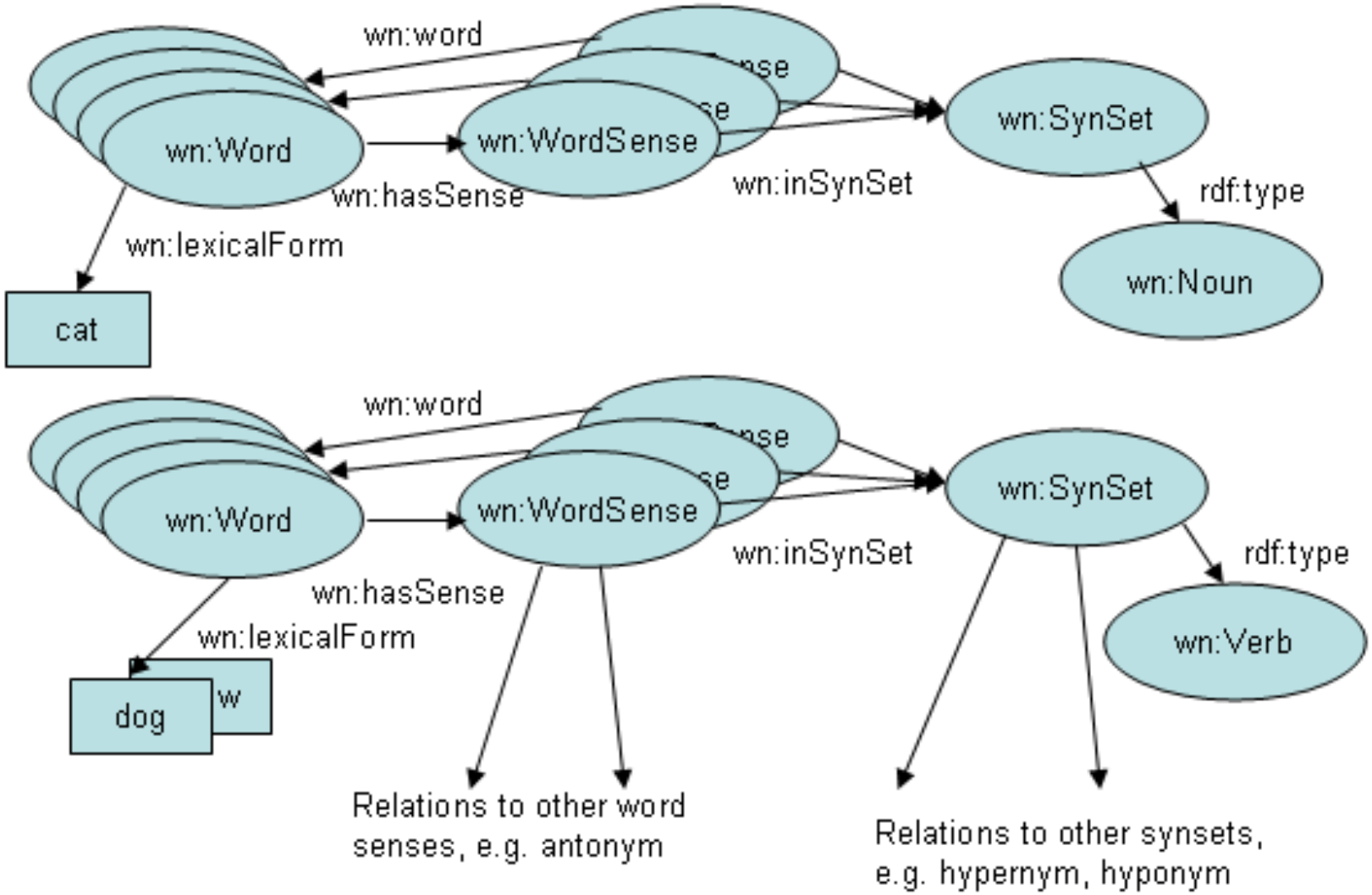
Digression: approaches to word meaning

1. Create a network of all possible word senses, with links between them (e.g. for hyponym, antonym). A word then has a number of these possible senses (WordNet) – this is an expensive undertaking.
2. Try to decompose word senses into complex expressions involving *primitive* concepts (Schank, Jackendoff) – only possible in limited sense areas.

In either case, the senses/concepts need to be related to domain objects (e.g. database fields)

[See later lecture on Semantics]

1. Network of words and senses



2. Decompose meaning into primitives

KILL =
CAUSE TO die =
CAUSE TO NOT LIVE

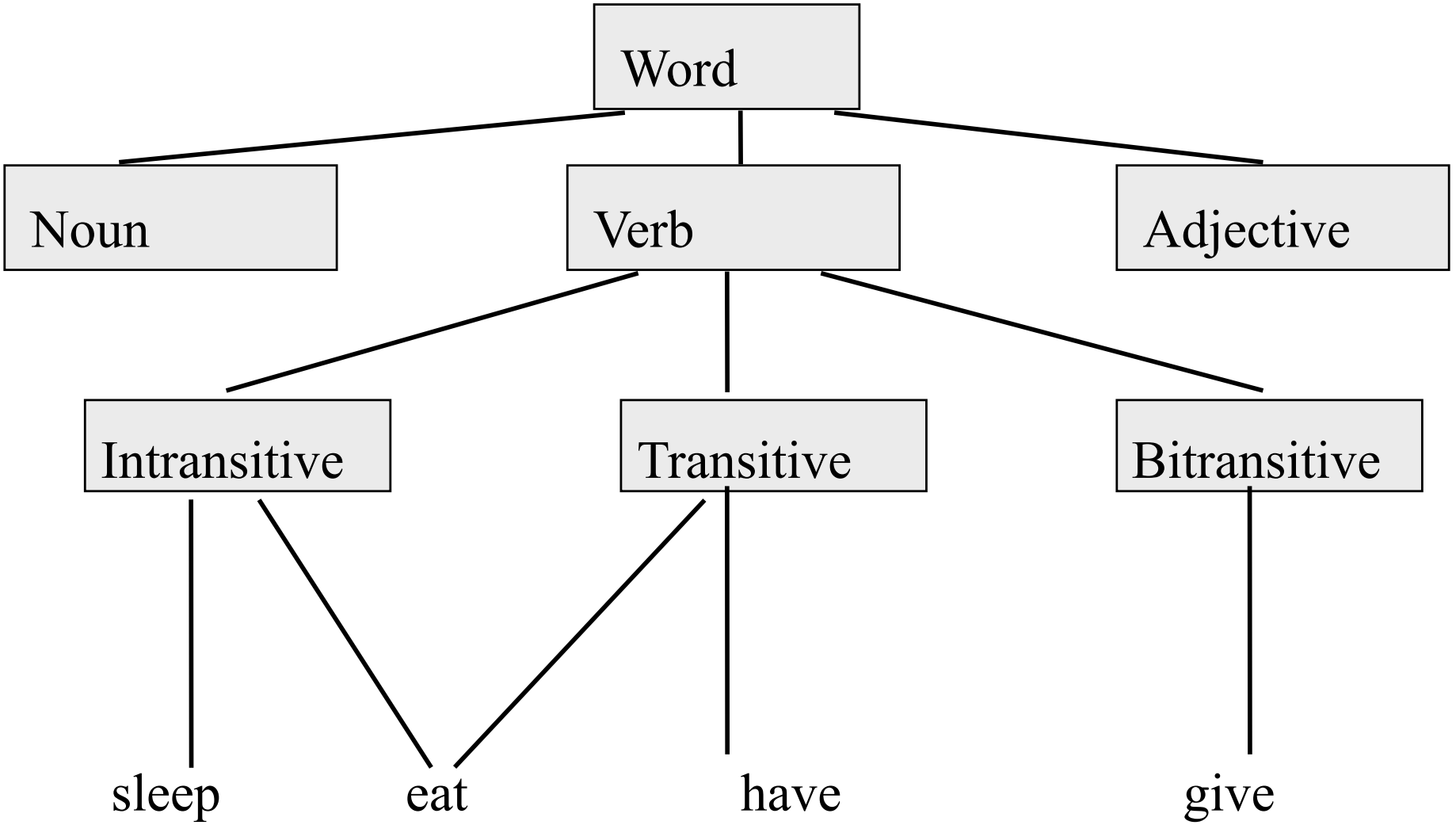
THROW =
project WITH hand =
(CAUSE to MOVE) WITH BODYPART

General inference rules can be formulated in terms of CAUSE, NOT, WITH etc.

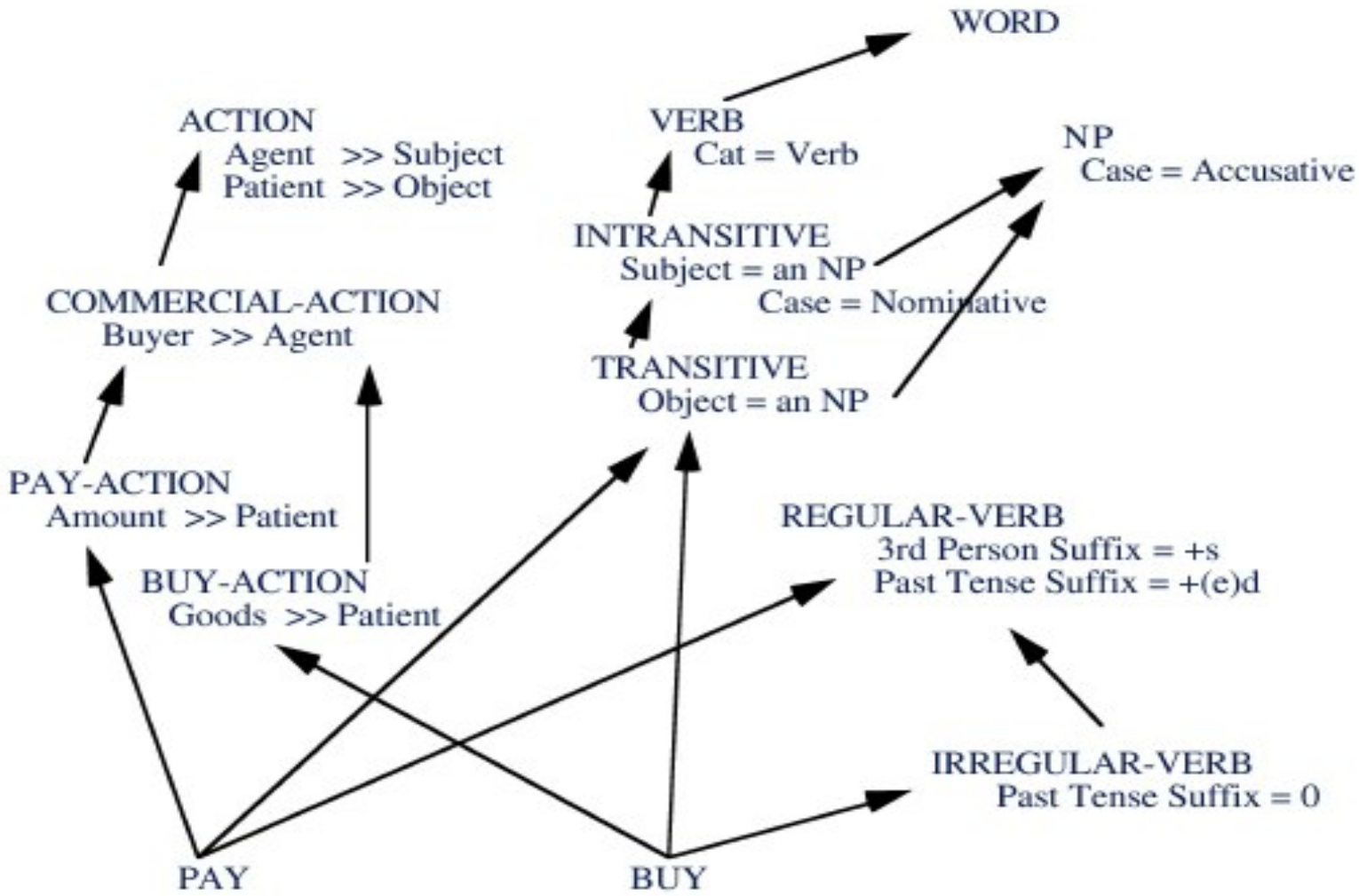
Lexicon Structure

- Object-oriented representation
  - » Noun, Verb, etc are classes
  - » Individual words are instances
  - » Slots (data members) hold information.
- Use of inheritance
  - » Members of a class inherit properties (e.g. values of slots, applicable rules)
  - » Multiple inheritance is common

Example: Lexicon Structure



Detailed Lexical information



Morphology

- Variations of a root form of a word
  - » prefixes, suffixes
- Inflectional morph - same core meaning
  - » plurals, past tense, superlatives
  - » part of speech unchanged
- Derivational morph - change meaning
  - » prefix *re* means do again: *reheat, resit*
  - » suffix *er* means one who: *teacher, baker*
  - » *part of speech changed*
  - » *e.g.* ***Disestablishmentarianism***

Noun Inflections

- Nouns are inflected to form plurals, usually by adding *s*
- Example
  - *chair - Tom has one chair*
  - *chairs - John has 2 chairs*
- Also possessive inflection with *'s*
  - *The boy's mother*

Adjective Inflections

- Adjectives are inflected to form comparative (*er*) and superlative (*est*) forms.
- Example:
  - *fast - A Sun is fast.*
  - *faster - A Sun is faster than a PC.*
  - *fastest - The Cray is the fastest computer*

Verb Inflections

- Tense and aspect
  - » infinitive (root) - *I like to walk to the store*
  - » past (*ed*) - *I walked to the store*
  - » past participle (*ed* or *en*) - *I have walked to the store*
  - » present participle (*ing*) - *I am walking to the store.*
- Agreement with subject
  - » pres/3sing (*s*) - *John walks to the store*

Spelling Changes

- Some spelling changes are automatically made when adding a suffix to a word.
- Delete final “e” when ending starts with a vowel
  - *write* + *ing* = *writing*, not *writeing*
- Change final “y” to “i”
  - *fry* + *ed* = *fried,* not *fryed*

Irregular forms

- Some words have irregular forms that must be looked up in a dictionary
  - » plural of *child* is *children*, not *childs*
  - » past of *eat* is *ate*, not *eated*
- Irregular forms are usually maintained when a prefix is added
  - » past of *rethink* is *rethought*

Tasks

- Morphological analysis
  - » Replace inflected forms by root+inflection
    - *The children ate apples* becomes
    - *The child+s eat+ed apple+s*
- Morphological generation
  - » Replace root+inflection by inflected forms
    - *The child+s eat+ed apple+s* becomes
    - *The children ate apples*

Implementation – 1: stemming

- Standard endings, spelling changes
  - » 2 pages of code
  - » Porter stemmer in Information Retrieval

conditon	suffix	replacement	examples
(*V*)	ed	null	plastered->plaster, bled->bled

- Dictionary of special cases
  - » 1500 special case rules (Sussex morpha)
- More complex processing is needed for languages with complex morphology.

Real Morphology (Turkish)

- Uygarlastiramadiklarimizdanmissinizcasina
  - *uygar -civilized*
  - *+las +BEC (become)*
  - *+tir +CAUS (cause to)*
  - *+ama +NEGABLE (not able)*
  - *+dik +PPART (past tense)*
  - *+lar +PL (plural)*
  - *+imiz +P1PL (first person plural - we)*
  - *+dan +ABL*
  - *+mis +PAST*
  - *+siniz +2PL*
  - *+casina +ASIF*
  - *"(behaving) as if you are among those whom we could not civilize/cause to become civilized"*

Implementation 2: analysis & generation

- Express possible word analyses as simple concatenations of morphemes, e.g. “in+probable+ly” (can express allowable combinations via a finite state automaton)
- Represent the principles of a particular spelling change (e.g. “in+p -> imp”) in a mapping between this and the surface form which enforces this but leaves everything else unchanged
- Mappings can be implemented by finite state transducers, which can efficiently test correctness.

Morphology as tape-tape mapping

TAPE 1	i	n	+	p	r	o	b	a	b	l	e	+	l	Y
TAPE 2	i	m		p	r	o	b	a	b				l	y

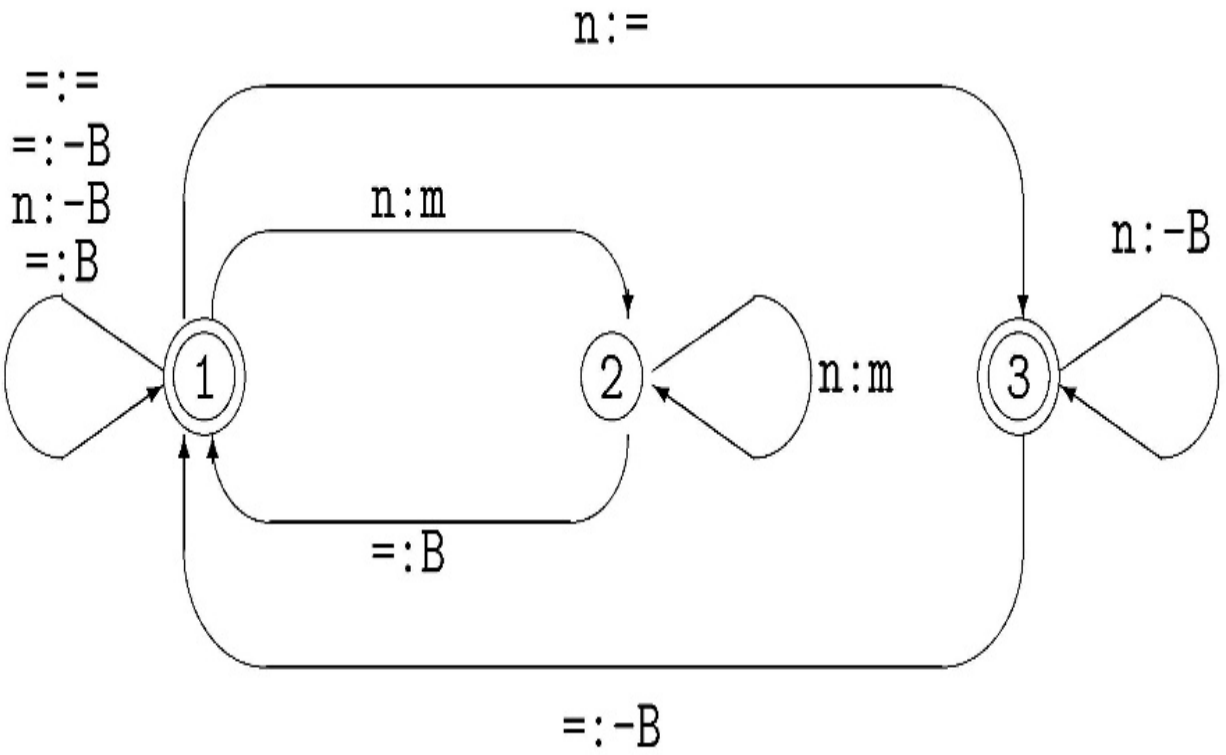
Different (partial) mappings involved:

- n_to_m: Knows about when to legally transform n to m
- y_to_i: Knows about when to legally transform y to i
- …

Finite State Transducers (FSTs)

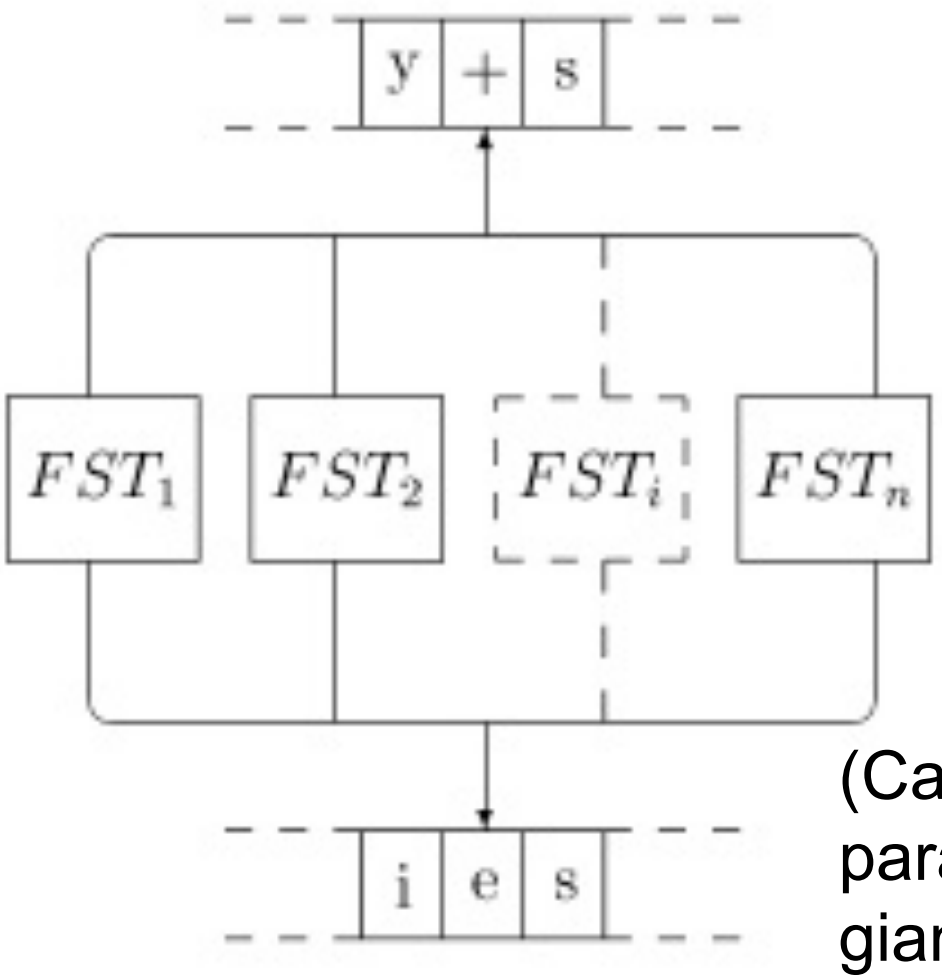
- A *finite state transducer* is like a finite state automaton, except that it accepts *two* tapes, rather than one.
- Each transition has a label **a:b** where a is a symbol to appear on the first tape and b on the second
- Abbreviations can be used to specify sets of symbols (the actual FST will have multiple transitions corresponding to each of these).
- FSTs can be used to express a mapping between the first tape and the second.
- FSTs: J&M (2nd Edition) section 3.4

FST – n to m before B (={b,p})



(Accepts a pair of tapes, and can be used to generate one from the other)

Parallel FSTs for morphology



(Can compile a set of parallel FSTs into one giant FST)

Summary

- The lexicon is a vital part of an NLP system
- Lexicons need to be organised properly to ease creation and maintenance (e.g. object oriented)
- Various information needs to be stored about a word
- Words belong to classes and change in form according to rules of morphology (2 kinds)
- Simple analysis of regular morphology is quite easy for English (Porter stemmer)
- Other languages or more complete coverage may require more sophisticated techniques (FSTs)