L90: Overview of Natural Language Processing

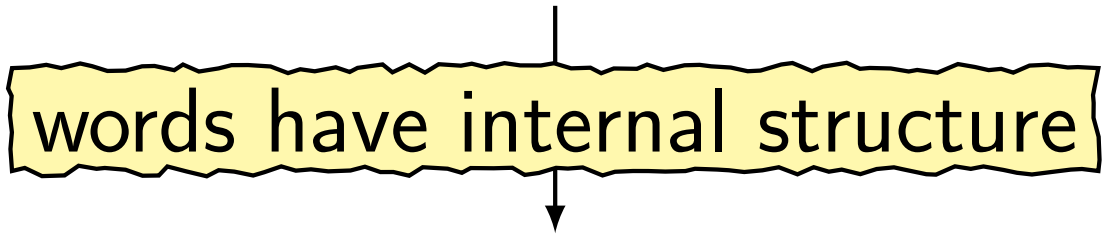
Lecture 2: Morphology and Finite State Techniques

Simone Teufel

Department of Computer Science and Technology
University of Cambridge

Michaelmas 2021/22

Some yinkish dripners blorked quastofically into the nindin with the pidibs



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

Lecture 2: Morphology and Finite State Techniques

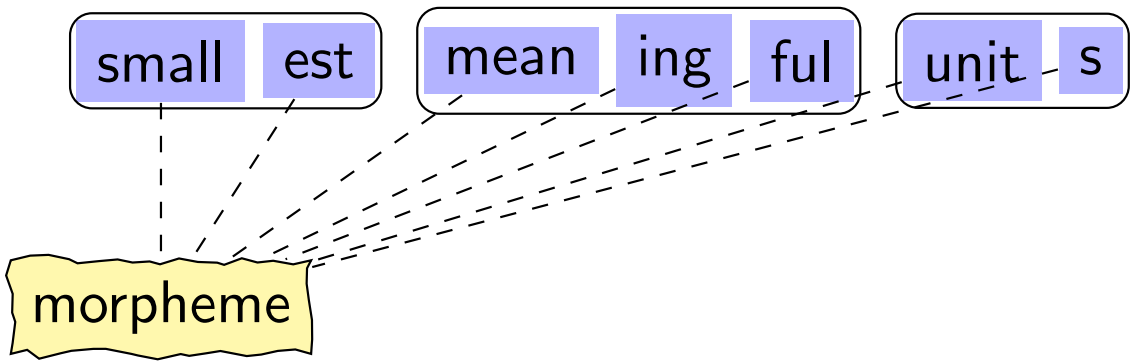
1. A brief introduction to morphology
2. Using morphology in NLP
3. Aspects of morphological processing
4. Finite state techniques

materials mostly by Ann Copestake and Weiwei Sun

Morphology

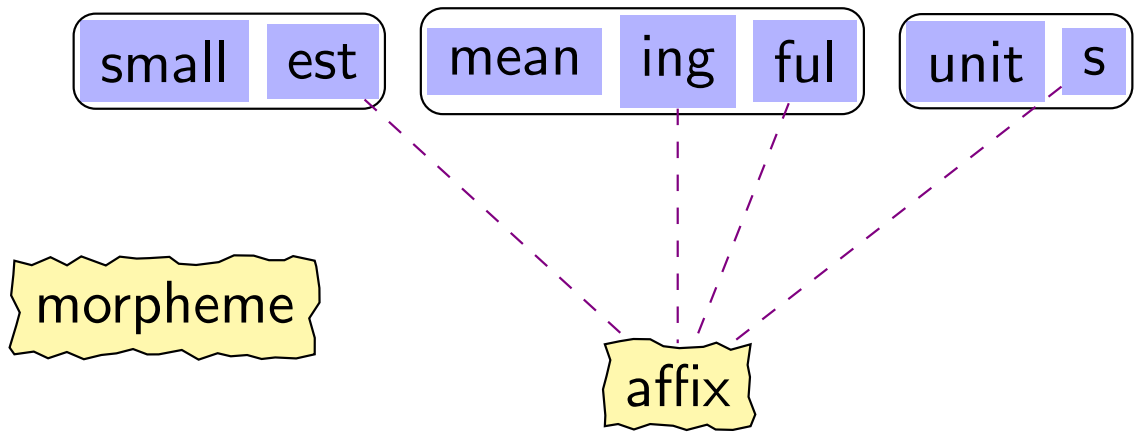
Morpheme

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).

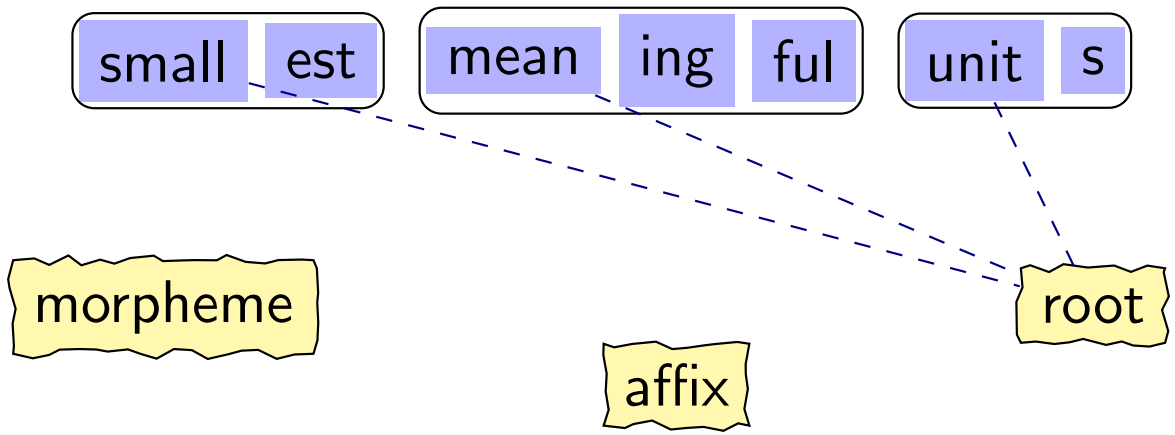


There are two kinds of morphemes:
Affix: morpheme which only occurs in conjunction with other morphemes.

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

Morpheme

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



There are two kinds of morphemes:

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

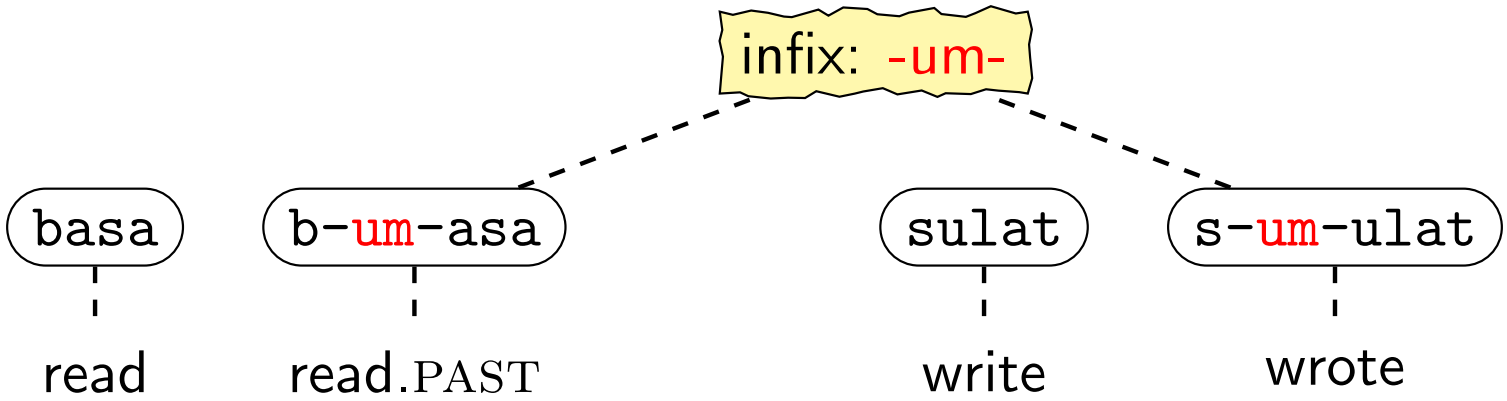
- suffix (unit*s*), prefix (*in*complete), infix, circumfix

Root: nucleus of the word that affixes attach too.

(Also referred to as bound and free morphemes)

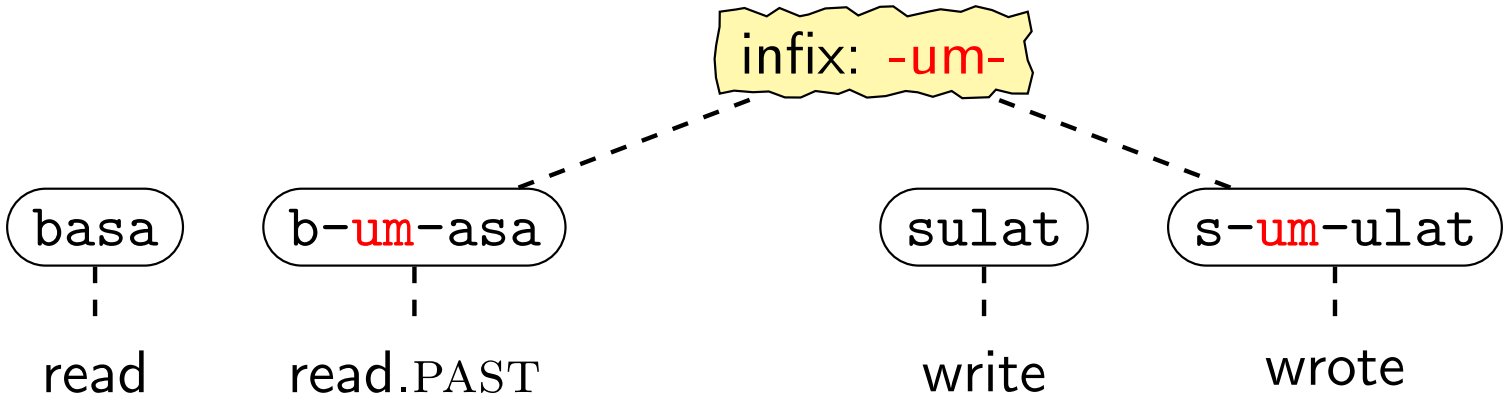
Infix

Tagalog (Philippines)



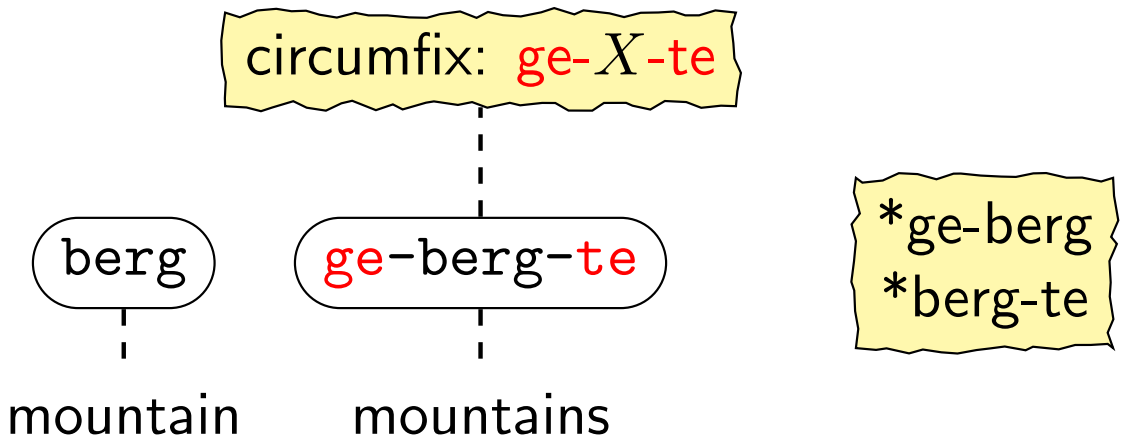
Infix

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
- Infixation pattern

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
- Infixation pattern

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

The infixation pattern is not (no longer) 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., *textee*, *?dropee*, *?escapee*, *?snoree*, **readee* (* and ?)
- tends to affect the *meaning* of the word, and may change part-of-speech
- tends to be more irregular; the meaning is more idiosyncratic and less compositional.

Compound and multiword expression

Root: nucleus of the word that affixes attach to.

Compounds contain more than one root.

(1) a. railway

b. beam-width

c. sunset

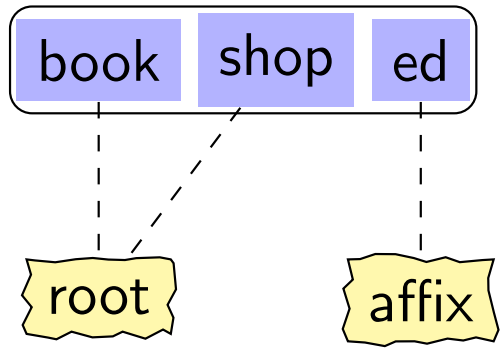
Multiword expression: combinations of two or more words that exhibit syntactic and semantic idiosyncratic behavior.

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 a secret)	*strong tea*

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

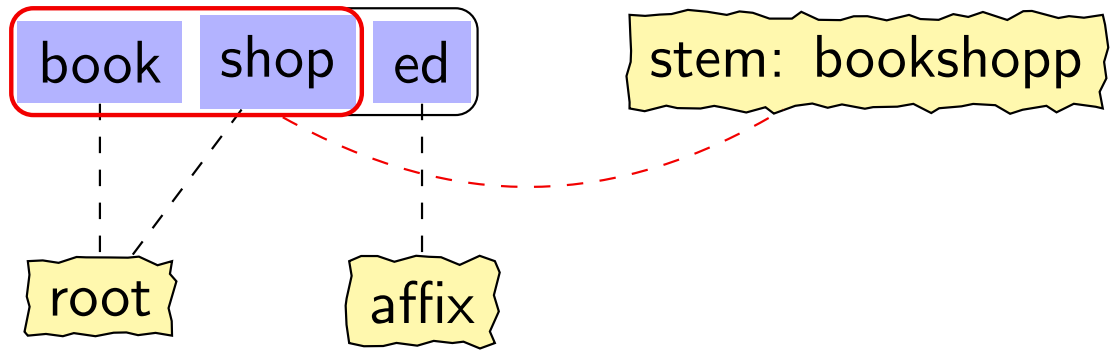
Stem: word without its inflectional affixes = root + all derivational affixes.

bookshopped



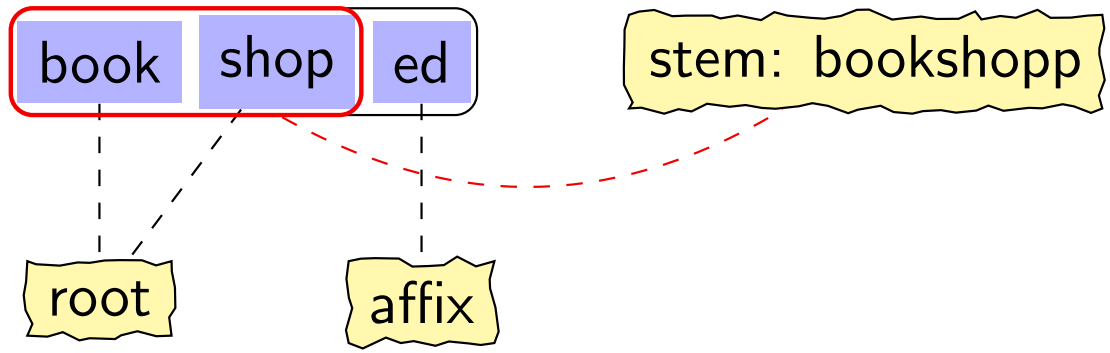
Stem: word without its inflectional affixes = root + all derivational affixes.

bookshopped



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

bookshopped



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

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

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

bookshop (cf. the stem—*bookshopp*)

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

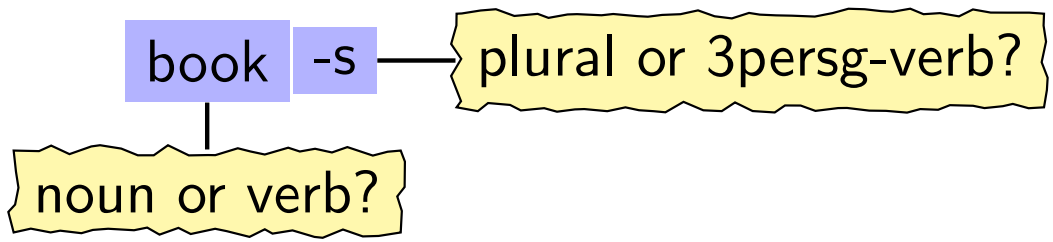
- talk-ed $\neq$ *ed-talk
- re-write $\neq$ *write-re
- un-kind-ly $\neq$ *kind-un-ly

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

- Many languages use exclusively suffixes and no prefixes.
- An example for such languages are postpositional and head-final languages such as Japanese ($\langle OV, OP \rangle$).
- Very few languages use only prefixes and no suffixes
- Prepositional and head-initial languages ($\langle VO, PO \rangle$) use both prefixes and suffixes.

Internal structure: ambiguity

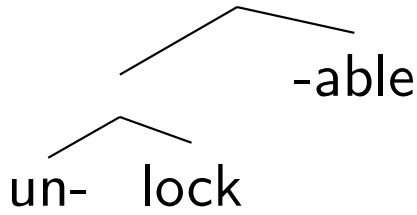
books



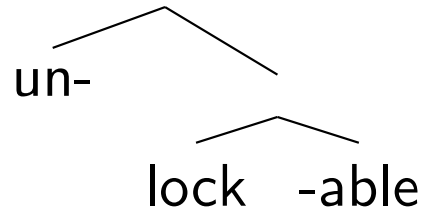
Internal structure: ambiguity

Structural ambiguity: different combinations of morphemes

unlockable



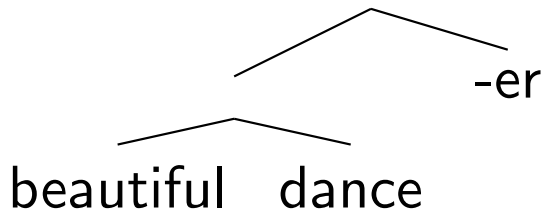
Capable of being unlocked.

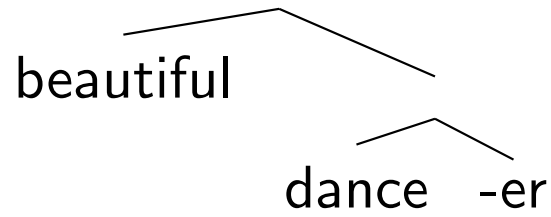


Not capable of being locked.

Cross word boundaries: syntax all the way down

beautiful dancer





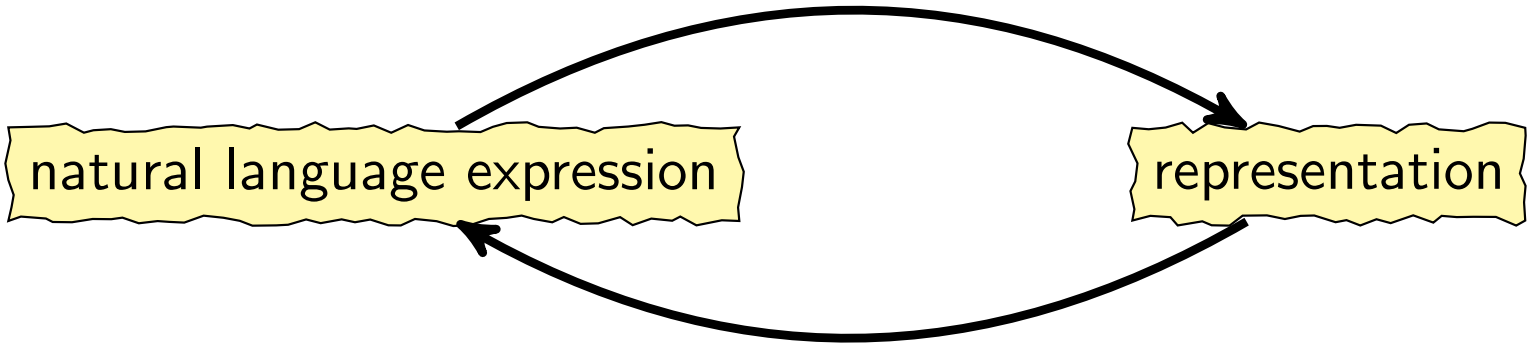
More about *beautiful dancer*: Larson (1998).

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

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

Using Morphology in NLP

Abstraction



Surface form $\rightarrow$ 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.

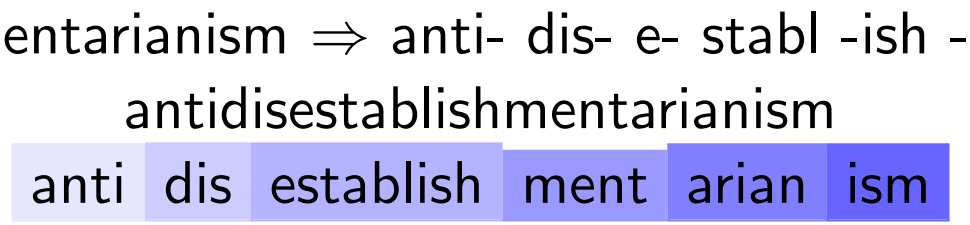
Computational tasks

		natural language 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 (subwords) *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 ⇒ anti- dis- e- stabl -ish -ment -arian -ism



en.wikipedia.org/wiki/Antidisestablishmentarianism

www.etymonline.com/word/antidisestablishmentarianism

important for some applications,
e.g. bioinformatics

Text normalization

- Distorted spelling/punctuation for emphasis
  goooooood Sunday morning !!!!!! (Good Sunday morning!)
- Phonetic spelling
  dat iz enuf (That is enough)
- Dropping vowels
  i hv cm to c my luv. (I have come to see my love.)
- Dropping verb
  I hv 2 go. Dinner w parents. (I have to go. Having 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`

Multiword expression and grammatical errors

(2) a. I tried to *take it on my stride* but I couldn't.
» take it in my stride

b. *By the other side*, I have never climbed a mountain but I always wanted to do it.
» On the other hand

c. However, I told my teacher that I am willing to *give a hand* next time.
» lend a hand

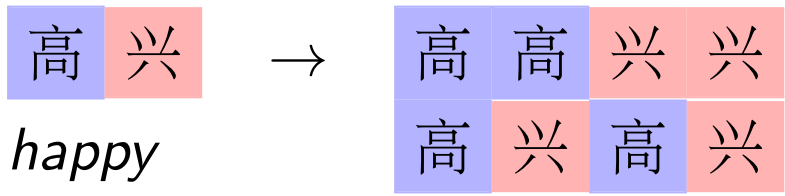
d. It is a *dream becames true* and was really unexpected for me!
» dream come true

from Shiva Taslimipoor

Aspects of Morphological Processing

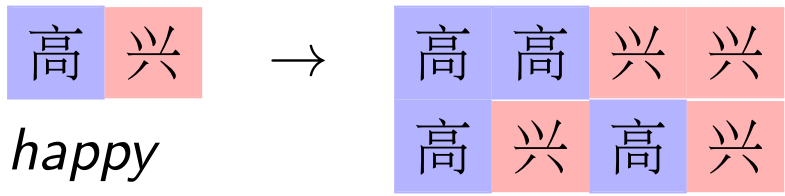
Cross-lingual variants

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



Cross-lingual variants

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



- 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`)

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* is the branch of linguistics that deals with the phonological representation of morphemes.
- 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

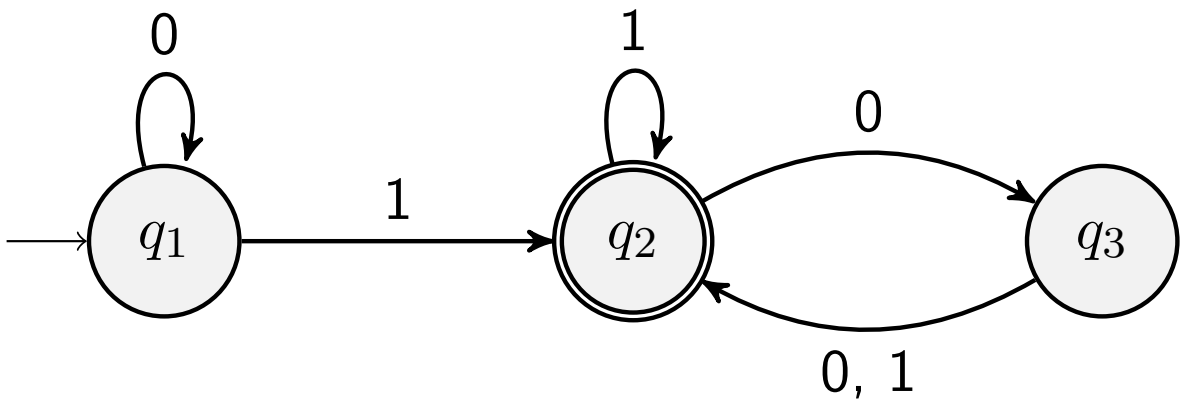
irregular forms, with associated information similar to that for affixes

began VERB.PAST begin

begun VERB.PSP begin

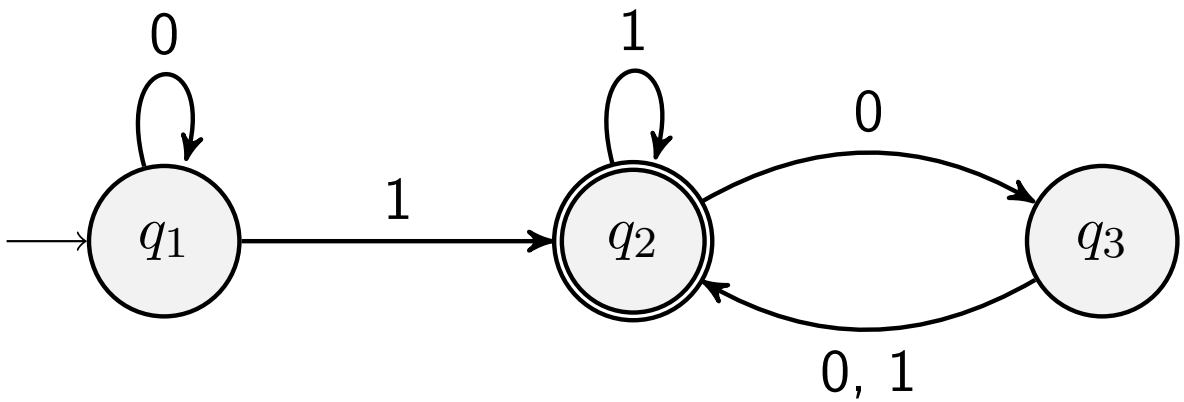
Finite State Techniques

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



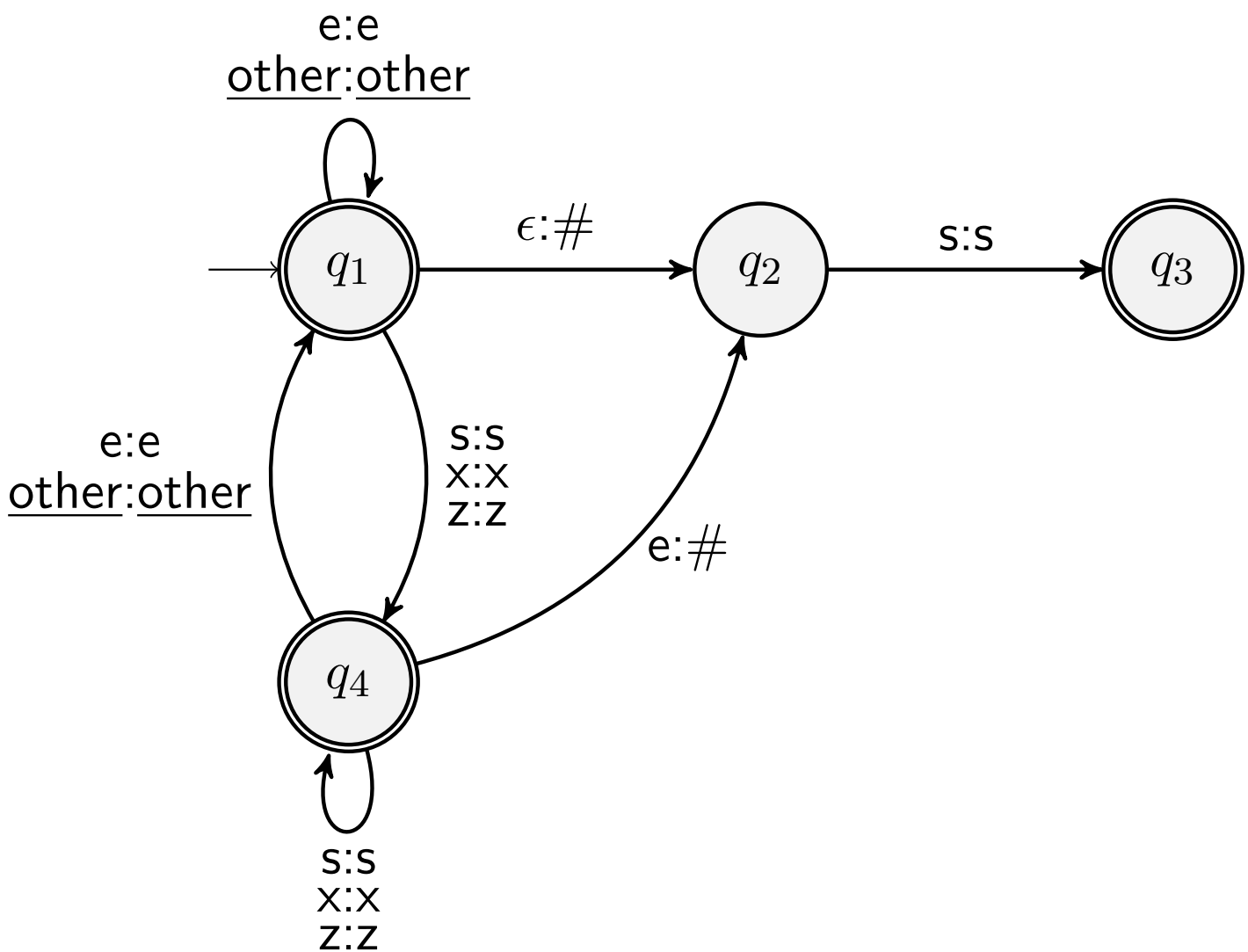
- Circles are states of the automaton.
- Arrows are called transitions.
- The automaton changes states by following transitions.
- The double circle indicates that this state is an accepting state. The automaton accepts the string if it ends in an accepting state.

- **Form Transformation**: augmenting transitions

  `input→input:output`

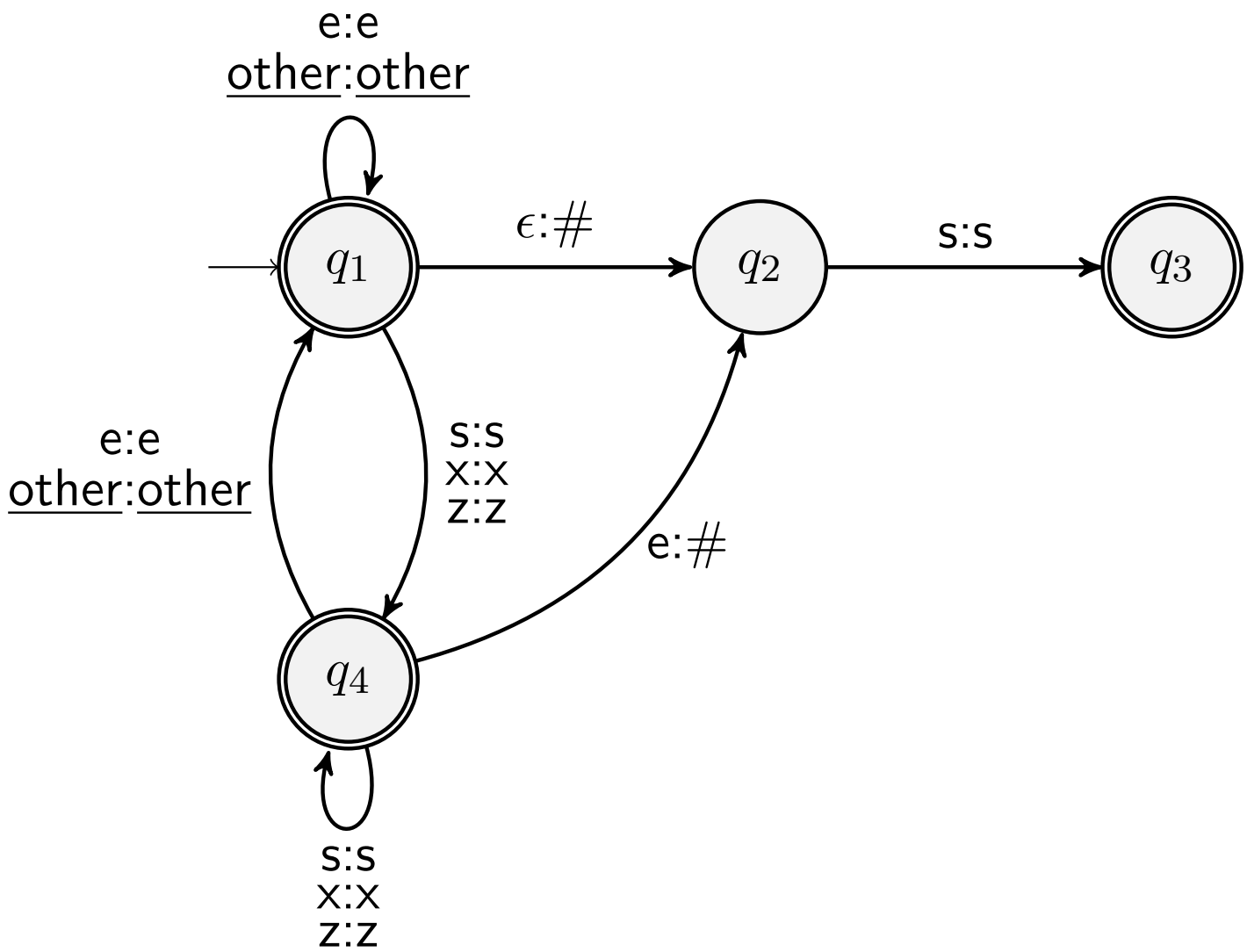
Finite state transducer

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



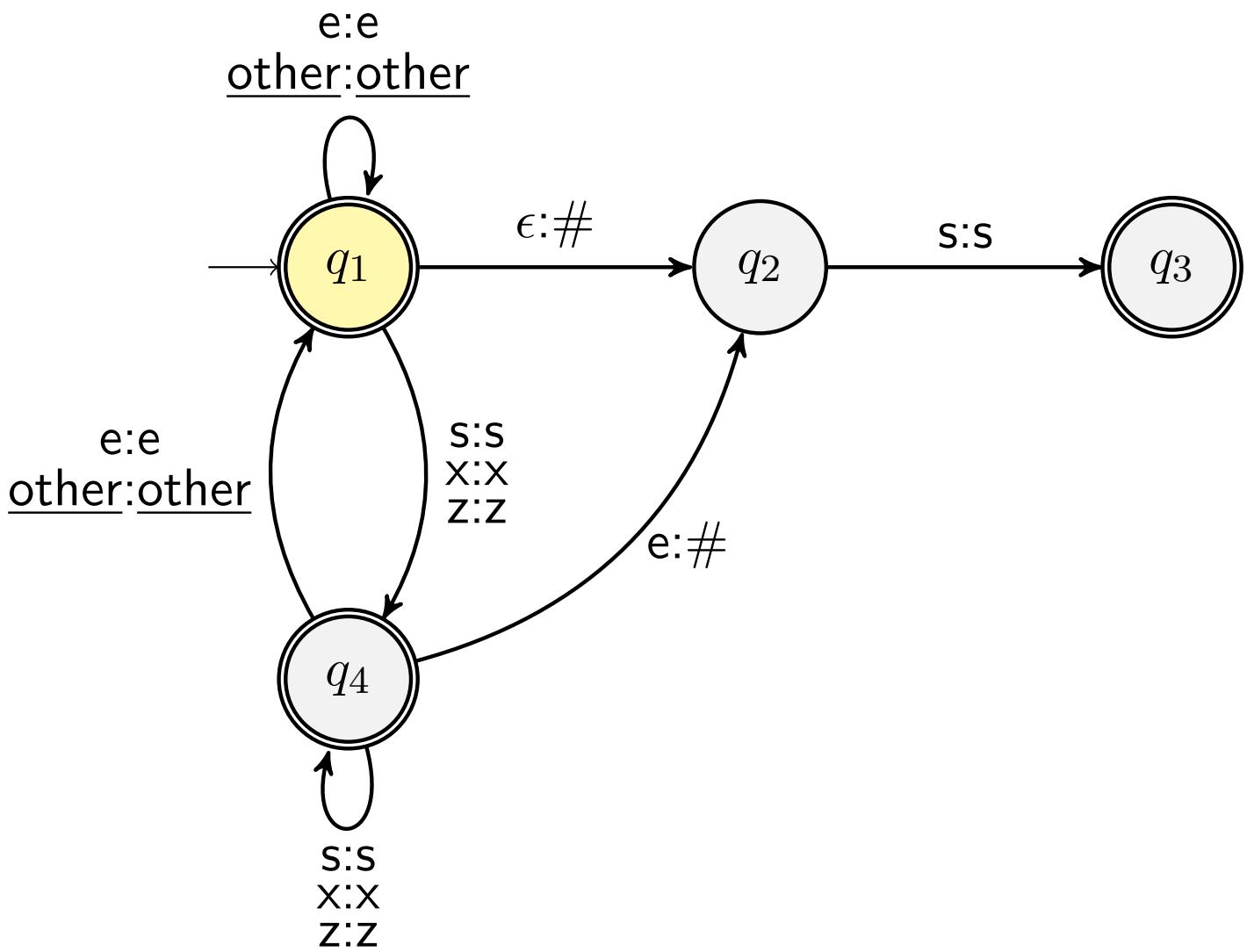
Analysing *boxes*

OUTPUT					
INPUT	b	o	x	e	s



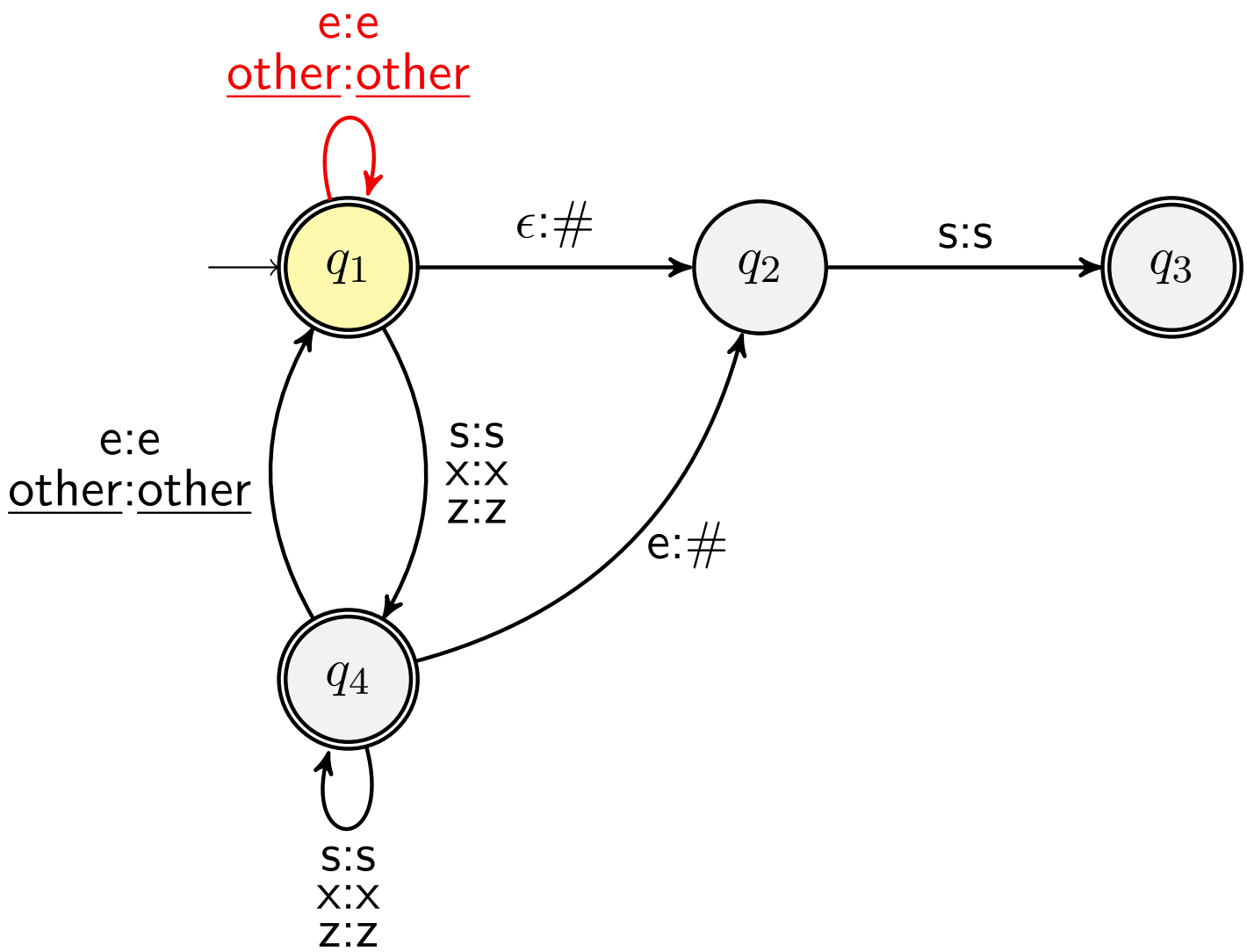
Analysing *boxes*

OUTPUT					
INPUT	b	o	x	e	s



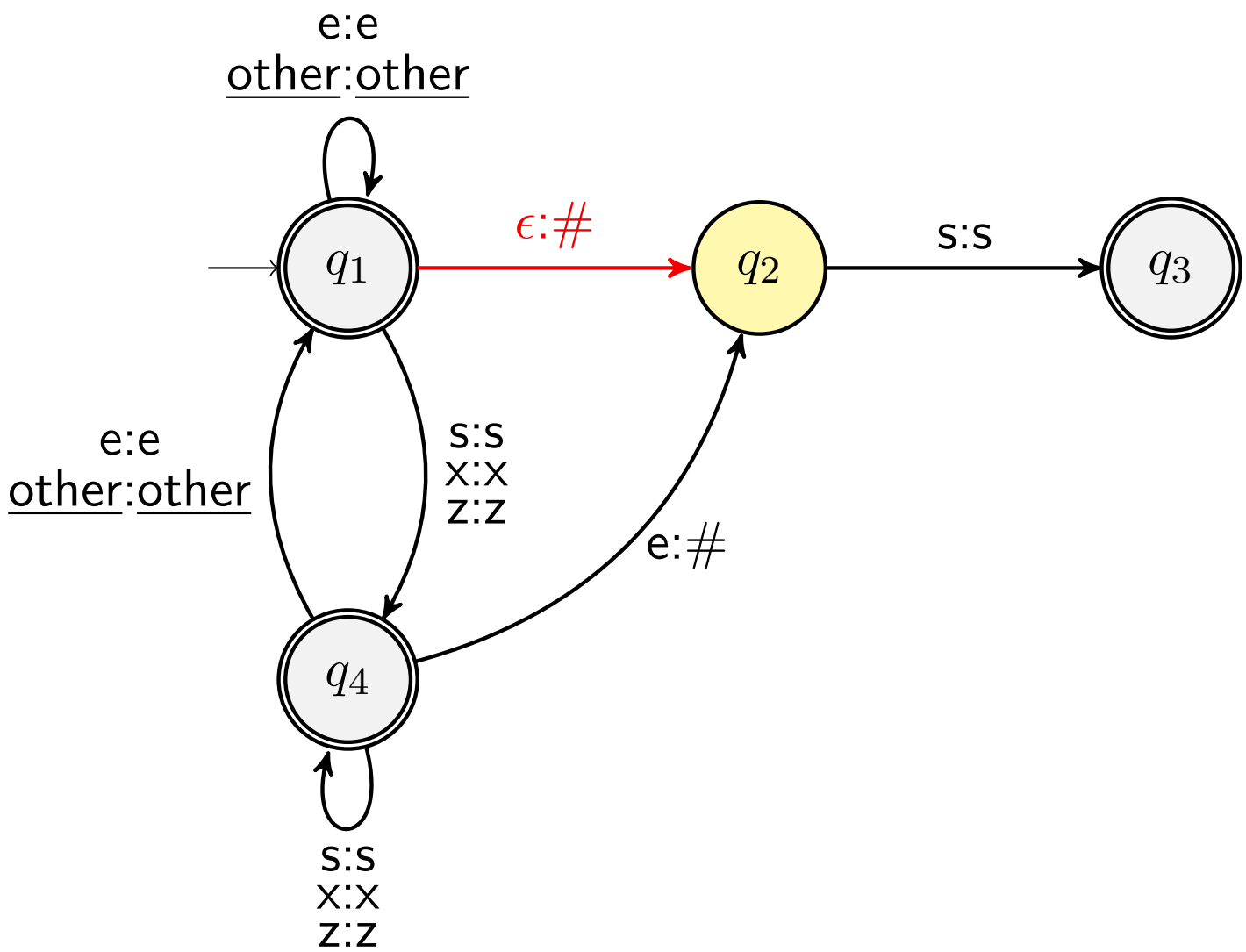
Analysing *boxes*

OUTPUT	b				
INPUT	b	o	x	e	s



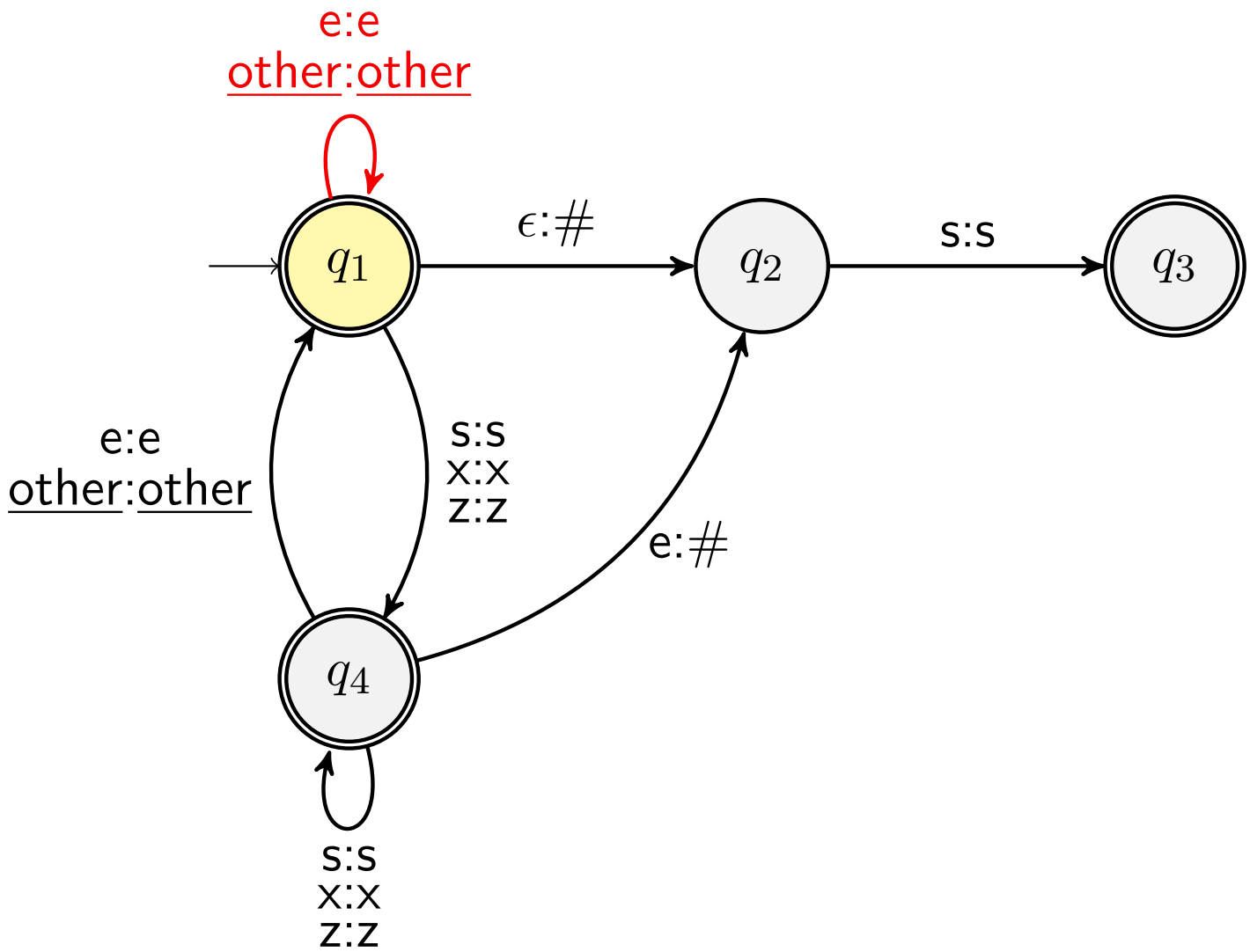
Analysing *boxes*

OUTPUT	b				
INPUT	b	o	x	e	s



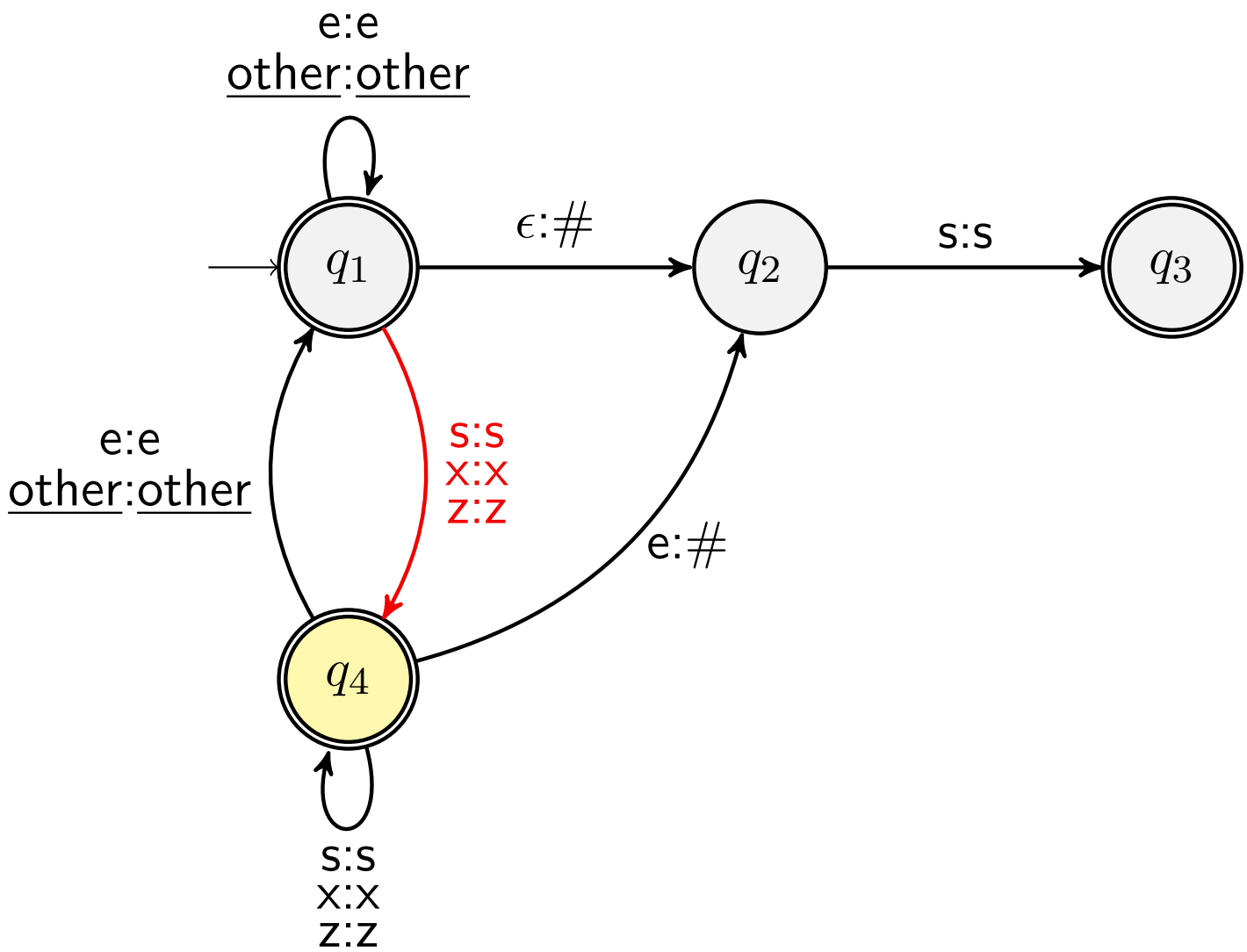
Analysing *boxes*

OUTPUT	b	o			
INPUT	b	o	x	e	s



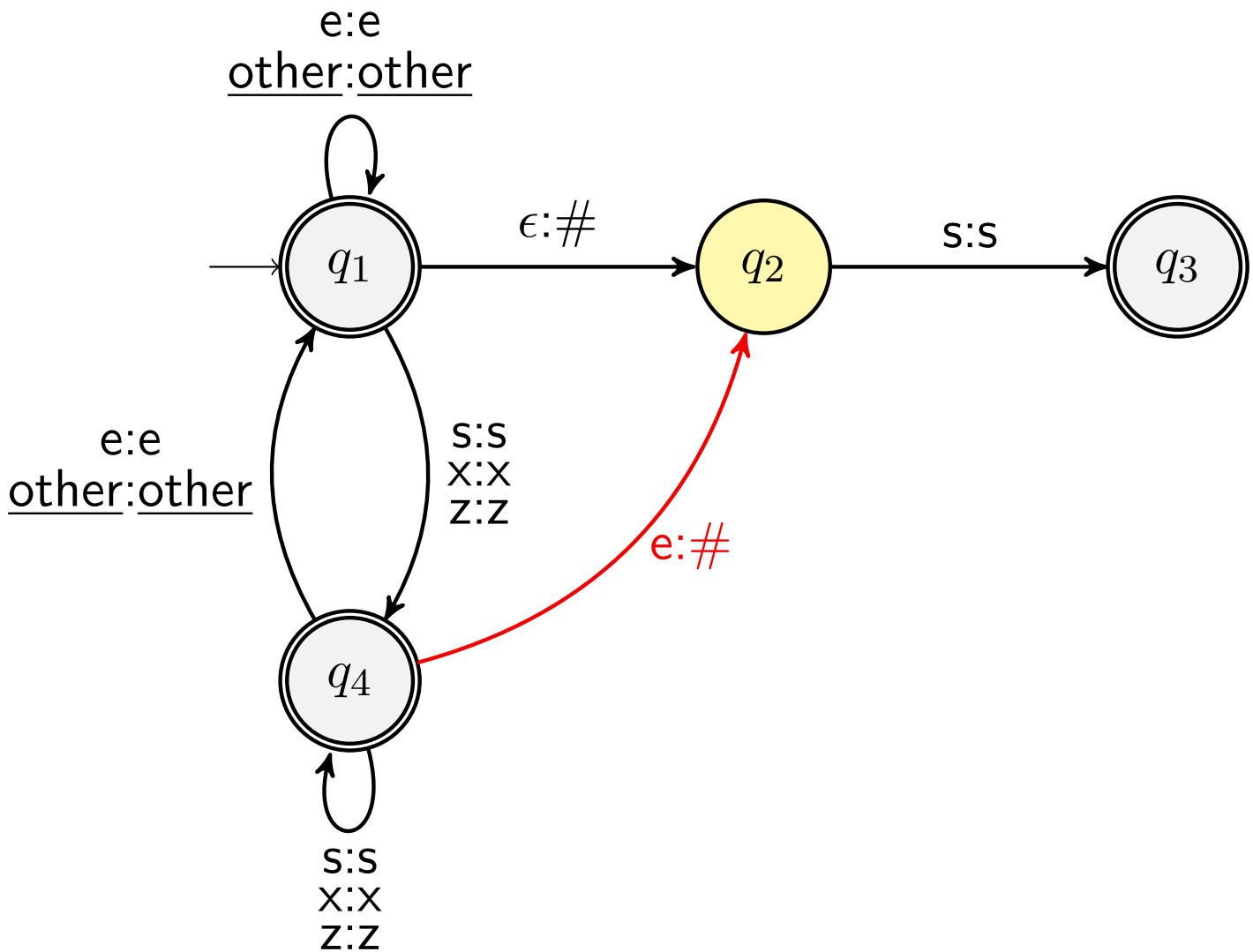
Analysing *boxes*

OUTPUT	b	o	x		
INPUT	b	o	x	e	s



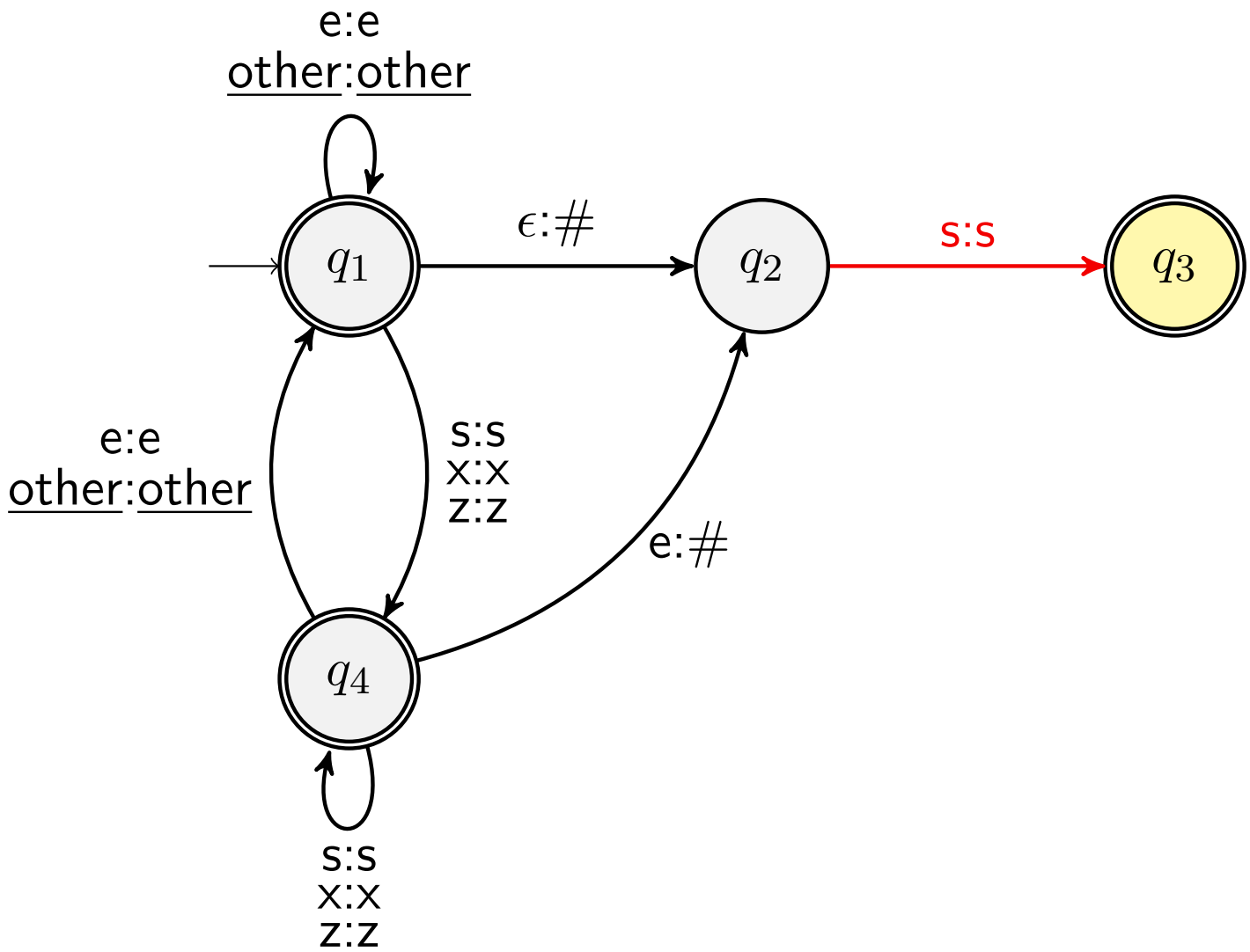
Analysing *boxes*

OUTPUT	b	o	x	#	
INPUT	b	o	x	e	s



Analysing *boxes*

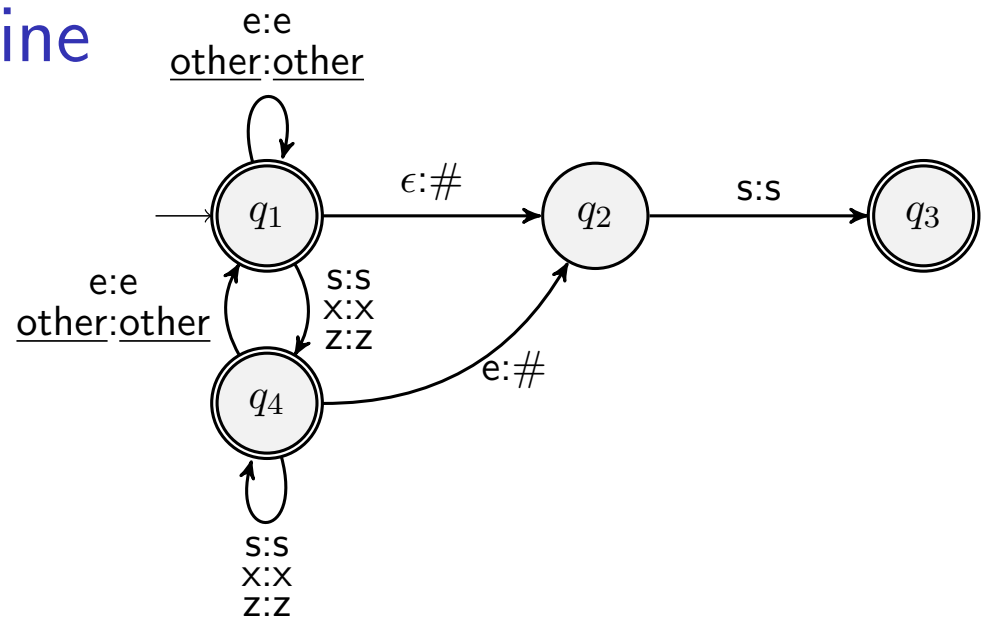
OUTPUT	b	o	x	#	s
INPUT	b	o	x	e	s



Finite-state techniques in NLP

- Some sub-languages are finite
- Morphology, named entities, numbers . . .
- Finite state technology can also provide partial grammars for text preprocessing, e.g. tokenization.

Reminder: 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 follows the transitions in turn.
- Input is accepted if this process ends up in an accepting state.

Reading

Required

- Ann Copestake's lecture notes. `https://www.cl.cam.ac.uk/teaching/1920/NLP/materials.html`
- 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`

Optional

* 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/`