Introduction to Morphology

Linguistics for Computer Scientists
Session 4

Antske Fokkens

Department of Computational Linguistics
Saarland University

03 October 2009

Outline

Outline

What is Morphology?

Morphology is the study of form and structure.

In linguistics, it generally refers to the study of form and structure of words.

What is morphology?

The term *Morphology* can refer to three different things

a. Description of the behaviour of morphemes and how they are combined.
b. Derivational, inflectional and compositional processes of word formation occurring in a specific language.
e.g. “German has a richer morphology than English”
c. Description of such word formation processes.

What are Morphemes?

- **Morphemes**
  - Morphemes are minimal meaning-bearing units:
    e.g. *talked* contains two morphemes: *talk* and *-ed* (past).
  - Form-function pairs (sound/sign-meaning)
  - Basic units of morphology

Morphemes are the “building stones” of phrases

Why study morphology? (1/2)

- One of the main properties of language are the sound/meaning pairs
- When analyzing language (or learning a foreign language), we can't simply list all expressions: there is an infinite number of them!
- So we compose expressions into smaller units: usually into phrases and words (syntax)

Why study morphology? (2/2)

- Can we use words as basic sound/meaning units?
  - Problems:
    1. Definition of words is unclear
    2. Words can be composed of many components that contribute to meaning and/or grammar

Several applications in Computational Linguistics benefit from morphological analysis (more later)

Words and Morphemes

There are two main usages of the term *word*:

1. Surface form (spoken or written representation)
2. Abstract form (lemma or dictionary entry,
   e.g. bare infinitives in English, nominative single form of nouns in Latin)

   The class of forms representing a word in different contexts is called a **lexeme**
   e.g. sing = {*sing, sings, sang, sung, singing*}

Based on Crysmann 2006

A definition of words?

Words can be described as units of language (either sequences of sounds, or signs) that function as meaning bearers. But this is a fuzzy notion, e.g.:

- *talked* in *she talked* expresses both “talking” and past tense.
- Is *more or less* one word, or are there three words?

A structuralist solution: **morphemes**

A language:

11-112 phonemes

↓

4,000-10,000 morphemes

↓

An infinite number of sentences

Morphs and Morphological Analysis

- The realisations of morphemes are called *morphs*:
  - e.g. English plural morpheme:
    [NUMBER pl]: -s, -es, -en, -∅
    boy-s, box-es, ox-en, sheep
  - These different realisations of the same morpheme are called **allomorphs**.
- **Morphological analysis**
  - Segmentation of expressions into basic units (mostly starting from word-level).
  - Classification of these basic units according to function.

Based on Crysmann 2006

Types of morphemes

- **Free Morphemes**
  Free morphemes can occur independently. Free morphemes are common in both English and German.

  e.g. *boy*, *sing*
- **Bound Morphemes**
  Bound morphemes must be attached to another morpheme, and cannot be used independently.

  e.g. [NUMBER pl] *-s* → *boys*

Based on Crysmann 2006

Types of bound morphemes

Typical bound morphemes are:

- **affixes** (*boy+s*, *talk+ed*)
- **clitics** (French: *je ne sais pas*, *je* and *ne* cannot occur without a verb)
- **roots** (Spanish *habl-* needs an ending indicating person, number, mode, etc.)

Formatives

Morphemes are form-meaning pairs, but not all segmental forms have an identifiable meaning:

- **Formatives** are forms without identifiable meaning

  e.g. Linking elements in German compounds:
  *Geburt+**s**+tag* (Birthday), *Schwan+**en**+hals* (swan neck).

Based on Crysmann 2006

Pseudo Morphemes

- **Pseudo-morphemes** or **cranberry morphemes** are special cases of formatives.
  They are segment-able part of a complex word, but do not have an independent meaning:

  e.g.
  - ***cran***+*berry*, ***rasp***+*berry*
  - *re*+***ceive***, *con*+***ceive***

Based on Crysmann 2006

Outline

Areas of Morphology

We distinguish:

- **Word forming**:
  - Derivational morphology
  - Compounding
- **Inflection**

Derivational Morphology

- allows to build complex words by combining bound and free morphemes.
- Derivational operations are per definition optional, i.e. not required by syntactic criteria.

Changes made by derivational morphemes

(a) semantics,
e.g. [*clear*] $\rightarrow$ [*un*+[*clear*]] = unclear

(b) syntactic category,
e.g. $[derive]_V \rightarrow [[[derive]_V + ation]_N + al]_{Adj}$ = derivational

(c) valency of a verb,
e.g. [*qaw*] 'it breaks' $\rightarrow$ [*t*+[*qaw*]] 'he breaks it' (Havasupai)

(d) several from the above, e.g. $[understand]_V \rightarrow$
$[[understand]_V + able]$ = understandable

Compounding

- allows to build complex words by juxtaposition of free morphemes.
  [[*sale*]+s+[*man*]], [[*dish*]+[*washer*]].
- Productive compounding results in an infinite lexicon.
  $\left\{\begin{array}{l}\textit{English}\\ \textit{German}\\ \textit{Havasupai}\end{array}\right\}\left\{\begin{array}{l}\textit{phonetics}\\ \textit{phonology}\\ \textit{morphology}\end{array}\right\}\left\{\begin{array}{l}\textit{teacher}\\ \textit{researcher}\\ \textit{student}\end{array}\right\}$

Based on Crysmann 2006

Inflectional Morphology (1/2)

- Inflection is required by syntactic criteria, e.g. an English verb must have tense.
- It marks grammatical (=morpho-syntactic) distinctions:
  - Conjugation (verbal categories):
    1. person, number, gender
    2. tense, aspect, mood, agreement
  - Declination (nominal categories)
    - case, number, gender, degree, definiteness

Based on Crysmann 2006

Inflectional Morphology (2/2)

- Meaning or, at least, the general concept is (generally) not changed, though *when, who* or *what* and sometimes *where, how* and *whether* may be specified by inflectional morphemes.
- There are bound and free inflectional morphemes:
  go [TENSE past]: *went*
  go [TENSE future]: *will go*

Based on Crysmann 2006

Inflection — paradigm

Inflectional morphology is typically organised in **paradigms**.

Paradigm

"A set of forms having the same root/stem, one of which must be selected in a certain syntactic environment" (definition based on [Crystal(1997)] (p. 277) and [Payne(1997)] (p. 26))

Paradigm - an example

For instance, German conjugation:

present	NUMBER		*past*	NUMBER	
	singular	*plural*		*singular*	*plural*
1.	dehn-e	dehn-en	1.	dehn-te	dehn-te-n
2.	dehn-st	dehn-t	2.	dehn-te-st	dehn-te-t
3.	dehn-t	dehn-en	3.	dehn-te	dehn-te-n

Taken from Crysmann 2006

Outline

Some Basic Notions

- **Root**: an unanalysable form, expressing the basic lexical content of a word. Also defined as 'what is left of a complex form when all affixes are stripped'.
- **Stem**: consists of at least a root.
  It can contain (an) derivational affix(es).
  In inflectional morphology, *stem* is generally defined as the root + a thematic vowel.
- **Base**: a form to which an affix may be added. A base may be simplex (root) or complex (root + affixes).

Morphological Processes

Bases can be altered by the following processes:

- Affixation
  - Prefixation
  - Suffixation
  - Circumfixation
  - Infixation
- Stem Modification
  - Substitution (vowel mutation, suppletion)
  - Subtraction
- Suprasegmental Modification
  - Tone
  - Stress

Affixation

- Affixes are bound morphemes
- Their position is fixed with respect to the base
  - a **prefix** precedes the base
    - ***im****-possible*
  - a **suffix** follows the base
    - *want-****ed***
  - a **circumfix** surrounds the base
    - ***ge****-dehn-****t***
  - an **infix** is placed within the base
    - *f-****um****-ikas* 'become strong', *fikas* 'be strong' (Bontok, Philippines)
- Affixation can be a recursive process
- **Prefixes** and **suffixes** are most frequent cross-linguistically

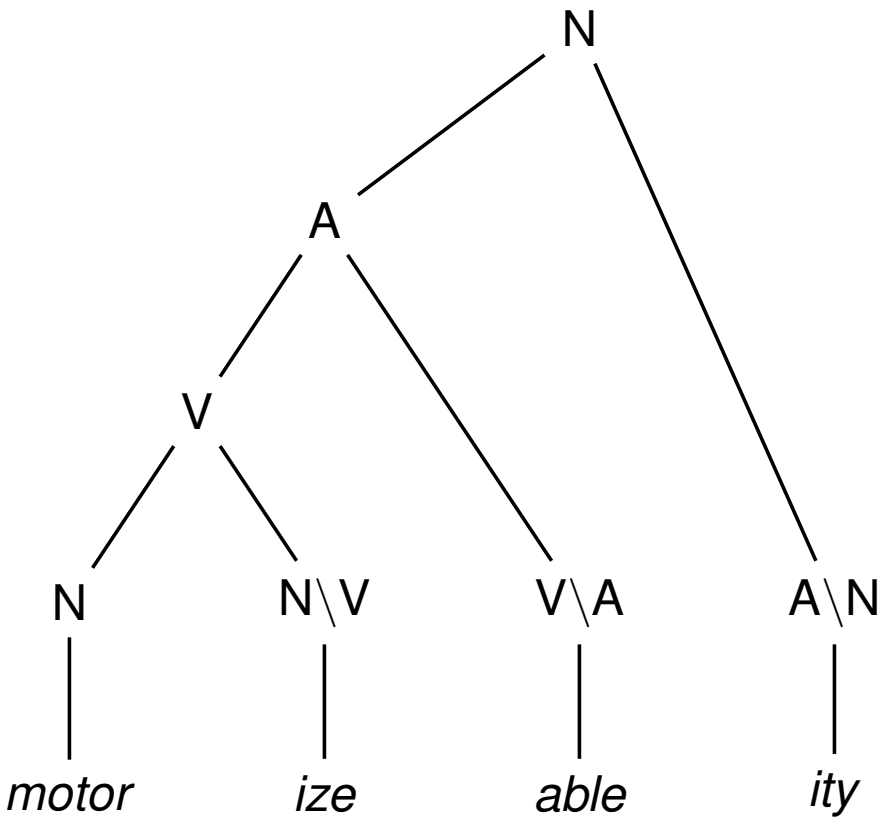
Affixation (cont)

- Words can have an internal structure (see next slide)
- The order of application can be significant, e.g.

  [in-[describe-able]], [[*in-describe]-able]
  [[un-do]-able] vs [un-[do-able]]
- Constraints on morpheme order are described by **morphotactics**
- Morphotactics can be determined by
  - word syntax (e.g. indescribable)
  - lexical strata
    - *non-im-partial* vs. *in-non-partial*

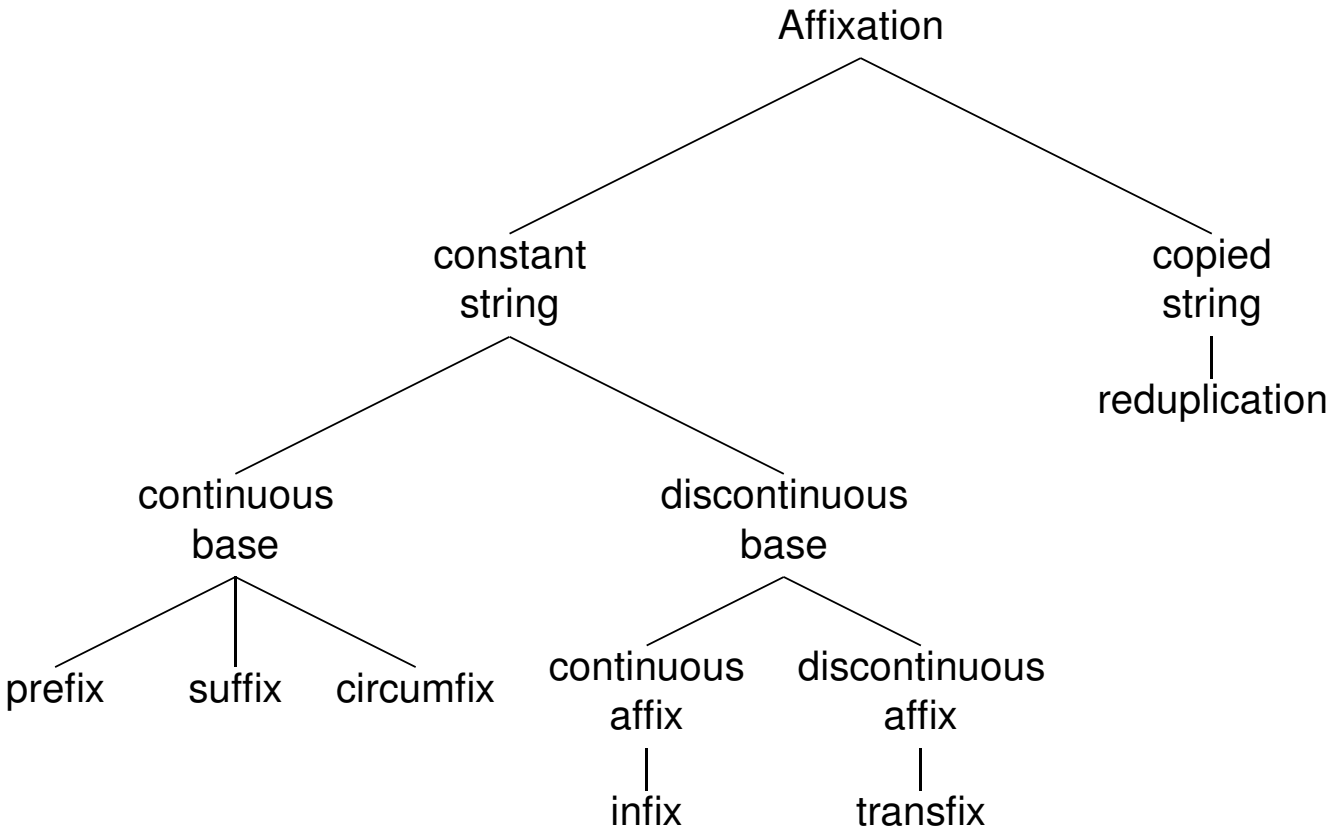
Based on Crysmann 2006

Internal structure of *motorizability*



(Sproat (1992), p. 84)

Types of affixational processes



(Crysmann 2006)

Infixation

- An **infix** is a continuous affix that attaches within the base
- Infixation is rare in European languages
- Infixation is often motivated by prosodic factors
  - Tagalog places affixes in the base to avoid closed syllables (i.e. syllables that end in a consonant)
    - *um-* + *sulat* → *sumulat*
    - *sulat* + reduplication: *susulat* and *sumusulat*
    - *um-* + *aral* → *umaral*
- Infixation can also be purely morphologically conditioned:
  - e.g. Udi (Nakh-Daghestanian, Azerbaijan) infixation:

Root	Transitive		Intransitive	
box	*bo-**ne**-x-sa*	boils	*box-**ne**-sa*	boils
uk	*u-**ne**-k-sa*	eats	*uk-**ne**-sa*	is edible

Based on Crysmann 2006

Transfixation

- A the segment of a **transfix** interleaves with the base's segment (i.e. both base and affix are discontinuous)
- Transfixation is common in Semitic languages (e.g. Arabic and Hebrew)
- The following forms are derived from the root *ktb* in Maltese

Transfix	Word	Gloss
-i-e-	*kiteb*	'he wrote'
-i–u	*kitbu*	'they wrote'
mi–u-	*miktub*	'written'
–ie-	*ktieb*	'book'
-o–a	*kotba*	'books'

Based on Crysmann 2006

Modification

- Morphological processes can effect stem internal segments
- The German vowel mutation ("umlaut" and "ablaut") are typical examples of such a process
- Umlaut:
  - Phonologically predictable segmental alternation (e.g. vowel fronting in German)

    $a \rightarrow ä$ (*Wald, Wälder* ("forest, forests"))
    $u \rightarrow ü$ (*Mutter, Mütter*, ("mother, mothers"))
    $o \rightarrow ö$ (*tot, Tödlich* ("dead, deadly"))
- Ablaut:
  - Phonologically unpredictable segmental alternation
    - *g**e**hen, g**i**ng, geg**a**ngen* vs *s**e**hen, s**a**h, ges**e**hen*

Based on Crysmann 2006

Example of a suprasegmental morpheme

- In Sabaot (Nilotic, Kenya & Uganda) uses advanced tongue root and normal vowels as morphemic contrast.
- This process may be applied to the entire word, as in the example below:

(1) *kɔ̀cmnyccncctɛ́*
ka – a – mnyaan – aa – tɛ – ATR
PAST-1SG-be.sick-STAT-DIR-IMPERF
"I went being sick (but I am not sick now)"

(2) *kɔ́ɔ́mnyɔ́ɔ́nɔ́ɔ́tɛ́*
ka – a – mnyaan – aa – tɛ
PAST-1SG-be.sick-STAT-DIR
"I became sick while going away (and I am still sick)"

(Payne 1997, p.29)

Suppletion

- Suppletion refers to 'stem replacement': a verb has more than one stem which are used in different contexts.
- In many European languages, suppletion occurs with the verb 'to be', e.g. in English, the verb uses three historically different roots:
  - *am*, *are*, *is*
  - *was*, *were*
  - *be*

(Payne, 1997)

Subtractive Morphology (1/2)

- **Subtractive morphology** means that part of the stem is omitted to mark a morphological process.
- For instance Koasati (a Muskogean language, spoken in the US):

Singular	Plural	Gloss
*pit**af**-fi-n*	*pit-li-n*	to slice up the middle
*las**ap**-li-n*	*las-li-n*	to lick something
*acokcan**a:**-kaln*	*acokcan-ka-n*	to quarrel with someone
*obakhit**ip**-li-in*	*obakhit-li-n*	to go backwards

Data taken from Sproat (1992)

Subtractive Morphology (2/2)

- The shape of the base cannot be predicted from the derived form
- Subtractive Morphology is problematic for theories assuming that morphology consists of the addition of morphemes

Based on Crysmann 2006

Reduplication

- Reduplicated morphemes are formed by reduplicating (part of) the base.
- In **total reduplication** the entire base is copied, though minor changes may occur, e.g. ([Kiparsky(1987)] (p. 115-117)
  - Indonesian:

orang	***orang*** *orang*
'man'	'men'

  - Javanese:

Base	Habitual-Repetitive	Gloss
bali	*bola bali*	'return'
udan	*udan udɛn*	'rain'

Based on Crysmann 2006

Suprasegmental Marking

- Stress
  - English verb-noun derivations:

Verb	Noun
pro**duce**	**pro**duce
per**mit**	**per**mit
im**port**	**im**port
in**sult**	**in**sult
dis**count**	**dis**count

- Tone
  - Chichewa:

Form	Tense/aspect
ndi-ná-fótokoza	simple past
ndi-na-fótókoza	recent past
ndí-nâ:-fótókoza	remote past
ndí-ma-fotokózá	present habitual
ndi-ma-fótókoza	past habitual

Morphophonological Processes (1/2)

- The environment of morphemes can influence their appearance (phonological and/or graphemic alternations)
- Morphophonological Alternations
  - Assimilation

    Homographic nasal assimilation
    iN+possible → ***im****possible*
    iN+complete → ***in****complete*
    iN+resistable → ***ir****resistable*
  - Epenthesis: *wish+s* → *wishes*
- Graphemic alternations:
  - *y* + *s* ∼ *ies*

Based on Crysmann 2006

Morphophonological Processes (2/2)

- The environment influencing the morpheme's form need not be directly adjacent to the morpheme
- Harmony rules impose identity of sound features (typically vowel features)

  E.g. Finnish vowel harmony

	low	mid	high
back vowels	a	o	u
front vowels	ä	ö	ü
neutral vowels		e	i

  - taivas + ta → taivasta (*taivastä)
  - lyhyt + ta → lyhyttä (*lyhytta)

(Morpho)phonological rules

- [Chomsky and Halle(1968)] propose phonological rules to derive “surface” morphemes in *The Sound Pattern of English* (SPE)
- They were formalized as (ordered) context-sensitive rewrite rules:

  a → b/v_w
  e.g. *iN-* → *im-*/_*m*

(Morpho)phonological rules

- There was a strong believe that related morphemes are all derived from the same **underlying representation**, even if this form never occurs on the surface (e.g. *divine* and *divinity* would come from the root *divIn*)
- The approach did not take general phonetic constraints within the language in account, nor did it address rules and tendencies in morpheme structures

Declination of *puella*

Latin declination of a noun of the first declination:

case	NUMBER	
	singular	*plural*
NOM	puella	puellae
GEN	puellae	puellarum
DAT	puellae	puellis
ACC	puellam	puellas
ABL	puella	puellis

Syncretism/exponence

We observe both:

- **syncretism**: the same form is used to express different feature combinations.
  e.g. in the declination of *puella*:
  - *-ae*: GEN or DAT singular, or NOM plural
  - *-a*: NOM or ABL singular
  - *-is*: DAT or ABL plural
- **exponence**: the relation between form and function is **m:n**:
  - **multi-exponence** (cumulation): one form expresses several functions.
    Here: *-am* expresses both accusative and singular
  - **Extended exponence**: in *ge-dehn-t*, *ge-* and *-t* express one function together.

Morphological Properties — Synthesis

Synthesis: the number of morphemes that tend to occur within a word.

- In **isolating** languages words tend to consist of only one morpheme. (e.g. Chinese languages)
- **Polysynthetic** languages are known for the large number of morphemes that may occur in a single word. For instance, the Quechua and Inuit languages. The following example is from Yup'ik:

(3) tuntussuqatarniksaitengqiggtuq
tuntu-ssur-qatar-ni-ksaite-ngqiggte-uq
reindeer-hunt-FUT-say-NEG-again-3gg-IND
'He had not yet said again that he was going to hunt reindeer'

([Payne(1997)], p. 28)

Morphological Properties — Fusion (1/2)

Fusion: the number of meaning units that are found in one morphological shape:

- **Agglutinative** languages have little fusion: each meaning component is represented by its own morpheme (e.g. Turkish).
- **Fusional** languages have morphemes that express many meaning units: e.g. *-ó* in Spanish *habló* expresses indicative mode, 3rd person, singular, past tense and perfect aspect.

Morphological Properties — Fusion (2/2)

In English, both examples of agglutinative morphemes, and fusional ones can be found:

- **agglutinative**: anti+dis+establish+ment+arian+ism
- **fusion**: vowel change in plural forming (*goose/geese*) and strong verbs (*sing/sang*).
  Individual morphemes (root and number/tense) cannot be segmented in chunks, therefore these forms are fusional.

Outline

Morphology in Computational Linguistics

Morphology related applications in computational linguistics are:

1. Analysing complex words, defining their component parts:

   anti+dis+establish+ment+arian+ism

2. Analysis of grammatical information, encoded in words:

   sings
   sing[PERSON 3, NUMBER singular,TENSE present]

Morphological Processing

- Inflection
  - lemmatisation/stemming
  - extraction of grammatical (morpho-syntactic) features (preprocessing for parsing)
  - State of the art: finite state technology (to be discussed)

Reduction of lexicon size (English 2:1, German 5:1, Finnish/Turkish >200:1) (Crysmann 2006)

Morphological Processing (cont)

- Derivational Morphology
  - Semi-productivity is still a challenge
    - Rule-based approaches tend to suffer from over-generation
- Compound Analysis
  - Important for languages with productive compounding
  - Additional task: bracketing

Why do we need morphology?

- For linguistic tools, such as parsers:
  significant reduction of lexicon size
- For statistical methods:
  reduces unseen data: in a morphologically rich language, many words will be found in each possible form, even in a large training corpus.

Machine translation runs into problems, in particular when translating from a morphologically poor to a morphologically rich language. This is expected to become a 'hot topic' in MT

State of the art: Finite State Transducers

Non-deterministic Finite Automata (NFA)

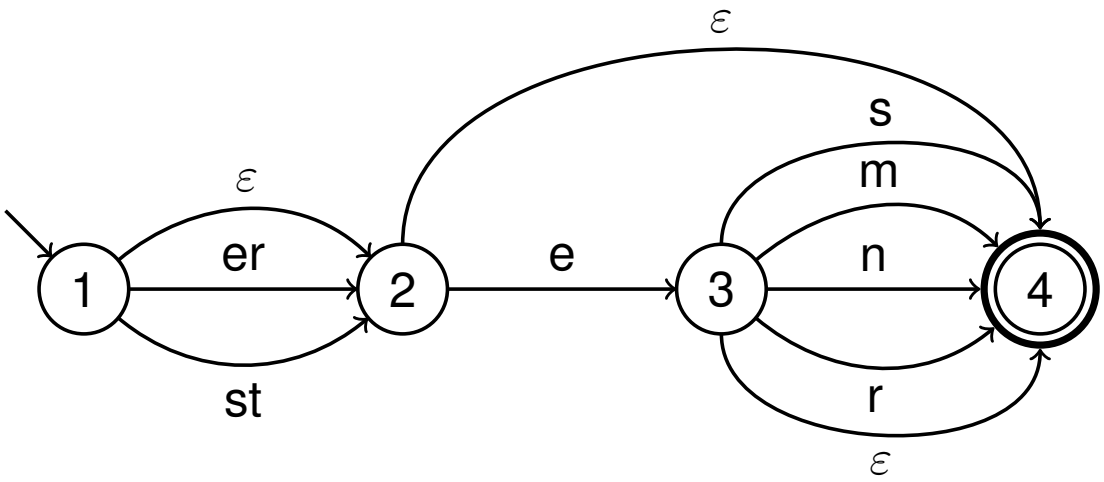
Definition

- A non-deterministic finite automaton is a quintuple (Q, Σ, δ, q_0, F), where
  - Q is a finite set of states
  - Σ is a finite set of symbols
  - δ is a transition function *delta* : $Q \times \Sigma \rightarrow Q$, such that for each $q_i \in Q$ and each $\sigma \in \Sigma$, there is a q_j such that $\delta(q_i, \sigma) = q_j$, where q_j is a non-final sink state, unless σ is licit at state q_i
  - $q_0 \in Q$ is a unique initial state
  - $F \subseteq Q$ is a set of final states
- At worse, a NFA's complexity is exponential at word length

Based on Crysmann 2006

An example of a NFA

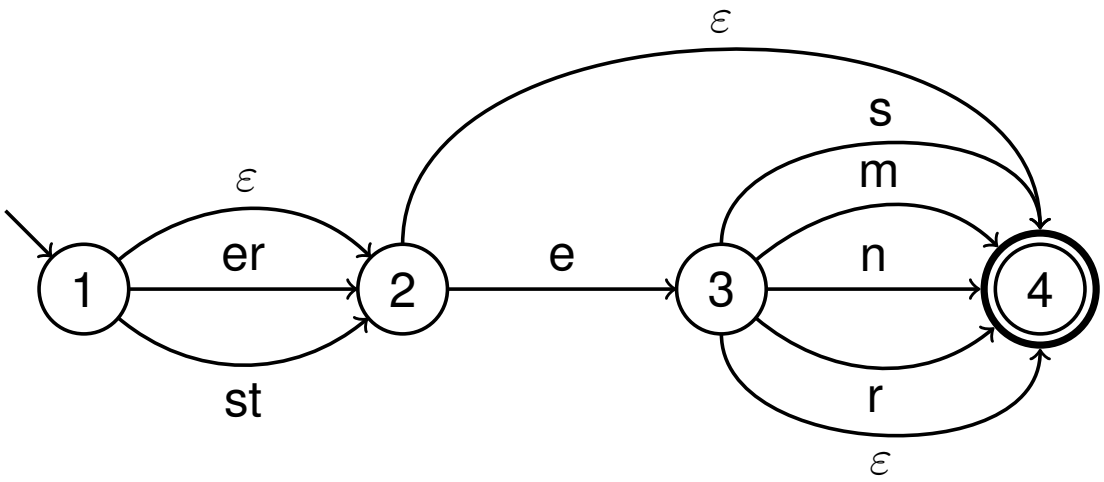
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

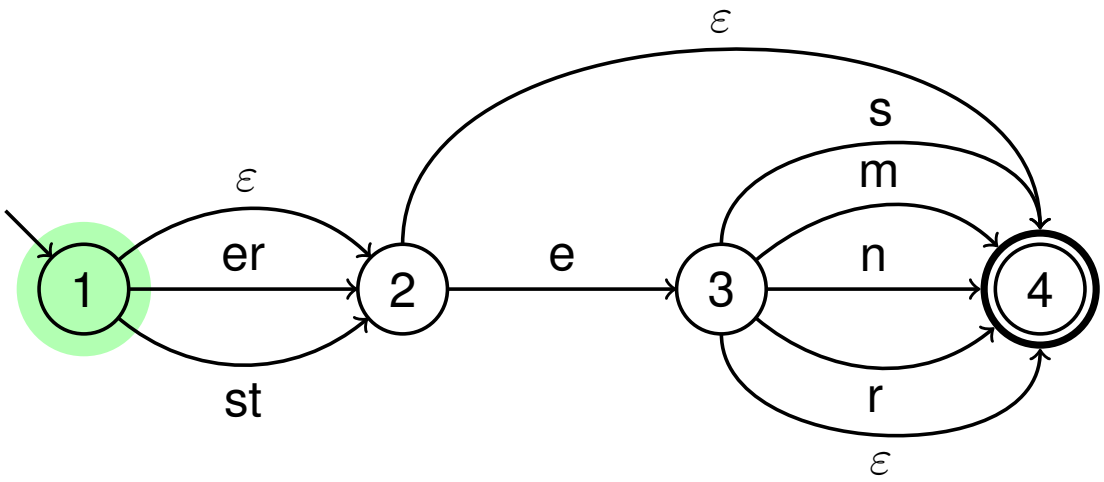
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

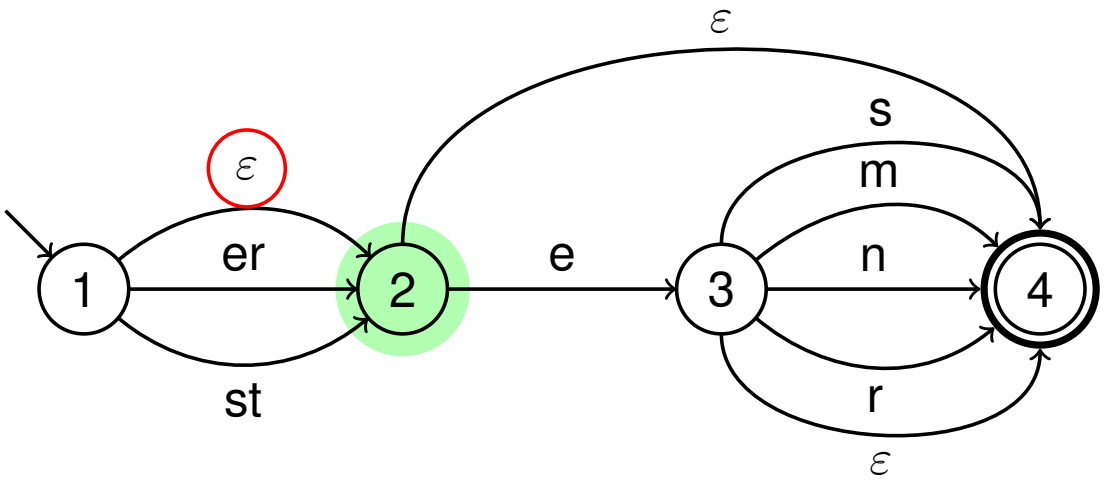
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

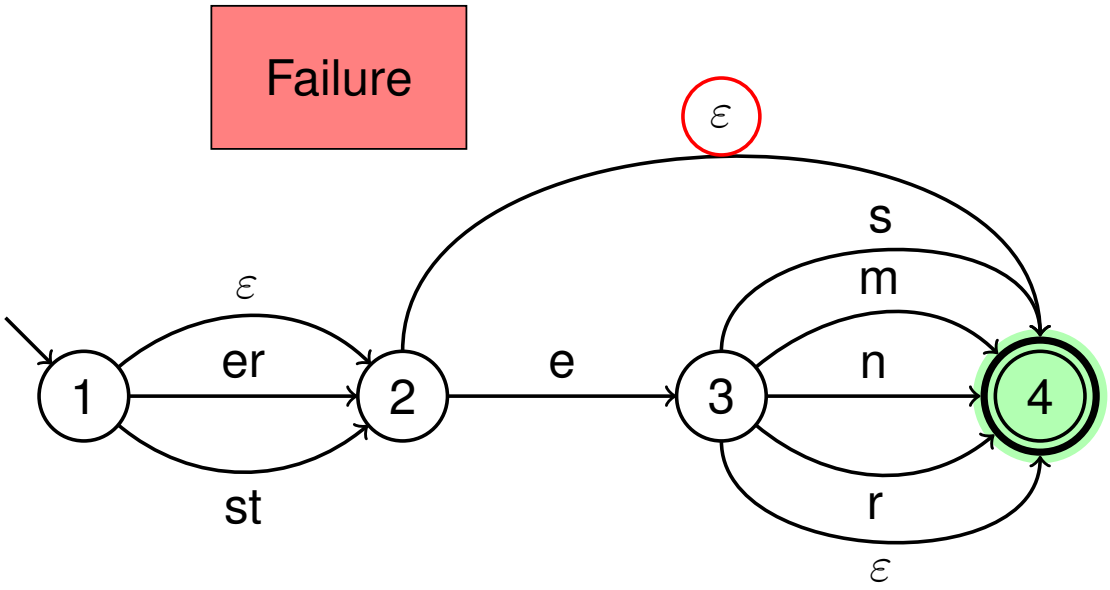
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

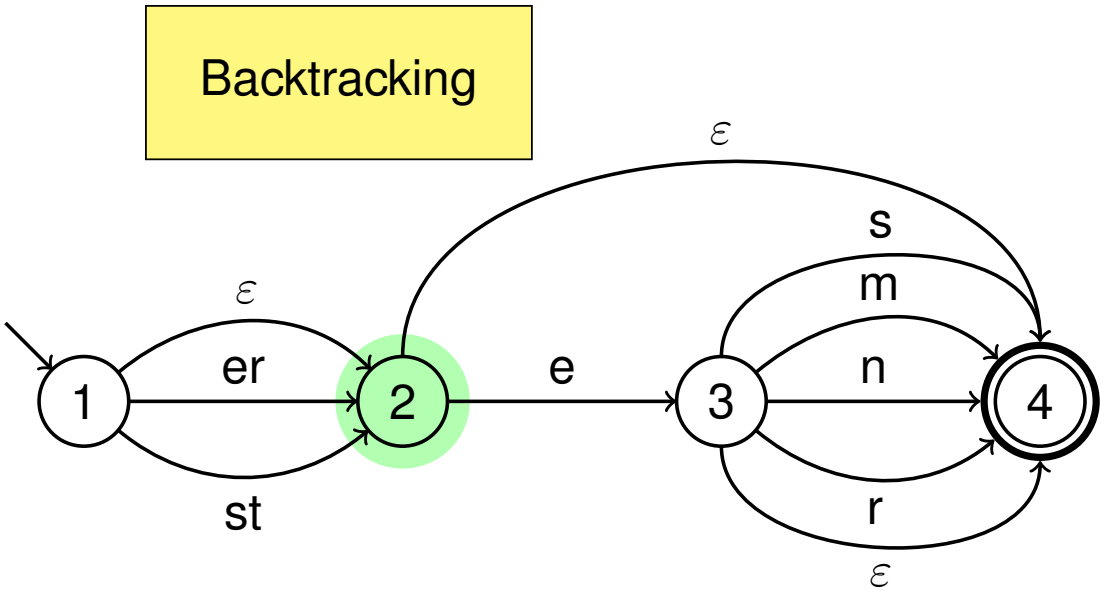
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

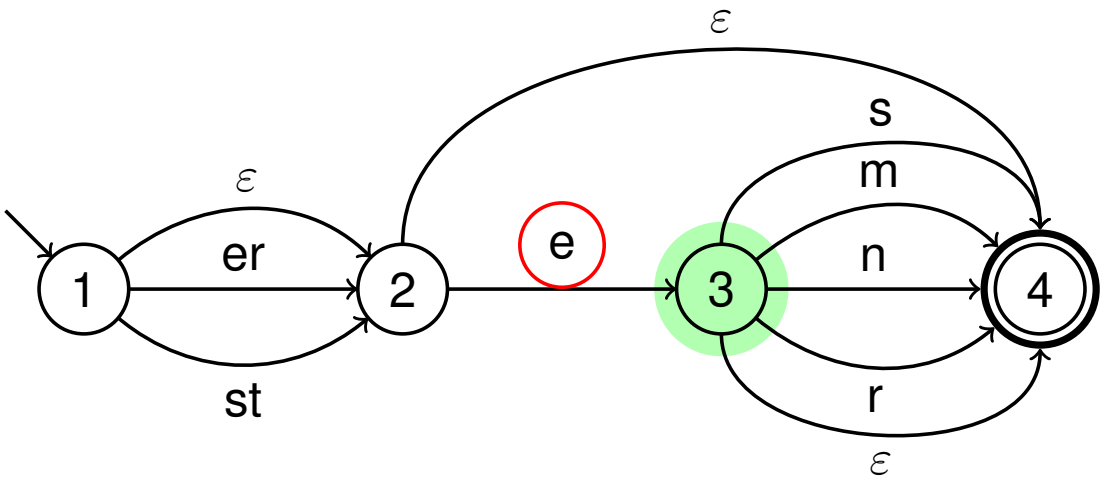
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

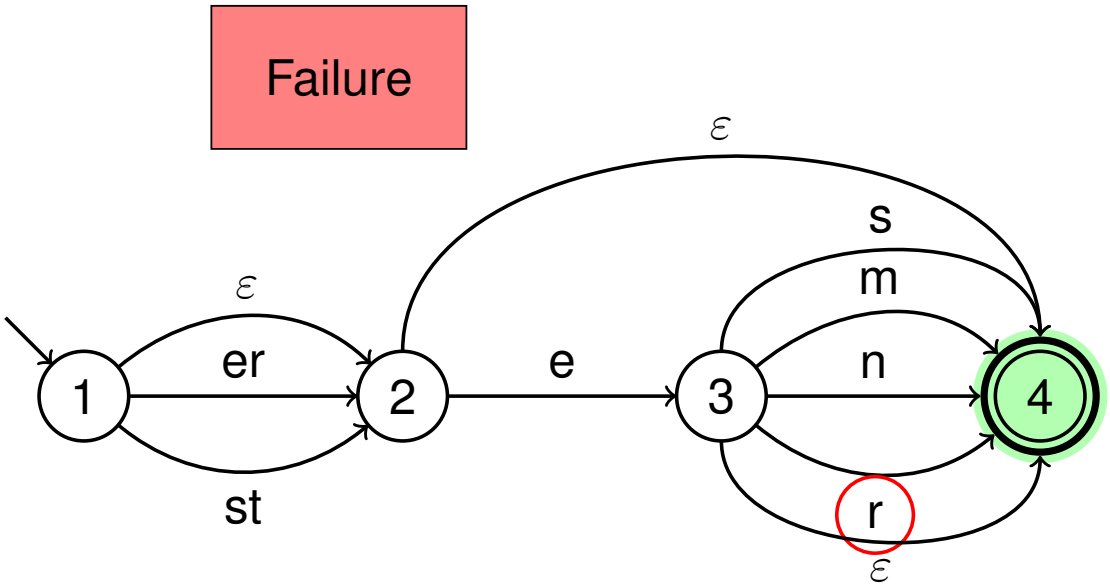
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

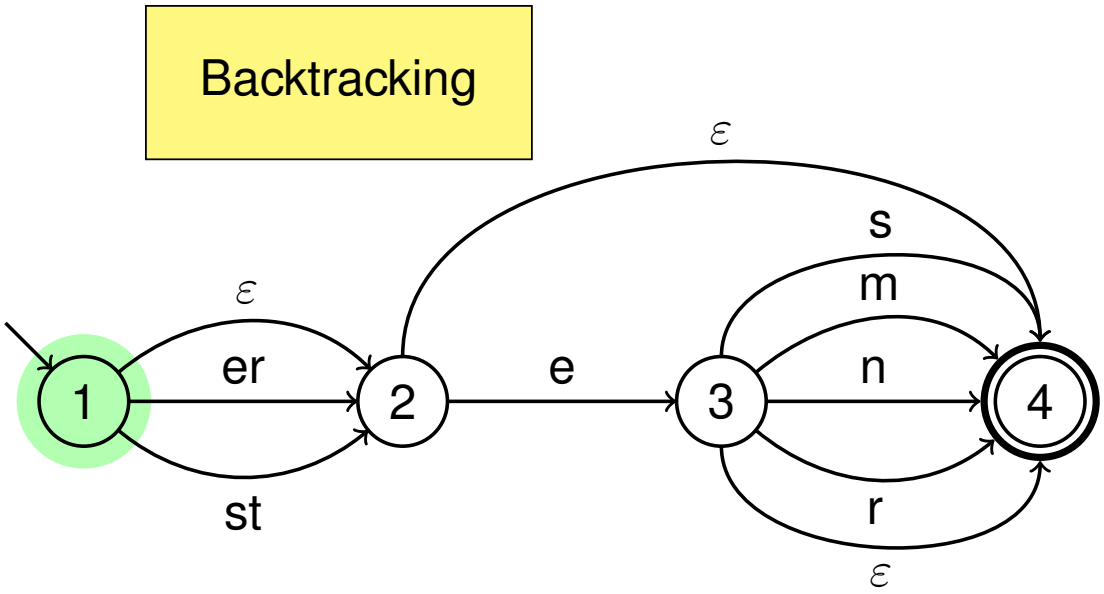
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

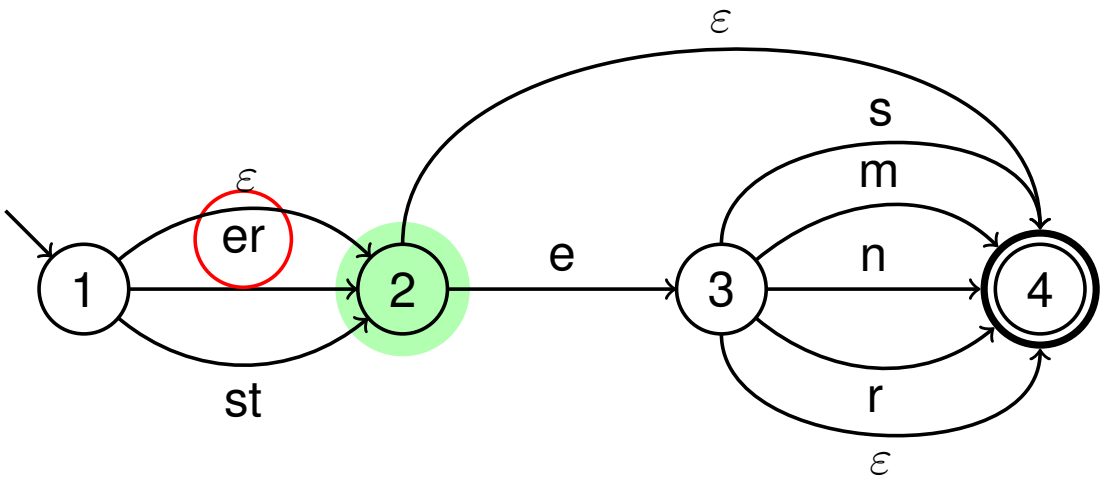
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

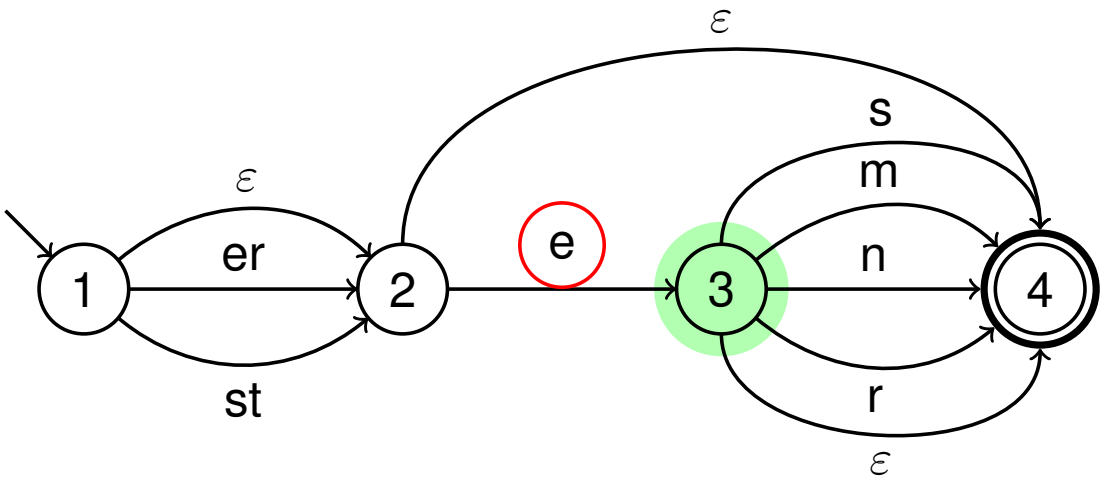
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

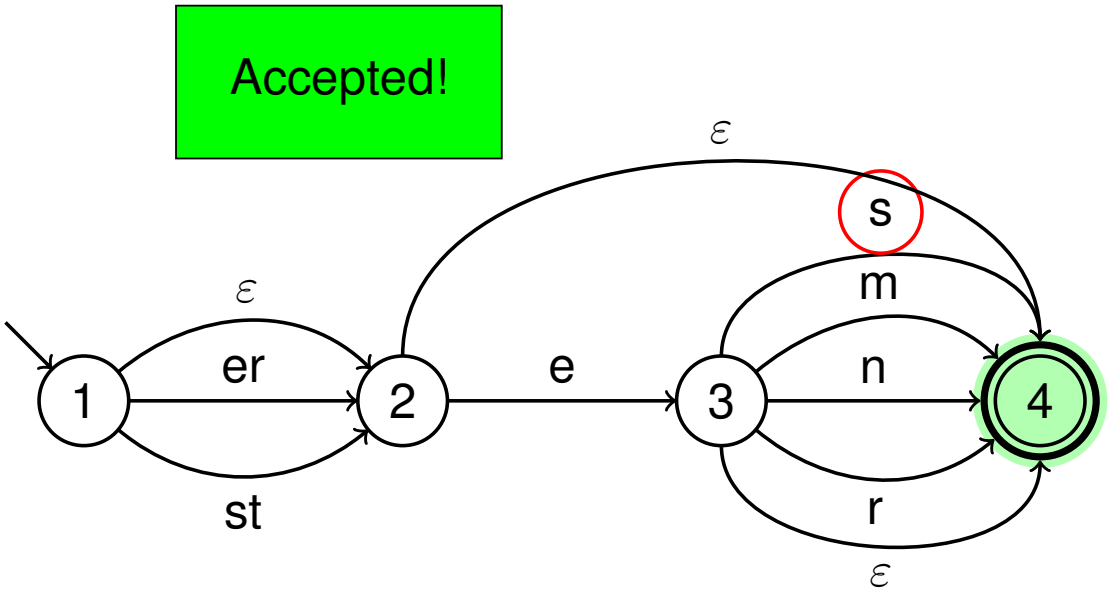
- German adjectives
- klein+ er+ es



Based on Crysmann 2006

An example of a NFA

- German adjectives
- klein+ er+ es



Based on Crysmann 2006

Deterministic Finite Automata (DFA)

- So what about the worse case exponential complexity of NFA?
- **Deterministic** Finite Automata (DFA) are linear at worse case
- For each NFA, there is always an equivalent DFA (Hopcroft and Ullman 1979)

DFA, Definition

- A deterministic finite automaton is a quintuple $(Q, \Sigma, \delta, q_0, F)$, where
  - Q is a finite set of states
  - Σ is a finite set of symbols
  - δ is a transition function $\delta : Q \times \Sigma \rightarrow Q$,
  - $q_0 \in Q$ is a unique initial state
  - $F \subseteq Q$ is a set of final states

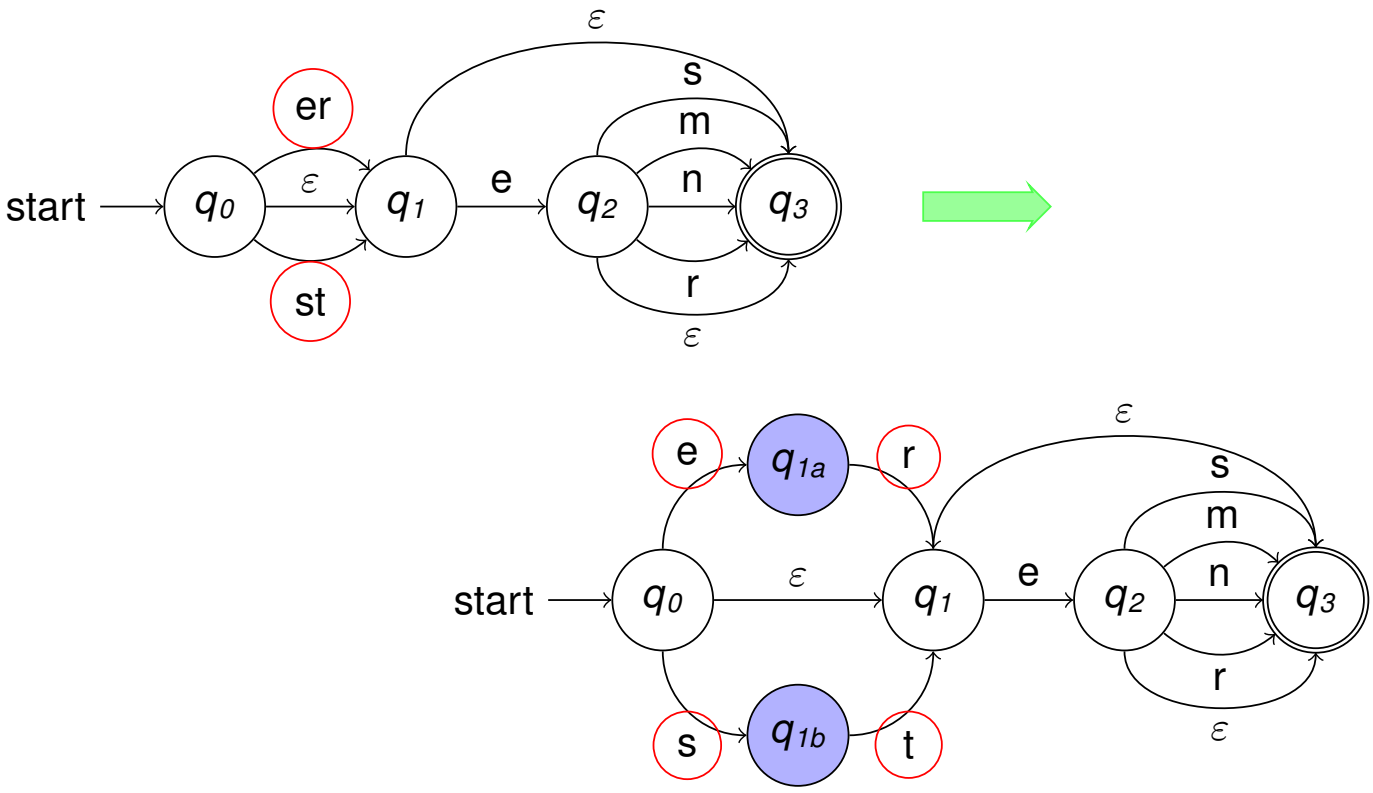
From NFA to DFA

For each **Nondeterminstic** finite state machine, there is an equivalent **deterministic** finite state machine

Step to take:

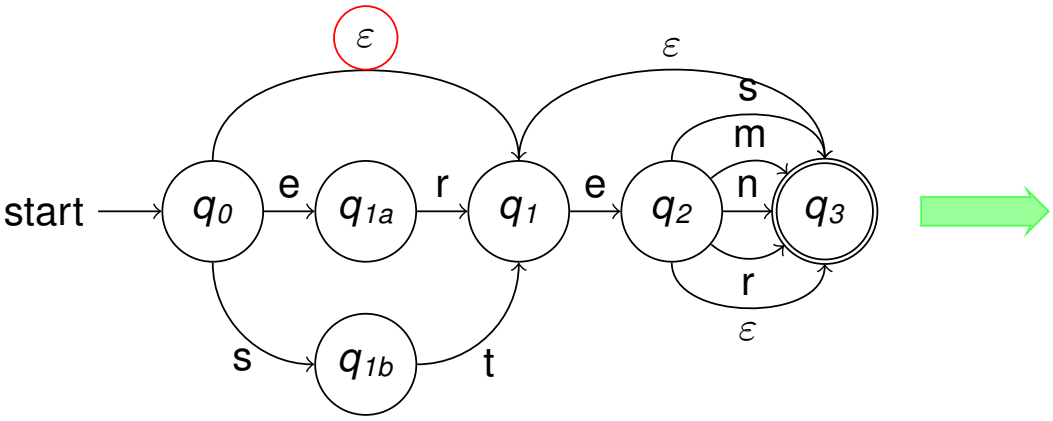
1. Expand edges that take more than one input character
2. Eliminate ε-edges (by adding alternative edges)
3. Construct power automaton (recursively combine states reached by the same input symbol)

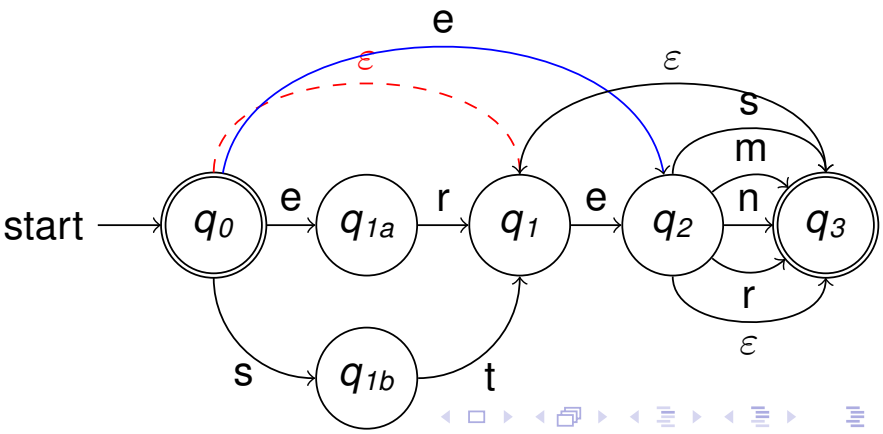
Expanding multiple symbol edges



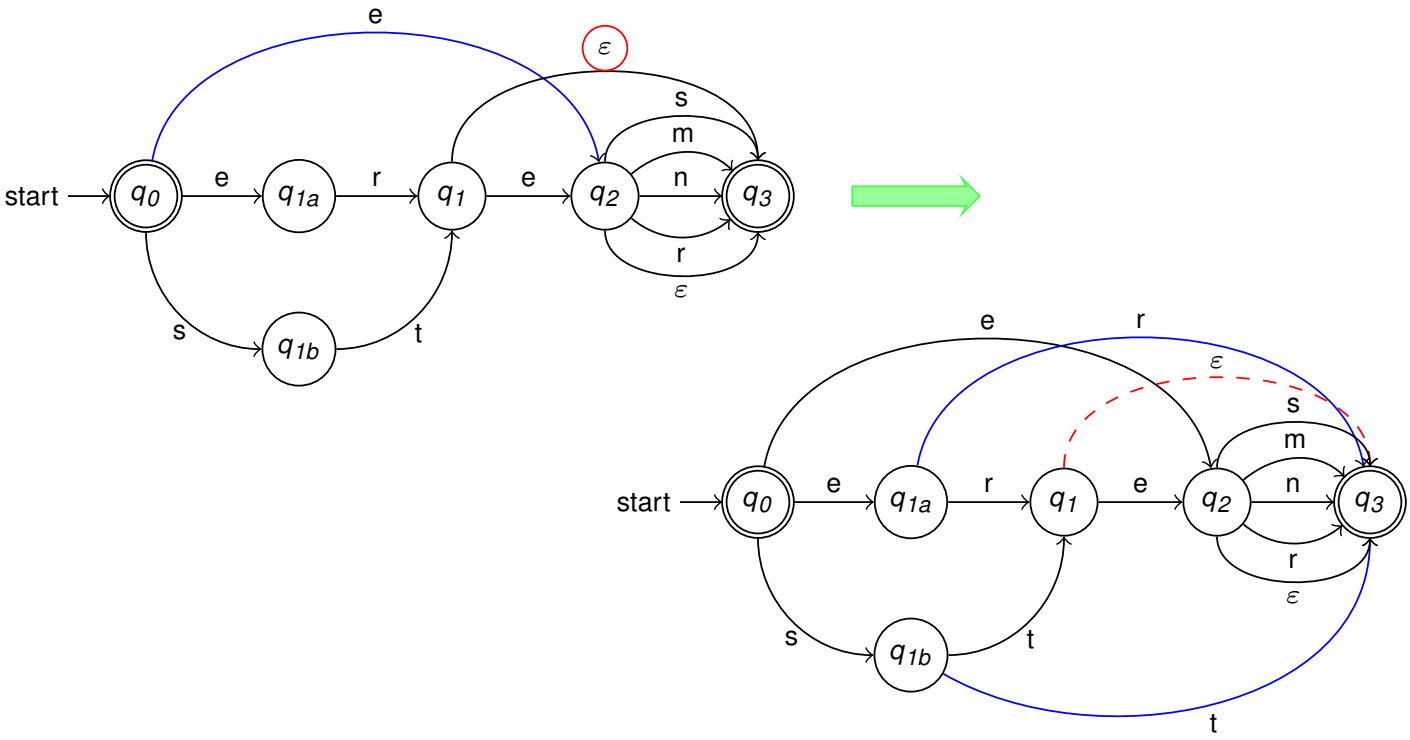
Based on Crysmann 2006

Eliminating ε-edges



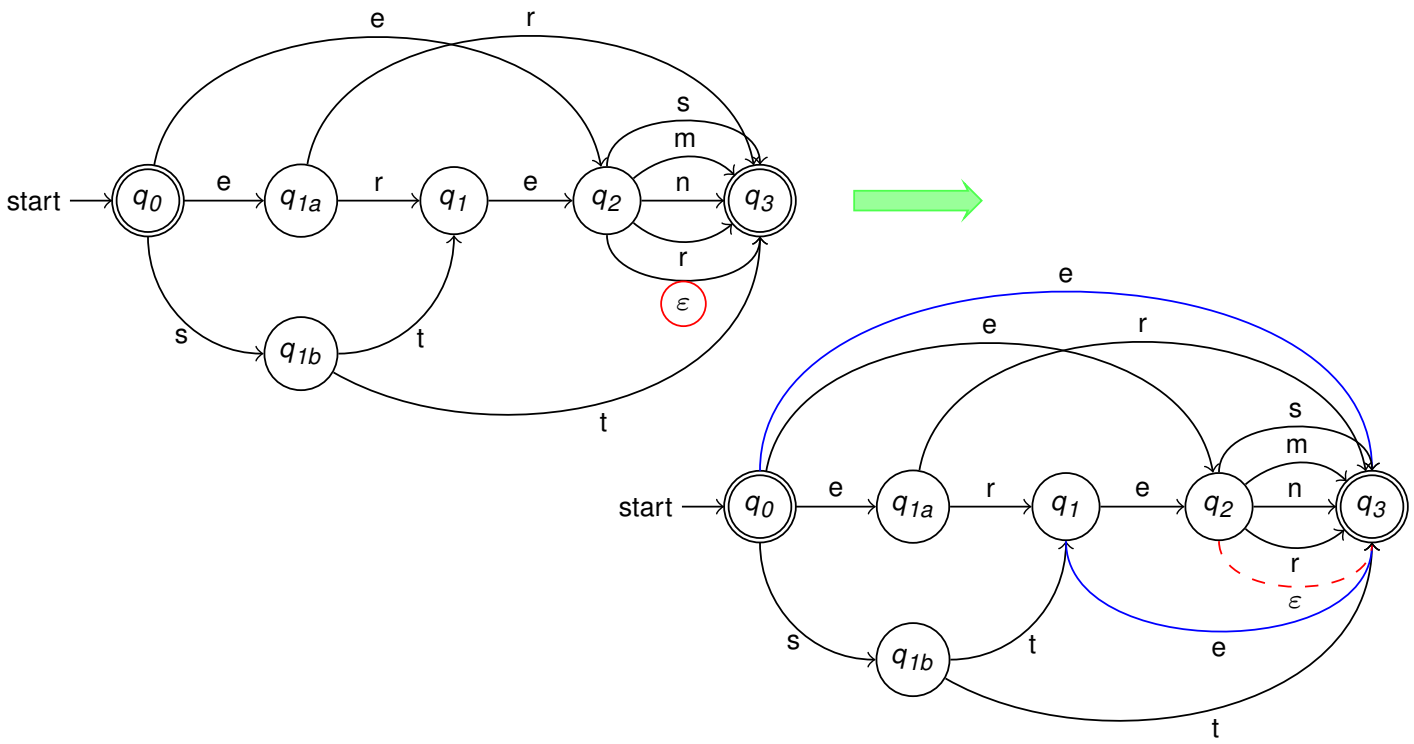


Elimination of ε edges



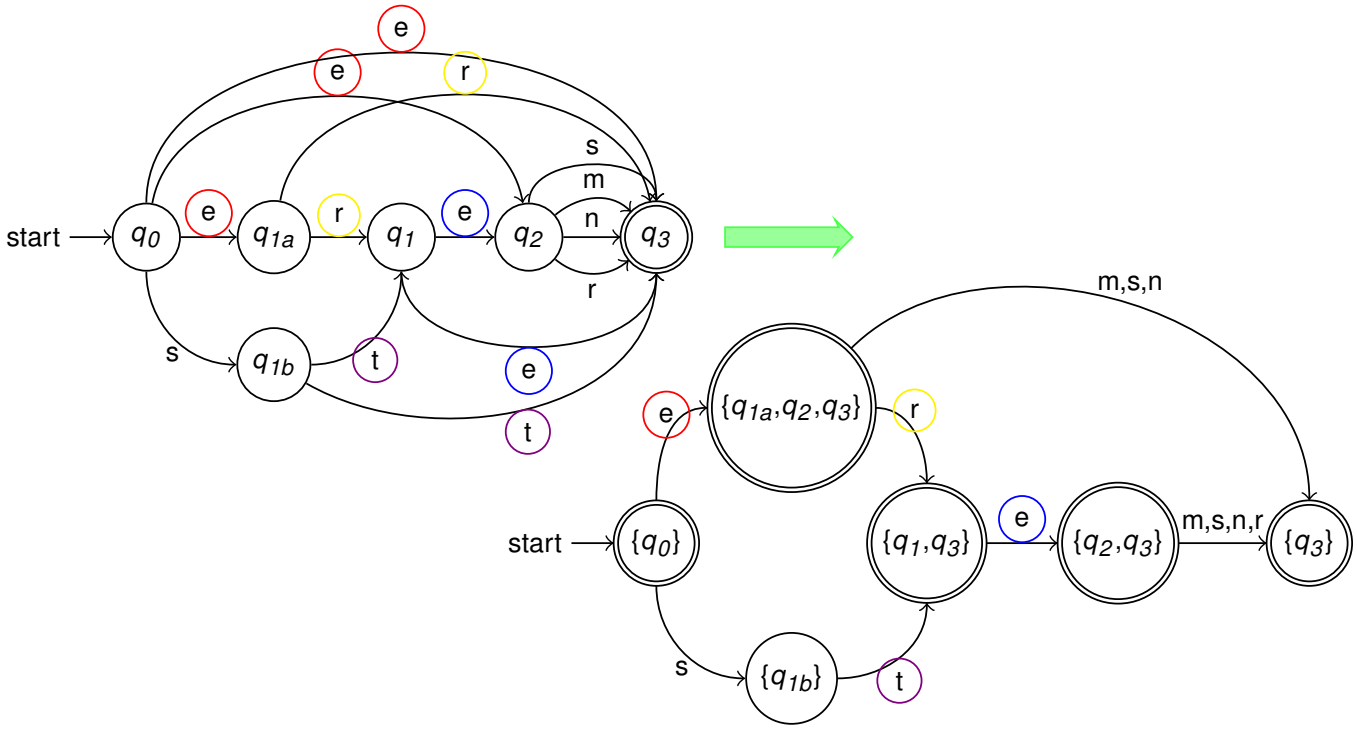
Based on Crysmann 2006

Elimination of ε edges



Based on Crysmann 2006

Constructing a power automaton



Based on Crysmann 2006

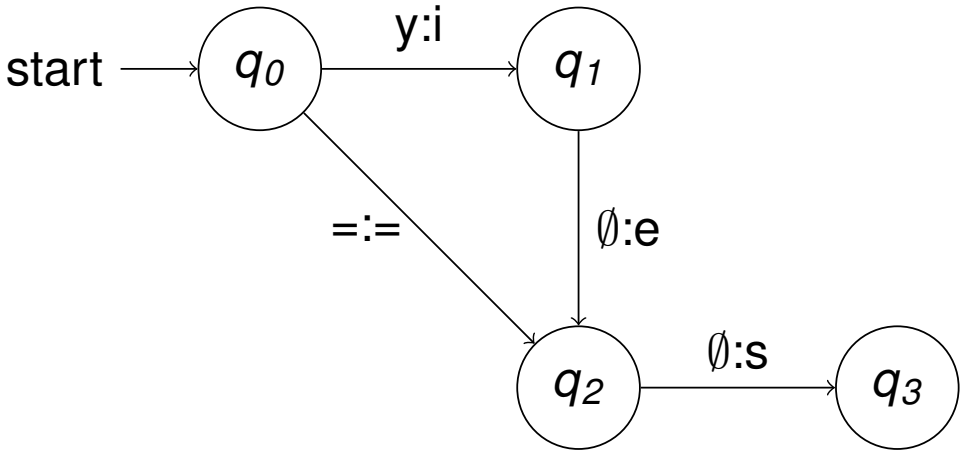
Finite State Transducers

- Finite State Transducers are variants of Finite State Machines that accepts language over symbol pairs (*a*:*a*,*a*:*c*) instead of single symbols
- Conventionally, left hand symbols correspond to lexicon input, and right-hand symbols to the surface string
- The $\emptyset$ can appear both on input string and output string, the symbol “=” (or @) stands for the ’any’ symbol
- FSTs can be used to implement phonological rules ([Johnson(1972)])

Based on Crysmann 2006

A Finite State Transducer

- $y + s \rightarrow ies$



Based on Crysmann 2006

Summary

- Morphemes are minimal sign/meaning pairs
- Morphological analysis plays a role in reduction of lexicon size, unknown word recognition, etc
- Several meaning units can be mapped in one morpheme (multi-exponence)
- Phenomena such as reduplication, syncretism, allomorphism, and morphophonological processes make that morphemes are not necessarily easily recognizable
- FSM forms the standard (basic) technique for morphological analysis

Bibliography I

Chomsky, Noam and Halle, Morris. 1968. The Sound Pattern of English.New York, USA: Harper and Row.

Crysmann, Berthold. 2006. Foundations of Language Science and Technology: Morphology. http://www.coli.uni-saarland.de/~hansu/courses/FLST05/schedule.html. Accessed on the 14th of August 2008.

Crystal, David. 1997. The Cambridge Encyclopedia of Language. Cambridge, UK: Cambridge University Press.

Johnson, C. Douglas. 1972. Formal Aspects of Phonological Description. The Hague, NL: Mouton.

Kiparsky, Paul. 1987. The Phonology of Reduplication.

Payne, Thomas E. 1997. Morphosyntax – a guide for field linguists. Cambridge, UK: Cambridge University Press.

Sproat, Richard. 1992. Morphology and Computation. Cambridge, USA. MIT Press.