Vorlesung CL

Morphologie, Endliche Automaten und ihre Verwendung in der morphologischen Verarbeitung

Hans Uszkoreit

WS 07/08

Das Morphem

Das Morphem ist die grundlegende Einheit der Morphologie.

Morpheme sind wie die Phoneme oder Lexeme abstrakte Einheiten, die in der Rede durch diskrete, d.h. voneinander deutliche abgrenzbare, Einheiten realisiert werden, und zwar in der mündlichen Sprache als Phonemfolgen, in der schriftlichen als Graphemfolgen. Diese Repräsentationseinheiten werden Morphe genannt.

Morpheme, die selbständig als Wörter vorkommen können heißen freie Morpheme.
Haus, Hund, Wiese, Katze, Baum bzw. *boy, book, sing* etc. sind freie Morpheme.

Morpheme, die nicht als selbständige Wörter vorkommen können, heißen gebundene Morpheme

Phänomenbereiche der Morphologie

Flexionsmorphologie:
Markierung von Tempus, Person, Kasus, Numerus...
geht-ging
der Mann-des Mannes

Derivationsmorphologie:
Bedeutungsverändernde Bildung von Wörtern aus einem Stamm-Morphem und einem Derivationsmorphem
klar - unklar
Sache - sächlich o. sachlich

Komposition:
Zusammensetzung von mehreren Stamm-Morphemen
Bauer u. Hof - Bauernhof
Sonne u. baden - sonnenbaden

Morphologische Verarbeitung

Eingabe:

Segmente (Buchstaben/Graphe, Phone/Allophone)

Verarbeitung:

Jedes Wort wird in seine Morpheme zerlegt. Dabei müssen phonologische Prozesse rückgängig gemacht werden. Zugriff auf die zugrundeliegenden Morphem-Einträge im Lexikon. Aufbau der internen Wortstruktur.

Ausgabe:

Repräsentationen der Wörter, die mit den für die syntaktische und semantische Verarbeitung relevanten Merkmalen versehen sind.

Prozesse

Konkatenation (von Morphen)

z.B. geh + st ⇔ gehst
ab + ge + frag + t + e ⇔ abgefragte

Nichtkonkatenative Phänomene

Veränderung des Stammvokals
z.B. Umlaut
Pluralbildung (Mutter ⇒ Mütter)
z.B. Ablaut
Tempusmarkierung (geb ⇒ gab)

Morphophonologie

Wenn Morpheme konkateniert werden, kann es (nicht nur an den Verbindungsstellen) zu systematischen phonologischen Änderungen kommen.

Die phonologische Gestalt des Stammes beispielsweise kann sich ändern, wenn ein bestimmtes Morphem angefügt wird.

Wenn z.B. das englische Pluralmorphem an den Stamm *wife* angefügt wird, wird das stimmlose /f/ durch das stimmhafte /v/ ersetzt: *wife + s* → *wives.* Ähnlich bei {/haus/}+{/iz/} → {/hauziz/}.

Epenthese

z.B. bad + st => badest

Elision

z.B. ras + st => rast

Merkmalsveränderungen an Phonemen

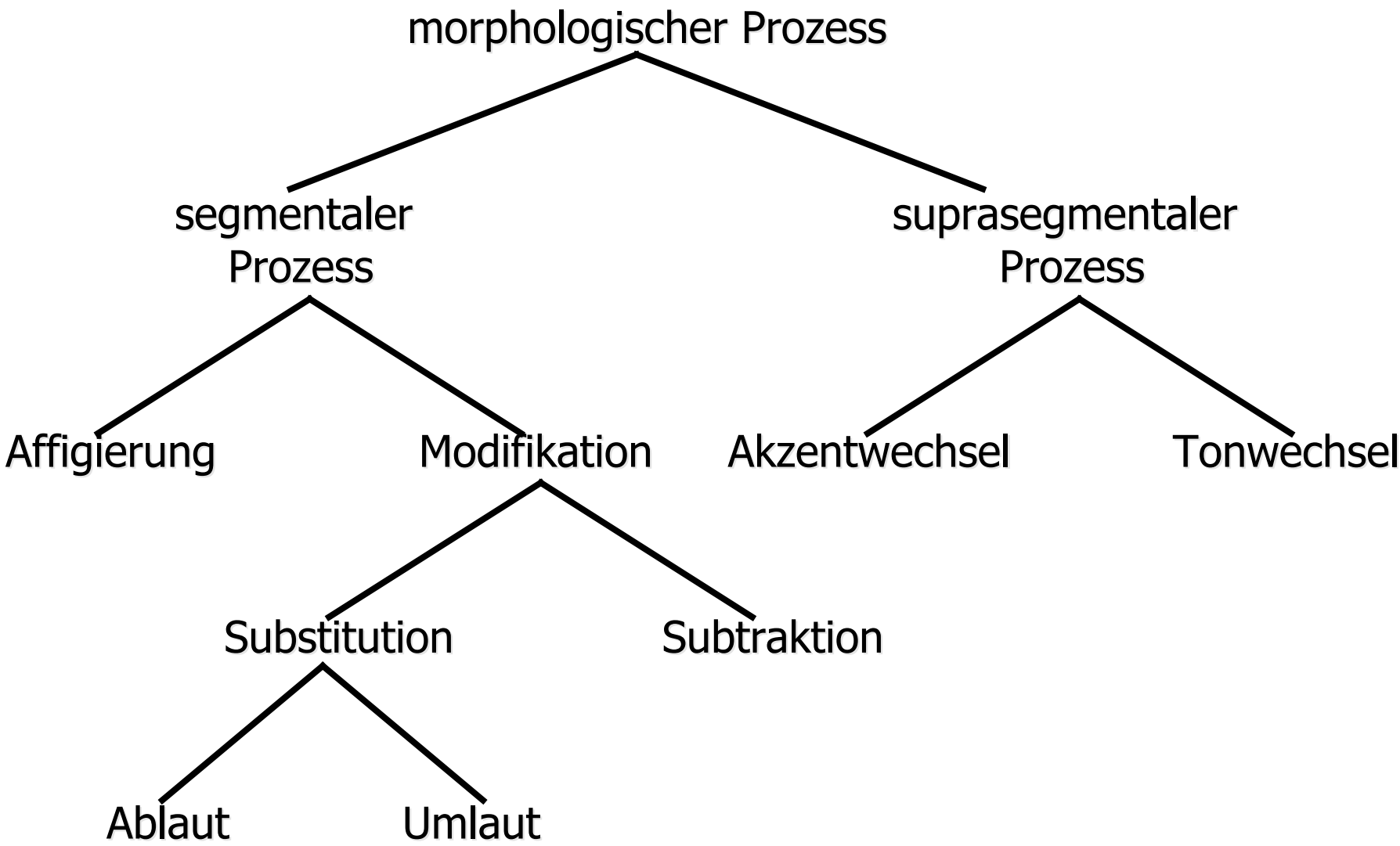
z.B. Auslautverhärtung
(stimmhafte Konsonanten werden am Silbenende stimmlos)

/bad/ => [bat]

Morphologische Prozesse

Wir können verschiedene Prozesse unterscheiden, mithilfe derer Wörter aus elementareren Elementen wie z.B. Morphemen konstruiert werden können.

- Segmental
  - Affigierung
  - Modifikation
    - Substitution von Segmenten oder Merkmalen
    - Subtraktion (Tilgung) von Segmenten
- Suprasegmental
  - Akzentwechsel
  - Tonwechsel


morphologischer Prozess
segmentaler Prozess
suprasegmentaler Prozess
Affigierung
Modifikation
Akzentwechsel
Tonwechsel
Substitution
Subtraktion
Ablaut
Umlaut

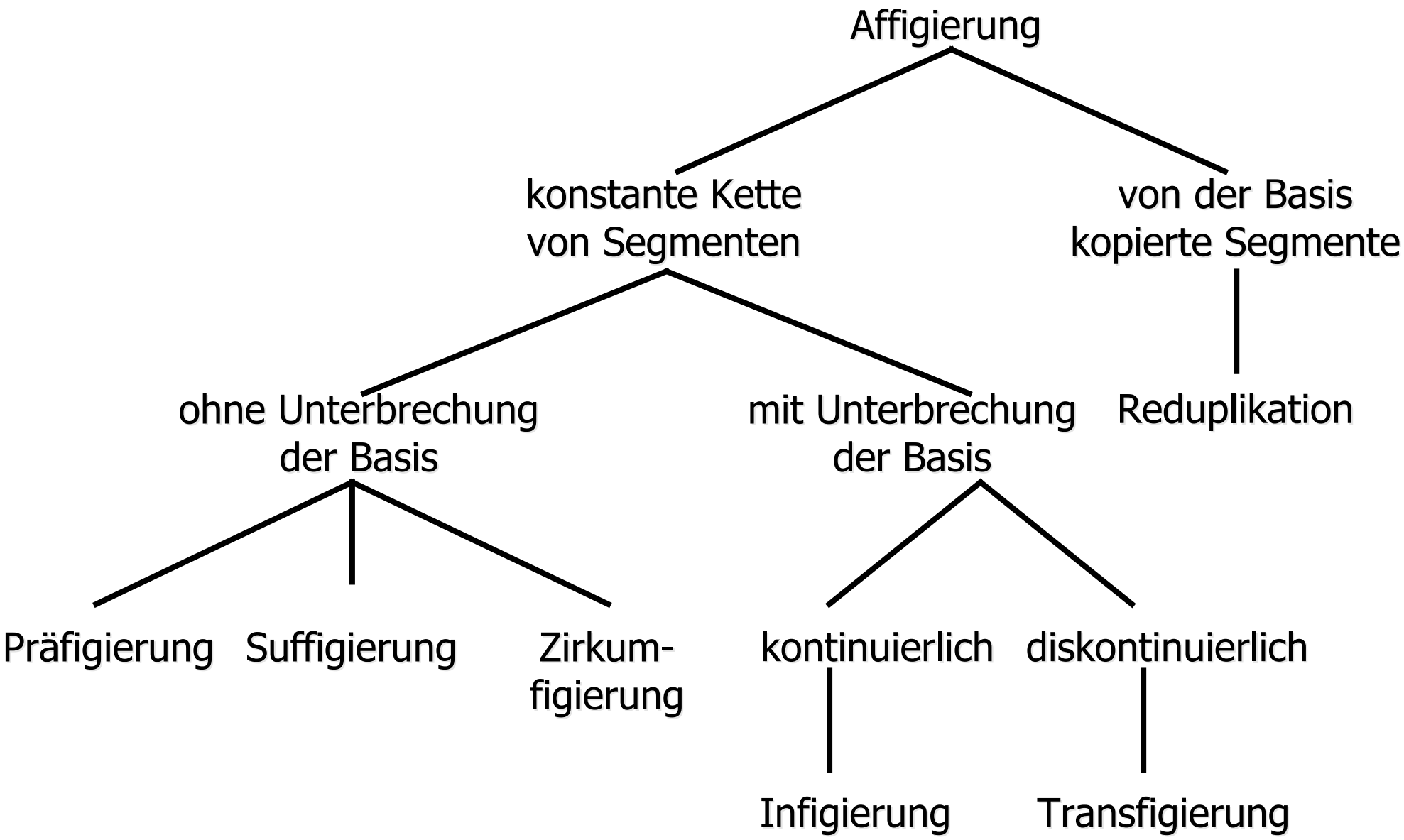
Affigierung

Der gebräuchlichste morphologische Prozeß im Deutschen und Englischen und vielen anderen Sprachen besteht im Anfügen eines Morphems an eine Wurzel oder einen Stamm:

trag + bar →tragbar

sing + ing → singing

Der morphologische Prozeß, durch den grammatische oder Lexikalische Information an einen Stamm angefügt wird, wird Affigierung genannt.


Affigierung
konstante Kette von Segmenten
von der Basis kopierte Segmente
ohne Unterbrechung der Basis
mit Unterbrechung der Basis
Reduplikation
Präfigierung
Suffigierung
Zirkum-figierung
kontinuierlich
diskontinuierlich
Infigierung
Transfigierung

Affigierung

Affigierung ist ein rekursiver Prozeß insofern er wiederholt auf seinen eigenen Output angewandt werden kann, zum Beispiel:

de + scribe → describe

describe + able → describable

in + describeable → indescribable

Die Reihenfolge, in welcher die Affixe an den Stamm angefügt werden, ist signifikant, d.h. Wörter können eine interne Struktur haben, die über die bloße sequentielle Anordnung hinaus geht:

[in[[de scribe] able]]

Affix

Affix ist der Sammelbegriff für alle Arten von Formativen, die nur in Verbindung mit einem anderen Morphem (der Wurzel oder dem Stamm) verwendet werden können, d.h. Affixe sind ein Typ gebundener Morpheme.

Affixe werden gewöhnlich in Klassen eingeteilt, je nach ihrer Position bezüglich der Wurzel oder des Stammes eines Wortes.

Ohne Unter-brechung der Basis

- Präfix
- Suffix
- Zirkumfix
- Superfix

Mit Unterbrechung der Basis

- Infix
- Transfix

Präfix - Präfigierung

Ein Präfix ist ein Affix, das am Anfang einer Wurzel oder Stammes angefügt wird.

Der Prozeß des Anfügens eines Präfixes wird Präfigierung genannt.

Der Prozeß der Präfigierung wird im Deutschen und Englischen häufig zur Bildung neuer Lexeme verwendet nicht jedoch zur Bildung von Flexionsformen:

un- + *glücklich* → *unglücklich*

un- + *happy* → *unhappy*

mini- + *computer* → *minicomputer*

Suffix - Suffigierung

Ein Suffix ist ein Affix, das am Ende einer Wurzel oder eines Stammes angefügt wird.

Der Prozeß des Anfügens eines Suffixes wird Suffigierung genannt.

Suffigierung wird im Deutschen und Englischen sehr häufig verwendet, sowohl zur Derivation neuer Lexeme als auch zum Ausdruck grammatischer Beziehungen:

heiter + *keit* → *Heiterkeit*

Kopf + *los* → *kopflos*

national + *-ise* → *nationalise*

generate + *-ion* → *generation*

Flexionsendungen wie dt. *-t , -st, -en* bzw. engl. -s, *-ed, -ing*

Zirkumfix - Zirkumfigierung

Ein Zirkumfix ist ein diskontinuierliches Affix, das um eine Wurzel oder einen Stamm gelegt wird, also aus einem präfigierenden und einem suffigierendem Teil besteht.

Der Prozeß des Anfügens eines Zirkumfixes wird Zirkumfigierung genannt.

Als Beispiel für eine Zirkumfigierung wird häufig die Bildung des Partizips im Dt. herangezogen: ge + worf + en, ge + mach + t. Da das ge- aber auch fehlen kann (bestell+t, be + komm + en) scheint das Suffix aber enger zum Stamm zu gehören. Vgl. aber die Pluralbildung im Toiko (Nord-Togo): bara 'Frau' vs. m+bara+m 'Frauen'

Infix - Infigierung

Ein Infix ist ein Affix das im Inneren einer Wurzel oder eines Stammes eingefügt wird.

Der Prozeß des Einfügens eines Infixes wird Infigierung genannt.

Infigierung ist in den europäischen Sprachen eine sehr seltene Erscheinung, findet sich jedoch häufig in asiatischen, amerikanischen und afrikanischen Sprachen. Historisch gesehen ist das *-n-* im deutschen *stand* (im Gegensatz zu *stehen)* ein Infix.

Tagalog:

sulat 'schreiben' – s-um-ulat 'schrieb' – s-in-ulat 'wurde geschrieben'

Transfix - Transfigierung

Ein Transfix ist ein mehrteiliges diskontinuierliches Affix, das mit einer Wurzel oder einem Stamm verzahnt wird.

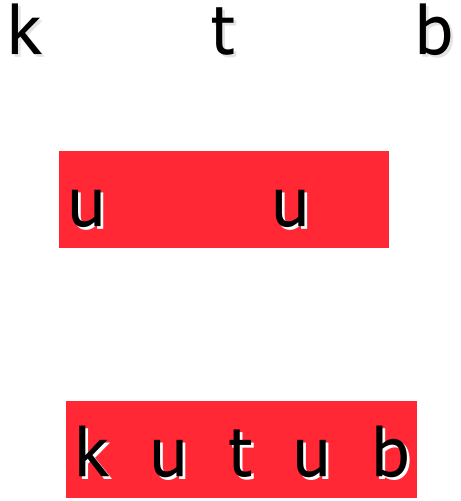
Der Prozeß des Einfügens eines Transfixes wird Transfigierung genannt.

Die Transfigierung ist ein häufig verwendetes Mittel zur Flexion in den semitischen Sprachen. Wortformen entstehen durch die Kombination von diskontinuierlichen konsonan-tischen Wurzeln, die die lexikalische Bedeutung kodieren, mit vokalischenTransfixen.

Transfix - Transfigierung

Die arabische Wurzel ktb bedeutet 'schreiben'. Daraus lassen sich z.B. im ägyptischen Arabisch Formen ableiten wie:

kátab 'er schrieb'
jíktib 'er wird schreiben'
maktúub 'geschrieben'
maktába 'Buchhandlung'
makáatib 'Buchhandlungen'
kitáab 'Buch'
káatib 'Schreiber'
kutub 'Bücher'



Reduplikation

Unter Reduplikation versteht man die Verdoppelung von an- oder auslautenden Teilen einer Wurzel oder eines Stammes zum Ausdruck morphosyntaktischer Kategorien

Gotisch:

haldan 'halten' – haihald

haitan 'heissen' – haihait

Latein:

tundo 'stoße' – tutudi 'stieß'

pello 'treibe' – pepuli 'trieb'

Maori:

tau 'Mann' – ta-tau 'Männer'

mero 'Junge' – me-mero 'Jungen'

Modifikation

Ein weiterer wichtiger morphologischer Prozeß ist die Modifikation, eine Veränderung in der Wurzel oder im Stamm eines Wortes.

Ein Beispiel dafür ist der Vokalwechsel zwischen den Singular- und Pluralformen vieler deutscher sowie einiger englischer Substantive:

dt. *Sohn ~ Söhne, Hut ~ Hüte, Lamm ~ Lämmer*

eng. *man ~ men, mouse ~ mice, goose ~ geese).*

Ein verbreiteter Vorgang dieser Art ist der Ablaut.

Ablaut

Ablaut nennt man den regelhaften Vokalwechsel in Wörtern des gleichen Lexems, der nicht phonologisch konditioniert ist.

Einschlägige Beispiele finden wir bei vielen sog. starken Verben

dt. *singen ~ sang ~ gesungen, finden ~ fand ~ gefunden, werden ~ ward ~ geworden,*

engl. *sing ~ sang ~ sung, find ~ found ~ found, give ~ gave ~ given* etc.

Der Ablaut ist vom Umlaut zu unterscheiden.

Umlaut

Umlaut ist eine Vokalalternation zwischen verwandten Vorderzungen- und Hinterzungenvokalen, die — zumindest historisch betrachtet — phonologisch konditioniert ist (regressive Assimilation unter dem Einfluß von /i, j/ in der Folgesilbe).

Wo jedoch die Bedingungsfaktoren verlorengegangen sind, muß Umlaut als ein morphologischer Prozeß aufgefaßt werden.

Beispiele:

Mutter ~ Mütter, Vater ~ Vater, Vogel ~ Vögel,

man ~ men, mouse ~ mice, fox ~ vixen (dt. Fuchs ~ Füchsin).

Subtraktion

Genus im französischen Adjektiv

maskulinum		femininum	
Schrift	Laut	Schrift	Laut
grand	[gʁã]	*grande*	[gʁãd]
petit	[pti]	*petite*	[ptit]
gris	[gʁi]	*grise*	[gʁiz]
gentil	[ʒãti]	*gentille*	[ʒãtij]

Orthographisch betrachtet werden die femininen Formen durch Anhängen von –e an den Stamm gebildet.
Phonologisch jedoch sind die maskulinen Formen durch Tilgung des Auslautkonsonanten abgeleitet.

Superfix (Suprafix) - Superfigierung

Ein Superfix (oder Suprafix) suprasegmentales Affix, das eine Wurzel oder einen Stamm überlagert.

Der Prozeß der Modifikation durch eine Superfix wird Transfigierung genannt. Manifestationen sind Akzentwech-sel und Tonwechsel zum Ausdruck grammatischer Bedeutungen:

Akzentwechsel (engl.):

prodúce (v.) vs. próduce (n), permít (v.) vs. pérmit (n.)
impórt (v.) vs. ímport (n.), insúlt (v.) vs. ínsult (n.), discóunt vs. díscount

Tonwechsel (Kanuri, Nigerien;´ = hoher Ton, ` = fallender Ton):

lezè (Konj.) vs. lezé (Opt.) 'gehen'
tussè (Konj.) vs. tussé (Opt) 'ruhen'

Konversion

Konversion ist ein besonderer Ableitungsprozeß, wobei ein Lexem in eine neue Lexemklasse überführt wird, ohne daß ein Affix angefügt wird.

Beispiele: Verb → Nomen:

schau-en → Schau

bau-en → Bau

fall-en → Fall

Da das Englische keine sehr ausgeprägte Flexion hat, ist die Konversion ein sehr verbreitetes Wortbildungsmittel;

Verb → Nomen: *smell, taste, hit, walk*

Adjektiv → Verb: *dirty, empty, lower*

Komposition

Komposition ist der morphologische Prozeß, durch den neue zusammengesetzte Lexeme durch die Kombination zweier oder mehrerer freier Formen gebildet werden.

Ein durch Komposition gebildetes Wort heißt Kompositum (engl. *compound*).

Beispiele:

dt. *Haus + Tür → Haustür, groß + Stadt → Großstadt,*

engl. *bed + room → bedroom, black + bird → blackbird,*
washing + machine → washing machine

Automaten

Automaten in der weiteren Bedeutung des Wortes sind ein zentrales Konzept aber nicht formal definiertes Konzept in der Informationsverarbeitung.

Wenn wir durch eine Sequenz von Handlungen bestimmte Effekte auslösen, dann ist das nicht unbedingt Informationsverarbeitung. Normalerweise wird mit einer jeden Handlung kausal eine Wirkung erzeugt.

Beispiel: ich öffne eine Tür mit zwei Schlössern, indem ich jedes Schloß mit einem Schlüssel aufschließe.

Wenn wir aber ein System haben, das so gebaut ist, daß es in Abhängigkeit von meinen Handlungen „Entscheidungen“ über seine Handlungen trifft, dann liegt Informationsverarbeitung vor. In diesem Fall bewirken meine Handlungen die Handlungen der Maschine nicht direkt kausal.

Automaten sind Systeme, die in Abhängigkeit von meinen Handlungen, bestimmte Aktionen ausführt bzw auslöst.

Damit wir von Automaten sprechen, muß es mindestens einen Entscheidungspunkt geben.

Automaten

Automaten werden verwendet, um Symbol- oder Handlungsabfolgen zu überprüfen, zu analysieren oder abzuarbeiten.

Sie werden z.B. bei der Verarbeitung von Benutzereingaben verwendet.

Dabei wird geprüft,

ob Eingaben korrekt sind und

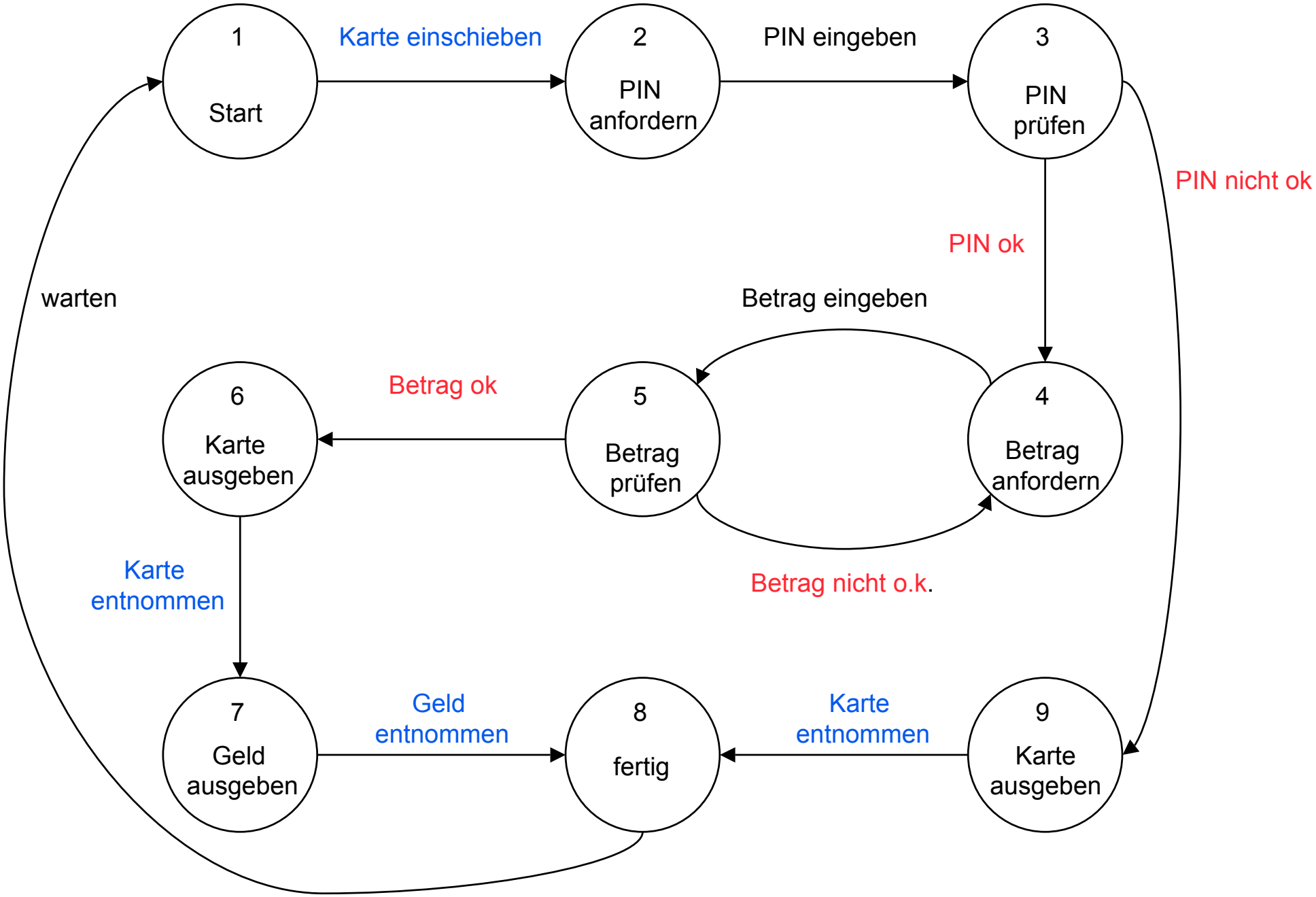
in der richtigen Reihenfolge bzw. im richtigen Kontext erfolgen.

Automaten spielen auch eine Rolle bei der Überprüfung bzw. Übersetzung von Programmier- oder Spezifikationssprachen in die internen Sprachen des Computers.

In der Sprachverarbeitung werden Automaten an vielen Stellen eingesetzt, so z.B. bei der morphologischen Analyse von Wörtern.

„Endlicher Geld-Automat“

Das Beispiel ist leicht abgewandelt von Prof. Dr. Reinhard Völler (Hamburg) übernommen.



Bestandteile eines Automaten:

endliche Menge von Zuständen

Startzustand, Endzustand (auch mehrere)

Eingaben: Worte einer Sprache

Ausgaben: Worte einer Sprache

Regeln für die Auswirkungen einer Eingabe

Wichtig: diese Regeln müssen vollständig sein, d.h. für jede Eingabe muß eine Reaktion des Automaten spezifiziert sein.

Notfalls wird ein Fehlerzustand definiert.

Endliche Automaten

Ein endlicher Automat ist ein mathematisches Modell eines Systems mit Ein- und Ausgaben.

Ein solches System befindet sich immer in einer aus einer endlichen Anzahl möglicher interner Konfigurationen.

Man sagt auch: das System befindet sich in einem Zustand .

Beispiele:

Ein Schaltkreis mit n Gattern befindet sich in einem von möglichen Zuständen.

Texteditoren oder lexikalische Analysatoren von Compilern kann man als endliche Automaten modellieren.

Auch ein Computer ist ein endlicher Automat. Allerdings ist dieses Modell wegen der großen Anzahl möglicher Zustände nicht besonders hilfreich.

Deterministischer endlicher Automat (DFSA)

Ein deterministischer endlicher Automat ist ein Fünftupel $A = (Z, E, \delta, z_0, F)$

Z	Menge der Zustände
E	Menge der Eingabesymbole
$\delta: Z \times E \rightarrow Z$	Zustandsübergangsfunktion
$z_0 \in Z$	Anfangszustand
$F \subseteq Z$	Menge der Endzustände

DFSA als Akzeptor

Für eine Eingabekette $w= e_1, e_2,...,e_n$ soll überprüft werden, ob sie durch einen Automaten A akzeptiert wird

Wir definieren uns zwei Variablen, q für den gegenwärtigen Zustand und e für das gegenwärtige Eingabesymbol

Wir setzen $q = z_0$ und wiederholen dann für jedes Symbol der Eingabekette e_i

$q := \delta (q, e_i)$

$e := e_{i+1}$

Wenn $e = e_n$ und $q \in F$ dann ist die Eingabekette durch A akzeptiert, ansonsten gilt sie als zurückgewiesen.

Beispiel 1

Namensliste:

Peter Müller

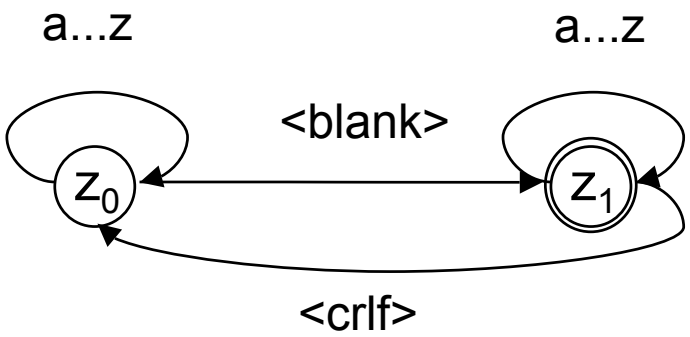
.

.

.

Doris Steckler

	z_0	z_1
a	z_0	z_1
.	z_0	z_1
.	z_0	z_1
.	z_0	z_1
z	z_0	z_1
<blank>	z_1	
<crlf>		z_0



	z_0	z_1
z_0	a...z	<blank>
z_1	<crlf>	a...z

Nichtdeterministischer endlicher Automat

Ein nichtdeterministischer endlicher Automat ist ein Fünftupel $A = (Z, E, \delta, z_0, F)$

Z	Menge der Zustände
E	Menge der Eingabesymbole
$\delta: Z \times E \rightarrow 2^Z$	Zustandsübergangsfunktion
$z_0 \in Z$	Anfangszustand
$F \subseteq Z$	Menge der Endzustände

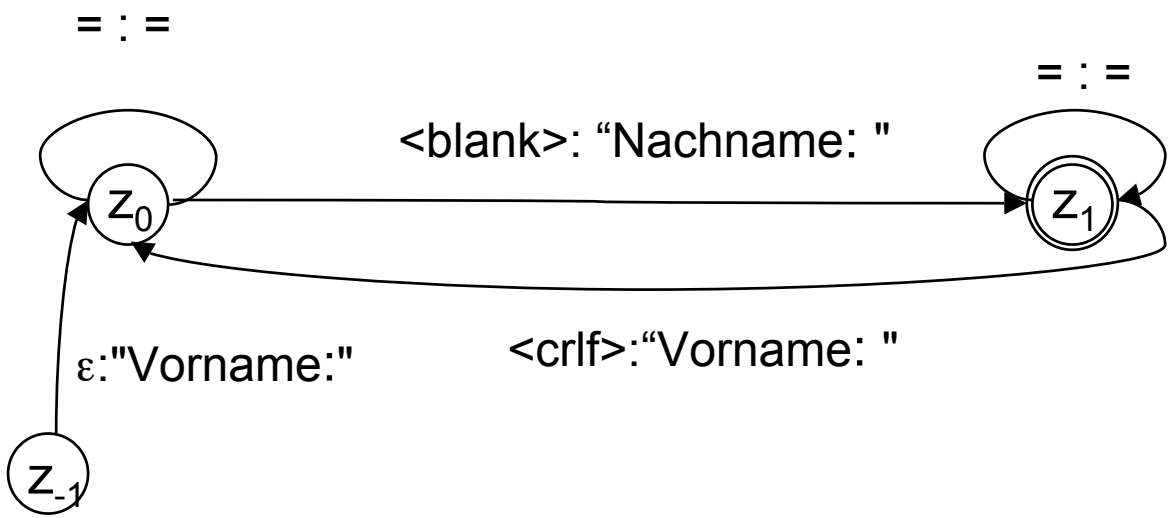
Automaten mit Ausgabe

Ein Mealey-Automat ist ein Sechstupel $A = (Z, E, A, \delta, z_0, \lambda)$

Z	Menge der Zustände
E	Menge der Eingabesymbole
A	Menge der Ausgabesymbole
$\delta: Z \times E \rightarrow 2^Z$	Zustandsübergangsfunktion
$\lambda: Z \times E \rightarrow A$	Ausgabefunktion
$z_0 \in Z$	Anfangszustand

Beispiel 2

	z_{-1}	z_0	z_1
ε : "Vorname:"		z_0	
= : =		z_0	z_1
<blank>: "Nachname:"		z_1	
<crlf> : "Vorname:"			z_0



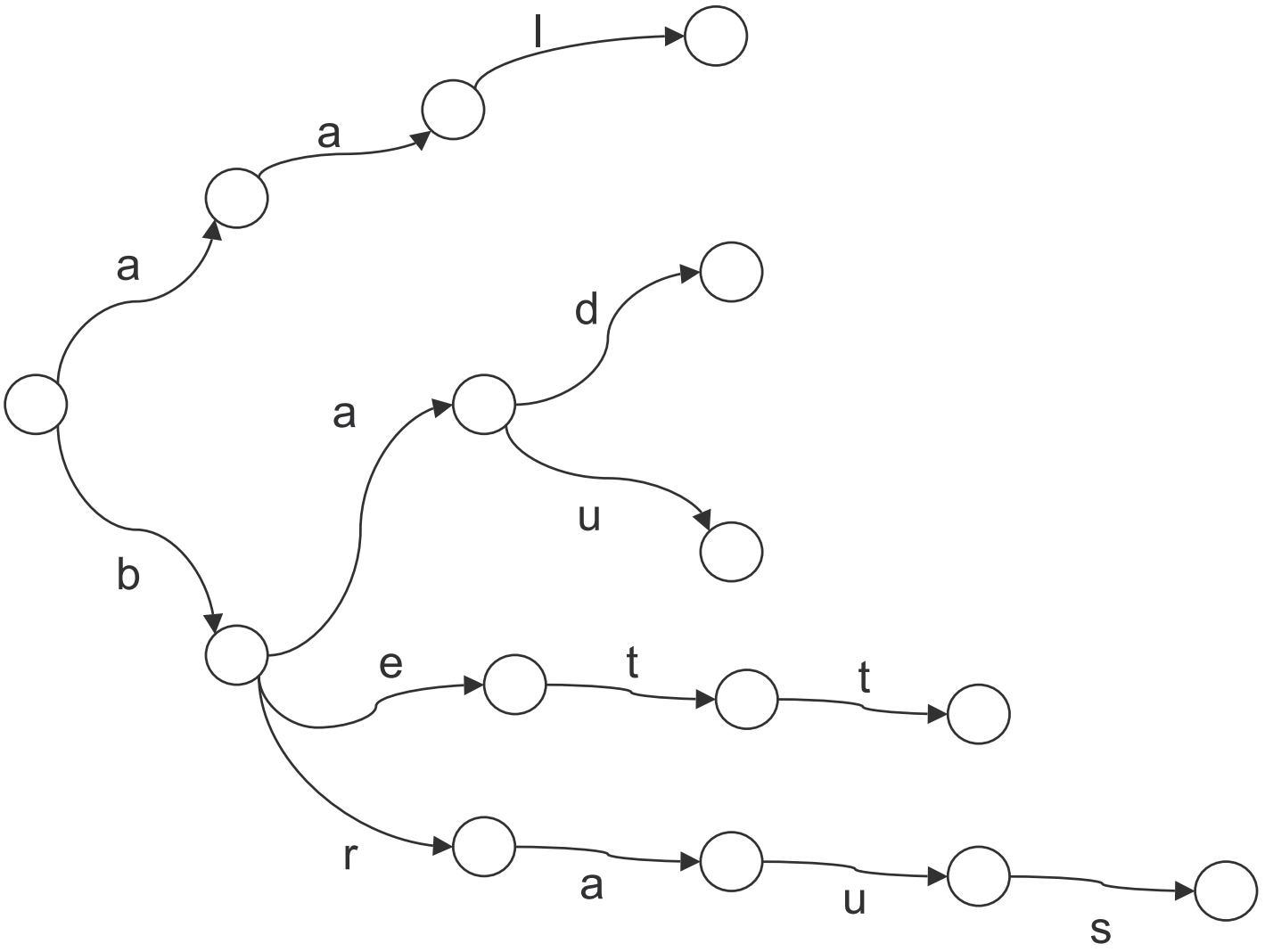
Beispiel 2

Vorname: Marion Nachname: Abbecker

Vorname: Klaus Nachname:Becker

Vorname: Günter Nachname:Bruck

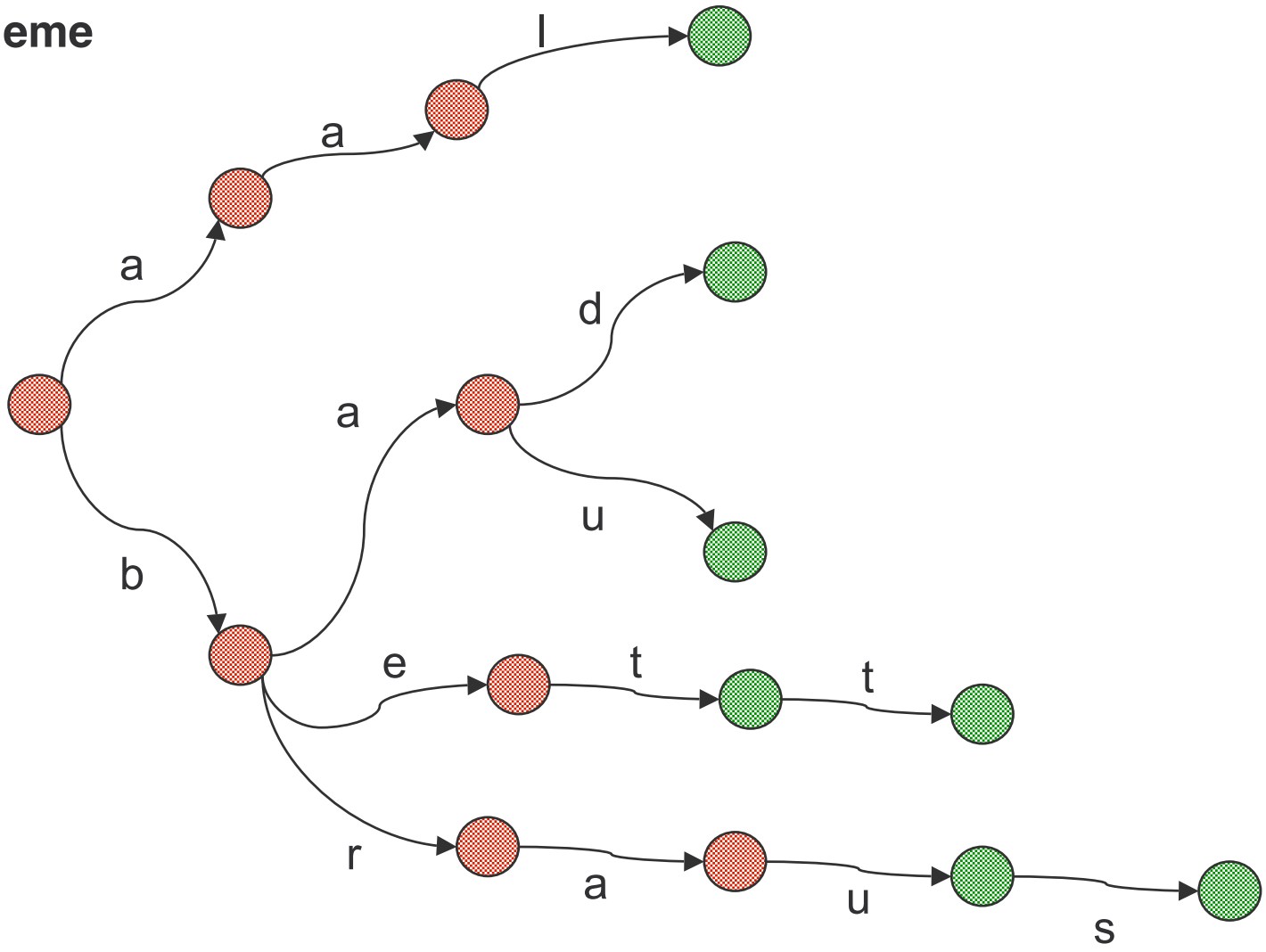
Buchstabenbaum



Buchstabenbaum

Verbstamm-Morpheme

aal
bad
bau
bet
bett
brau
braus

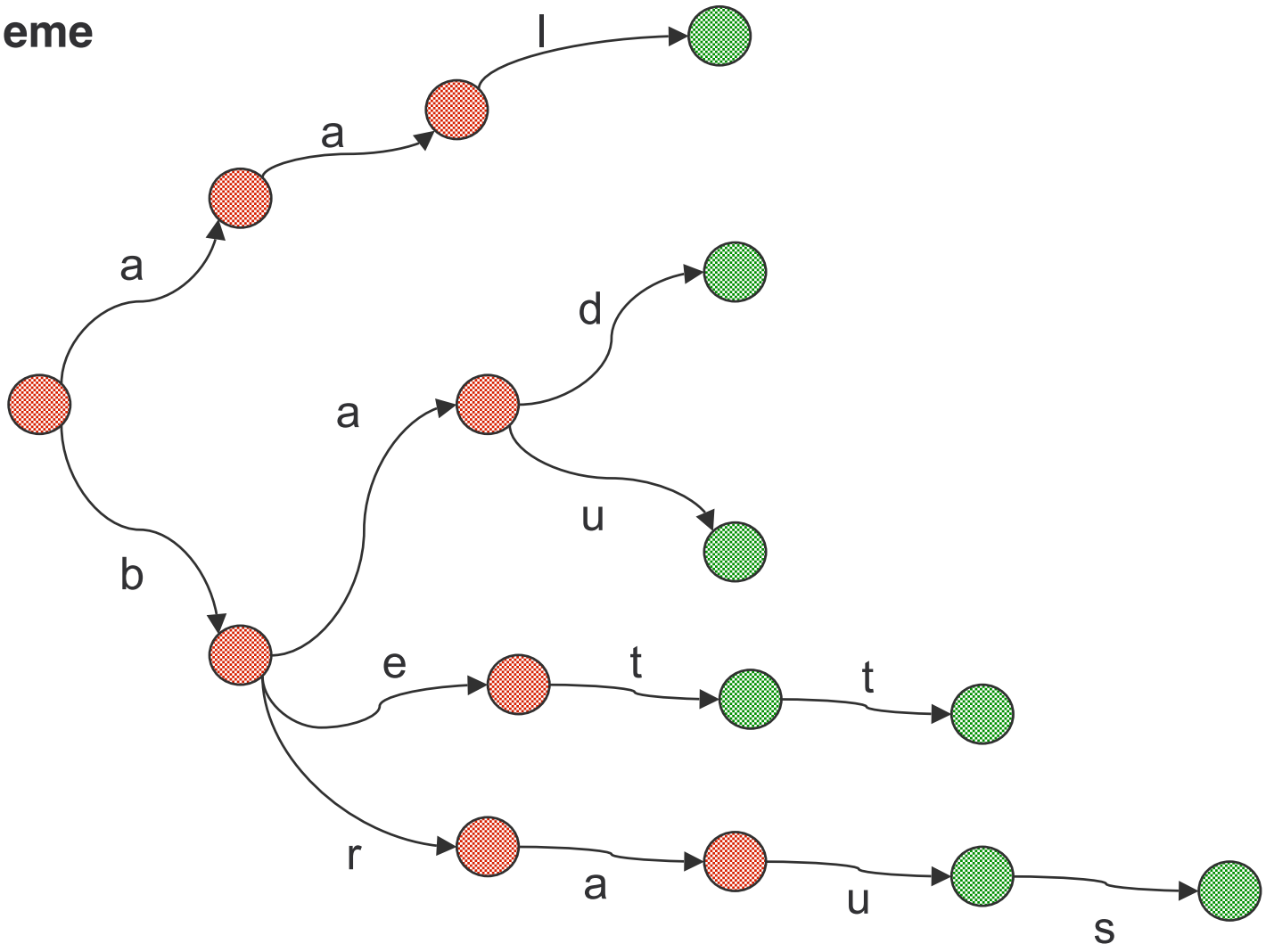


Buchstabenbaum

VSTEM: aal

Verbstamm-Morpheme

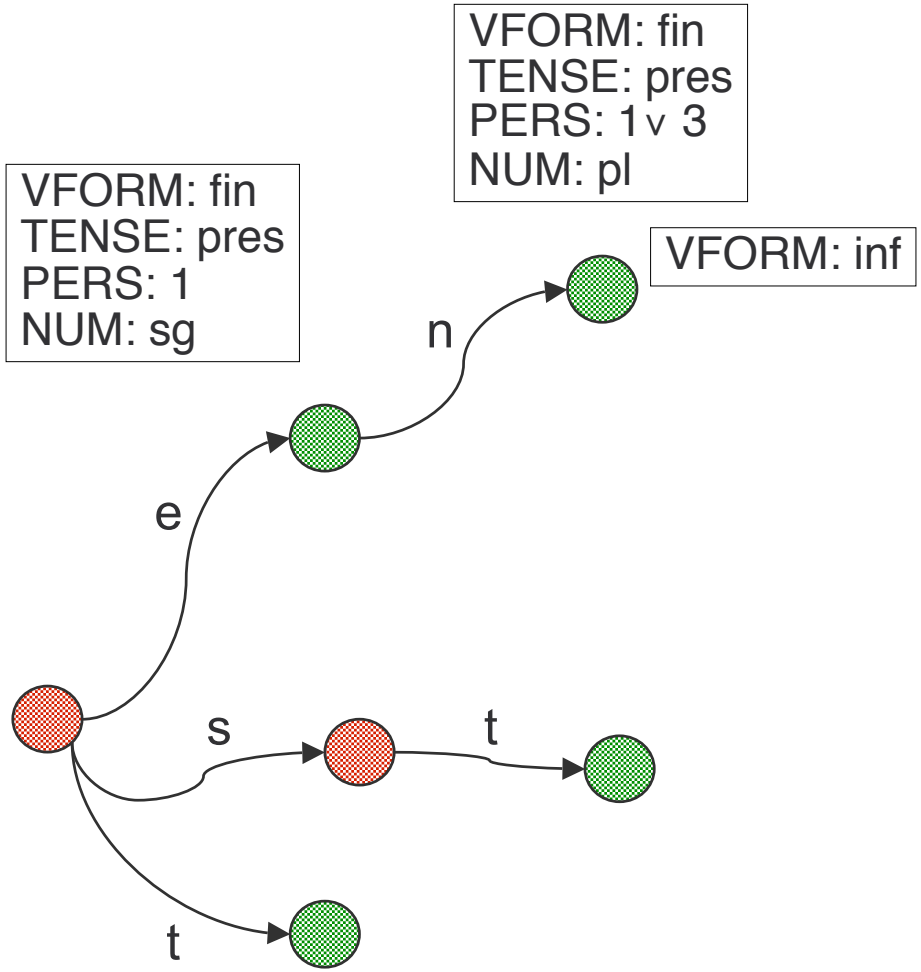
aal
bad
bau
bet
bett
brau
braus



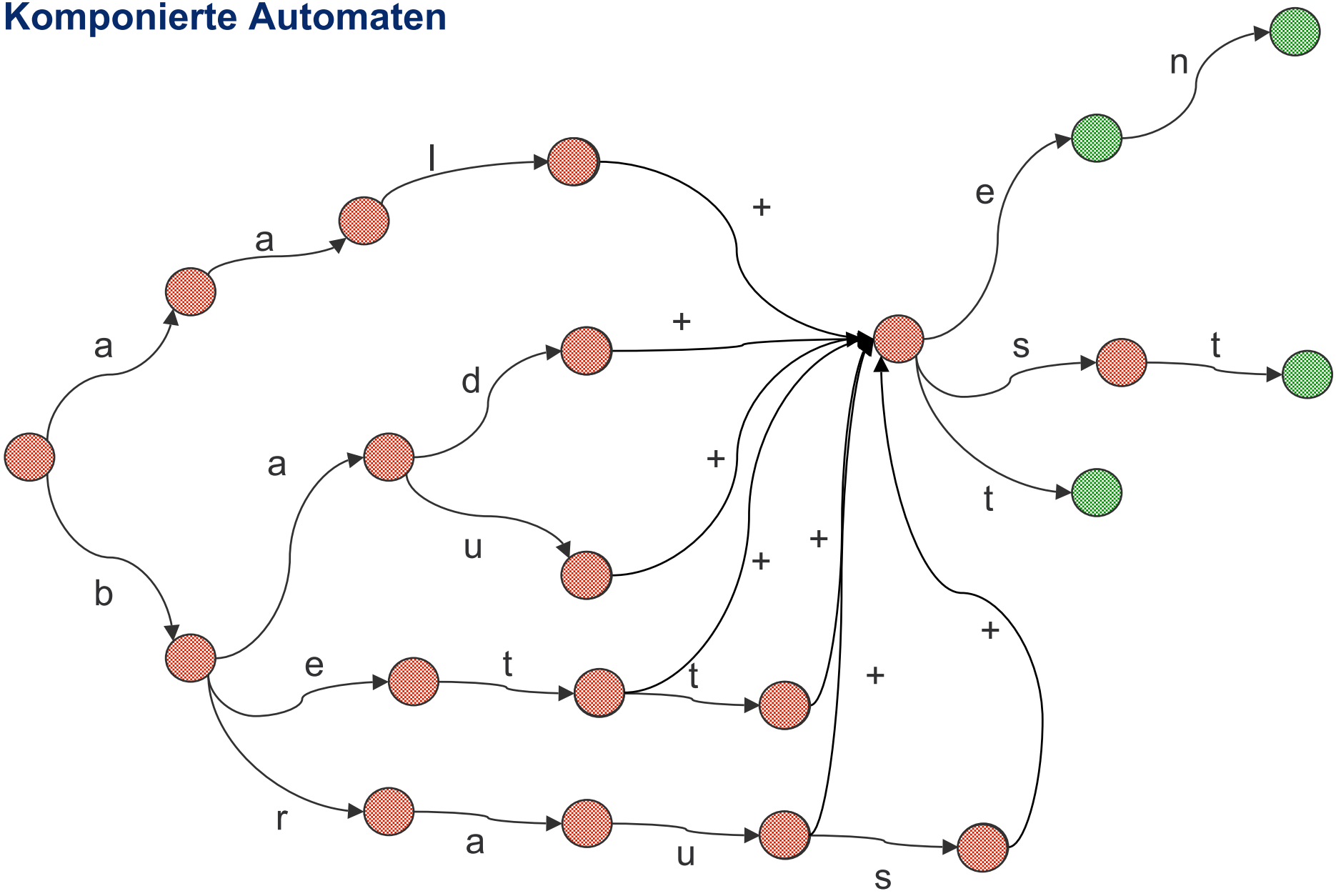
Buchstabenbaum

Verbsuffix-Morpheme

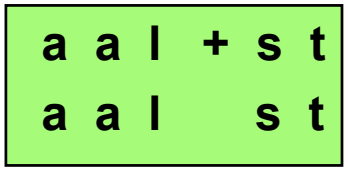
-e
-en
-st
-t

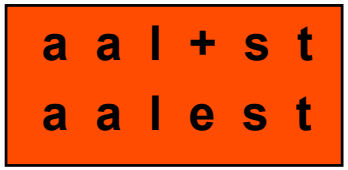


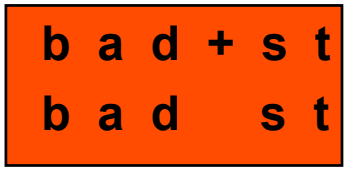
Komponierte Automaten

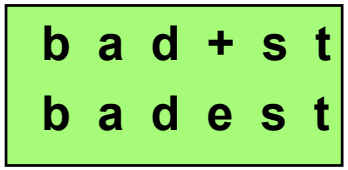


Beispiele





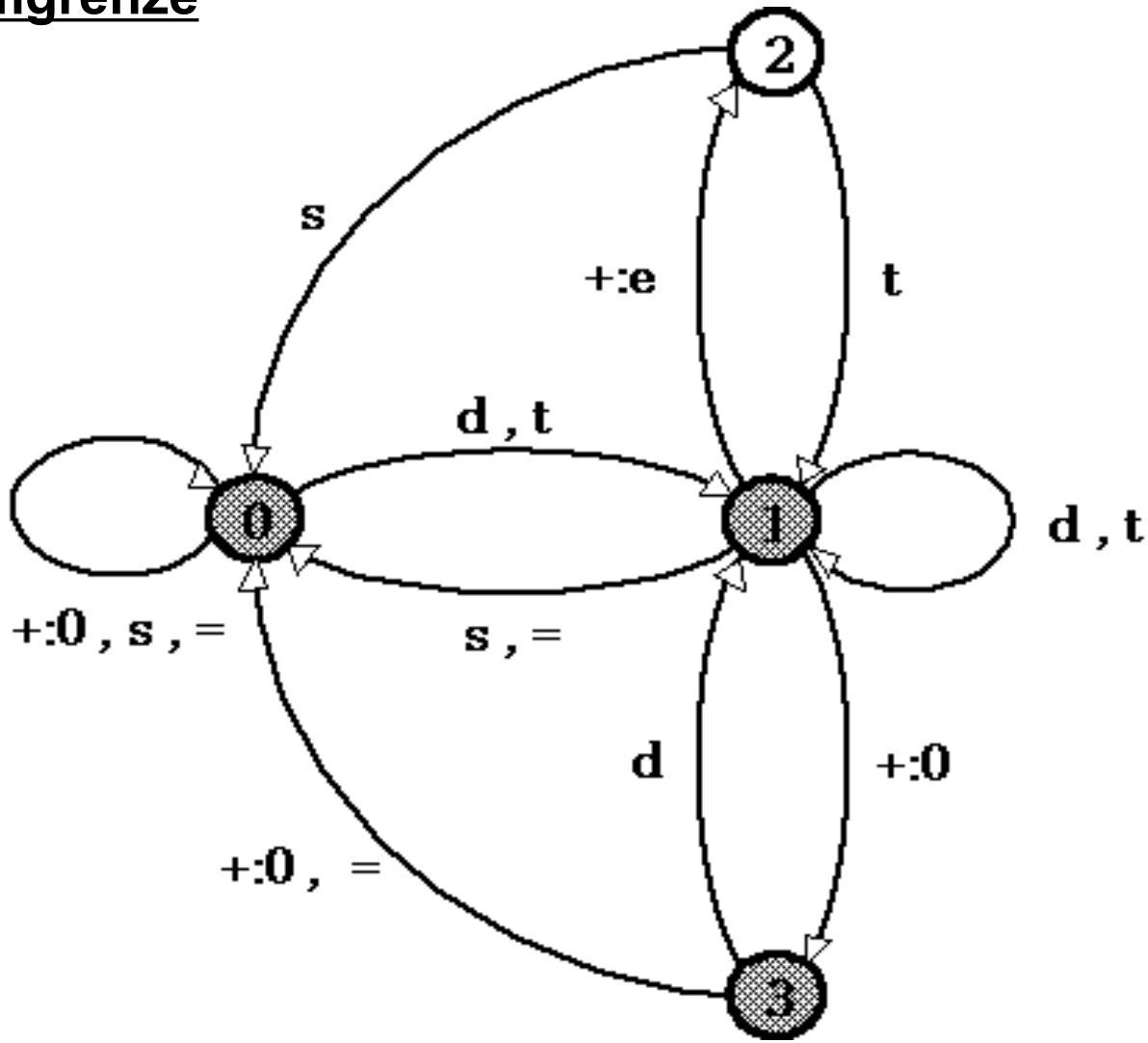




Zwei-Ebenen Morphologie

e-Epenthese an der Morphemgrenze (vereinfacht)

Regel: +:e <=> {d, t} _ {s, t}



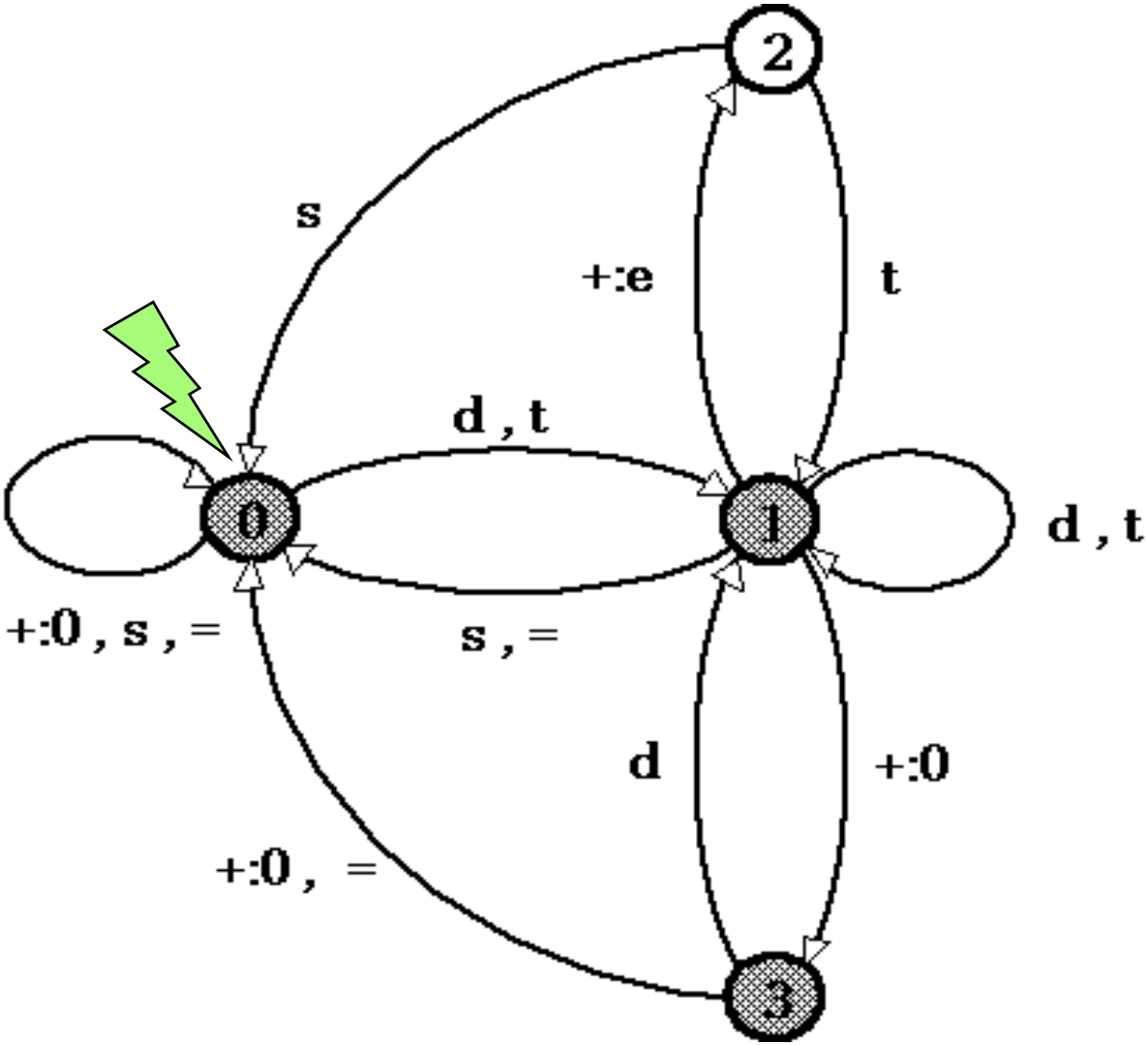
Übergangstabelle:

	0	1	2	3
+:e	–	2	–	–
+:0	0	3	–	0
d:d	1	1	–	1
s:s	0	0	0	–
t:t	1	1	1	–
=:=	0	0	–	0

Abgewandeltes Beispiel von Prof. Dr. Harald Trost (Wien)

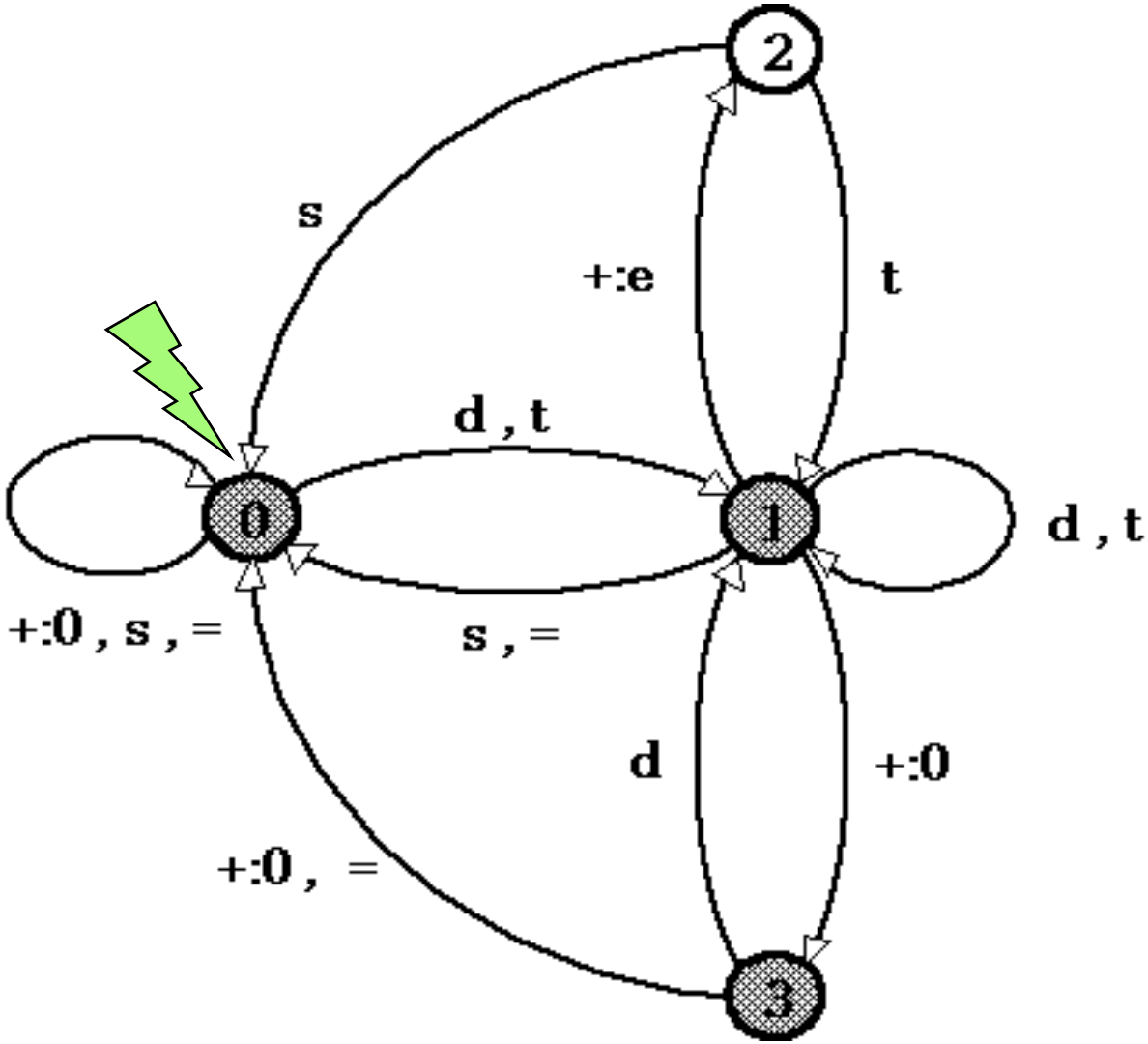
b a d + s t

b a d s t



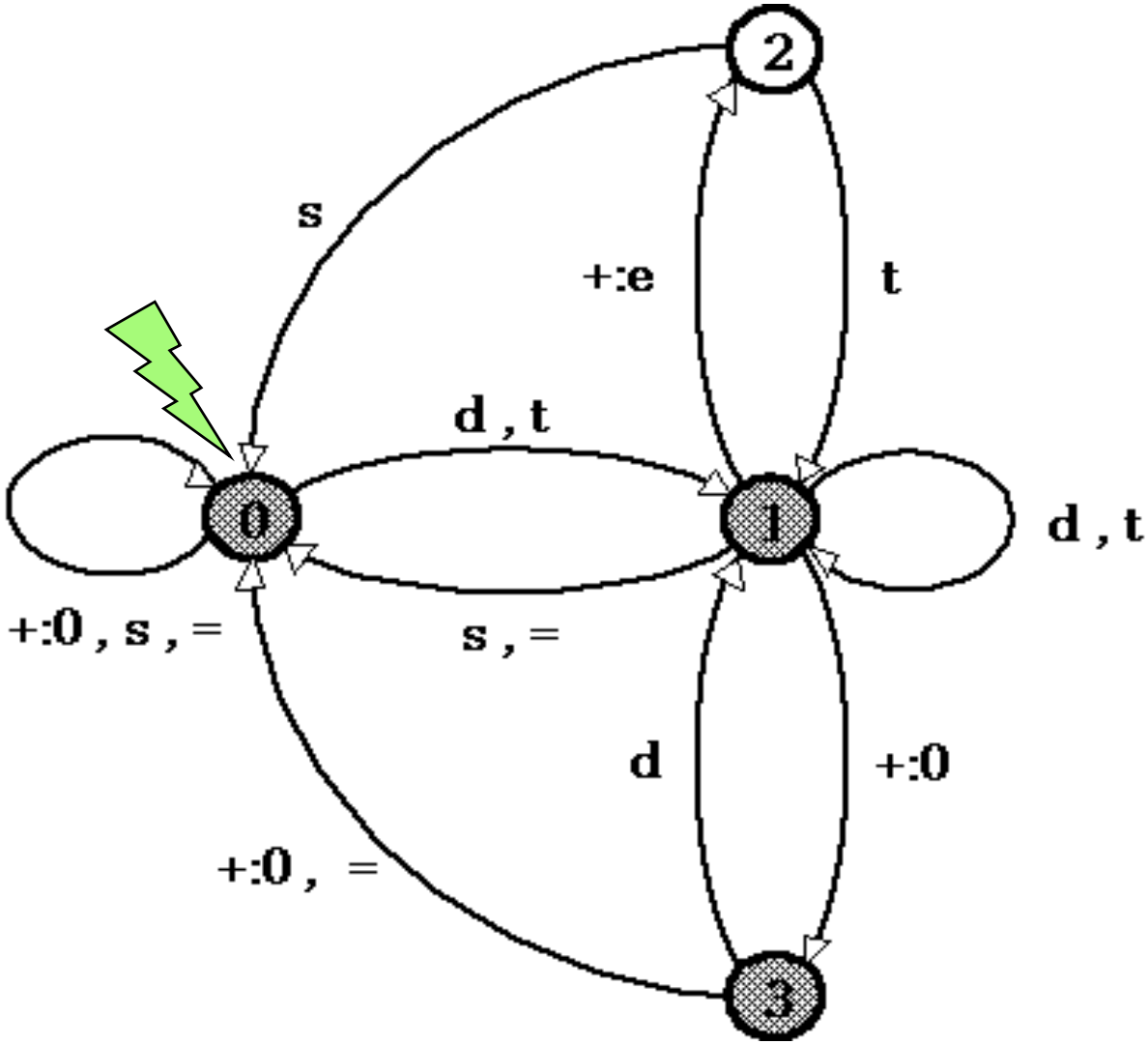
b a d + s t

b a d s t

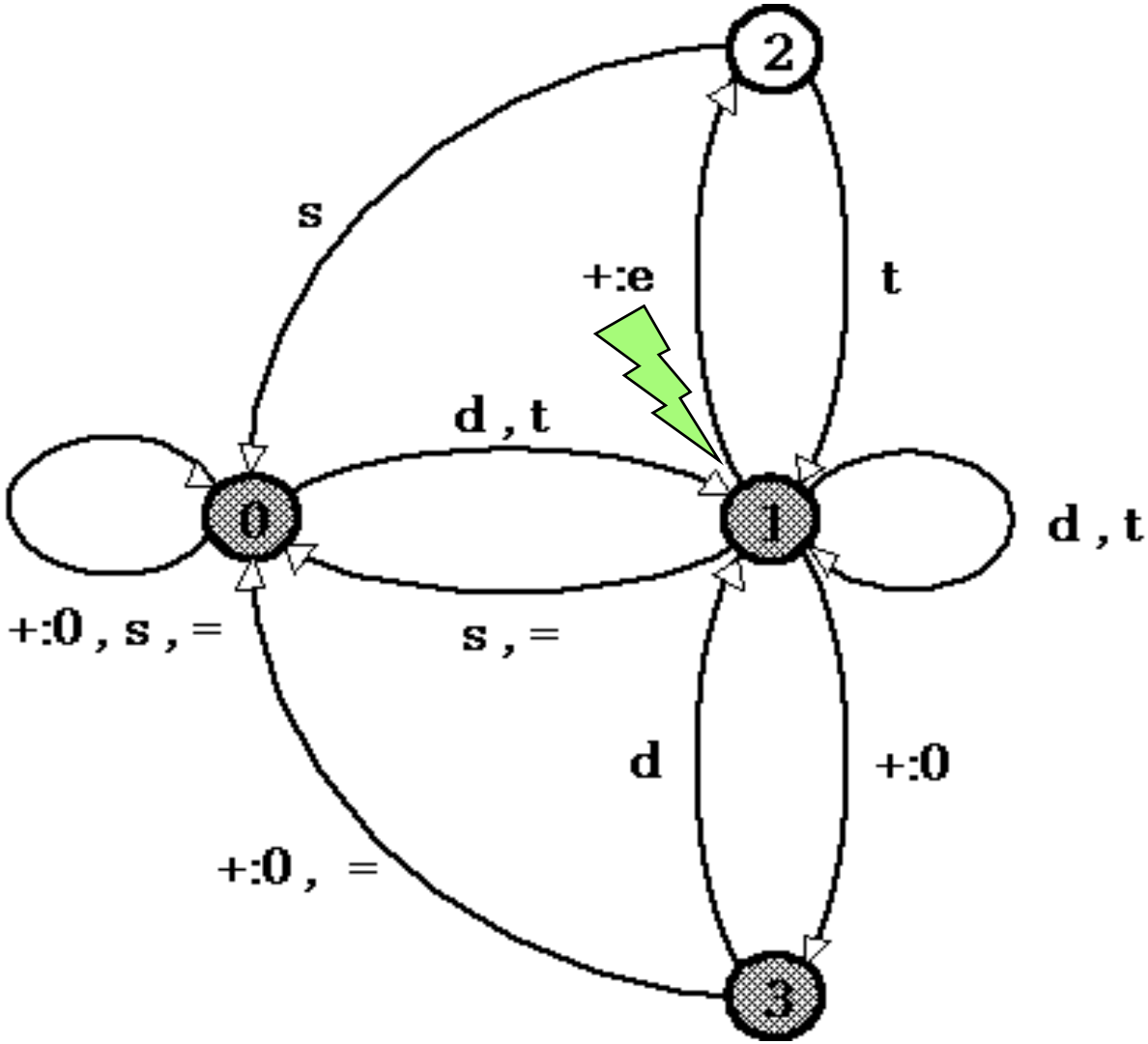


b a d + s t

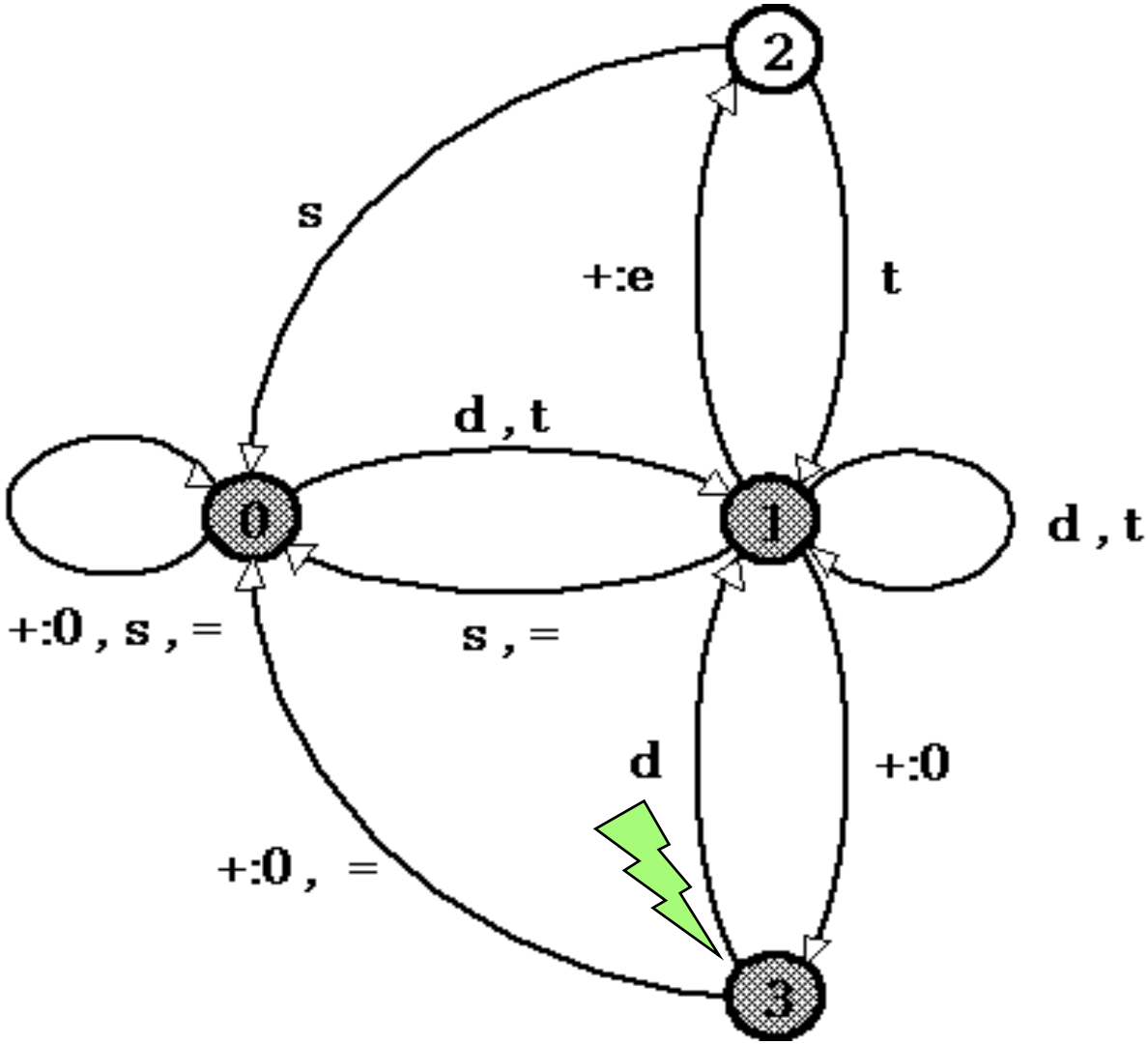
b a d s t



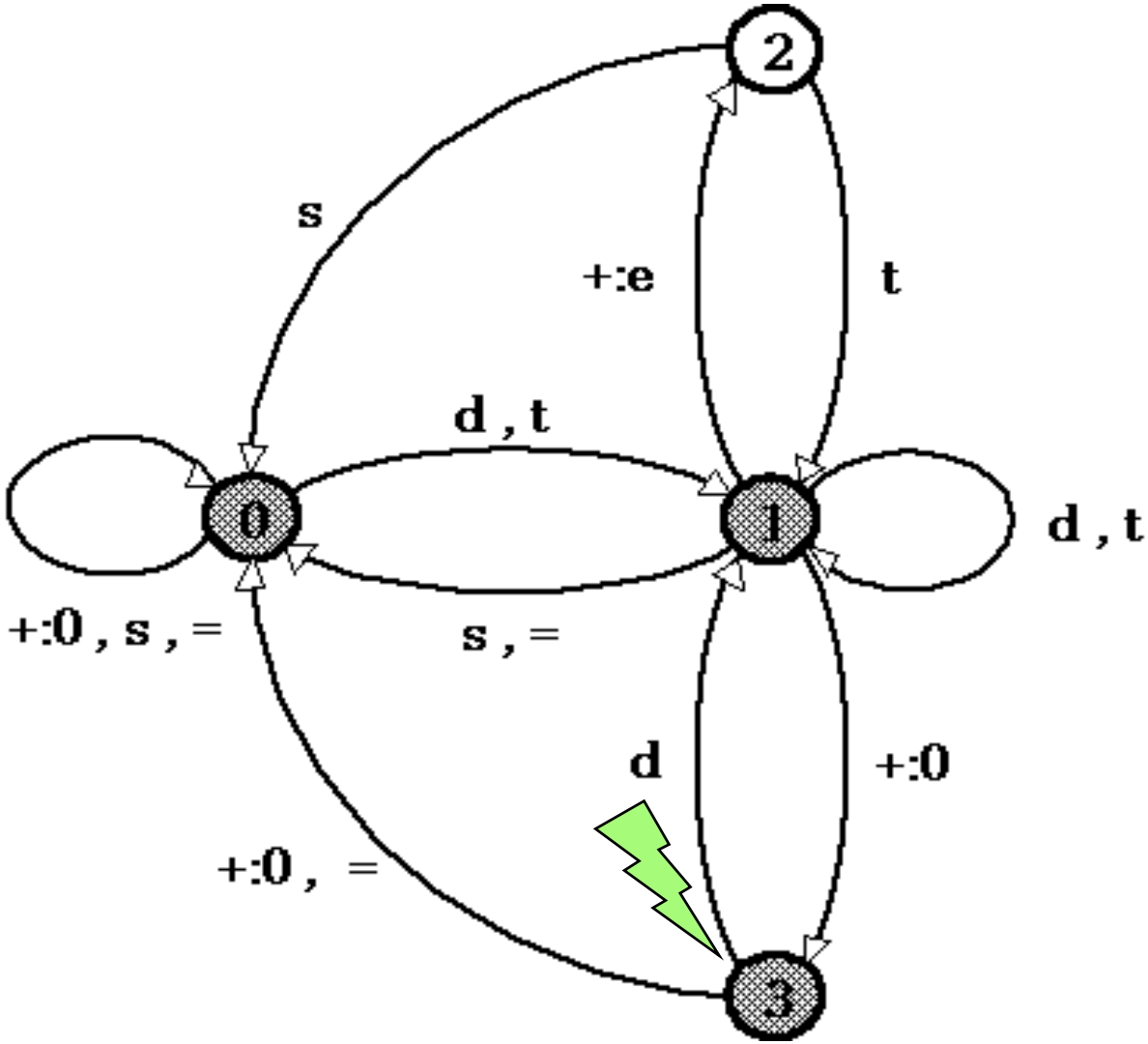
b a d + s t
b a d s t



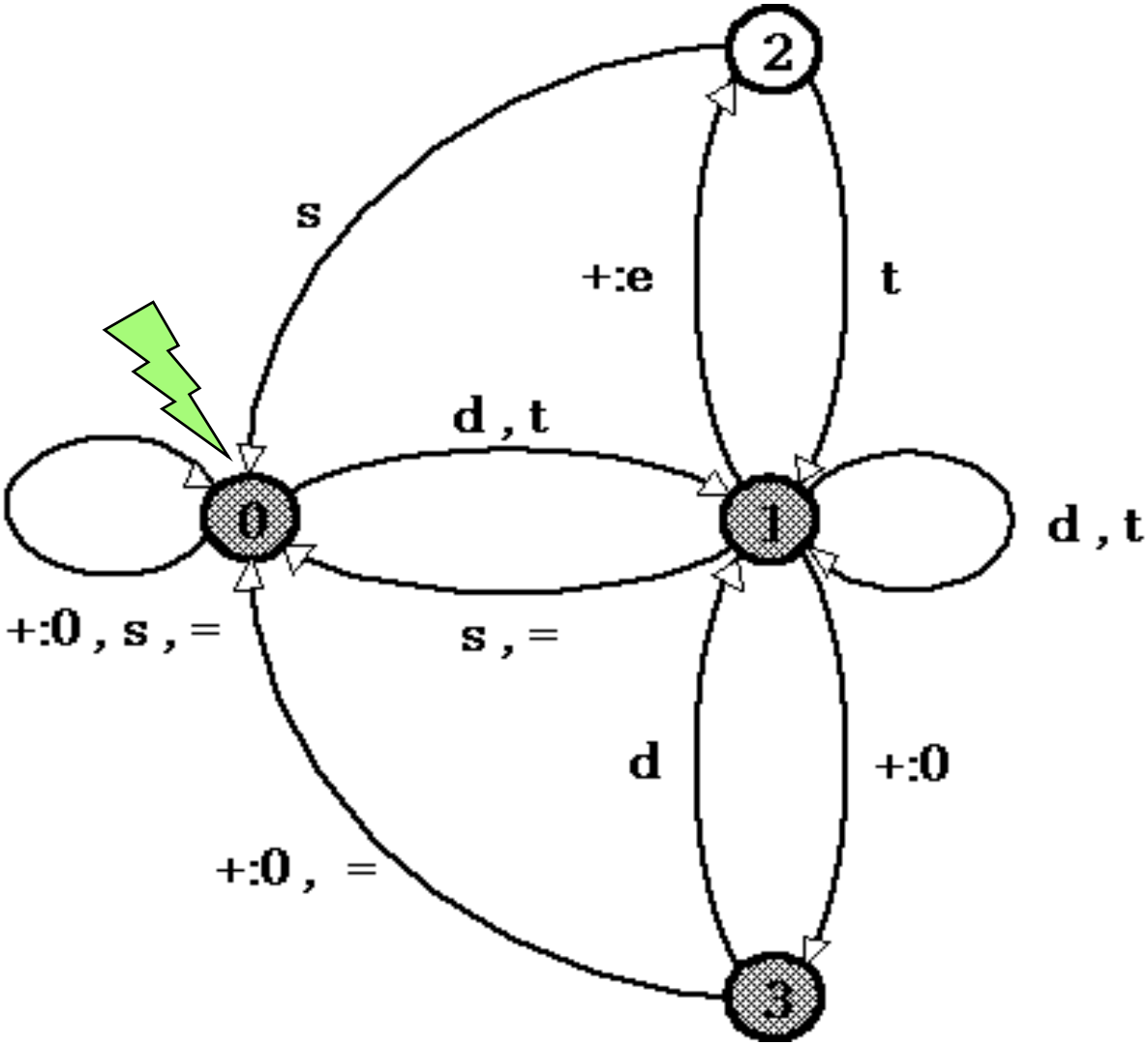
b a d + s t
b a d s t



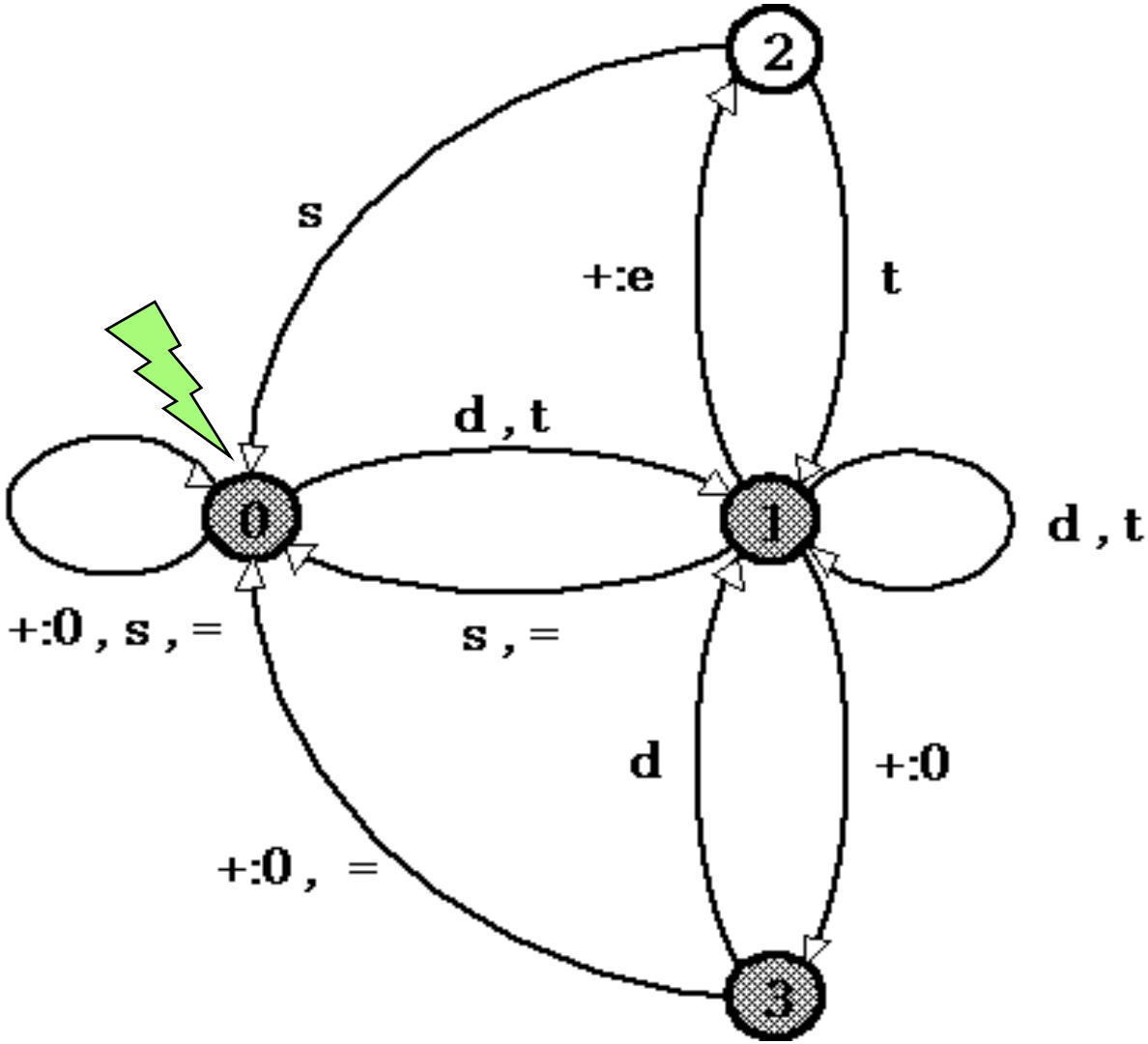
b a d + s t
b a d s t



b a d + s t
b a d e s t

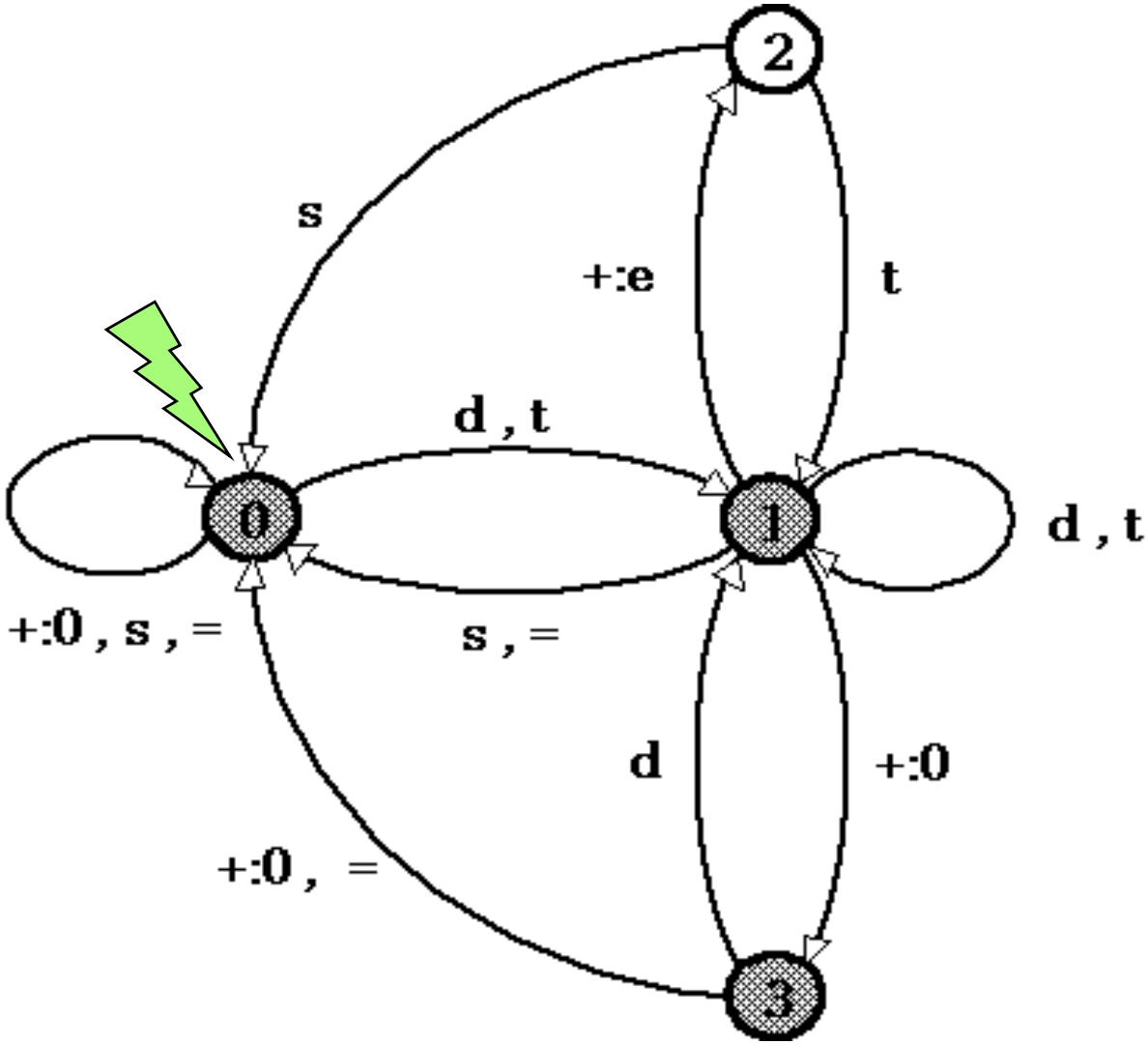


b a d + s t
b a d e s t

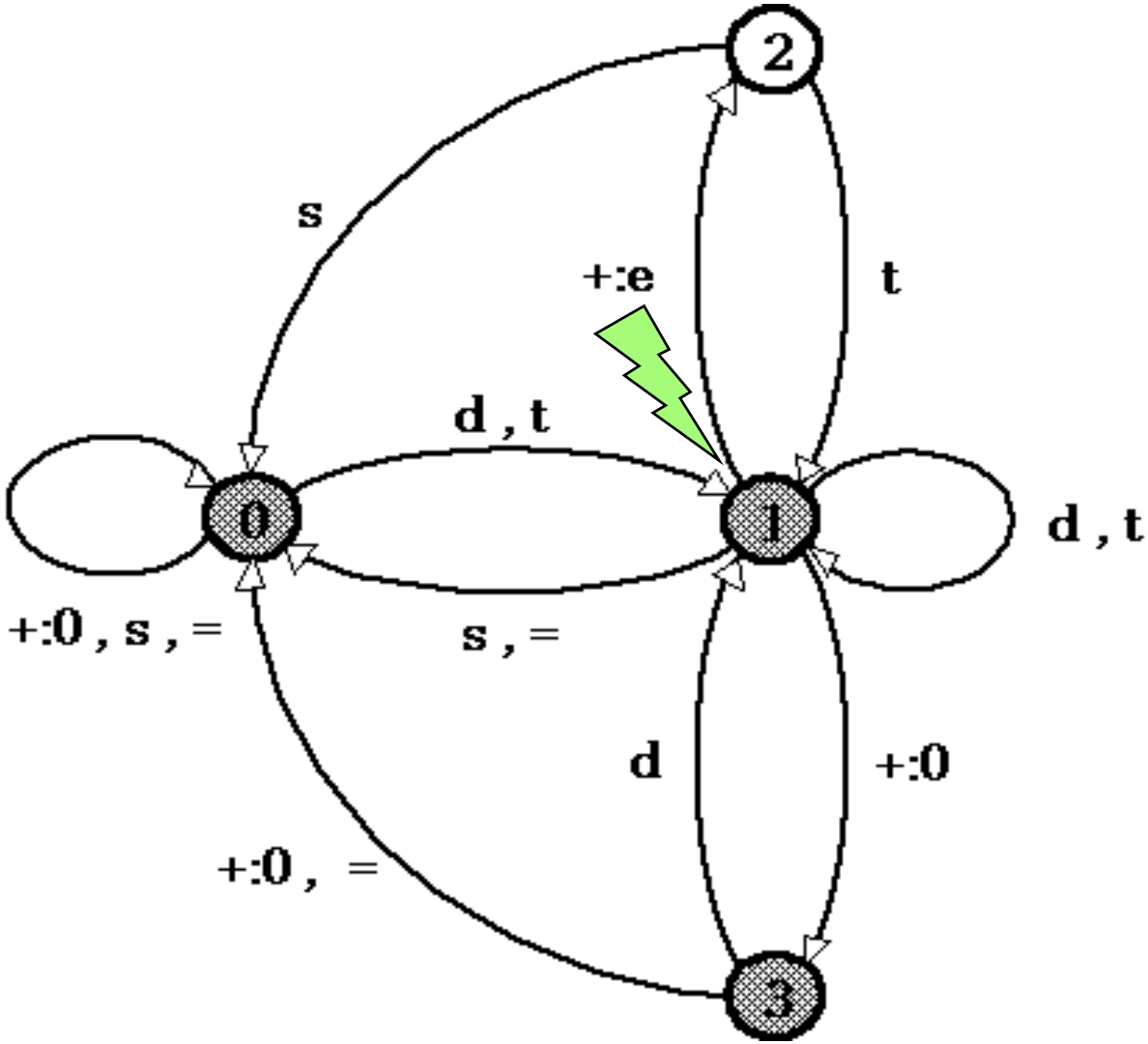


b a d + s t

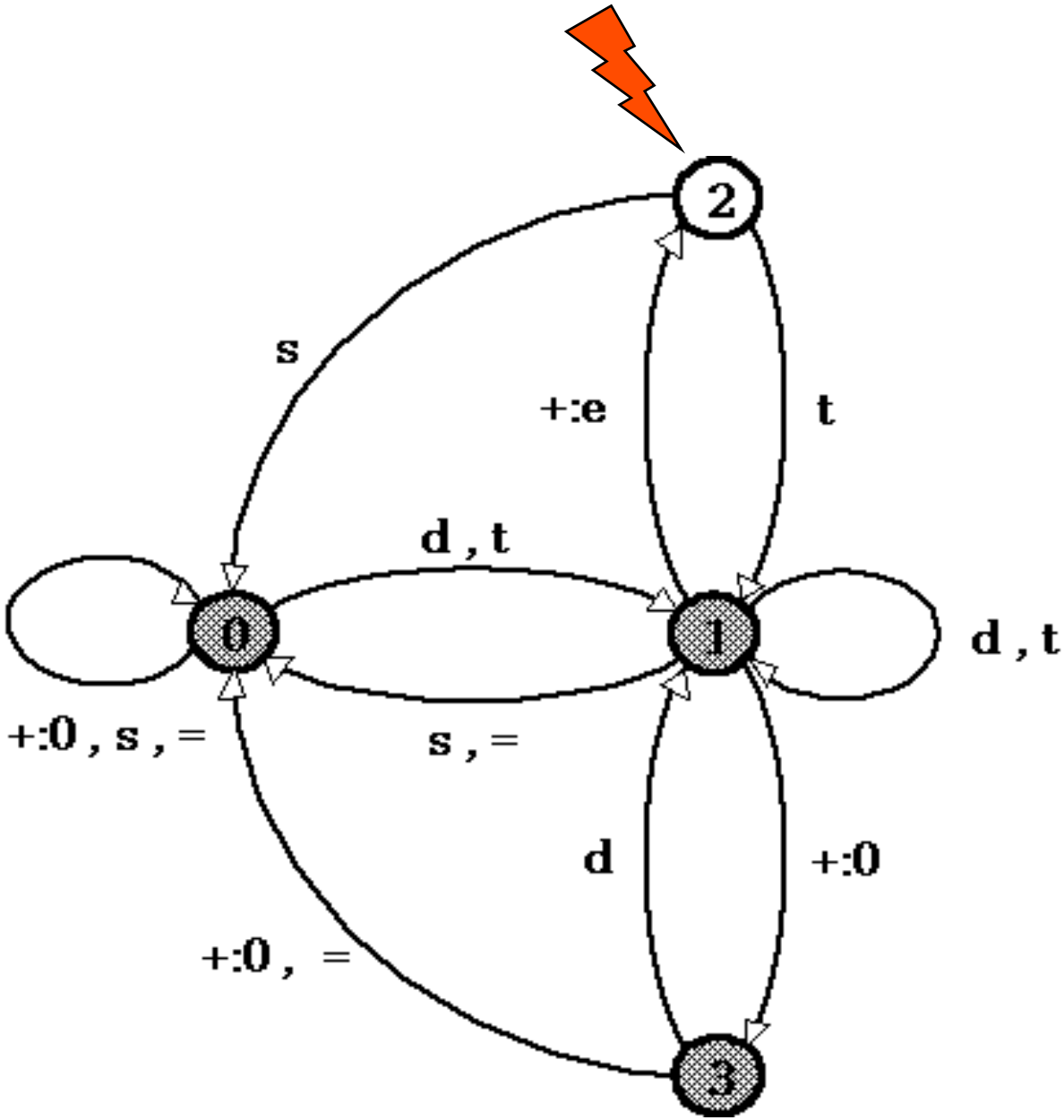
b a d e s t



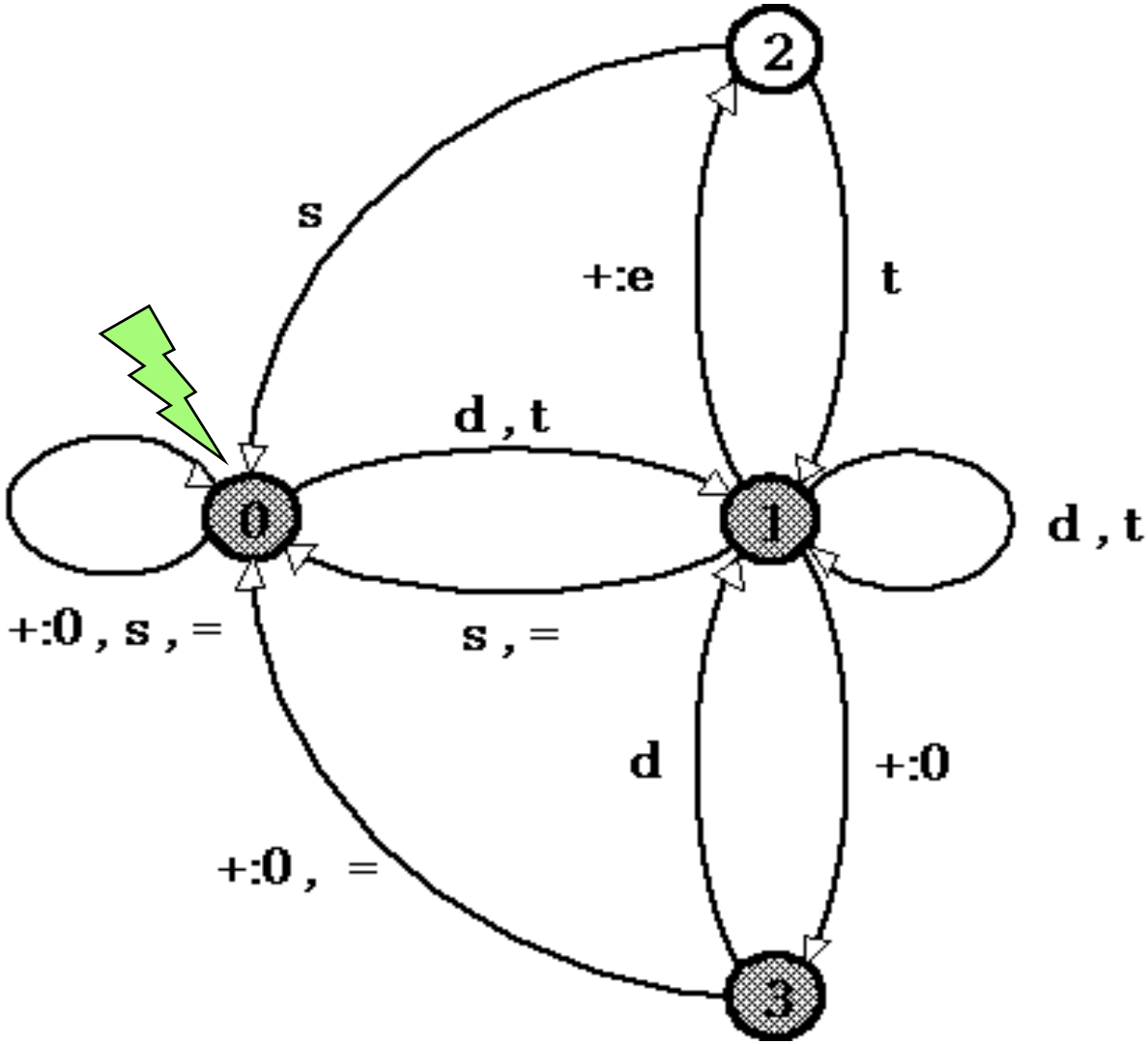
b a d + s t
b a d e s t



b a d + s t
b a d e s t

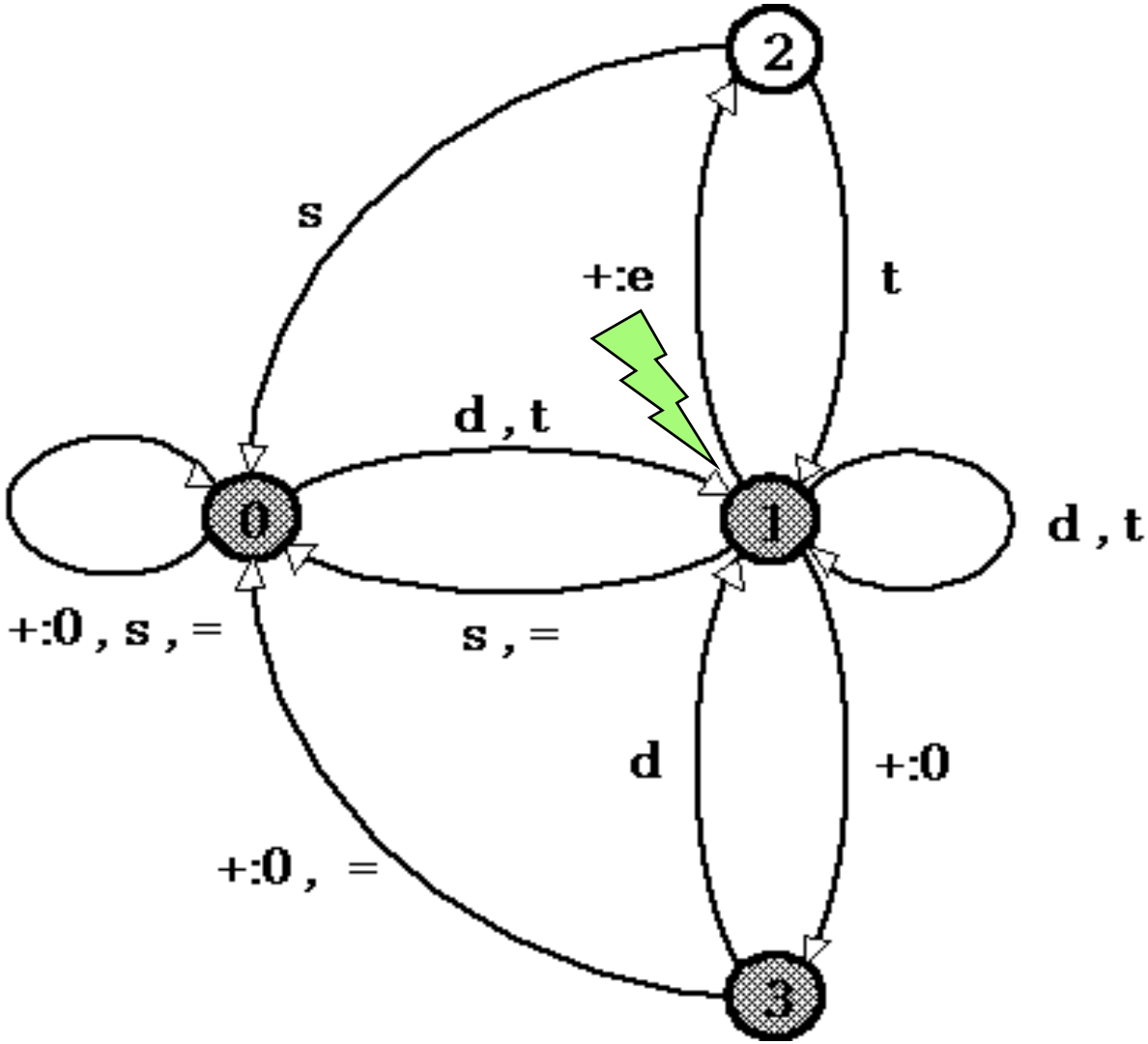


b a d + s t
b a d e s t



b a d + s t

b a d e s t



Übung

Entwerfen Sie einen Transduktor, der Umlaute (ä,ö,ü) in (ae, oe, ue) umwandelt.

Entwerfen Sie nun einen Automaten, der die umgekehrte Umwandlung vornimmt

ae --> ä, oe --> ö, ue --> ü. Wie verhindert Ihr Automat die folgenden Umwandlungen:

Poebene --> Pöbene, Frauen --> Fraün?

Entwerfen Sie einen Transduktor, der im Englischen aus "kiss+s" "kisses" und aus "wish+s" "wishes" macht.