Some yinkish dripners blorked quastofically into the nindin with the pidibs

words have internal structures

. . . dripn+ER+S blork+ED quastofical+LY into the nindin with the pidib+S

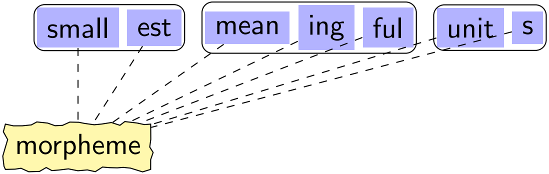
Lecture 2: Morphology and Finite State Techniques

1. A brief introduction to morphology
2. Using morphology in NLP
3. Aspects of morphological processing
4. Finite state techniques

materials mostly by Ann Copestake

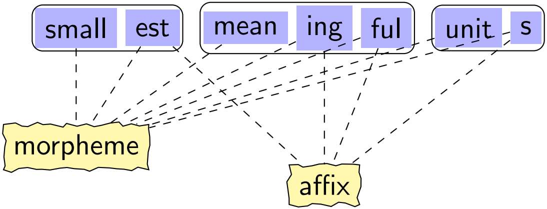
Morpheme

Morphemes are the *smallest meaningful units* of language. Words are composed of morpheme(s).



Morpheme

Morphemes are the *smallest meaningful units* of language. Words are composed of morpheme(s).



Affix: morpheme which only occurs in conjunction with other morphemes.

- suffix (unit*s*), prefix (*in*complete), infix, circumfix

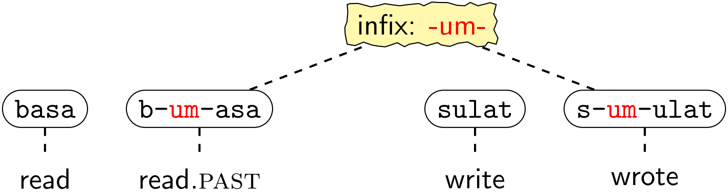
Infix

Tagalog (Philippines)

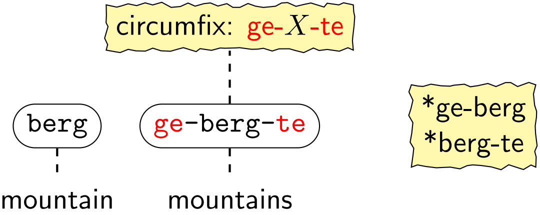
infix: -um-

basa	b-um-asa	sulat	s-um-ulat
read	read.PAST	write	wrote

Infix — Tagalog (Philippines)



Circumfix: occur on both sides — Dutch collectives



Source: J Hana & A Feldman. *ESSLLI 2013: Computational Morphology*.

`http://ufal.mff.cuni.cz/~hana/teaching/2013-esslli/`

Productivity

Productivity: whether affix applies generally, whether it applies to new words

- s*i*ng, s*a*ng, s*u*ng
- r*i*ng, r*a*ng, r*u*ng

Productivity

Productivity: whether affix applies generally, whether it applies to new words

- s*i*ng, s*a*ng, s*u*ng
- r*i*ng, r*a*ng, r*u*ng

- But, *ping*, ping*ed*, ping*ed*

This infixation pattern is not productive: *sing*, *ring* are *irregular*

Inflection and derivation

Inflection creates new forms of the same word

- e.g. *bring*, *brought*, *brings*, *bringing*
- generally fully productive (modulo irregular forms)
- tends to affect only its *syntactic function*

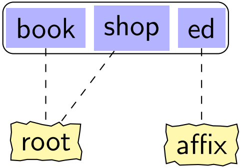
Derivation creates new words

- e.g. *logic*, *logical*, *illogical*, *illogicality*, *logician*, etc.
- generally semi-productive: e.g., *escapee*, *textee*, *?dropee*, *?snoree*, **cricketee* (* and ?)
- tends to be more irregular; the meaning is more idiosyncratic and less compositional.
- tends to affects the *meaning* of the word, and may change part-of-speech

Root: nucleus of the word that affixes attach too.

Compounds contain more than one root.

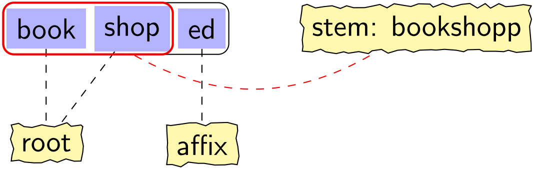
bookshopped



Root: nucleus of the word that affixes attach too.

Compounds contain more than one root.

bookshopped

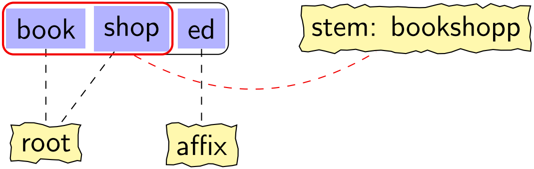


Stem: word without its inflectional affixes = root + all derivational affixes.

Root: nucleus of the word that affixes attach too.

Compounds contain more than one root.

bookshopped



Stem: word without its inflectional affixes = root + all derivational affixes.

Lexeme: the set of all forms related by inflection (but not derivation).

{*bookshops*, *bookshopped*, *bookshopping*, ...}

Lemma: the *canonical/base/dictionary/citation* form of a lexeme chosen by convention.

bookshop (cf. the stem—*bookshopp*)

Compound and multiword expression

Compound

(1) a. beam-width

b. sunset

Multiword expression

Combinations of two or more words that exhibit syntactic and semantic idiosyncratic behavior.

Fixed		**(Syntactically) flexible**
by and large		*put on the clothes* *put the clothes on*
Non-compositional	**Semi-compositional**	**Compositional**
kick the bucket	*spill the beans* (reveal the secret)	*strong tea*

Etymology

slither, *slide*, *slip* etc have some what similar meanings; but *sl-* is not a morpheme.

slith, *slid* and *slip* are historically related.

See www.etymonline.com/word/slide

Internal structure: order

The order of morphemes matters

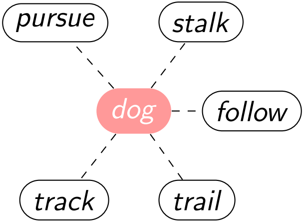
- talk-ed $\neq$ *ed-talk
- re-write $\neq$ *write-re
- un-kind-ly $\neq$ *kind-un-ly

Suffixing is more frequent than prefixing and far more frequent than infixing/circumfixing

- Postpositional and head-final languages use suffixes and no prefixes. cf. harmonic order: $\langle VO, PO\rangle$, $\langle OV, OP\rangle$
- Prepositional and head-initial languages use not only prefixes but also suffixes.
- Many languages use exclusively suffixes and no prefixes
- Very few languages use only prefixes and no suffixes

Internal structure: ambiguity

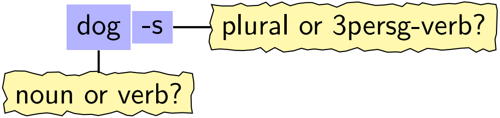
dog



Source of photo: `commons.wikimedia.org/w/index.php?curid=73851814`

Morpheme ambiguity: stems and affixes may be individually ambiguous.

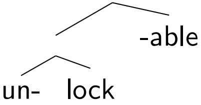
dogs



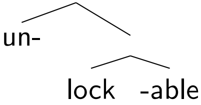
Internal structure: ambiguity

Structural ambiguity: different combinations of morphemes

unlockable



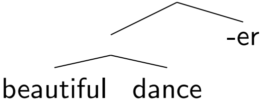
Capable of being unlocked.

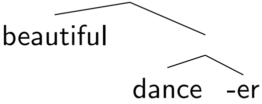


Not capable of being locked.

Cross word boundaries: syntax all the way down

beautiful dancer



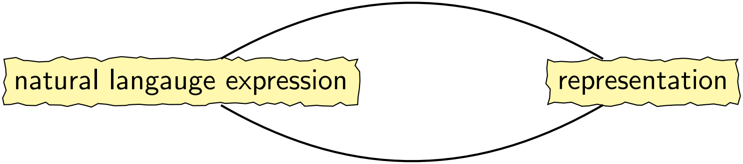


More about *beautiful dancer*: Larson (1998).

`semantics.uchicago.edu/kennedy/classes/f11/na/docs/larson08.pdf`

More about *unlockable*: `en.wiktionary.org/wiki/unlockable`

Abstraction



Surface form → Abstraction

- Indefinite article: *an* orange, *a* building
- Negation: *un*happy, *in*complete, *im*possible, *ir*rational
- Irregular: *sing*, *sang*, *sung*

The same morpheme may have different variants, which are called *allomorphs*. Allomorphs have the same function but different forms.

Computational tasks

	natural langauge expression		representation $\mathcal{R}$
LEMMATIZATION	word *saw*	→	lexeme {*see*, *saw*}
TAGGING	contextualized word *saw @ J saw M*	→	contextualized tag ⟨*see*, VERB.PAST⟩
SEGMENTATION	word *meaningful*	→	morphemes (subwods) *mean+ing+ful*
GENERATION	word *saw*	←	*abstract word* ⟨*see*, VERB.PAST⟩

compiling a full-form lexicon, stemming for Information Retrieval, preprocessing for parsing, . . .

Segmentation

antidisestablishmentarianism $\Rightarrow$ anti- dis- e- stabl -ish -ment -arian -ism

antidisestablishmentarianism

anti	dis	establish	ment	arian	ism

en.wikipedia.org/wiki/Antidisestablishmentarianism

www.etymonline.com/word/antidisestablishmentarianism

Bioinformatics

Text normalization

- Not using any punctuation at all
  Eh speak english mi malay not tt good (Eh, speak English! My Ma-lay is not that good.)
- Using spell-ing/punctuation for emphasis
  goooooood Sunday morning !!!!!! (Good Sunday morning!)
- Using phonetic spelling
  dat iz enuf (That is enough)
- Dropping vowel
  i hv cm to c my luv. (I have come to see my love.)
- Introducing local flavor
  yar lor where u go juz now (yes, where did you go just now?)
- Dropping verb
  I hv 2 go. Dinner w parents. (I have to go. Have dinner with parents.)

Examples are from Aw et al. (2005). `https://www.aclweb.org/anthology/P06-2005.pdf`

More: `noisy-text.github.io/norm-shared-task.html`

Cross-lingual variants

- The phones making up a morpheme don't have to be contiguous, e.g. in Hebrew,

Root	Pattern	PoS	Phonological Form	Gloss
ktb	**C**a**C**a**C**	v	katav	'wrote'
ktb	hi**CC**i**C**	v	hixtiv	'dictated'
ktb	mi**CC**a**C**	n	mixtav	'a letter'
ktb	**CC**a**C**	n	ktav	'writing, alphabet'

from E. Bender's tutorial (`faculty.washington.edu/ebender/papers/100things.pdf`)

Cross-lingual variants

- The phones making up a morpheme don't have to be contiguous, e.g. in Hebrew,

Root	Pattern	PoS	Phonological Form	Gloss
ktb	**C**a**C**a**C**	v	katav	'wrote'
ktb	hi**CC**i**C**	v	hixtiv	'dictated'
ktb	mi**CC**a**C**	n	mixtav	'a letter'
ktb	**CC**a**C**	n	ktav	'writing, alphabet'

from E. Bender's tutorial (`faculty.washington.edu/ebender/papers/100things.pdf`)

- English morphology is essentially concatenative
  cf. duplication in Chinese, e.g.

高兴	→	高高兴兴
happy		高兴高兴

Spelling rules

- 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)

  morphophonology
- In English, description is independent of particular stems/affixes

Lexical requirements for morphological processing

Knowledge

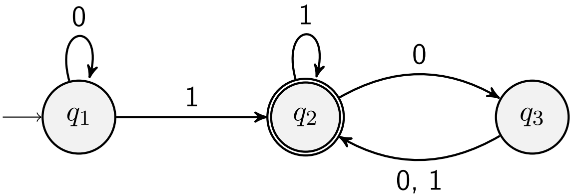
affixes, plus the associated information conveyed by the affix

- -ed VERB.PAST
- -ed VERB.PSP
- -s NOUN.PLURAL

irregular forms, with associated information similar to that for affixes

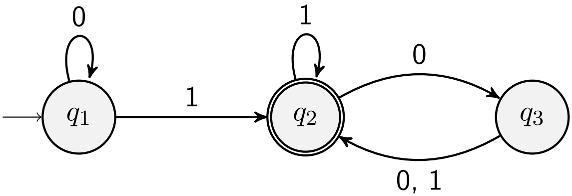
- began VERB.PAST begin
- begun VERB.PSP begin

Automata



- Circles are states of the automaton.
- Arrows are called transitions.
- The automaton changes states by following transitions.
- The double circle indicates that this state is an accepting state. The automaton accepts the string if it ends in an accepting state.

Automata



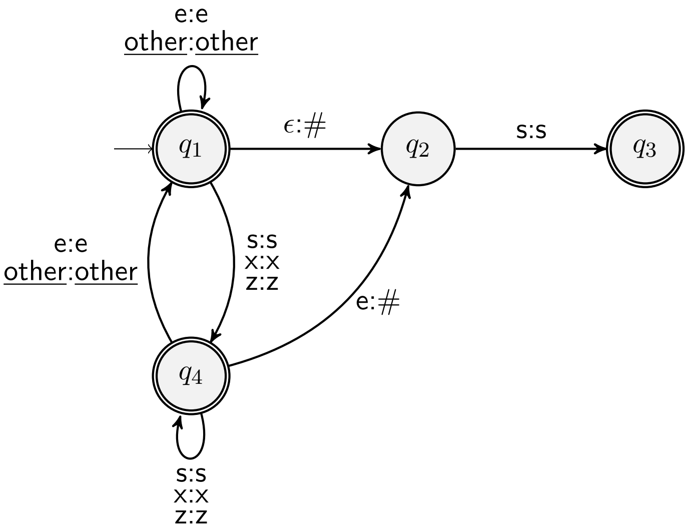
- Circles are states of the automaton.
- Arrows are called transitions.
- The automaton changes states by following transitions.
- The double circle indicates that this state is an accepting state. The automaton accepts the string if it ends in an accepting state.

- **Form Transformation**: agumenting transitions

  `input`$\rightarrow$`input:output`

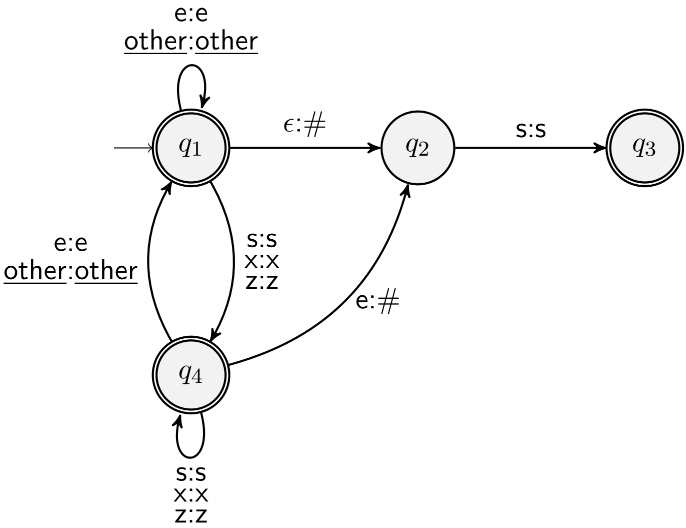
Finite state transducer

- *cakes* → *cake#s*
- *boxes* → *box#s*



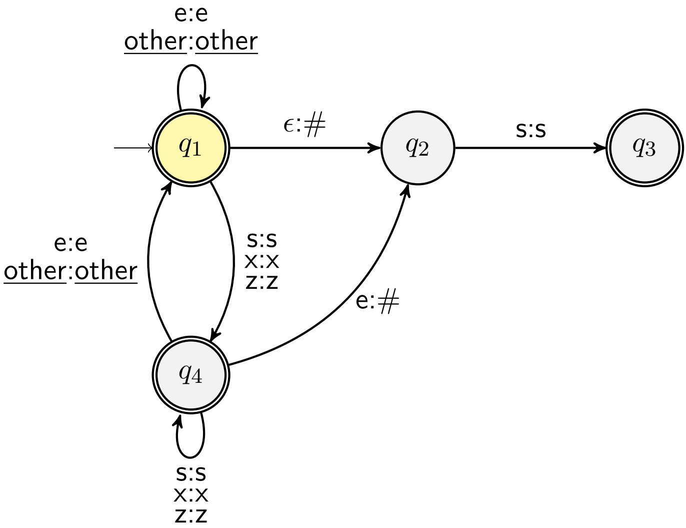
Analysing *boxes*

OUTPUT					
INPUT	b	o	x	e	s



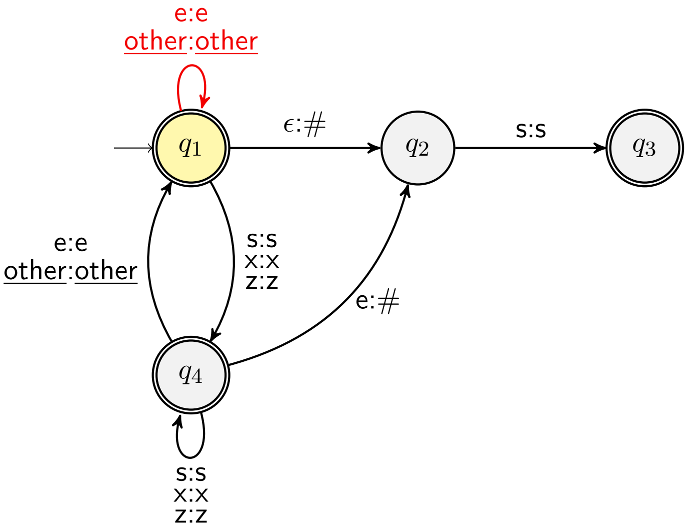
Analysing *boxes*

OUTPUT					
INPUT	b	o	x	e	s



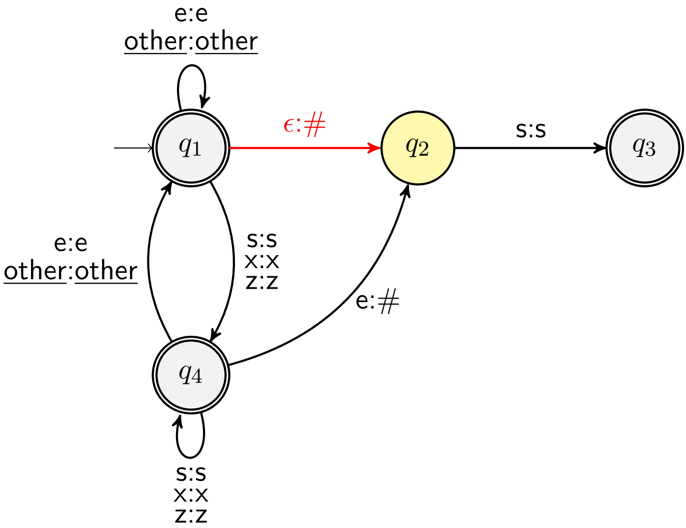
Analysing *boxes*

OUTPUT	b				
INPUT	b	o	x	e	s



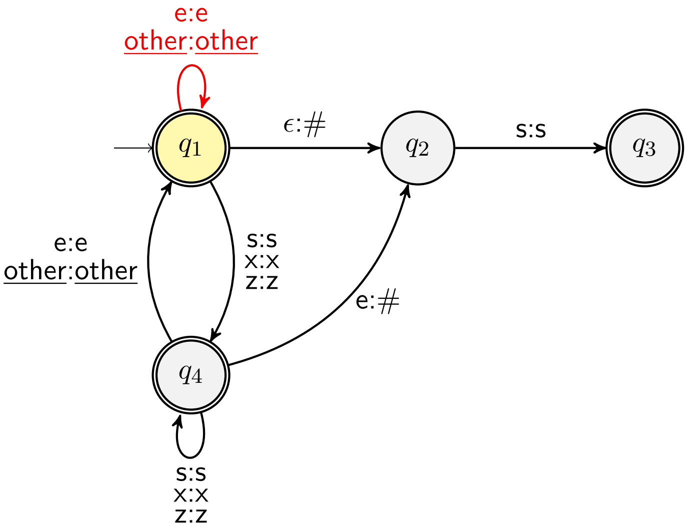
Analysing *boxes*

OUTPUT	b				
INPUT	b	o	x	e	s



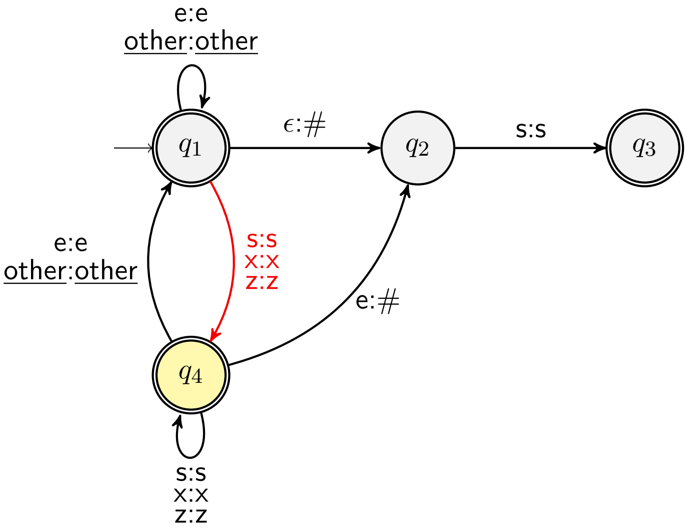
Analysing *boxes*

OUTPUT	b	o			
INPUT	b	o	x	e	s



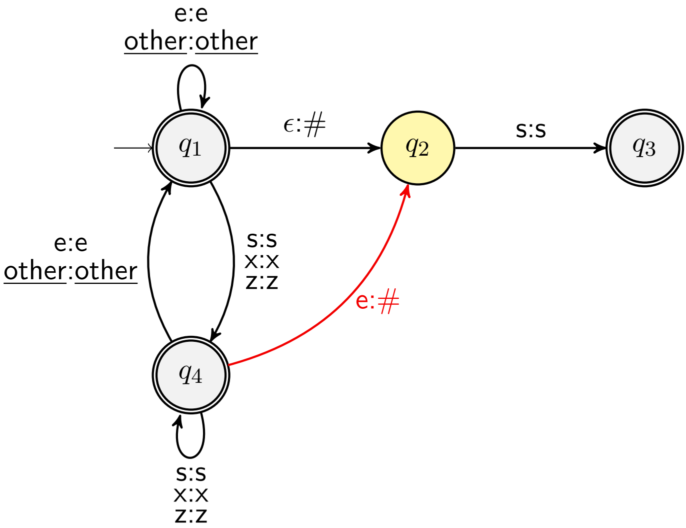
Analysing *boxes*

OUTPUT	b	o	x		
INPUT	b	o	x	e	s



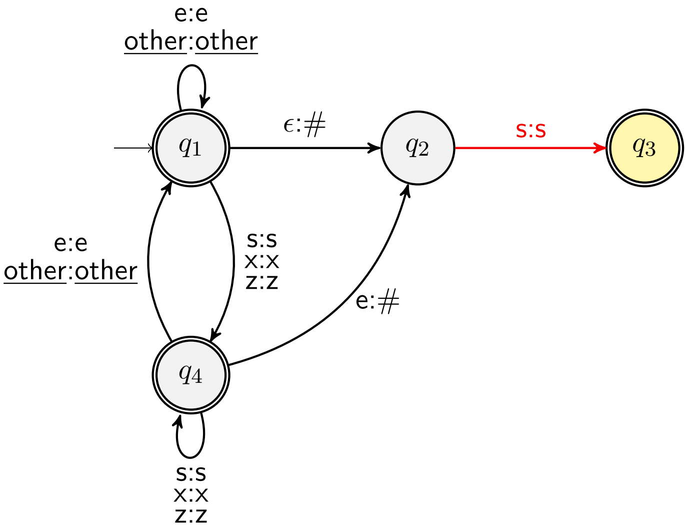
Analysing *boxes*

OUTPUT	b	o	x	#	
INPUT	b	o	x	e	s

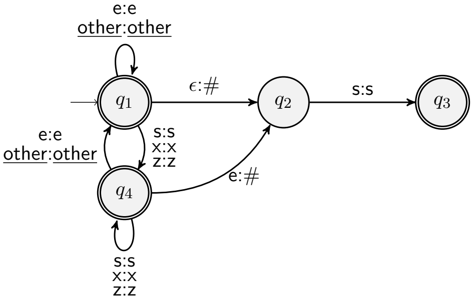


Analysing *boxes*

OUTPUT	b	o	x	#	s
INPUT	b	o	x	e	s

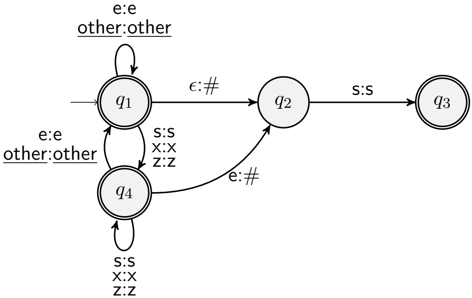


Finite-state machine



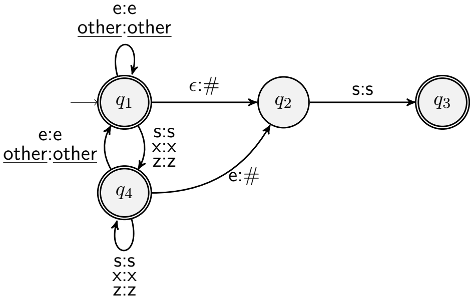
- A symbolic system that can recognize or transform forms.

Finite-state machine



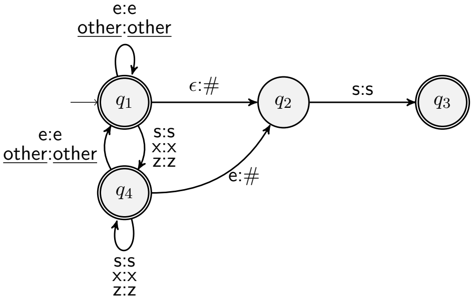
- A symbolic system that can recognize or transform forms.
- An automaton remembers only a finite amount of information.

Finite-state machine



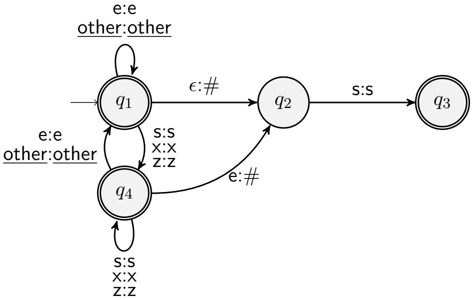
- A symbolic system that can recognize or transform forms.
- An automaton remembers only a finite amount of information.
- Information is represented by its states.

Finite-state machine



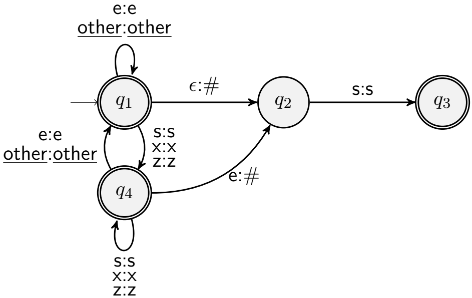
- A symbolic system that can recognize or transform forms.
- An automaton remembers only a finite amount of information.
- Information is represented by its states.
- State changes in response to inputs and may trigger outputs.

Finite-state machine



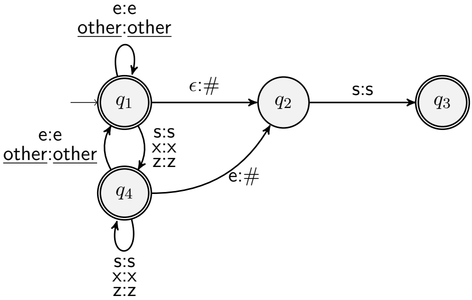
- A symbolic system that can recognize or transform forms.
- An automaton remembers only a finite amount of information.
- Information is represented by its states.
- State changes in response to inputs and may trigger outputs.
- Transition rules define how the state changes in response to inputs.

Finite-state machine



- A symbolic system that can recognize or transform forms.
- An automaton remembers only a finite amount of information.
- Information is represented by its states.
- State changes in response to inputs and may trigger outputs.
- Transition rules define how the state changes in response to inputs.
- Given a sequence of input symbols, a recognition process starts in the start state and follow the transitions in turn. Input is accepted if this process ends up in an accepting state.

Finite-state machine



- A symbolic system that can recognize or transform forms.
- An automaton remembers only a finite amount of information.
- Information is represented by its states.
- State changes in response to inputs and may trigger outputs.
- Transition rules define how the state changes in response to inputs.
- Given a sequence of input symbols, a recognition process starts in the start state and follow the transitions in turn. Input is accepted if this process ends up in an accepting state.
- Partial grammars for text preprocessing, tokenization, named entity recognition etc.

Readings

- Ann's lecture notes
- E. Bender. 100 Things You Always Wanted to Know about Linguistics But Were Afraid to Ask. NAACL-HLT 2012 tutorial. `faculty.washington.edu/ebender/papers/100things.pdf`

* J. Hana & A. Feldman. Computational Morphology. ESSLLI 2013 course. `ufal.mff.cuni.cz/~hana/teaching/2013-esslli/`
* M. Mohri. Finite-State Transducers in Language and Speech Processing. CL 1997 paper. `www.aclweb.org/anthology/J97-2003/`