Introduction to morphology

- *morpheme*: the minimal information carrying unit
- *affix*: morpheme which only occurs in conjunction with other morphemes
- words are made up of a *stem* (more than one in the case of compounds) and zero or more affixes. e.g., *dog* plus plural suffix +*s*
- affixes: prefixes, suffixes, infixes and circumfixes
- in English: prefixes and suffixes (prefixes only derivational morphology)
- *productivity*: whether affix applies generally, whether it applies to new words

Inflectional morphology

- e.g., plural suffix +*s*, past participle +*ed*
- sets slots in some *paradigm*
- e.g., tense, aspect, number, person, gender, case
- inflectional affixes are not combined in English
- generally fully productive (modulo irregular forms)

Derivational morphology

- e.g., *un*-, *re*-, *anti*- etc
- broad range of semantic possibilities, may change part of speech
- indefinite combinations (e.g., *antiantidisestablishmentarianism*)
- generally semi-productive
- zero-derivation (e.g. *tango*, *waltz*)

Internal structure and ambiguity

Stems and affixes can be individually ambiguous: e.g. *dog* (noun or verb), +*s* (plural or 3persg-verb)

Structural ambiguity:

- *unionised* could be *union -ise -ed* or *un- ion -ise -ed*
- *un- ion* is not a possible form
- *un-* is ambiguous:
  - with verbs: means ‘reversal’ (e.g., *untie*)
  - with adjectives: means ‘not’ (e.g., *unwise*)
- internal structure of *un- ion -ise -ed* has to be *(un- ((ion -ise) -ed))*

Spelling rules

- English morphology is essentially concatenative
- irregular morphology — inflectional forms have to be listed
- regular phonological and spelling changes associated with affixation, e.g.
  - -*s* is pronounced differently with stem ending in *s*, *x* or *z*
  - spelling reflects this with the addition of an *e* (*boxes* etc)
- in English, description is independent of particular stems/affixes

e-insertion

e.g. *box*$\hat{}$*s* to *boxes*

$$\varepsilon \rightarrow \mathrm{e}/ \left\{ \begin{array}{c} \mathrm{s} \\ \mathrm{x} \\ \mathrm{z} \end{array} \right\} \hat{}\ \_\ s$$

- map ‘underlying’ form to surface form
- mapping is left of the slash, context to the right
- notation:

_	position of mapping
ε	empty string
^	affix boundary — stem ^ affix

- corresponds to a finite state transducer

Applications of morphological processing

- compiling a full-form lexicon
- ‘stemming’ for IR
- lemmatization (often inflections only): finding stems and affixes as a precursor to parsing
- generation

Morphological processing may be **bidirectional**: i.e., parsing and generation.

```
sleep + PAST_VERB <-> slept
```

Lexical requirements for morphological processing

- affixes, plus the associated information conveyed by the affix

```
ed PAST_VERB
ed PSP_VERB
s PLURAL_NOUN
```

- irregular forms, with associated information similar to that for affixes

```
began PAST_VERB begin
begun PSP_VERB begin
```

- stems with syntactic categories (plus more) two stage processing, filter results (see lecture 5)
  e.g., *feed* analysed as *fee ^ ed*

Mongoose

A zookeeper was ordering extra animals for his zoo. He started the letter:

> “Dear Sir, I need two mongeese.”

This didn’t sound right, so he tried again:

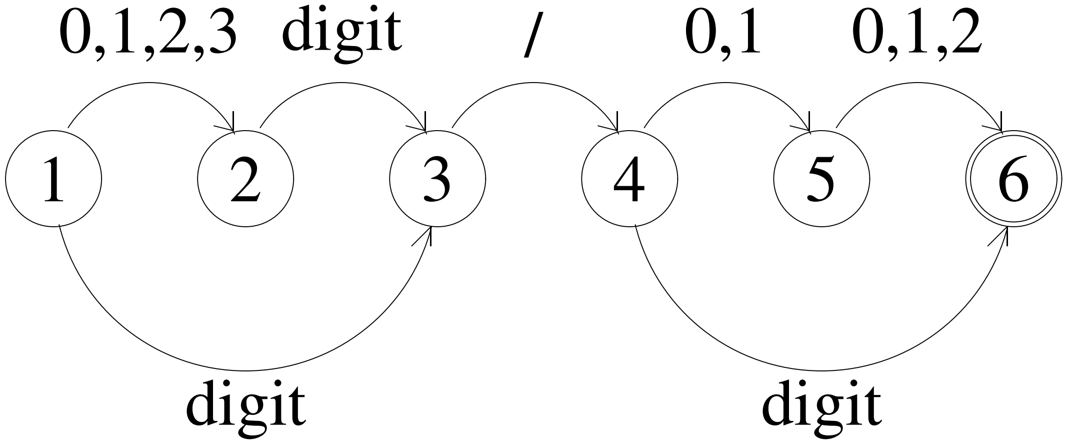
> “Dear Sir, I need two mongooses.”

But this sounded terrible too. Finally, he ended up with:

> “Dear Sir, I need a mongoose, and while you’re at it, send me another one as well.”

Finite state automata for recognition

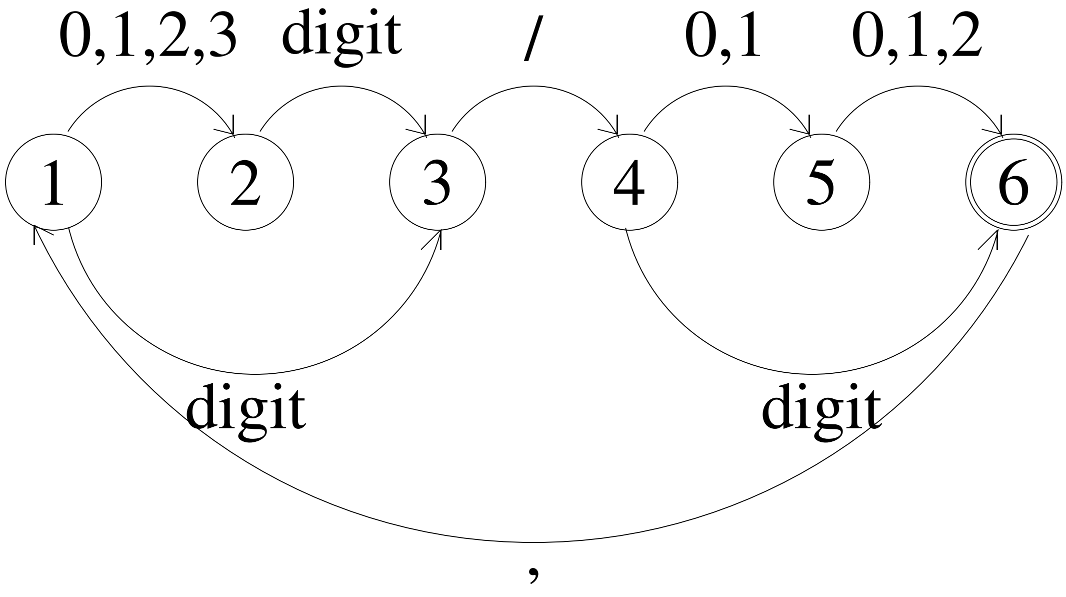
day/month pairs:



- non-deterministic — after input of ‘2’, in state 2 and state 3.
- double circle indicates accept state
- accepts e.g., 11/3 and 3/12
- also accepts 37/00 — overgeneration

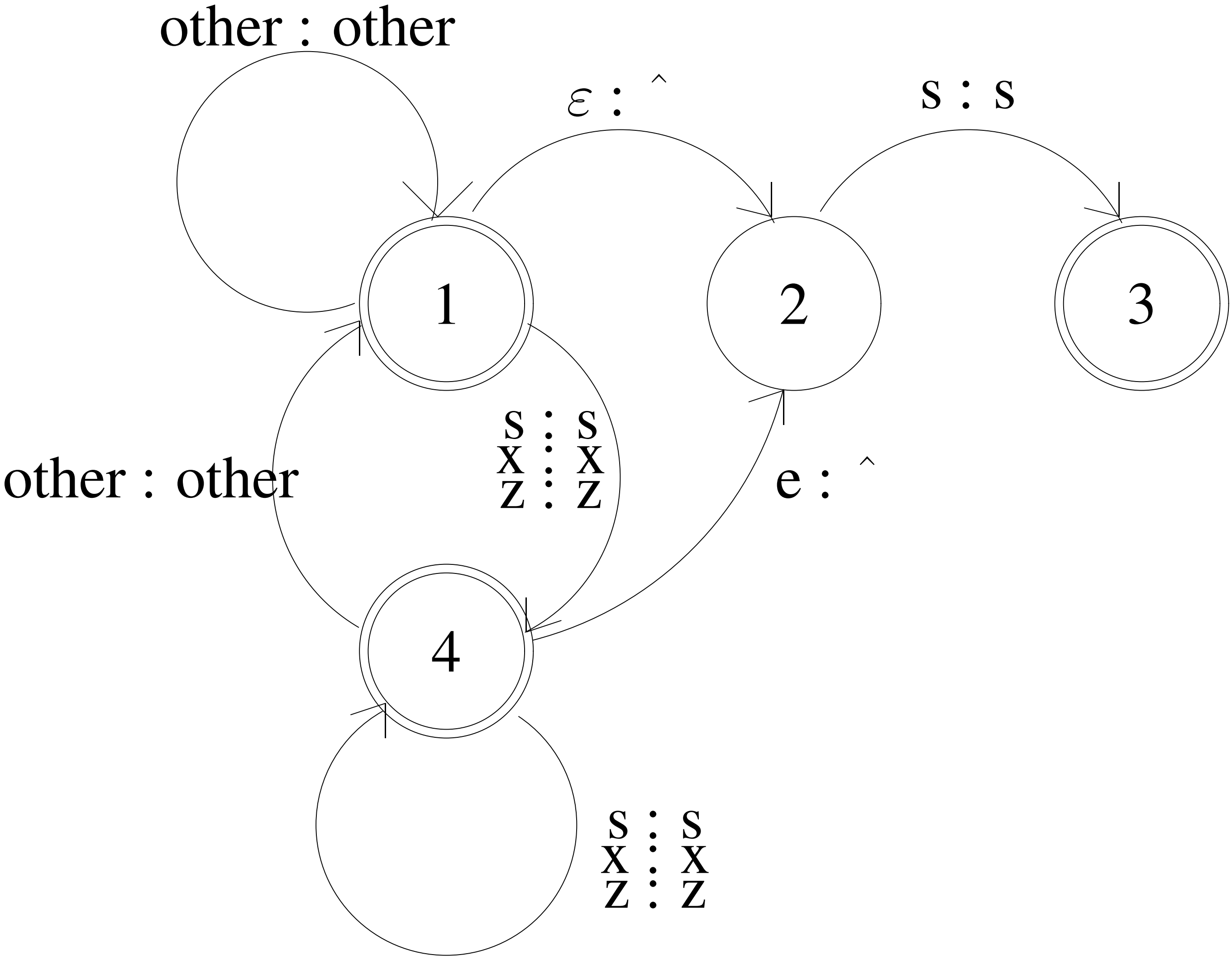
Recursive FSA

comma-separated list of day/month pairs:



- list of indefinite length
- e.g., 11/3, 5/6, 12/04

Finite state transducer

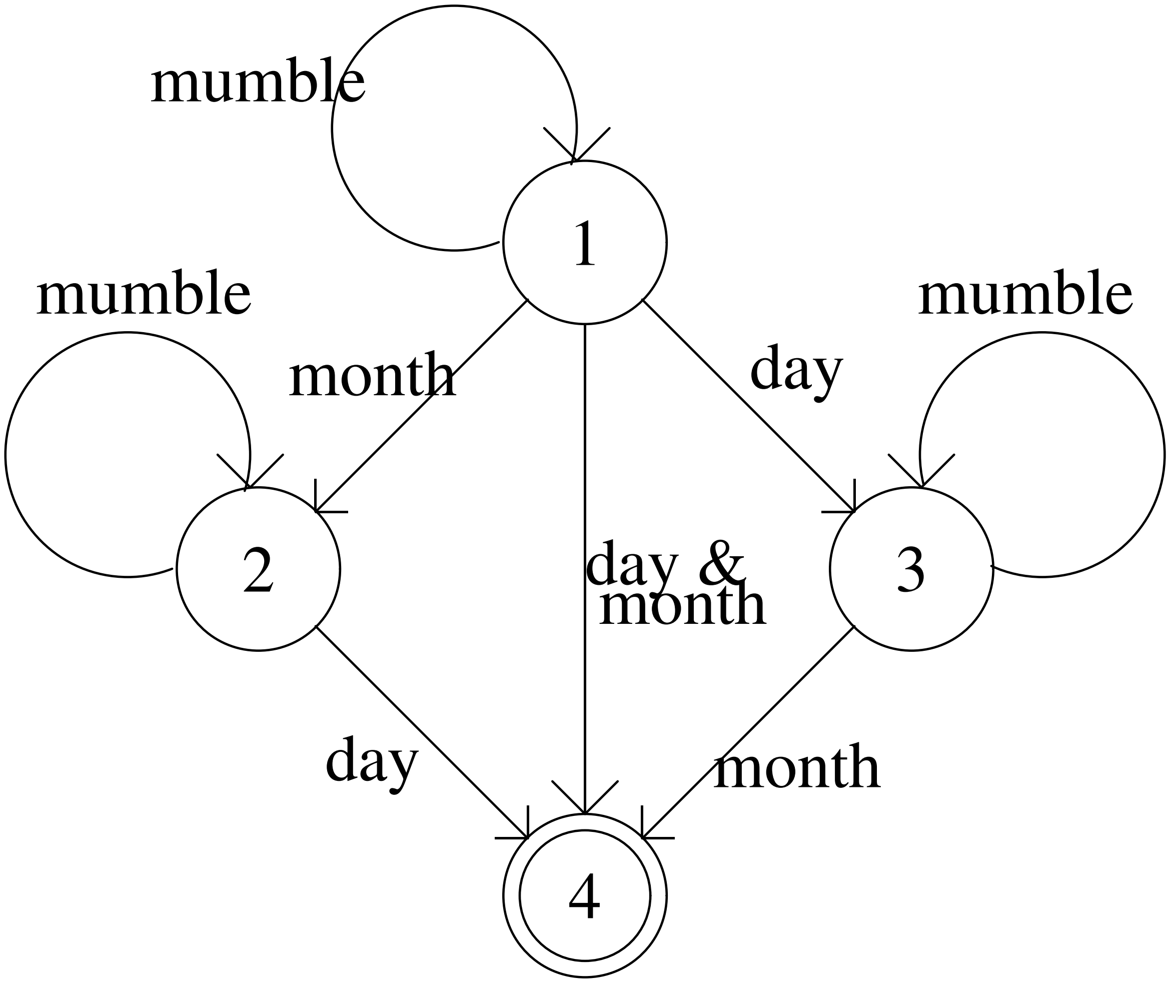


- surface : underlying
- c a k e s ↔ c a k e ^ s
- b o x e s ↔ b o x ^ s

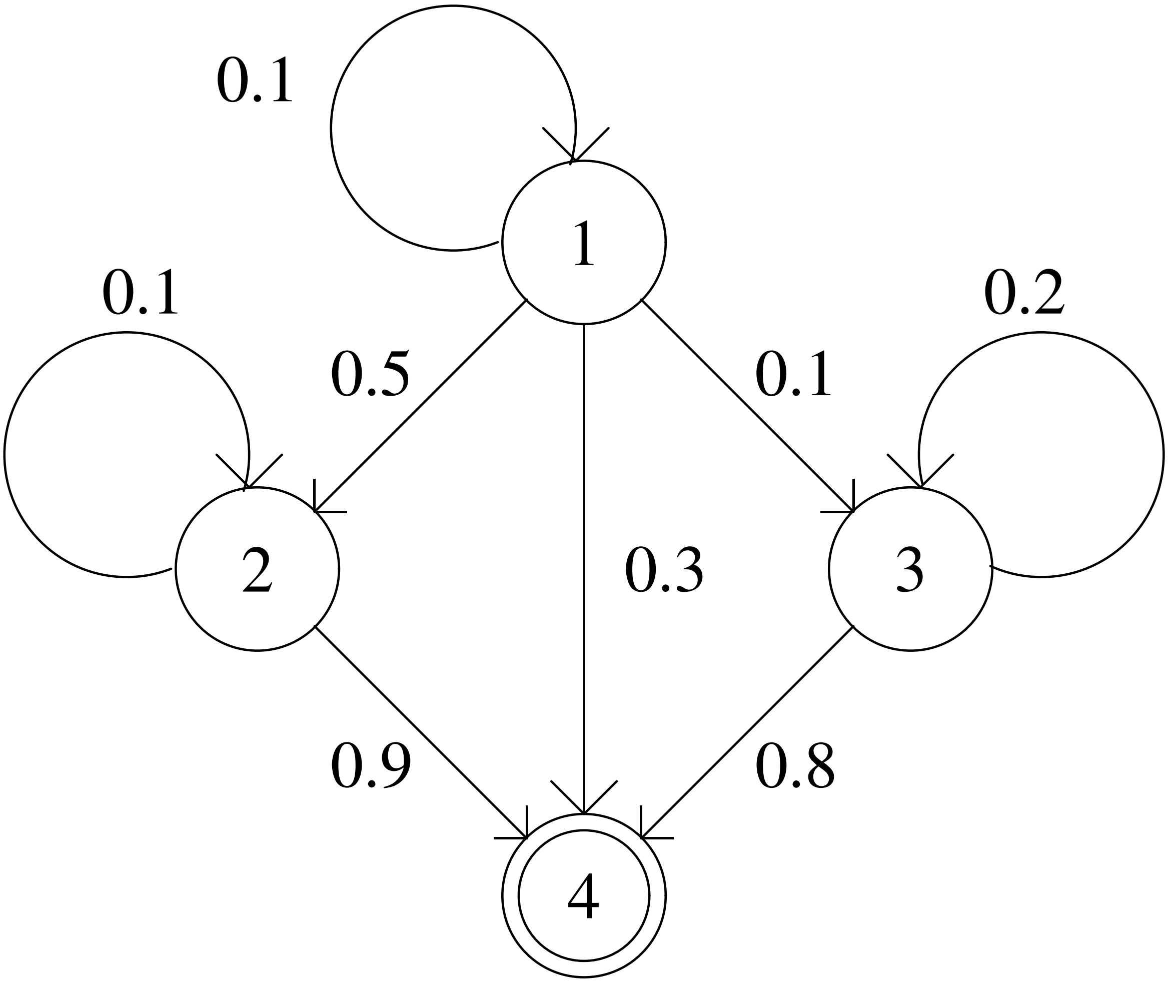
Some other uses of finite state techniques in NLP

- Grammars for simple spoken dialogue systems (directly written or compiled)
- Partial grammars for named entity recognition
- Dialogue models for spoken dialogue systems (SDS)
  e.g. obtaining a date:
  1. No information. System prompts for month and day.
  2. Month only is known. System prompts for day.
  3. Day only is known. System prompts for month.
  4. Month and day known.

Example FSA for dialogue



Example of probabilistic FSA for dialogue


0.1
1
0.1
0.2
0.5
0.1
2
0.3
3
0.9
0.8
4