# Foundations of Language Science and Technology: Morphology

**Berthold Crysmann**
crysmann@dfki.de

# Overview

- Basic terminology
- Subdomains of morphology: inflection, derivation, compounding
- Morphological processes
- Morphophonology
- Finite State Morphology

# Introduction

- **Morphology**
  - Subdiscipline of linguistics concerned with the internal structure of words
- **Major applications of morphology in computational linguistics**
  - Parsing of complex word forms into their component parts

    *antidisestablishmentarianism*
    anti+dis+establish+ment+arian+ism

  - Analysis of grammatical information encoded in word forms

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

# Words

- **Notion of word is ambiguous**
  - Word form (surface form)
  - Abstract notion (lemma or citation form, typically found in dictionaries)

    e.g. bare/infinitival form for verbs, nominative singular for nouns

- **Lexeme**
  - Class of equivalent forms that represent a word in different syntactic contexts

    e.g. sing = {*sing, sings, sang, sung, singing*}

# Morphemes

- **Morpheme**
  - Basic unit of morphology
  - Term introduced by structuralism
  - Abstract notion of a *minimal content-bearing unit*
  - Pairing of form and function
  - Surface realisation of abstract morphemes are called *morphs*

    e.g. English plural morpheme:

    [NUMBER pl]: *-s, -es, -en, -0, ...*

    *boy+s, match+es, ox+en, sheep*

  - Morphological analysis
    - segmentation into basic units
    - classification of units according to function

# Types of morphemes

- **Free morphemes**
  - In English or German, many morphemes can be used as independent words

    e.g. *boy, sing*

- **Bound morphemes**
  - Cannot be used independently

    *-s* [NUMBER pl] as in *boys*

  - Affixes are prototypical bound morphemes

# Formatives

- **Segmentable forms need not have a depictable meaning**

  e.g. linking element  in German compounds

  *Geburt*+***s***+*tag, Schwan*+***en***+*hals,*

- **Forms without any identifiable meaning are called *formatives***
- **Pseudomorphemes (“cranberry morphemes”)**
  - Special case of formatives
  - Examples:
    - ***cran***+*berry,* ***rasp***+*berry* etc.
    - *re*+***ceive***, *con*+***ceive***, *per*+***ceive***
  - Segmentable part of complex form cannot be assigned a constant meaning

# Areas of morphology 1

- **Inflection (Formenlehre)**
  - Marking of grammatical (=morphosyntactic) distinctions
  - Declination
    - Nominal categories (nouns, determiner, adjectives, pronominals)
    - Dimensions: case, number, gender, degree, definiteness
  - Conjugation
    - Verbal categories
    - Dimensions: Tense, aspect, mood, agreement
  - Distribution of forms conditioned by syntactic context
  - Inflectional marking by bound (synthetic) and free morphemes (analytic)

    *gehen* [TENSE past]: *ging*
    *gehen* [TENSE future]: *wird gehen*

- **Word formation**

# Inflectional morphology - Paradigms

- **Inflected forms of a lexeme can be organised in paradigms**
- **Inflectional features and their values define cells of a paradigm**
- **Cells are filled by the *exponents* of a morphological feature combinations**

| *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 |

- **Syncretism**
  - Different feature combinations can be expressed by the same form
  - Syncretism can cut across inflectional dimensions
- **Relation between form and function is m:n**
  - Multiple exponence (cumulation)
    - Morpheme *-e* expresses person, number and tense distinction
  - Extended exponence: *ge-dehn-t*

Source: Berthold Crysmann 2006

# Areas of morphology 2

- Inflection
- **Word formation**
  - Derivation
    - build complex words by combination of a free morphemes with bound morphemes

      e.g. $[[[derive]_V + ation]_N + al]_A$ = *derivational*

    - Changes semantics
    - May change syntactic category
  - Compounding
    - build complex words by juxtaposition of free morphemes
    - Productive compunding implies inifinite lexicon

      $[Flektion]_N + s + [morphologie]_N$ = *Flektionsmorphologie* `inflectional morphology'
      *[[sale]+s+[man]]* = *salesman*, *[[dish] [washer]] ]* = *dish washer*

    - Compounds are referential islands

Source: Berthold Crysmann 2006

# Morphological processes

- **Segmental processes**
  - Affixation
  - Modification
    - Substitution of segments (umlaut, ablaut, suppletion)
    - Subtractive morphology (deletion of segments)
- **Suprasegmental**
  - Stress
  - Tone

# Affixation

- **Recursive process**
- **Affixes are bound morphemes**
- **Affixes are positionally fixed with respect to the base**
  - prefix
    - un+happy
  - suffix
    - happy+ly
- **Root**
  - Part of a morphologically complex form after all affixes are stripped
- **Stem**
  - Root + thematic vowel in inflectional morphology
- **Base**
  - Part of a morphologically complex form to which an affix can be added
  - A base may be simplex (i.e. a root) or complex (root + affixes)

Source: Berthold Crysmann 2006

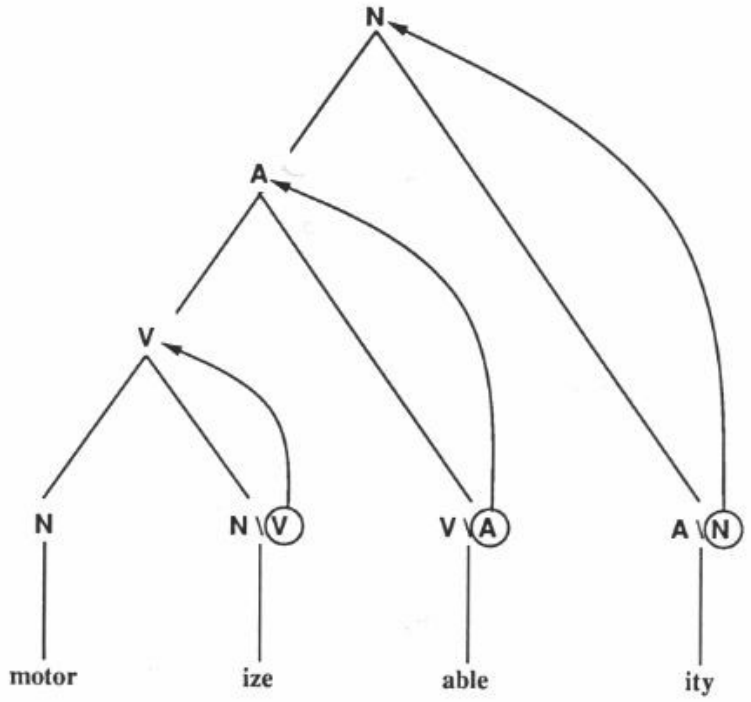
# Affixation

- **Order of application is meaningful**

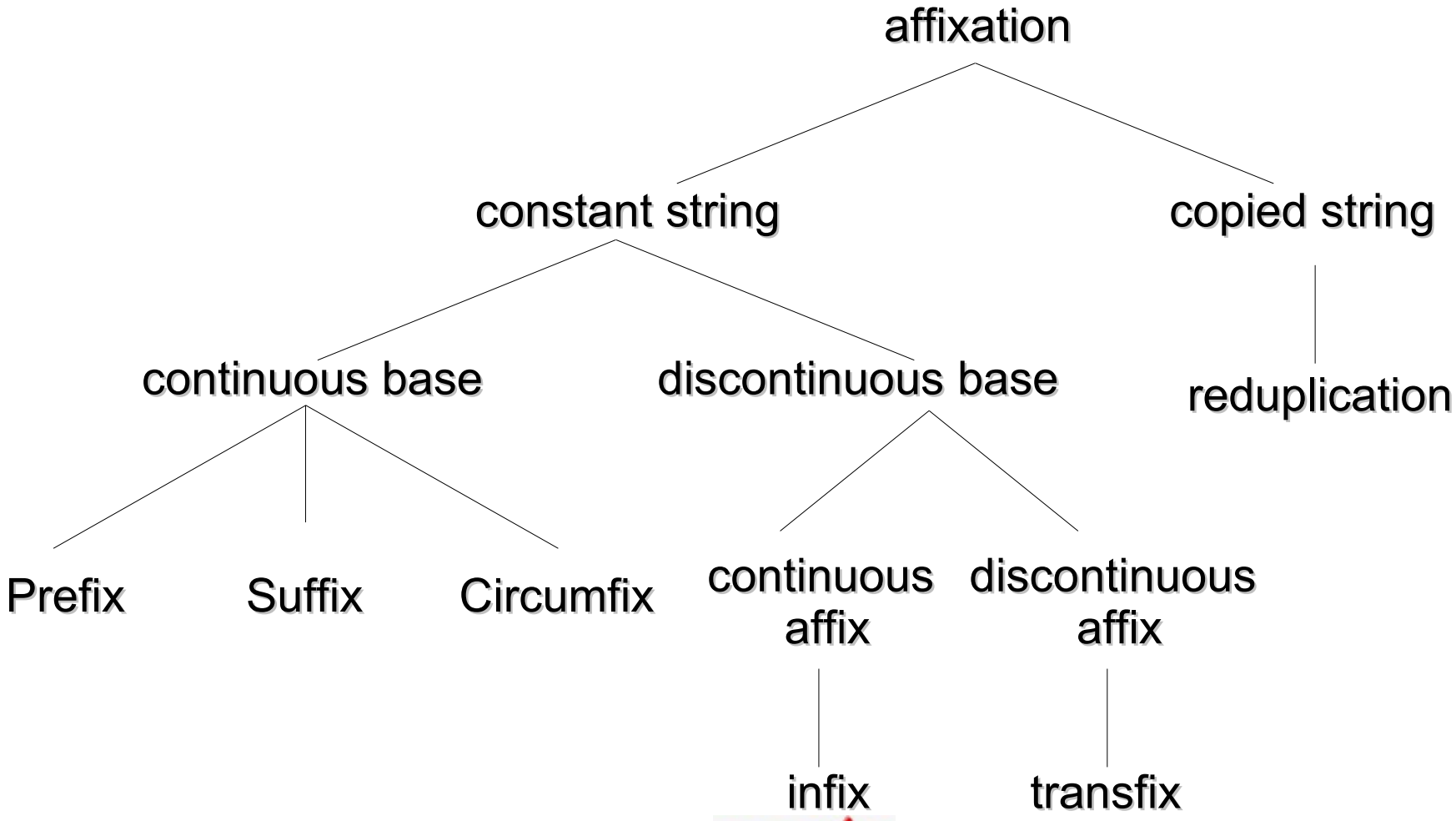
  **[in [[describe] able]]**

- **Words can have internal structure**
- **Morphotactics describes constraints on morpheme order**
- **Morphotactics can be determined by**
  - word syntax
  - non-syntactic factors, e.g. lexical strata

    e.g.: *non-impartial* vs. **in-non-partial*



# Types of affixation processes




Source: Berthold Crysmann 2006

# Prefixation, Suffixation, Circumfixation

- Prefixation and Suffixation are crosslinguistically predominant affixation processes
- In English and German, most inflectional and derivational affixes are suffixes
- In Bantu languages, such as Swahili, prefixation is dominant
- Circumfixation can be described as simultaneous addition of pre- and suffixes
- Ex: German regular past participles

  *ge+arbeit+et* `worked'

Source: Berthold Crysmann 2006

# Infixation

- **Infixes are affixes which are inserted into the base, thereby leading to discontinuous bases**
- **The infix itself is continuous**
- **Infixation is rare in European languages**
- **Infixation can be motivated by prosodic factors**
  - e.g. Tagalog *um* + *sulat* = *s-um-ulat,* (vs. *um* + *aral* = *um-aral*)
  - Avoidance of closed syllables (consonant-final syllables)
  - Prosodic conditioning of infixation extensively studied in Optimality Theory (McCarthy and Prince)
- **Infixation can also be purely morphologically conditioned**
  - e.g. Udi infixation (Harris 1997)

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

Source: Berthold Crysmann 2006

# Transfixation

- **Transfixation is an affixation where the segmental material of root and affix gets interleaved**
  - i.e. both the root and the affix are discontinuous
- **Transfixation is widely attested in Semitic languages, e.g. Arabic and Hebrew**
- **Ex.: forms of the Arabic root *ktb***

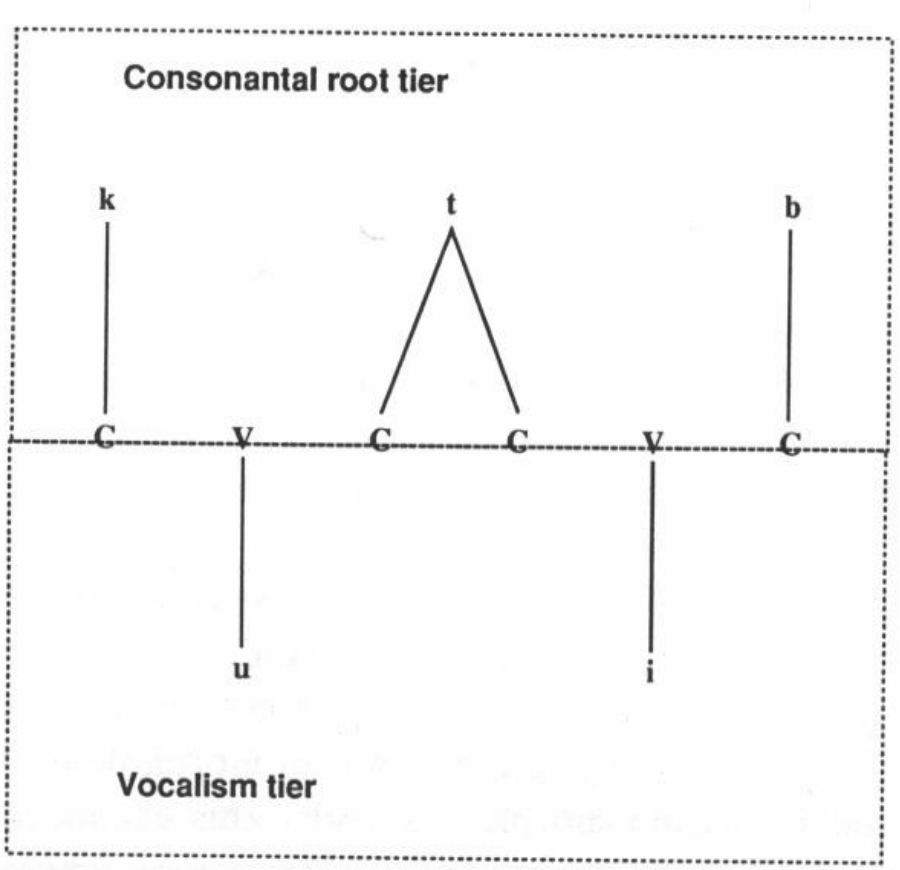
| Binyan | ACT (a) | PASS (u i) | Template | Gloss |
|---|---|---|---|---|
| I | *katab* | *kutib* | CVCVC | write |
| II | *kattab* | *kuttib* | CVCCVC | cause to write |
| III | *kaatab* | *kuutib* | CVVCVC | correspond |

- **Theoretically modeled by means of multidimensional representations (Autosegmental Phonology), associating consonantal and vocalic tiers to a CV skeleton**

Source: Berthold Crysmann 2006

# Transfixation

- **Theoretically modeled by means of multidimensional representations (Autosegmental Phonology), associating consonantal and vocalic tiers to a CV skeleton**




Source: Berthold Crysmann 2006

# Modification

- **Morphological process affects stem-internal segments**
- **Typical examples include “ablaut” and “umlaut” in German and English**
- **Umlaut:**
  - Phonologically predictable segmental alternation (e.g. fronting in German): a → ä, o → ö, u → ü
  - *Mutter* (sg)→ *Mütter* , *Wald* (sg)→ *Wälder* (pl), *Tod* (N)→ *tödlich* (A)
  - Umlaut in German is morphologically conditioned: e.g. Futter (sg)
- **Ablaut:**
  - Phonologically unpredictable segmental alternation
  - *g**e**hen – g**i**ng – geg**a**ngen* vs. *s**e**hen – s**a**h – ges**e**hen*

# Subtractive morphology

- **Process which marks morphological category by removing segments from the base**
- **Shape of the base cannot be predicted from the shape of the derived form**
- **Subtractive morphology presents severe foundational problem for morpheme-based theories of inflection and derivation**
- **Ex: Koasati**

| singular | plural | gloss |
|---|---|---|
| *pit**af**+fi+in* | *pit+li+n* | to slice up the middle |
| *las**ap**+li+n* | *las+li+n* | to lick something |
| *acokcan**a:**+ka+n* | *acokan+ka+n* | to quarrel with someone |

Source: Berthold Crysmann 2006

# Suprasegmental marking

- **Stress shift**
  - English verb-noun derivation:

    prod**u**ce (V) – pr**o**duce (N)
    perm**i**t (V) – p**e**rmit (N)
    imp**o**rt (V) – **i**mport (N)
    ins**u**lt (V) – **i**nsult (N)
    disc**ou**nt (V) – d**i**scount (N)

- **Tone**
  - Kanuri (North-eastern Nigeria)
    lezè (subjunctive) – lezé (optative) 'gehen'
    tussè (subjunctive) – tussé (optative) 'ruhen'

# Reduplication

- **Morphological process where (part of) the base is copied**
- **Often used to express categories such as plurality, iterativity, habituality etc.**
- **Total reduplication**
  - entire base is copied, e.g. Indonesian
    *orang* `man' – *orang orang* `men'
  - redup[lication can interact with segmental changes, e.g. Javanese
    *bali `return'* – *bola+bali* `return repeatedly/habitually'
- **Partial reduplication**
  - segmental material is partially copied, typically, a prosodic constituent, like a syllable or a foot, e.g. Yidin$^{y}$

    | | | |
    |---|---|---|
    | *mulari* | *mula+mulari* | `initiated man' |
    | *gindalba* | *gindal+gindalba* | `lizard' |
- **Autosegmental Phonology assumes affixation of CV templates and spreading (copying) of segments to skeleton slots**

# Morphophonology

- **Morphological process can trigger phonological or graphemic alternations**
- **Phonological alternations at the juncture between morphemes are highly frequent (internal Sandhi**
- **Sandhi can also occur at word boundaries (external sandhi)**
- **Morphophonological alternations**
  - Assimilation
    - Homorganic nasal assimilation
      *iN+possible* = *impossible* [imp...]
      *iN+complete* = *incomplete* [iŋk...]
    - Voicing assimilation
      cat+s = [...ts]
      dog+s = [...gz]
  - Epenthesis: *wish+s = wishes [wišiz]*
  - Deletion
- **Graphemic alternations**
  - *y + s ~ ies*

Source: Berthold Crysmann 2006

# Harmony processes

- **Phonological processes can also apply long-distance**
- **Harmony processes require identity of segments (typically vowels) with respect to some feature**

  **E.g. Finnish front/back vowel harmony**

  **[back +] vowels: a, u, o**
  **[back - ] vowels: ä, y, ö**
  **neutral vowels: i, e**

  **taivas (NOM) – taivas+ta (PART) – *taivas+tä**
  **lyhyt (NOM) – lyhyt+tä (PART) – *lyhyt+ta**

- **Number of interacting harmony processes highly restricted**
  - typically 1, at most 2 (Warlpiri)
  - Low number may be correlated with set of distinct features (Koskenniemi)

Source: Berthold Crysmann 2006

# (Morpho)phonology in Generative Grammar

- First formalisation of phonological rule systems goes back to Chomsky & Halle (1968)'s SPE model
- Phonological rules were context-sensitive rewrite rules of the general form:

  $a \rightarrow b / v \_ w$

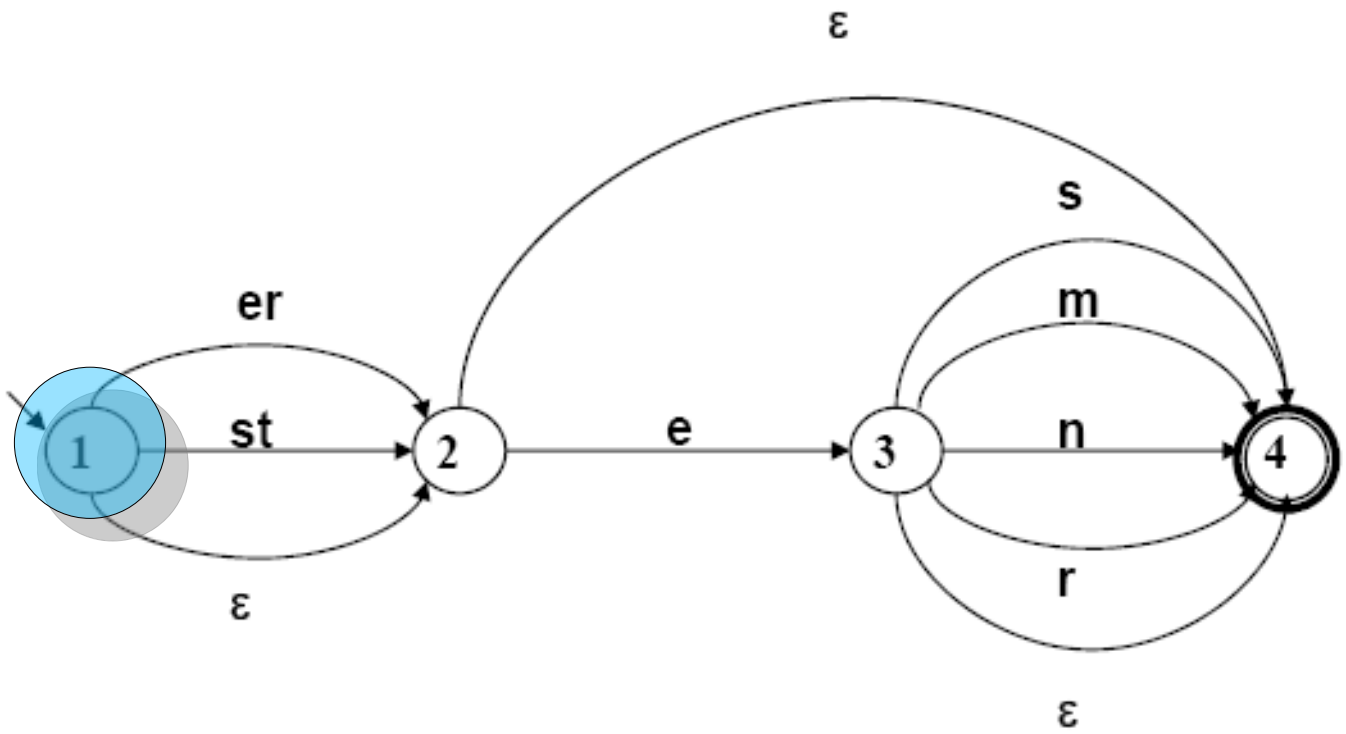
- Generative model derives surface from by successive rule application to an abstract underlying form
- Rules were assumed to be ordered
- Johnson (1972) observed that the full generative power was hardly ever used in actual phonological descriptions

# Automata - NFAs

- **Definition**
  - A nondeterministic finite state automaton is a quintupel A= (Q, E, $\delta$, $q_0$, F), with
  - Q: a finite set of states
  - $\Sigma$: a set of input characters (an alphabet)
  - $q_0 \in Q$: an initial state
  - $F \subseteq Q$: a set of final states
  - $\delta$: a transition **relation** Q x $\Sigma^*$ x Q
- Worst case complexity of NFAs is exponential to word length

# NFAs: Example automaton

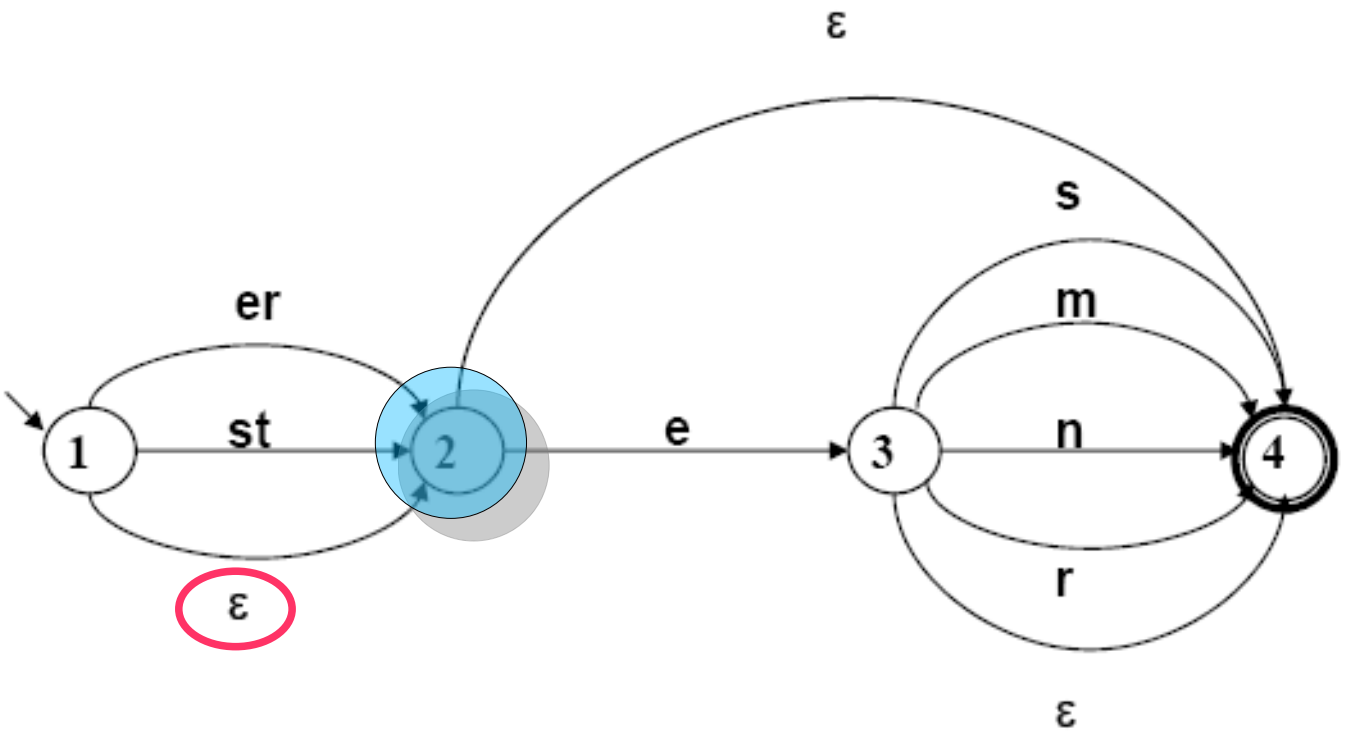
- klein + er +es




Source: Berthold Crysmann 2006

# NFAs: Example automaton

- klein + er +es

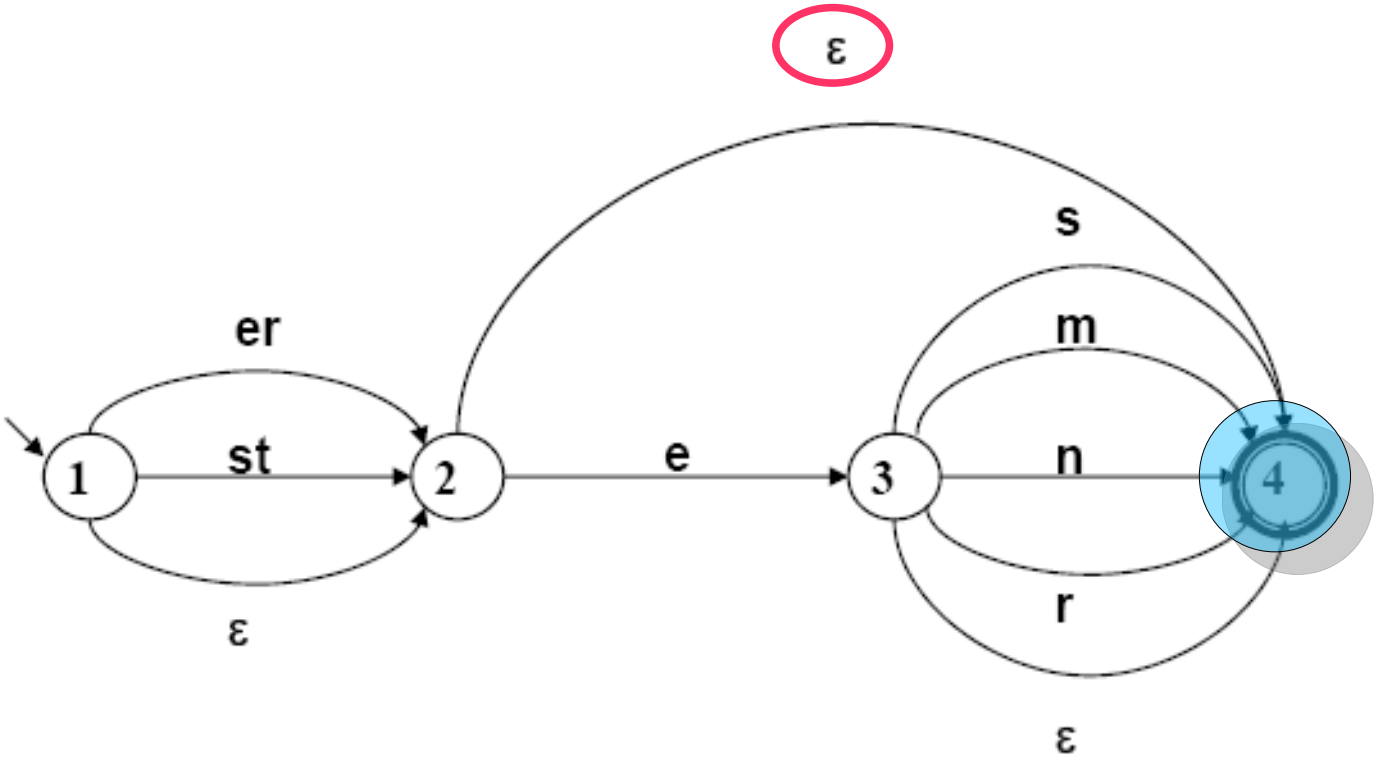



Source: Berthold Crysmann 2006

# NFAs: Example automaton

- klein + er +es

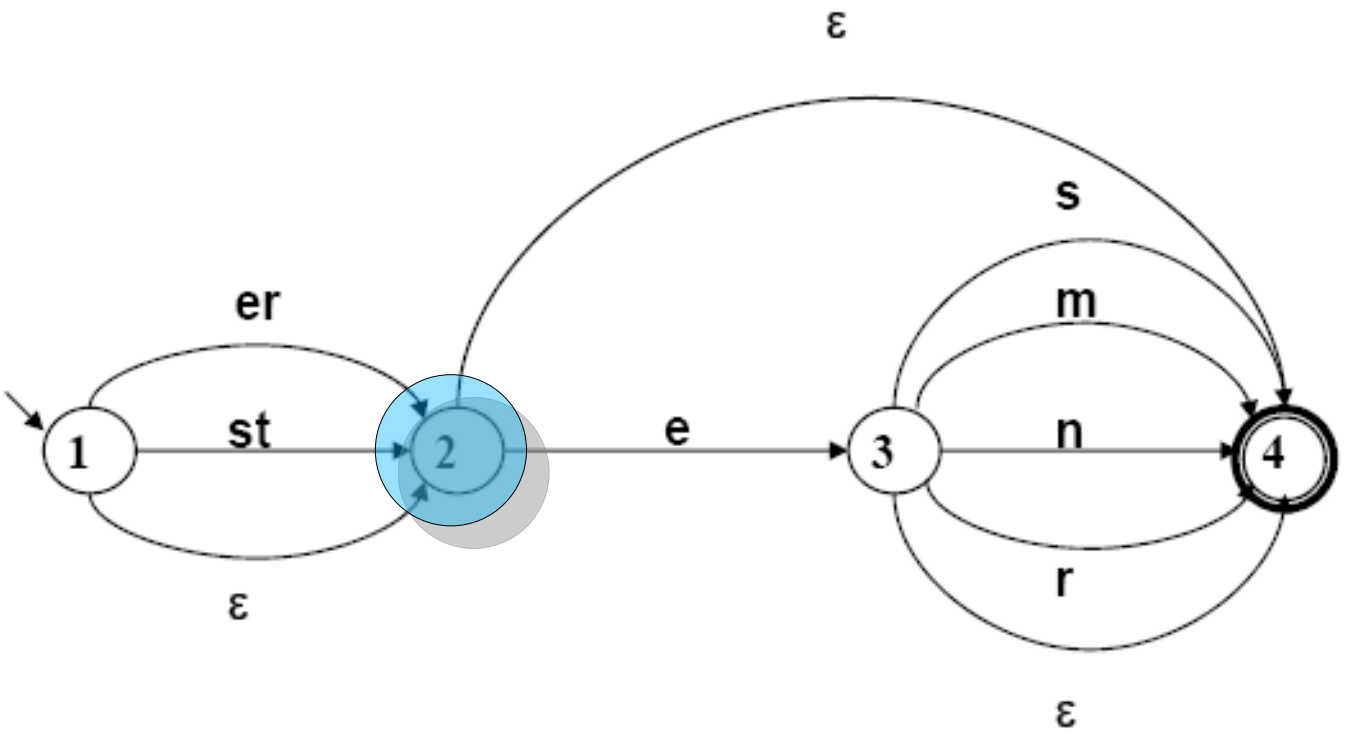
Failure



# NFAs: Example automaton

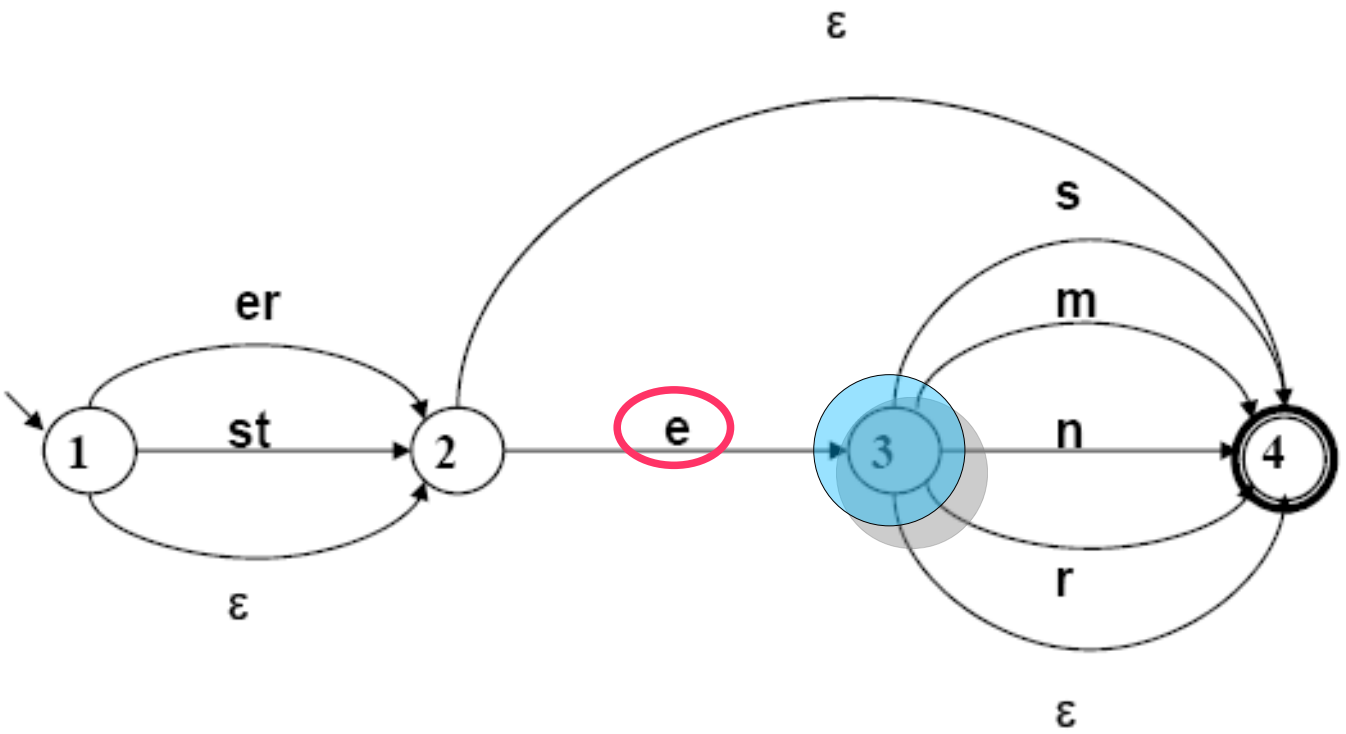
- klein + er +es

Backtracking



# NFAs: Example automaton

- klein + er +es

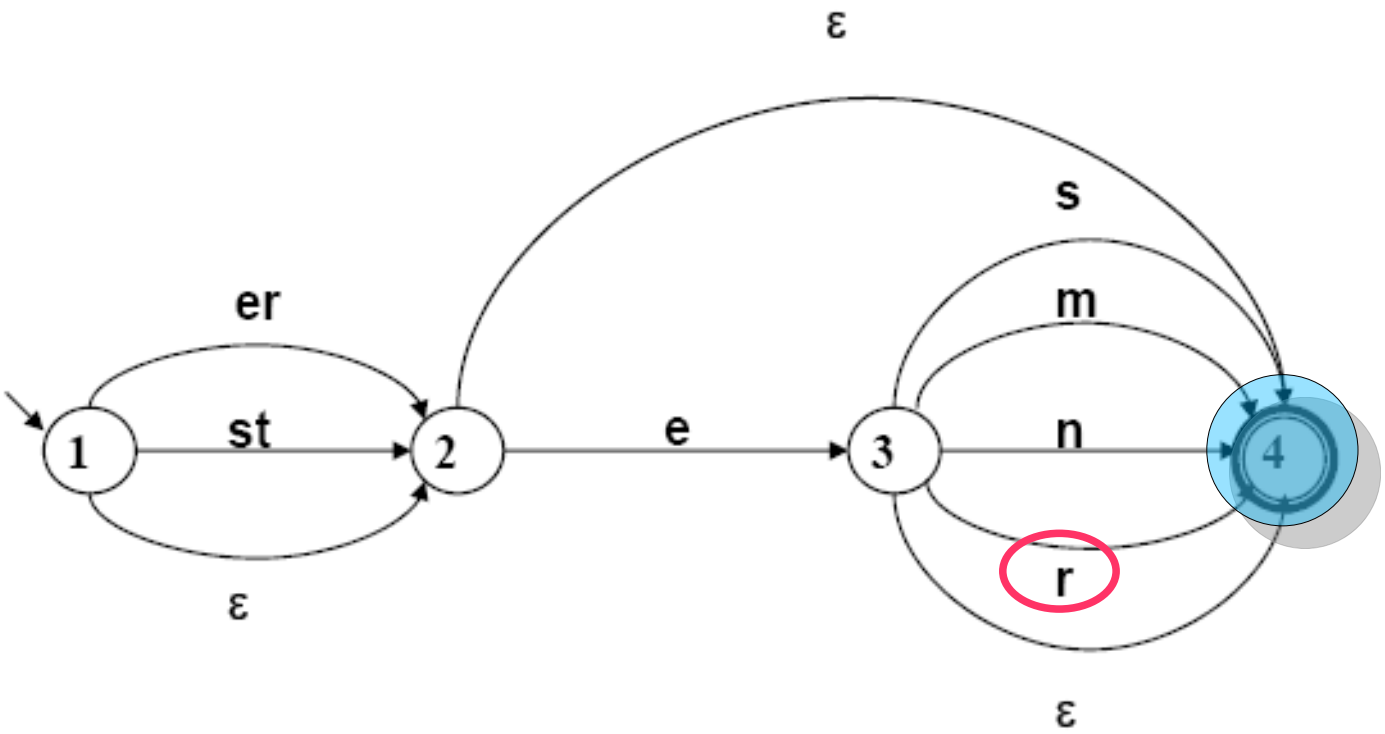



Source: Berthold Crysmann 2006

# NFAs: Example automaton

- klein + er +es

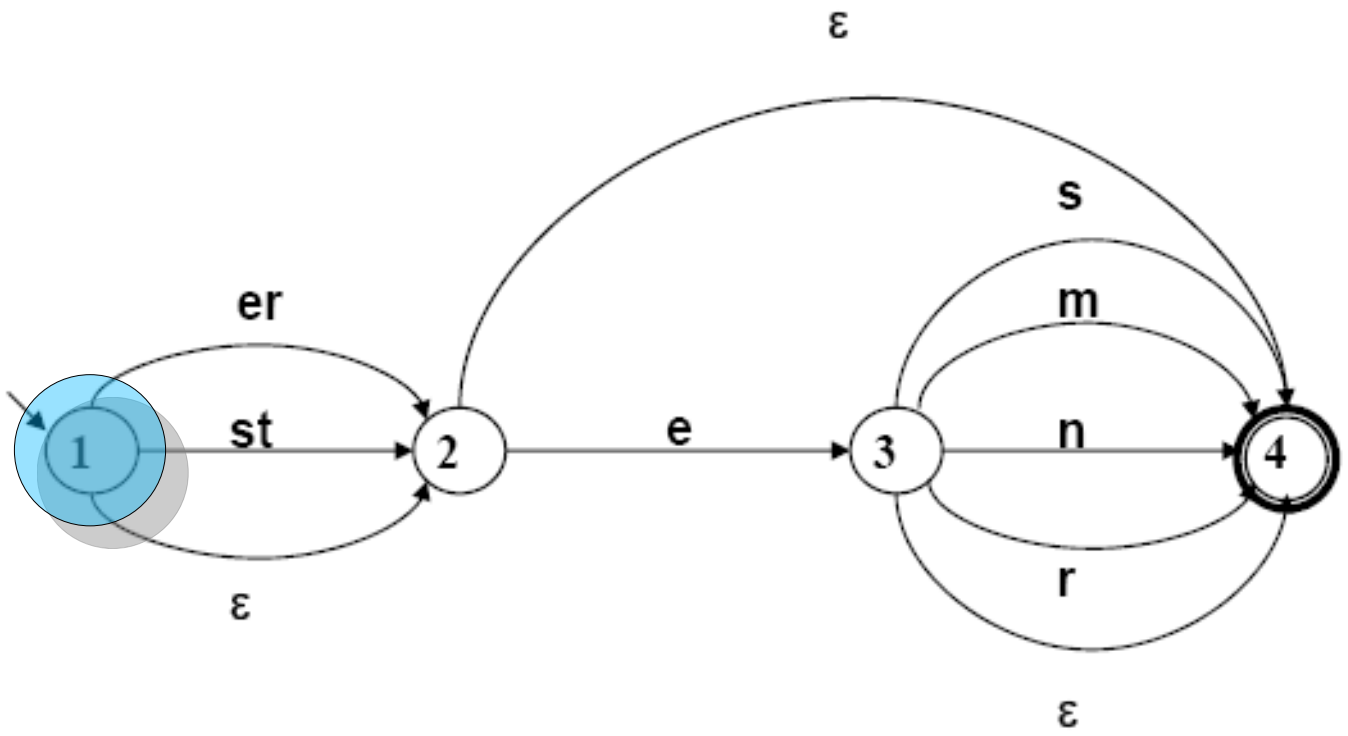
Failure



# NFAs: Example automaton

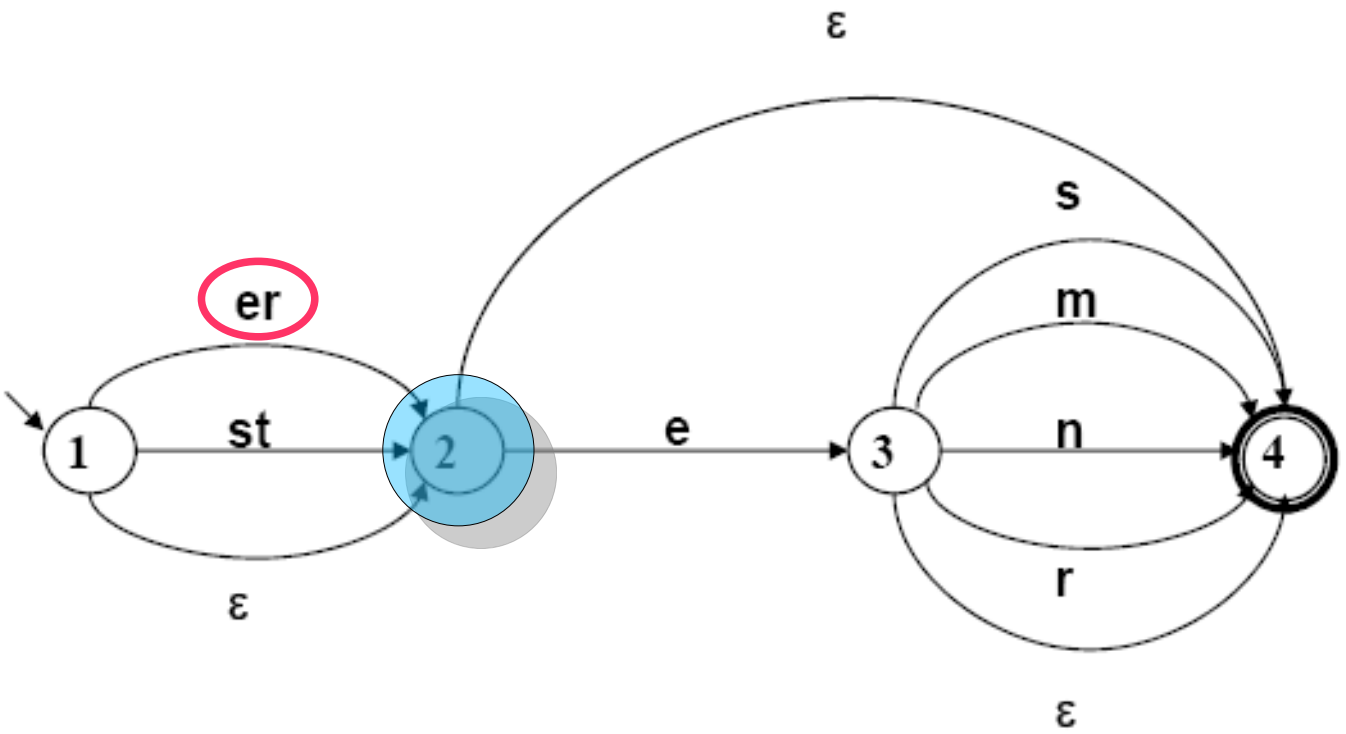
- klein + er +es

Backtracking



# NFAs: Example automaton

- klein + er +es

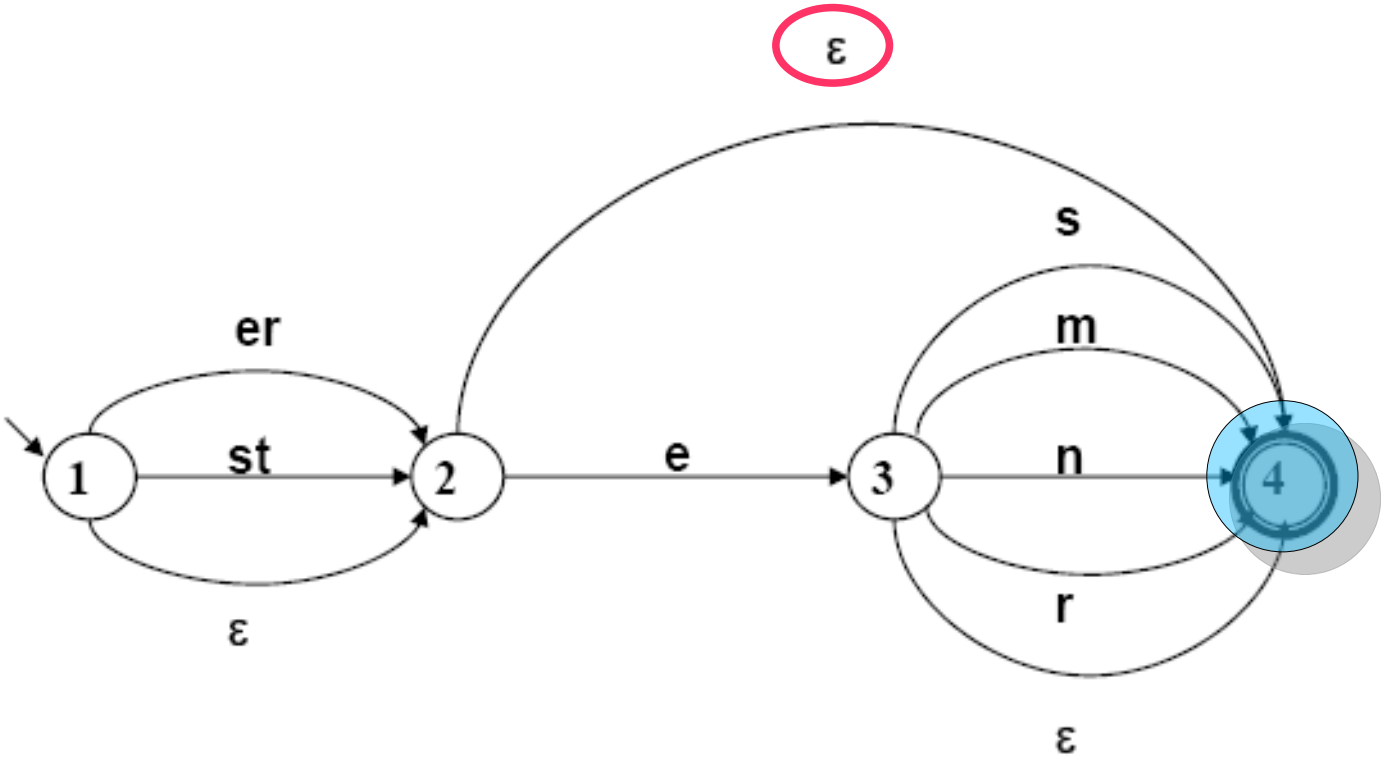



Source: Berthold Crysmann 2006

# NFAs: Example automaton

- klein + er +es

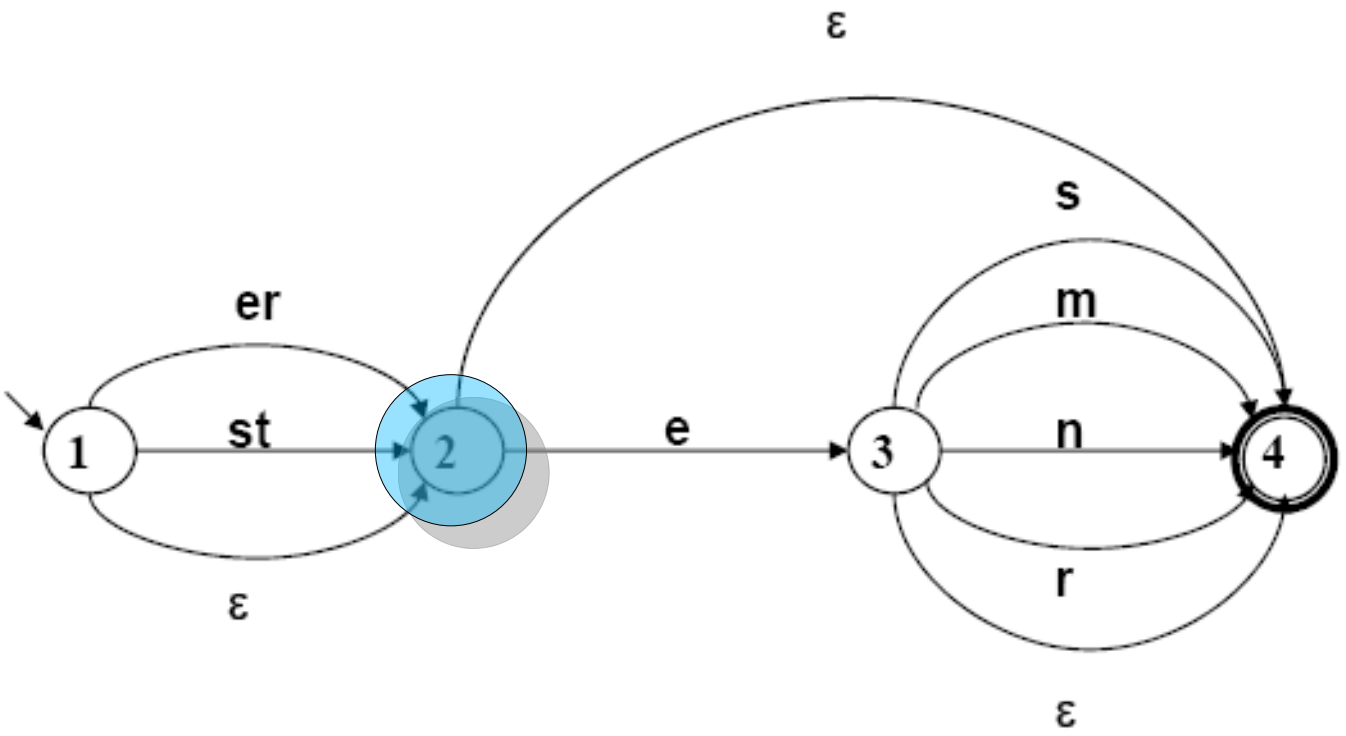
Failure



# NFAs: Example automaton

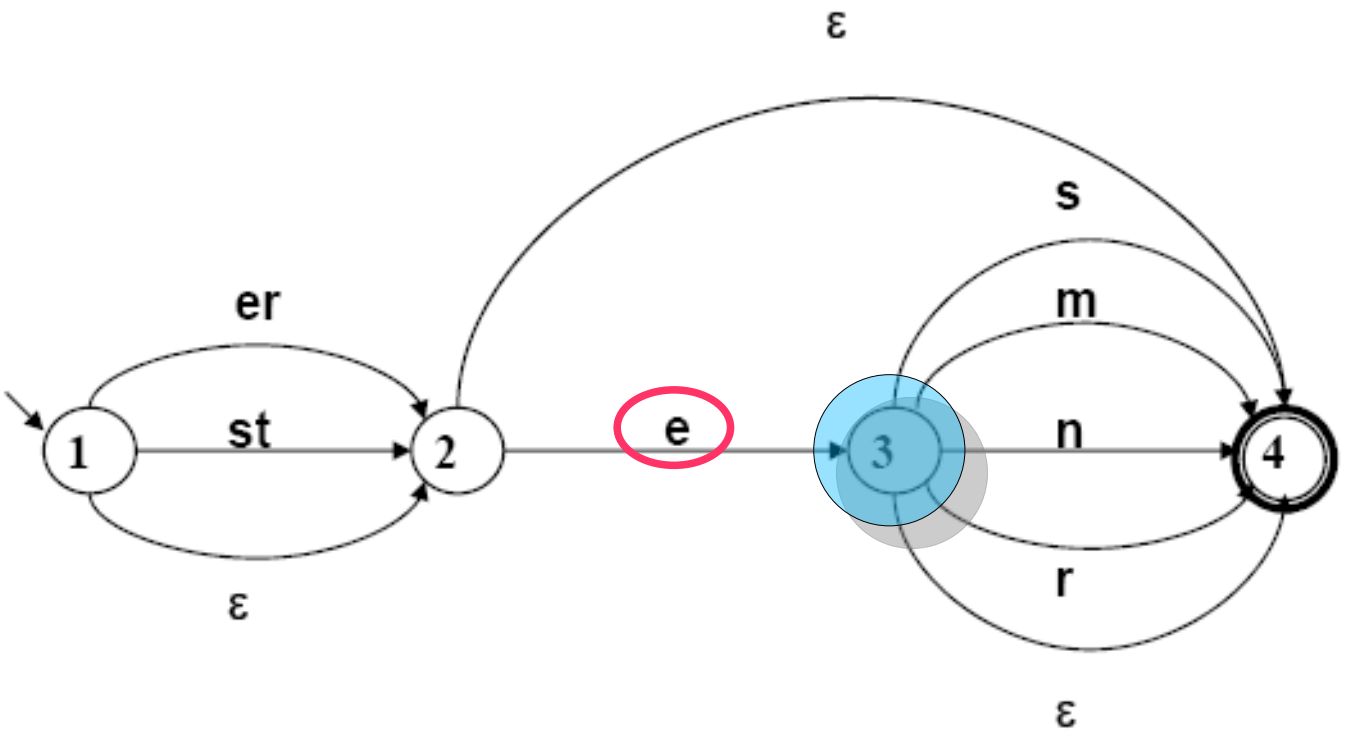
- klein + er +es

Backtracking



# NFAs: Example automaton

- klein + er +es

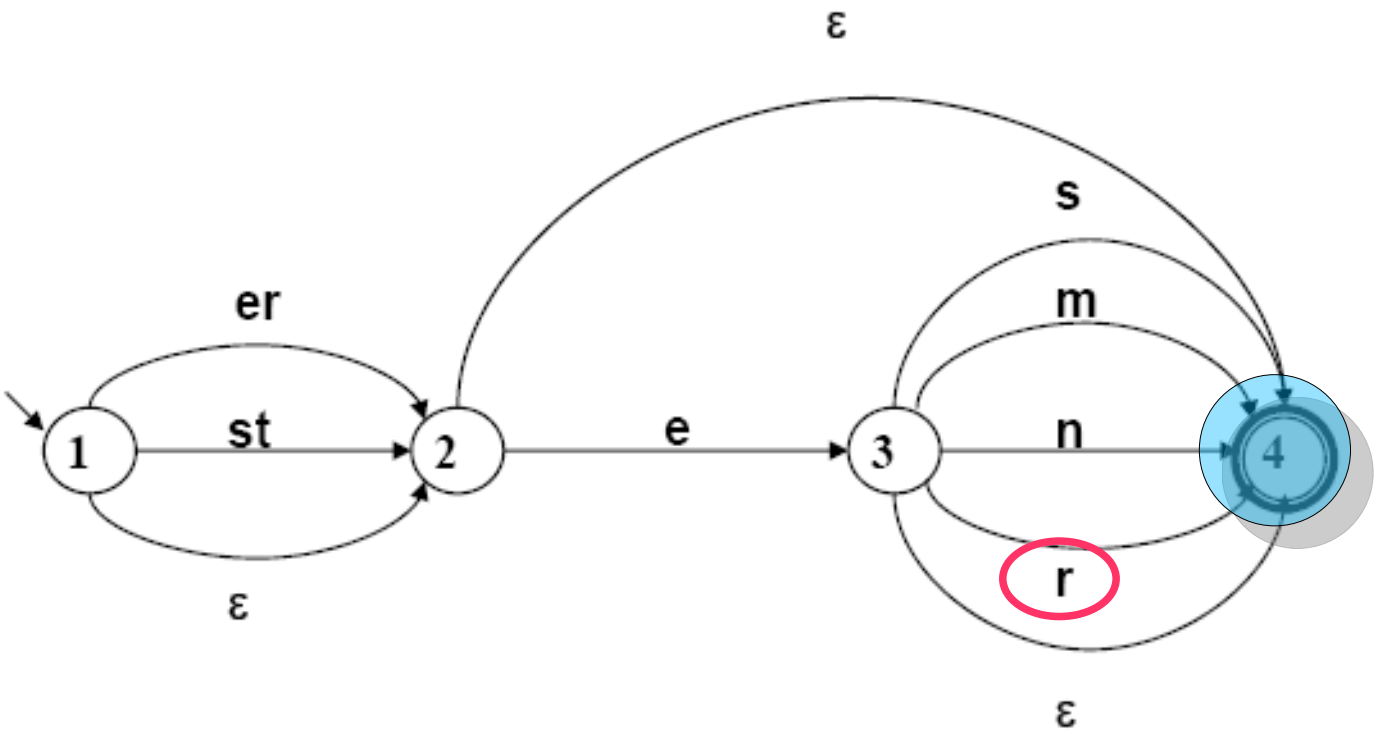



Source: Berthold Crysmann 2006

# NFAs: Example automaton

- klein + er +es

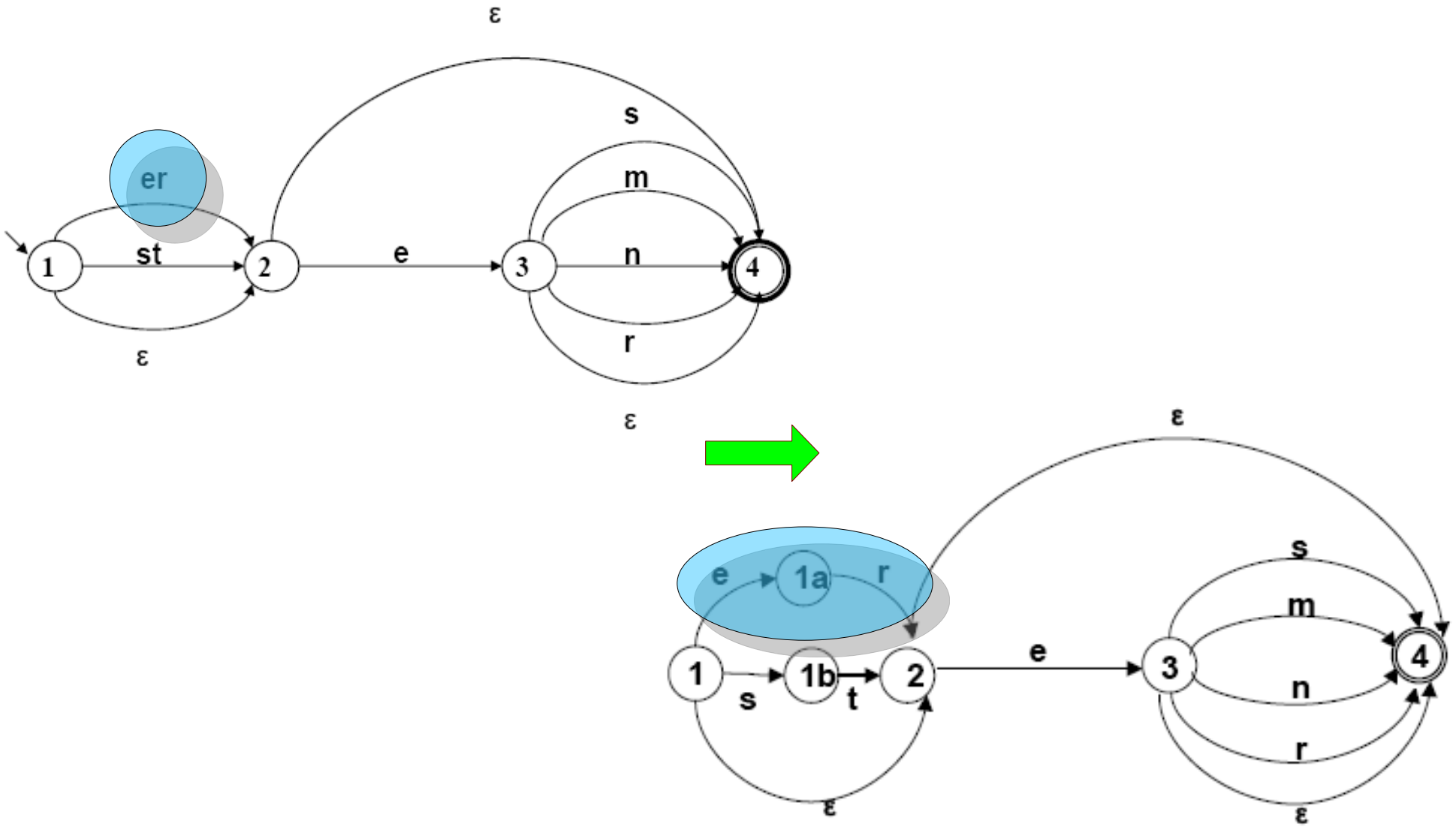
At last!



# Automata - DFAs

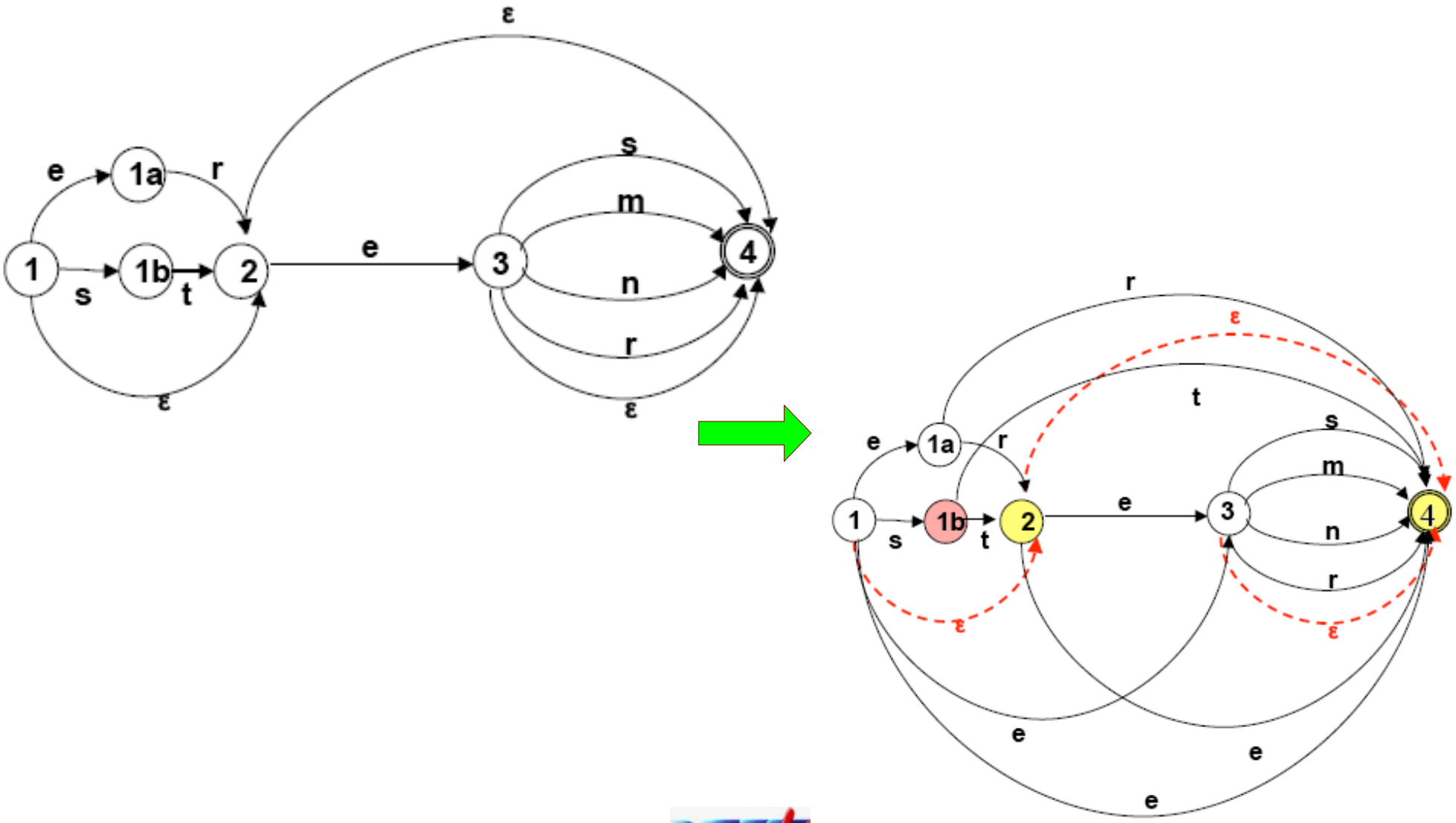
- **Definition**
  - A deterministic finite state automaton is a quintupel A= (Q, E, $\delta$, $q_0$, F), with
  - Q: a finite set of states
  - $\Sigma$: a set of input characters (an alphabet)
  - $q_0 \in Q$: an initial state
  - $F \subseteq Q$: a set of final states
  - $\delta$: a transition **function** $Q \times \Sigma \rightarrow Q$
- **For every NFA there is always an equivalent DFA (Hopcroft & Ullman 1979)**
  - Algorithm for determinisation involves
    - expansion of edges consuming more than 1 input character
    - elimination of ε-transitions (insertion of additional edges)
    - construction of power automaton (recursively combine states reached by same input symbol into new state)
  - Worst case complexity for DFAs is linear

# NFA to DFA conversion: expand multiple character transitions

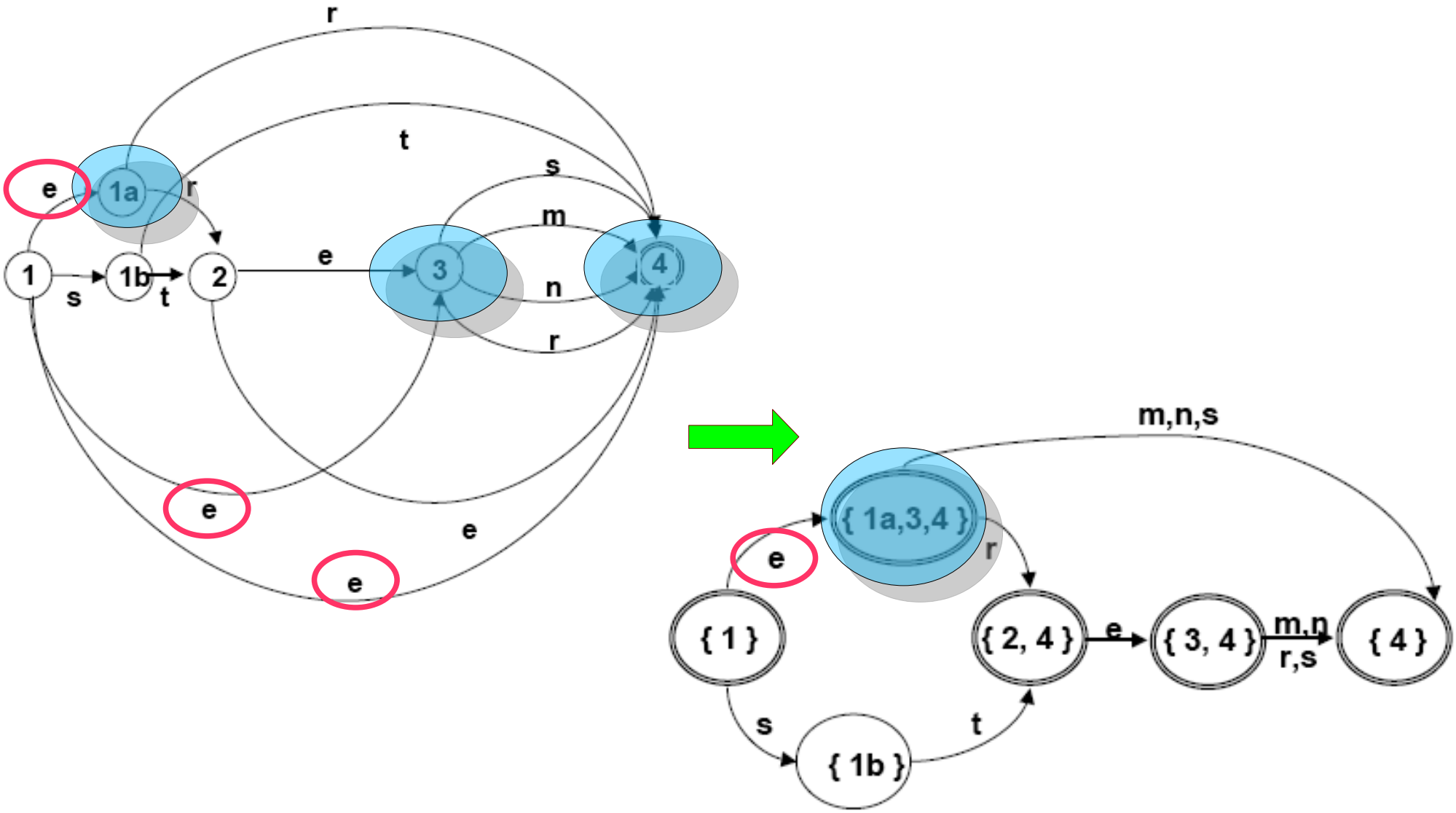



Source: Berthold Crysmann 2006

# NFA to DFA conversion: ε-elimination

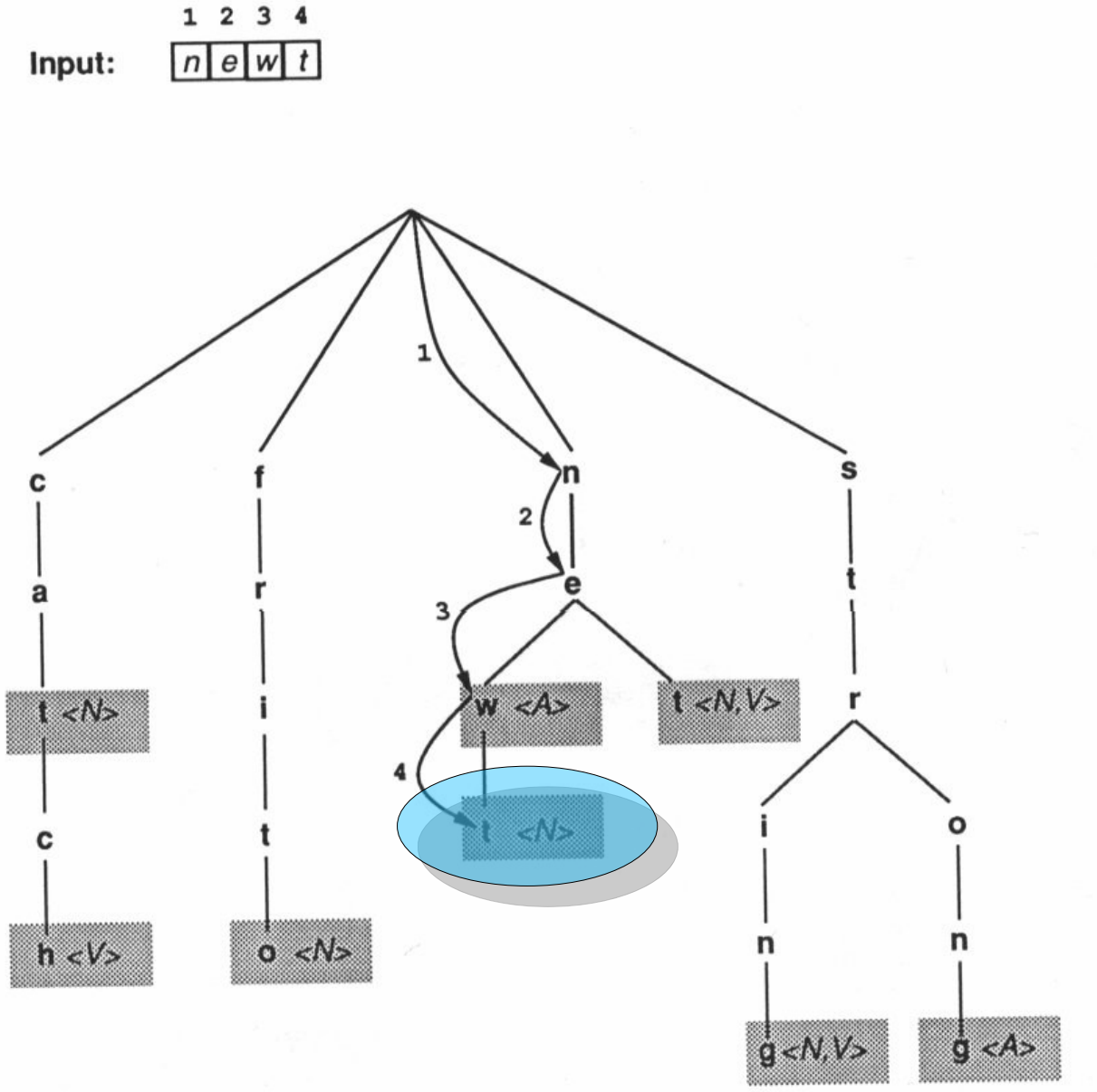


# NFA to DFA conversion: construct power automaton



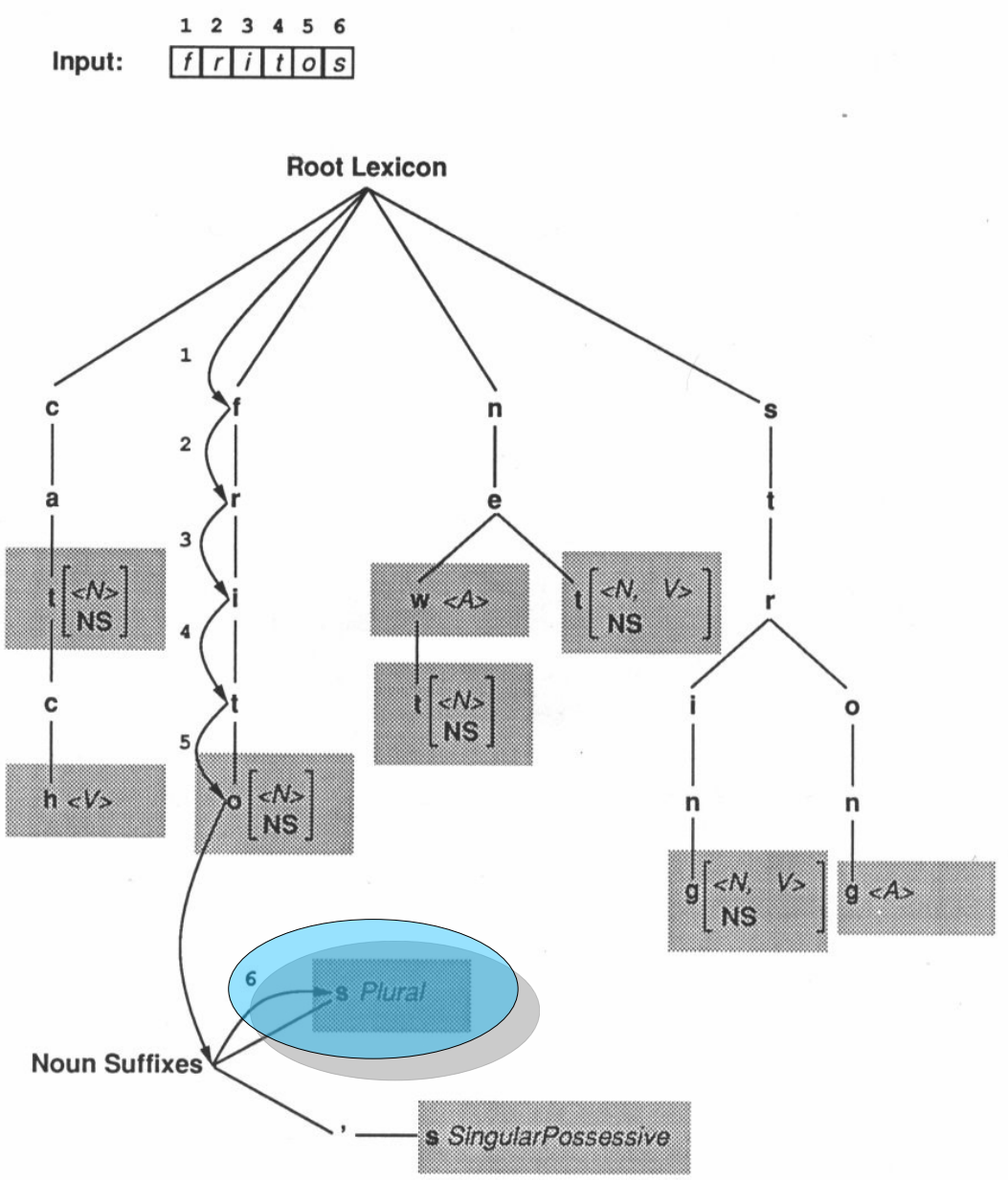
# Lexicon

- NL lexica can be efficiently encoded a letter trees (tries)
- Final states can be associated with featural annotations
- Lookup cost is proportial to string length (linear)



# Lexicon

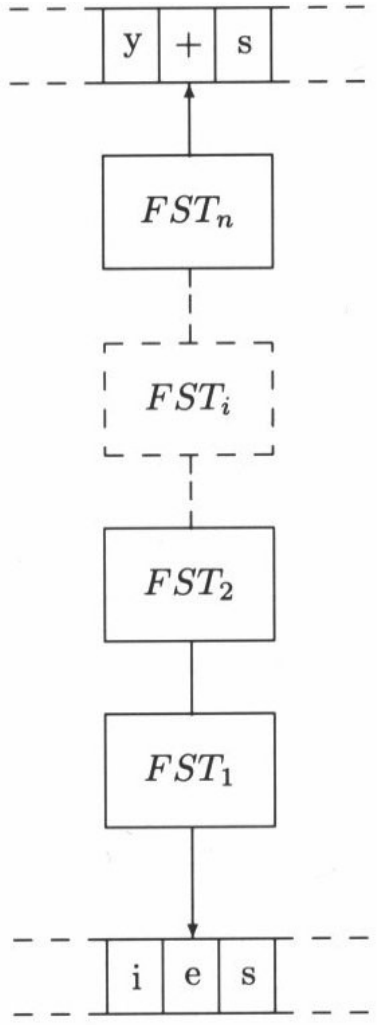
- Stem lexica can be combined with suffix lexica using continuation classes
- Stem and affix lexica can be compiled (concatenation) into a single automaton




Source: Berthold Crysmann 2006

# Finite state transducer

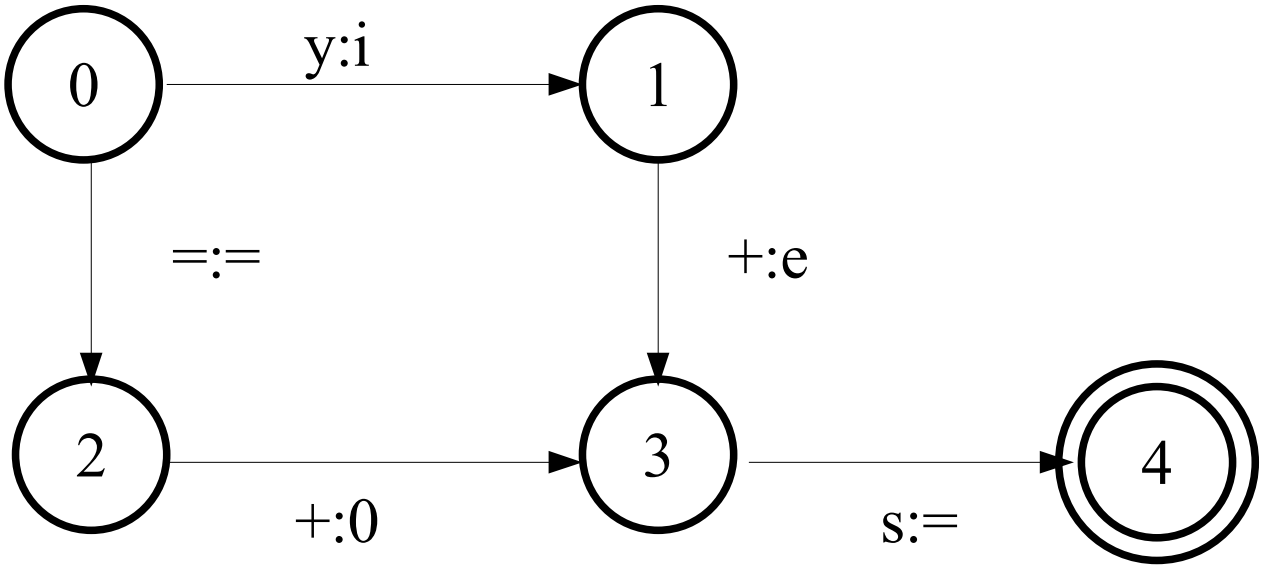
- **FSTs are finite state machines that accept languages of symbol *pairs* m:n**
- **By convention, left hand symbols correspond to the lexical tape, right hand symbols to the surface tape**
- **Kay & Kaplan (1983) suggest cascaded *finite state transducers (FSTs)* as a model for SPE-style phonology**
- **Cascade of transducers can be composed into a single FST**
- **Resulting FST can be minimised**
- **In contrast to FSAs, FSTs cannot always be processed deterministically, unless when running as an acceptor**
  - Ex: ((x:a)* a:a) | ((x:b)* b:b) running as a generator
- **Solution: compose phonological FST cascade with lexicon**




Source: Berthold Crysmann 2006

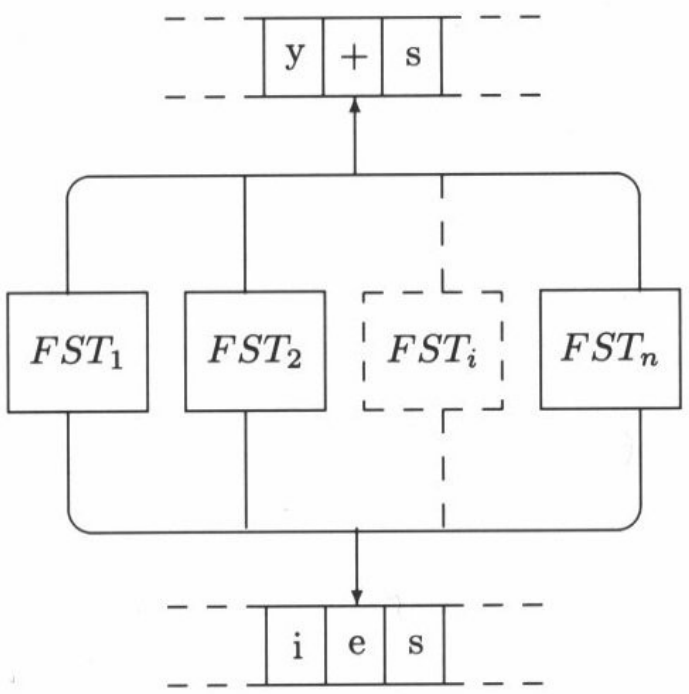
# Example

- **An FST for English *y+s ~ ies***



# Koskenniemi's Two-Level Morphology

- Two-level model of morphology employs parallel transducers instead of cascades
- Model directly relates lexical tape to surface tape
- Lexicon encoded as FSAs with continuation classes
- Parallel transducers efficiently processed even without composition
- Rule interaction must be taken care of in individual transducers




Source: Berthold Crysmann 2006

# Rule interaction

- **Example from Arabic (simplified)**
- **Rules:**
  - Glide deletion: {w,y} → 0 / V _ V
  - Vowel assimilation: V → i / _ i
- **Application: Glide deletion feeds vowel assimilation**
  - Example: *quwila* → *quila* → *qiila*
- **FST cascade: Assimilation**

| State: | V:i | i:i | =:= |
|---|---|---|---|
| 1: | 2 | 1 | 1 |
| 2: | 2 | 1 | 0 |

- **Two Level: Assimilation**

| State: | V:i | i:i | =:0 | =:= |
|---|---|---|---|---|
| 1: | 2 | 1 | 1 | 1 |
| 2: | 2 | 1 | 2 | 0 |

**Glide deletion:**

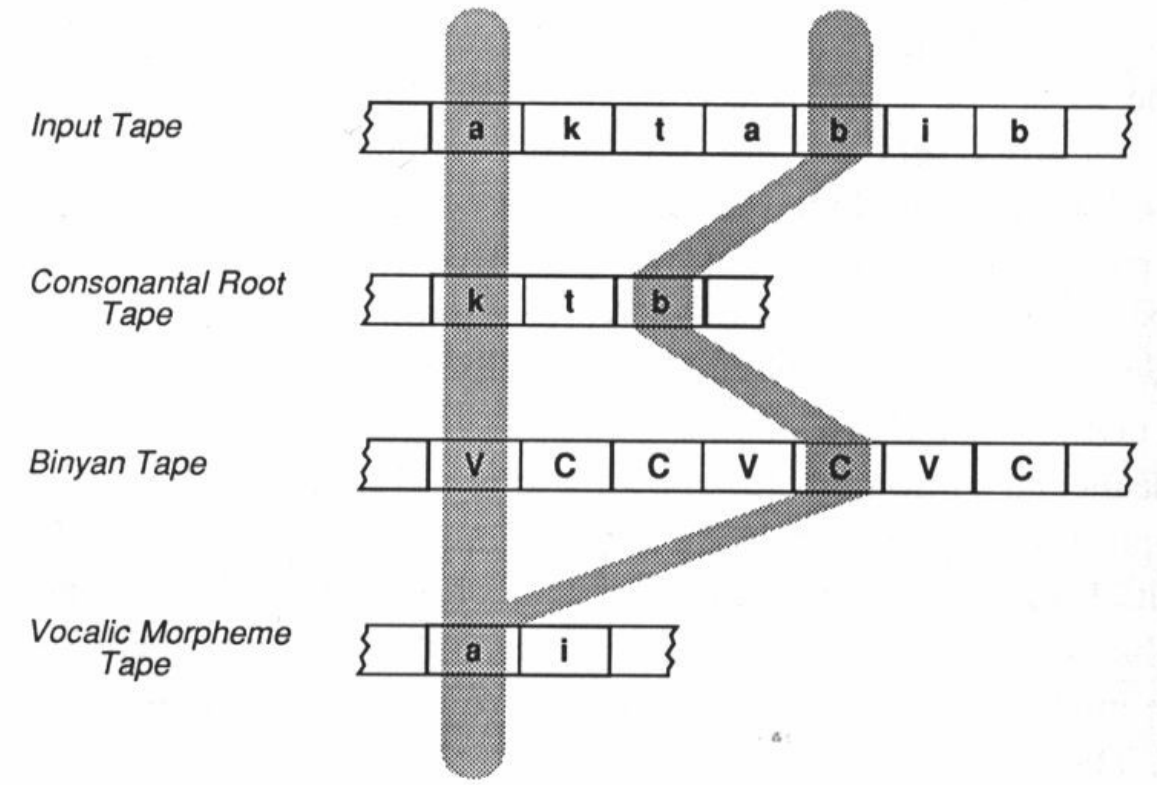
| State: | V:V | {w,y:0} | =:= |
|---|---|---|---|
| 1: | 2 | 0 | 1 |
| 2: | 2 | 3 | 1 |
| 3: | 2 | 0 | 0 |

- **Two level rule must explicitly refer to deleted segments in assimilation automaton**

Source: Berthold Crysmann 2006

# Root & Pattern Morphology

- Kay 1987 proposes finite state approach to Arabic
- Transducer involves 4 tapes operating parallel
- Surface tape and CV tape are always advanced together (step lock)
- CV tape controls alignment of input tape with root and vowel tapes
- Beesley (1989) uses different lexica for consonantal roots and vocalised CV templates




Source: Berthold Crysmann 2006

# Reduplication

- **Reduplication goes well beyond finite state machines**
- **However:**
  - at least partial reduplication can be approximated
  - enumerate all affix shapes derivable from the shape of (a portion of) the base
  - ensure that affix segments are matched to the segments of the base
- **Antworth (1990) implements a finite state machine for Tagalog CV-reduplication**
  - *pili* *pi+pili* `choose'
  - *tahi* *ta+tahi* `sew'
  - *kuha* *ku+kuha* `take'
- **Identity of vowel and consonants are implemented as separate machines, which are then intersected**
- **Sproat (1992) estimates that application to more complex reduplicants (e.g. Warlpiri CV(CC)V) will lead to a machine with more than 14,000 states**

Source: Berthold Crysmann 2006

# Morphological processing systems

- **Inflection:**
  - lemmatisation/stemming
  - extraction of grammatical (morphosyntactic) features (preprocessing for parsing)
  - reduction in lexicon size (1:2 for English, 1:5 for German, >1>200 for Finnish/Turkish)
  - Finite state technology is state of the art
- **Derivational morphology**
  - Semi-productivity and semantic opaqueness still pose problems
  - Rule-based approaches may suffer from overgeneration
  - Lexicalisation of complex forms useful
- **Compound analysis**
  - indispensible for languages with productive compounding (e.g. German)
  - Issues:  bracketing

Source: Berthold Crysmann 2006