Einführung in die Computerlinguistik
Morphologie

Hinrich Schütze

Center for Information and Language Processing

2018-10-22

Die Grundfassung dieses Foliensatzes wurde von Dr. Benjamin Roth erstellt. Fehler und Mängel sind ausschließlich meine Verantwortung.

Outline

Outline

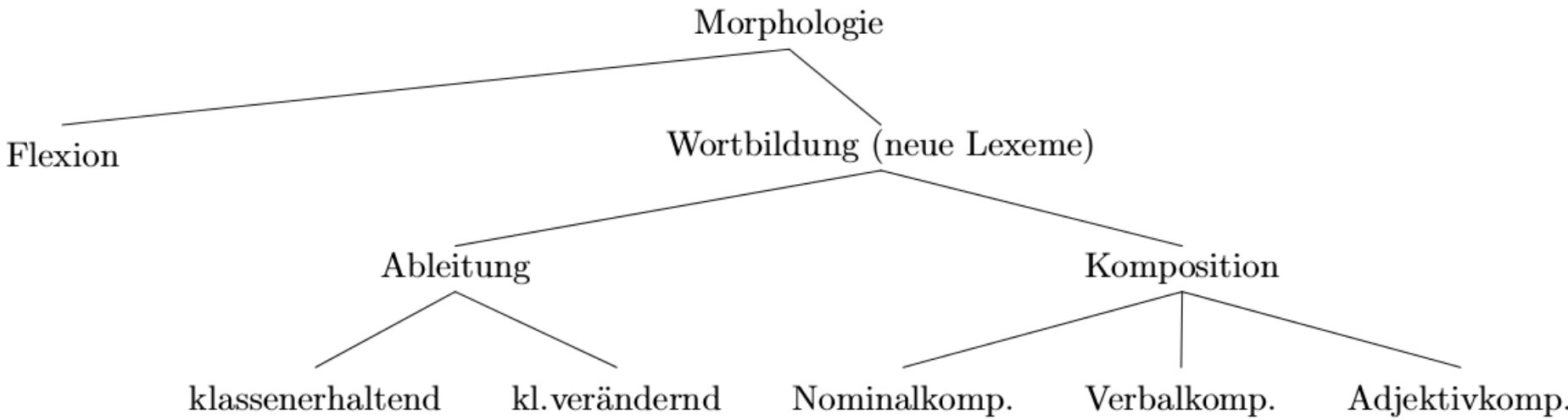
Der Begriff der Morphologie

- *griech.* morphe (Form, Gestalt) + logos (Sinn, Lehre) $\Rightarrow$ Formenlehre
- Aufbau von Wörtern aus kleinsten bedeutungstragenden Einheiten
- interne Struktur der Wörter

Die zwei Bereiche der Morphologie

- Flexionsmorphologie (Wortformbildung):
  - Markierung von Tempus, Person, Kasus, Numerus, ...
  - Aufbau von Wortformen aus Basis und Flexionsendung (Wort als Flexionsparadigma):
    der Mann – des Mannes
  - Aber:
    geht – ging
- Wortbildungslehre:
  - Derivationsmorphologie: Bedeutungsverändernde Bildung von Wörtern aus einer Basis und einem Derivationsmorphem.
    Beispiele: klar – unklar
    Sache – sächlich / sachlich
  - Komposita: Zusammensetzung von mehreren Teilen
    Beispiele: Bauer + Hof – Bauernhof
    Sonne + baden – sonnenbaden

Morphologie: Übersicht



Outline

Begriff “Morphem”

- die elementaren Einheiten des Wortes
- die kleinsten sprachlichen Einheiten, die Bedeutung haben
- abstrakte Einheiten, die durch Repräsentationseinheiten realisiert werden, und zwar
  - in der gesprochenen Sprache als Phonemfolgen.
  - in der Schrift als Graphemfolgen.

Beispiel: Formen des Morphems /ʃtuːl/

- Stuhl
  -s
  Stühl
  -e
  -en
- Formen: {Stuhl, Stuhls, Stühl, Stühle, Stühlen}
- Prinzip:
  Kombination von wenigen Elementen ergibt viele neue Elemente
  stuhl_morphem + plural_suffix *(Stühle)*
  fahren_morphem + gast_morphem *(Fahrgast)*

Infinite use of finite means

Wilhelm von Humboldt

Das Verfahren der Sprache ist aber nicht bloß ein solches, wodurch eine einzelne Erscheinung zu Stande kommt; es muss derselben zugleich die Möglichkeit eröffnen, eine unbestimmbare Menge solcher Erscheinungen und unter allen, ihr von dem Gedanken gestellten Bedingungen hervorzubringen. Denn sie steht ganz eigentlich einem unendlichen und wahrhaft gränzenlosen Gebiete, dem Inbegriff alles Denkbaren gegenüber. Sie muss daher von endlichen Mitteln einen unendlichen Gebrauch machen, und vermag dies durch die Identität der Gedanken- und Spracheerzeugenden Kraft.

Infinite use of finite means

Wilhelm von Humboldt

Das Verfahren der Sprache ist aber nicht bloß ein solches, wodurch eine einzelne Erscheinung zu Stande kommt; es muss derselben zugleich die Möglichkeit eröffnen, eine unbestimmbare Menge solcher Erscheinungen und unter allen, ihr von dem Gedanken gestellten Bedingungen hervorzubringen. Denn sie steht ganz eigentlich einem unendlichen und wahrhaft gränzenlosen Gebiete, dem Inbegriff alles Denkbaren gegenüber. Sie muss daher von endlichen Mitteln einen unendlichen Gebrauch machen, und vermag dies durch die Identität der Gedanken- und Spracheerzeugenden Kraft.

Arten von Morphemen

- Im Deutschen und Englischen können viele Morpheme selbständig als Wörter verwendet werden. Solche Morpheme heißen frei.
- Freies Morphem: Morphem, welches ohne Vorhandensein anderer Morpheme ein Wort bilden kann.
  Beispiele:
  {Garten}, {Zwerg}, {book}, {sing}
- Gebundenes Morphem: Morphem, welches nicht selbständig ein Wort bilden kann.
  Beispiele:
  {ge-} *(geschlafen)*
  {-s} als Genitiv Singular *(Peters)*
- Fugenelemente sind keine Morpheme, weil sie keine identifizierbare Bedeutung tragen.
  Beispiele: *(Krankheit)*-s-*(zeichen)* *(Schwan)*-en-*(hals)*

Morphem: Mel'čuks formale Definition (vereinfacht!)

Ein Morphem ist ein nicht-leeres Set von allen Wortformen $m_1, m_2, \ldots m_n = \{m_i\}$, die folgende drei Bedingungen erfüllen:

- Die Bedeutung von allen m_i ist identisch.
- Alle m_i gehören zur gleichen Klasse von Wortformen, d.h. alle m_i sind entweder Wurzeln oder Affixe.
- Die m_i stehen in komplementärer Distribution, die durch allgemeine Regeln beschrieben werden kann. Affixe können auch identische Distribution haben.
- Beispiel: "Stuhl-", "Stühl-"

Allomorph des Morphems {M}: alle Wortformen m_i, die zu {M} gehören

Morphem: Mel'čuks formale Definition (vereinfacht!)

Ein Morphem ist ein nicht-leeres Set von allen Wortformen $m_1, m_2, \ldots m_n = \{m_i\}$, die folgende drei Bedingungen erfüllen:

- Die Bedeutung von allen m_i ist identisch.
- Alle m_i gehören zur gleichen Klasse von Wortformen, d.h. alle m_i sind entweder Wurzeln oder Affixe.
- Die m_i stehen in komplementärer Distribution, die durch allgemeine Regeln beschrieben werden kann. Affixe können auch identische Distribution haben.
- Beispiel: "Stuhl-", "Stühl-"

Allomorph des Morphems {M}: alle Wortformen m_i, die zu {M} gehören

Morphem: Mel'čuks formale Definition (vereinfacht!)

Ein Morphem ist ein nicht-leeres Set von allen Wortformen $m_1, m_2, \ldots m_n = \{m_i\}$, die folgende drei Bedingungen erfüllen:

- Die Bedeutung von allen m_i ist identisch.
- Alle m_i gehören zur gleichen Klasse von Wortformen, d.h. alle m_i sind entweder Wurzeln oder Affixe.
- Die m_i stehen in komplementärer Distribution, die durch allgemeine Regeln beschrieben werden kann. Affixe können auch identische Distribution haben.
- Beispiel: "Stuhl-", "Stühl-"

Allomorph des Morphems {M}: alle Wortformen m_i, die zu {M} gehören

Morphem: Mel'čuks formale Definition (vereinfacht!)

Ein Morphem ist ein nicht-leeres Set von allen Wortformen $m_1, m_2, \ldots m_n = \{m_i\}$, die folgende drei Bedingungen erfüllen:

- Die Bedeutung von allen m_i ist identisch.
- Alle m_i gehören zur gleichen Klasse von Wortformen, d.h. alle m_i sind entweder Wurzeln oder Affixe.
- Die m_i stehen in komplementärer Distribution, die durch allgemeine Regeln beschrieben werden kann. Affixe können auch identische Distribution haben.
- Beispiel: "Stuhl-", "Stühl-"

Allomorph des Morphems {M}: alle Wortformen m_i, die zu {M} gehören

Morphem: Mel'čuks formale Definition (vereinfacht!)

Ein Morphem ist ein nicht-leeres Set von allen Wortformen $m_1, m_2, \ldots m_n = \{m_i\}$, die folgende drei Bedingungen erfüllen:

- Die Bedeutung von allen m_i ist identisch.
- Alle m_i gehören zur gleichen Klasse von Wortformen, d.h. alle m_i sind entweder Wurzeln oder Affixe.
- Die m_i stehen in komplementärer Distribution, die durch allgemeine Regeln beschrieben werden kann. Affixe können auch identische Distribution haben.
- Beispiel: "Stuhl-", "Stühl-"

Allomorph des Morphems {M}: alle Wortformen m_i, die zu {M} gehören

Komplementäre Distribution

Zwei sprachliche Elemente kommen nie in gleicher Umgebung vor, d.h. ihre Vorkommen schließen sich gegenseitig aus.

Morphem: Mel'čuks formale Definition (vereinfacht!)

Ein Morphem ist ein nicht-leeres Set von allen Wortformen $m_1, m_2, \ldots m_n = \{m_i\}$, die folgende drei Bedingungen erfüllen:

- Die Bedeutung von allen m_i ist identisch.
- Alle m_i gehören zur gleichen Klasse von Wortformen, d.h. alle m_i sind entweder Wurzeln oder Affixe.
- Die m_i stehen in komplementärer Distribution, die durch allgemeine Regeln beschrieben werden kann. Affixe können auch identische Distribution haben.
- Beispiel: "Stuhl-", "Stühl-"

Allomorph des Morphems {M}: alle Wortformen m_i, die zu {M} gehören

Outline

Wortstruktur

- Derivation und Flexion: Eine analysierbare Wortform kann rekursiv als Kombination eines Affixes und eines Restes, der Basis aufgefasst werden (die ggf. weiter analysierbar ist).
- Komposition: Eine analysierbare Wortform kann rekursiv als Kombination zweier Teile aufgefasst werden (die ggf. weiter analysierbar sind).

Wortstruktur: Affix und Basis

- Ein Affix ist ein gebundenes Morphem, welches verschiedene Basen in analoger Weiser modifiziert (*reihenbildend*).
  Unterkategorien:
  - Flexionsaffix
  - Derivationsaffix
- Jede Form, an die ein Affix angefügt werden kann, heißt Basis.
  Unterkategorien:
  - Die meisten Wurzeln sind Basen: "Tisch" $\rightarrow$ "Tisch-e"
  - Wenige Wurzeln sind nicht Basen: "kunter" (in "kunterbunt")
  - Nichtwurzeln, die Basen sind: "untouchable", "unklar", "sachlich"

Affix, Basis, Wurzel

- Ein Flexionsaffix geht eine Verbindung mit einer Basis ein. Es markiert grammatische Funktionen wie Tempus, Person, Kasus oder Numerus.
- Ein Derivationsaffix geht eine Verbindung mit einer Basis ein. Es verändert die Bedeutung der Basis.
- Wurzel: Teil der übrig bleibt, nachdem alle Affixe einer Basis entfernt wurden.

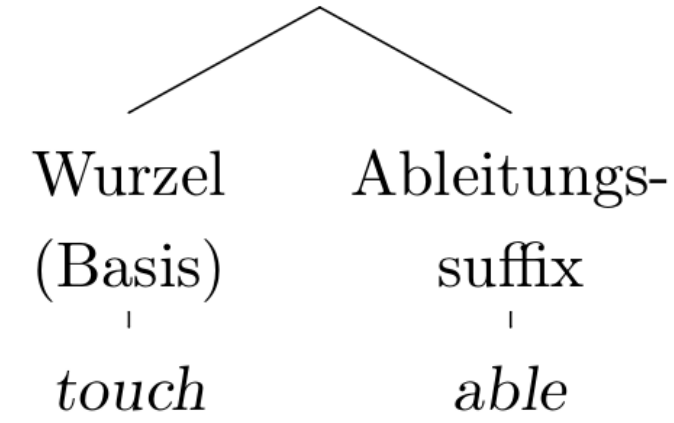
Affixe

Affixe werden gewöhnlich in drei Klassen eingeteilt, je nach ihrer Position bezüglich der Basis:

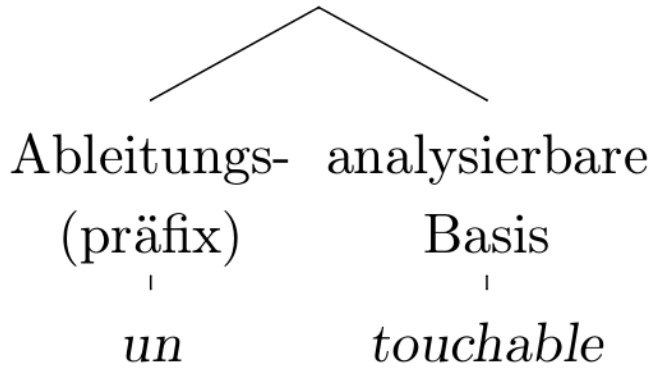
- Präfix; Beispiele:
  auf- *(auf-merk-sam)*
  un- *(un-glück-lich)*
- Suffix; Beispiele:
  -sam *(auf-merk-sam)*
  -keit *(heiter-keit)*
  engl. -ize *(nation-al-ize)*
- Infix (selten in europ. Sprachen); z.B. engl. {-bloody-} *(fan-bloody-tastic)*, deutsch {-ge-} *ein-ge-schoben*
- Zirkumfix (selten in europ. Sprachen); z.B. Ge-e *(Ge-zerr-e)*

Beispiel: Teilanalysen

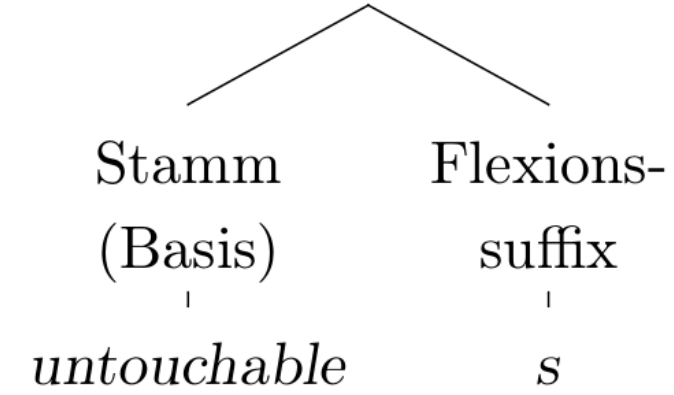
1.



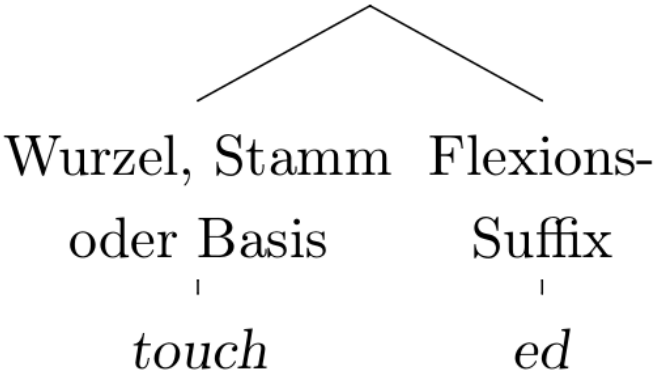
2.



3.



4.



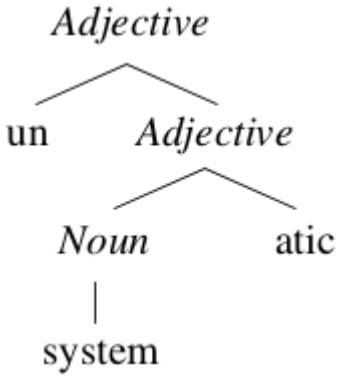
Zugrundeliegende Regeln der Wortstruktur

(Vgl auch: "The Hierarchical Structure of Words" in: Fromkin: *An introduction to language.*)

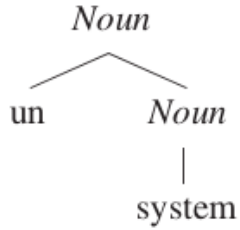
- Wörter haben eine interne Struktur, die durch Regeln der Zusammensetzung festgelegt wird.
- Beispielregel: Die meisten Affixe können nur mit Basen bestimmter Wortarten kombiniert werden.

Bestimmung der Wortstruktur: “unsystematic”

- Das Suffix “-atic” kann mit Nominalbasen kombiniert werden.
- Das Präfix “un-” kann mit Verb- und Adjektivbasen kombiniert werden, aber nicht mit Nominalbasen.
- Daraus ergibt sich folgende Analyse:



- Der folgende Teilbaum ist nicht möglich, weil “un-” nicht mit Nomen kombiniert werden kann:



Bestimmung der Wortstruktur: Wie?

- Testen, ob die Teilanalysen für sich genommen mögliche Basen ergeben
- Wenn mehrere Analysen möglich sind, entscheidet man sich für diejenige, deren Bedeutung am plausibelsten ist (d.h. bei der die internen Teile sinnvoll interpretiert werden können).
- Flexionsaffixe werden immer als letzes angehängt.

Outline

Flexionsmorphologie: Konstruktion von Paradigmen

- z.B. *Tag*:

	sg	pl
nom	–	e
gen	es	e
dat	–	en
akk	–	e

- Generalisierung der Paradigmen $\rightarrow$ lexikalische Kategorie

Flexionsmorphologie: Konstruktion von Paradigmen

- Flexionsendungen haben unspezifisches Bedeutungspotential , vgl. *Mensch*:

	sg	pl
nom	–	en
gen	en	en
dat	en	en
akk	en	en

Flexionsmorphologie: Synkretismus

- Wie wir gesehen haben, kommt es bei flektierenden Sprachen häufig vor, dass verschiedene Kategorien durch die gleiche Form repräsentiert werden. Man bezeichnet dies als Synkretismus.
- Definition: Synkretismus
  Die Tatsache, dass innerhalb eines Paradigmas verschiedene grammatische Kategorien durch die gleiche Form repräsentiert werden, nennt man Synkretismus.
- In *"He came"* und *"He has come"* haben das Präteritum und das Partizip Perfekt von "come" verschiedene Formen. In *"He tried"* und *"He has tried"* haben sie die gleiche Form, es handelt sich um einen Fall von Synkretismus.

Flexionsmorphologie: Beispiel

- Starke Adjektivflexion (Wortformen bei Vorkommen ohne Artikel)

	Singular			Plural
	Maskulinum	Neutrum	Femininum	
Nominativ	-er	-es	-e	-e
Akkusativ	-en			
Dativ	-em		-er	-en
Genitiv	-en			-er

klein+er	klein+e	klein+es	klein+e
klein+es/en	klein+er	klein+es/en	klein+er
klein+em	klein+er	klein+em	klein+en
klein+en	klein+e	klein+es	klein+e
klein+er+er	klein+er+e	klein+er+es	klein+er+e
klein+er+es/en	klein+er+er	klein+er+es/en	klein+er+er
klein+er+em	klein+er+er	klein+er+em	klein+er+en
klein+er+en	klein+er+e	klein+er+es	klein+er+e
klein+st+ er	klein+st+e	klein+st+ es	klein+st+e
klein+st+es/en	klein+st+er	klein+st+es/en	klein+st+er
klein+st+em	klein+st+er	klein+st+em	klein+st+en
klein+st+en	klein+st+e	klein+st+es	klein+st+e

Konstruktion von Paradigmen

- Im Deutschen insgesamt sechs Phoneme für Flexions-Suffixe: /e, m, n, r, s, t/
- Durch die Morphologie markierte Merkmale (sprachabhängig):
  - Numerus-Systeme: auch z.B. zusätzlich Dual
  - Genus-Systeme: auch z.B. belebt-unbelebt; nur mask-fem
  - Kasus: große Differenzen in Sprachen
  - Person: Sprecher, Angesprochener, Besprochenes
  - Tempus: sprachspezifisch (Anzahl und Arten)

Konjugation, Deklination, Komparation

Traditionell unterscheidet man folgende Flexionsprozesse:

- Konjugation: bezeichnet die morphologische Kennzeichnung nach Person, Tempus, Aktiv/Passiv (*"Genus verbi"*), Numerus, Aspekt, Modus (*Indikativ/Konjunktiv/Imperativ*),...
  $\Rightarrow$ Verben
- Deklination: bezeichnet Flexion nach Kasus, Genus, Numerus
  $\Rightarrow$ Adjektiv, Substantiv, Pronomen, Artikel
- Komparation: Graduierung und Vergleich.
  $\Rightarrow$ Adjektiv

Übung: Abweichungen von diesem Paradigma?

klein+er	klein+e	klein+es	klein+e
klein+es/en	klein+er	klein+es/en	klein+er
klein+em	klein+er	klein+em	klein+en
klein+en	klein+e	klein+es	klein+e
klein+er+er	klein+er+e	klein+er+es	klein+er+e
klein+er+es/en	klein+er+er	klein+er+es/en	klein+er+er
klein+er+em	klein+er+er	klein+er+em	klein+er+en
klein+er+en	klein+er+e	klein+er+es	klein+er+e
klein+st+ er	klein+st+e	klein+st+ es	klein+st+e
klein+st+es/en	klein+st+er	klein+st+es/en	klein+st+er
klein+st+em	klein+st+er	klein+st+em	klein+st+en
klein+st+en	klein+st+e	klein+st+es	klein+st+e

Outline

Derivationsmorphologie: Wiederholung

- Neue Wortform aus Basis + Derivationsuffix
- Ändert sich die syntaktische Wortart, spricht man von klassenenverändernder Derivation.
- Derivationsuffix:
  - neue Bedeutung
  - reihenbildend (modifiziert analog)

Komposition

- Definition: Rekursive Kombination
- Fugenelemente sind keine Morpheme.

- Nominalkomposition: [Straße]n[bahn] [Sprech][übung]
- Verbalkomposition: [press][schweißen] [stand][halten]
- Adjektivkomposition: [alt][ehrwürdig] [taub][stumm] [treff][sicher]
- Klammerung zur Darstellung der rekursiven Struktur ("Verschachtelung"):
  → [[Straße]n[bahn]][fahrerin]
- *Mädchenhandelsschule*
  → [Porzellan][[eier][korb]] vs. [[Porzellan][eier]][korb]?
- falsche Trennung erschwert Segmentieren beim Lesen , z.B.: Talent-wässerung , Gebirg-stier , Wachs-tube , Tau-schwert , Mais-turm , Rohr-ohr-zucker

Inflection-derivation continuum

	Inflection	Derivation
(i)	relevant to the syntax	not relevant to the syntax
(ii)	obligatory expression of feature	not obligatory expression
(iii)	unlimited applicability	possibly limited applicability
(iv)	same concept as base	new concept
(v)	relatively abstract meaning	relatively concrete meaning
(vi)	compositional meaning	possibly non-compositional meaning
(vii)	expression at word periphery	expression close to the base
(viii)	less base allomorphy	more base allomorphy
(ix)	no change of word-class	sometimes changes word-class
(x)	cumulative expression possible	no cumulative expression
(xi)	not iterable	possibly iterable

Table 5.5 A list of properties of inflection and derivation

Inflection-derivation continuum

Language	Formation	Example	cum	obl	new	unl	cm
English	3rd singular	*walk/walks*	I	I	I	I	I
English	noun plural	*song/songs*	D	I	I	I	I
Spanish	diminutive	*gato/gatito*	D	D	I	I	I
English	repetitive	*write/rewrite*	D	D	D	I	I
English	female noun	*poet/poetess*	D	D	D	D	I
English	action noun	*resent/resentment*	D	D	D	D	D

Note: cum= cumulative expression; obl = obligatory; new = new concept; unl = unlimited applicability; cm = compositional meaning.

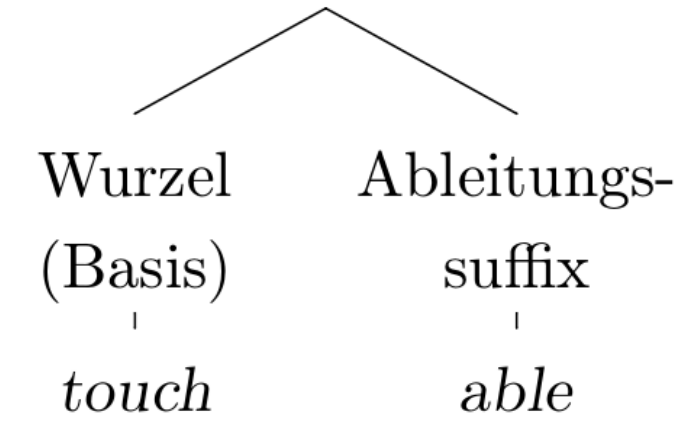
Table 5.6 A continuum from inflection to derivation

Übung: Morphologische Analysebäume

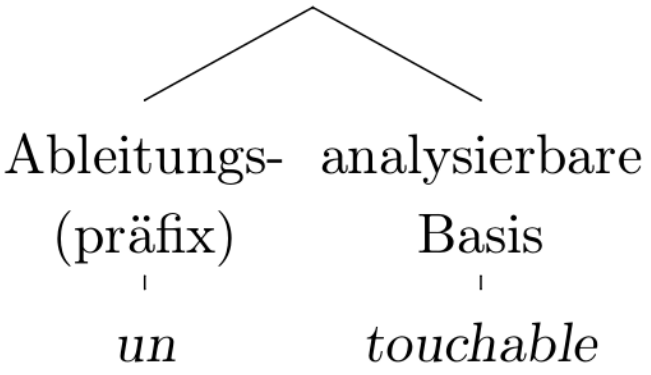
- haut
- Haut
- Bahnhof
- Gutshof
- Holzgeigenkasten
- unlockable
- verhaut

Beispiel: Teilanalysen

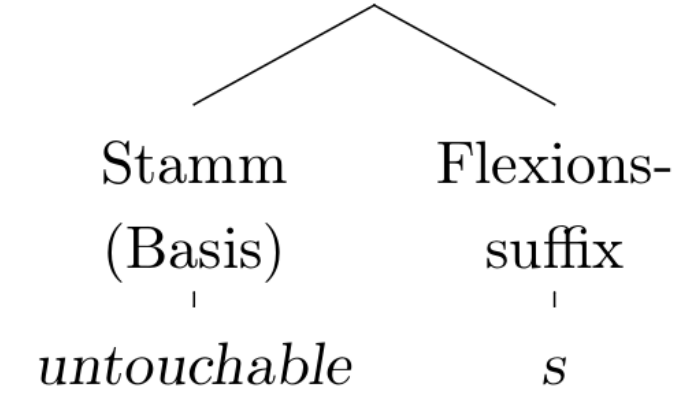
1.



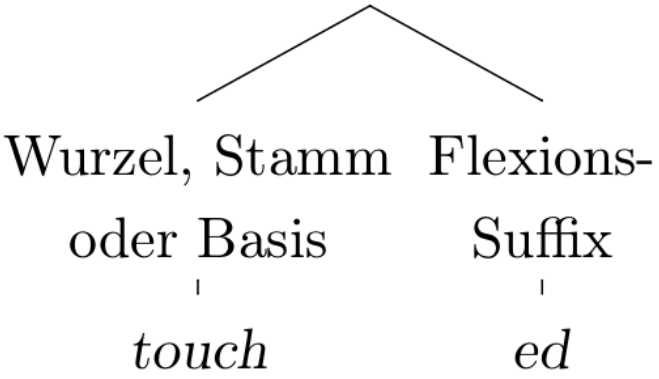
2.



3.



4.



Übung: Morphologische Analysebäume

- haut
- Haut
- Bahnhof
- Gutshof
- Holzgeigenkasten
- unlockable
- verhaut

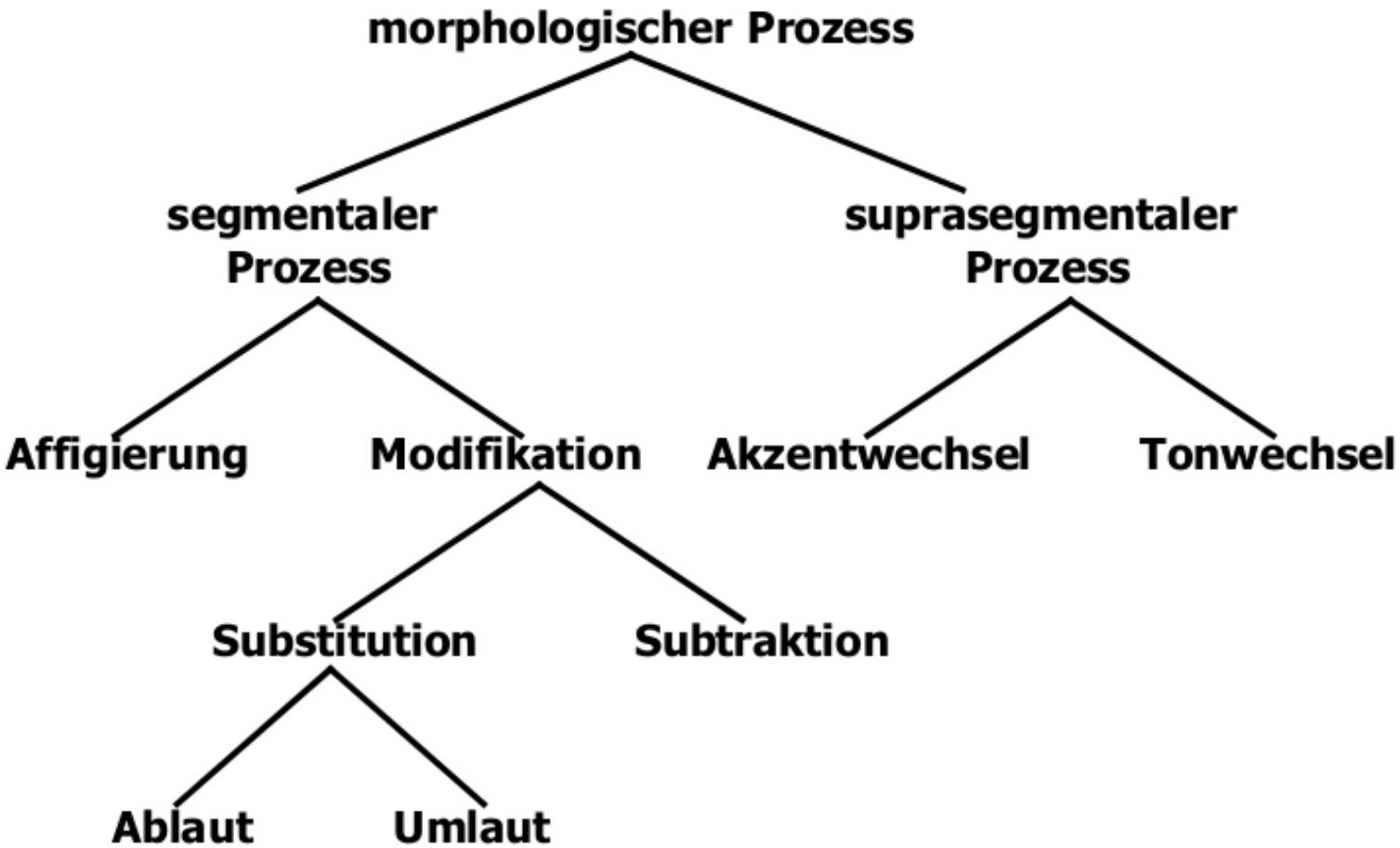
Outline

(Weitere) Morphologische Prozesse

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

- Affigierung (schon behandelt)
- Modifikation (Ablaut: *sing-sang-gesungen*; Umlaut: *Maus-Mäuse*)
- Subtraktion (Tilgung) von Segmenten oder Merkmalen
  - *Omnibus* $\Rightarrow$ *Bus*
  - phonologisch im Französischen: gris - /gʁi/ (maskuline Form durch Tilgung des */z/* gebildet)
- Suprasegmental (nicht auf orthographischer/phonemischer Ebene)
  - Akzentwechsel (pro'duce (v.) vs. 'produce (n), per'mit (v.) vs. 'permit (n.); analog: import, insult, discount . . .)
  - Tonwechsel (in tonalen Sprachen, z.B. Kanuri, saharanische Sprache)

Morphologische Prozesse



Volksetymologie

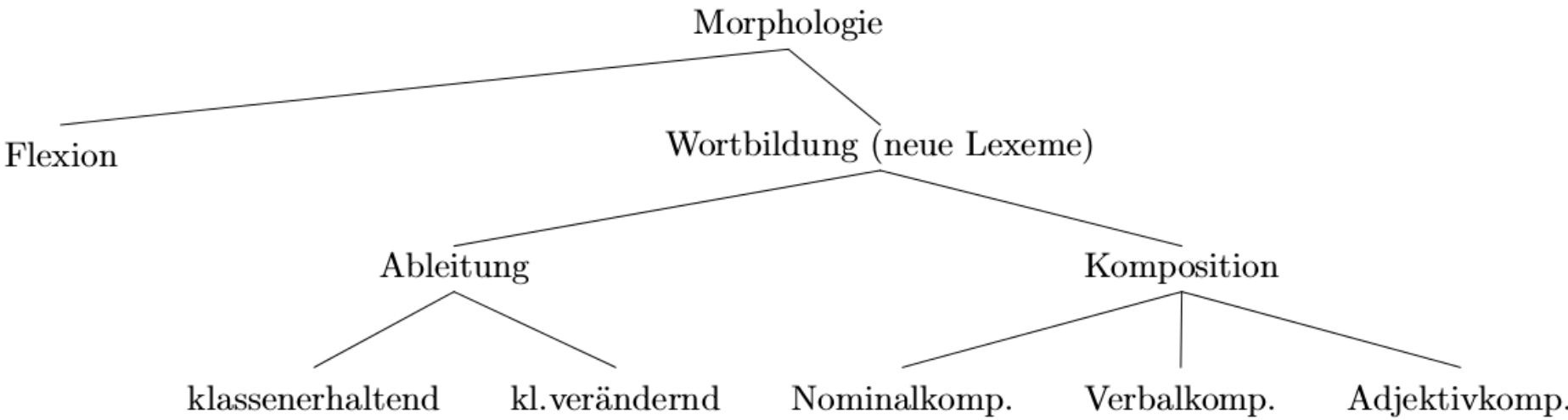
Oft werden Lehnwörter fremder Sprachen durch Kombination ähnlichklingender (und teils bedeutungsähnlicher) Morpheme nachgebildet:

- Hängematte: von Taino/Haiti hamaka (Schlafnetz)
- Vielfraß: von altnorwegisch fjeldfross (Gebirgskater)
- Messner: von lateinisch mansionarius (Aufseher des Gotteshauses)
- Quäntchen: von lateinisch quintus (ein Fünftel)
- Tollpatsch: von Ungarisch talpas (Fußsoldat)

Grammatikalisierung

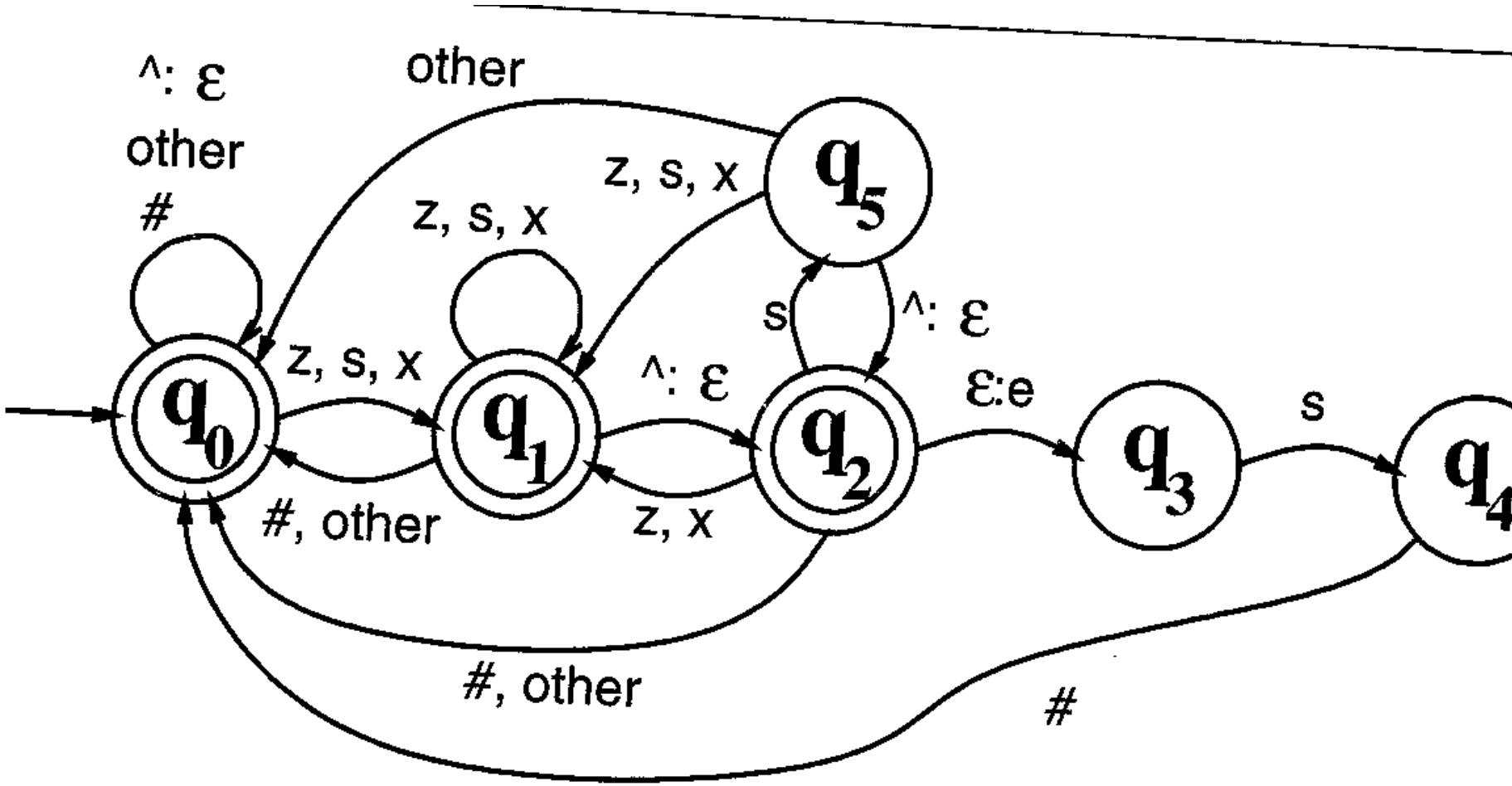
- Morphem verliert lexikalische Bedeutung und Freiheit in der Stellung
- Inhaltswörter werden zu Funktionswörtern
- freie Morpheme zu gebundenen
- z.B. dt. Präteritum-Affix *-t-*: urspr. *sagen-tat* $\rightarrow$ *sagte*
- Suffixe *-keit, -heit, -tum, -lich*: urspr. eigenständige Wörter:
  - *-lich*: ahd. 'Körper, Gestalt'
  - *-keit, -heit*: 'Art, Weise'
  - *-tum*: 'Würde, Stand'
- derzeit im Übergang: *Weise, frei, voll, mäßig, Zeug, Werk*
- ähnlich *bekommen: Er bekommt etwas geschenkt*, aber auch *Er bekam den Zahn gezogen*
- franz. *ne - pas*: (nicht) $\leftarrow$ 'keinen Schritt'

Morphologie: Übersicht

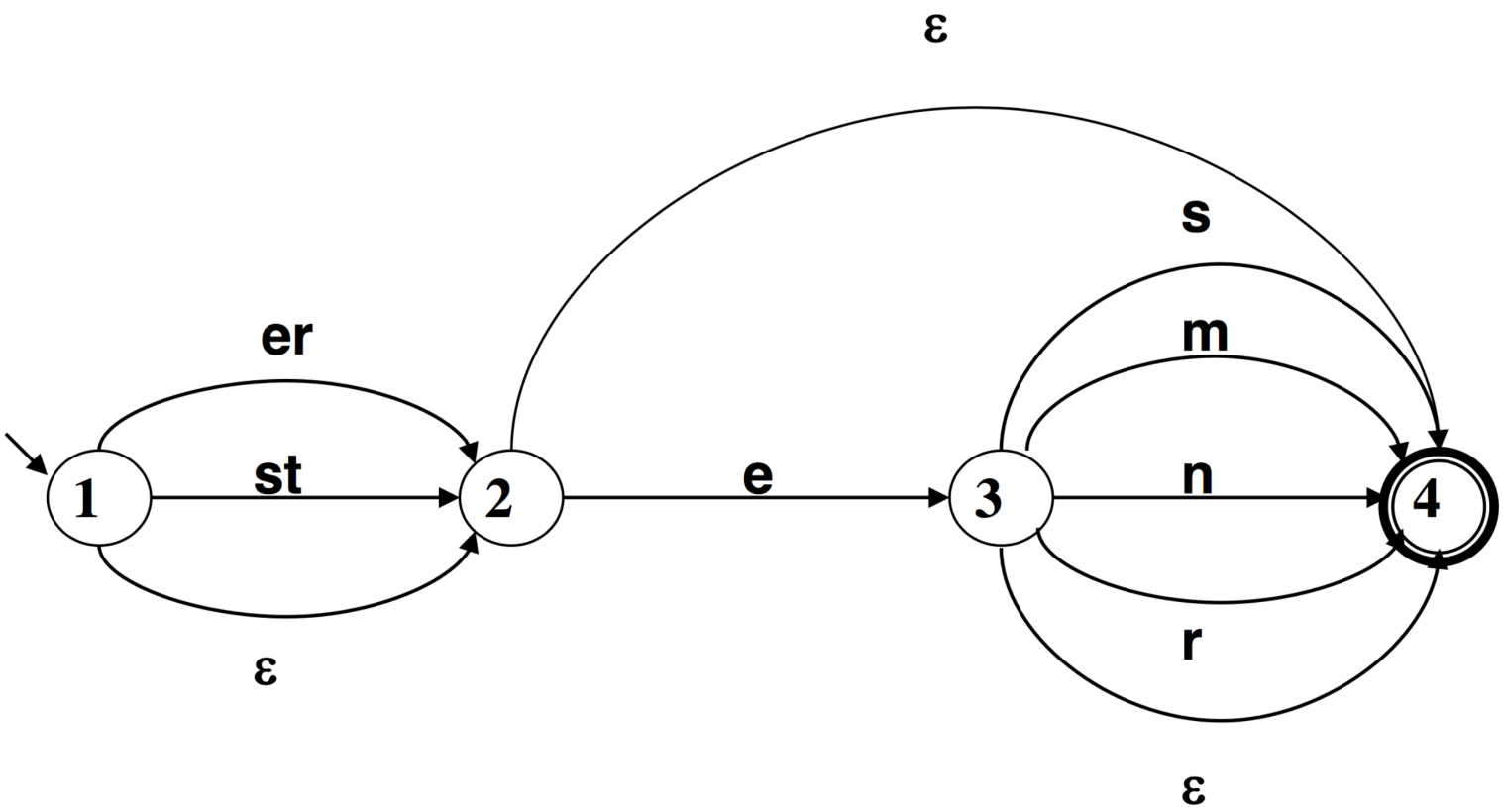


Outline

Finite state automaton

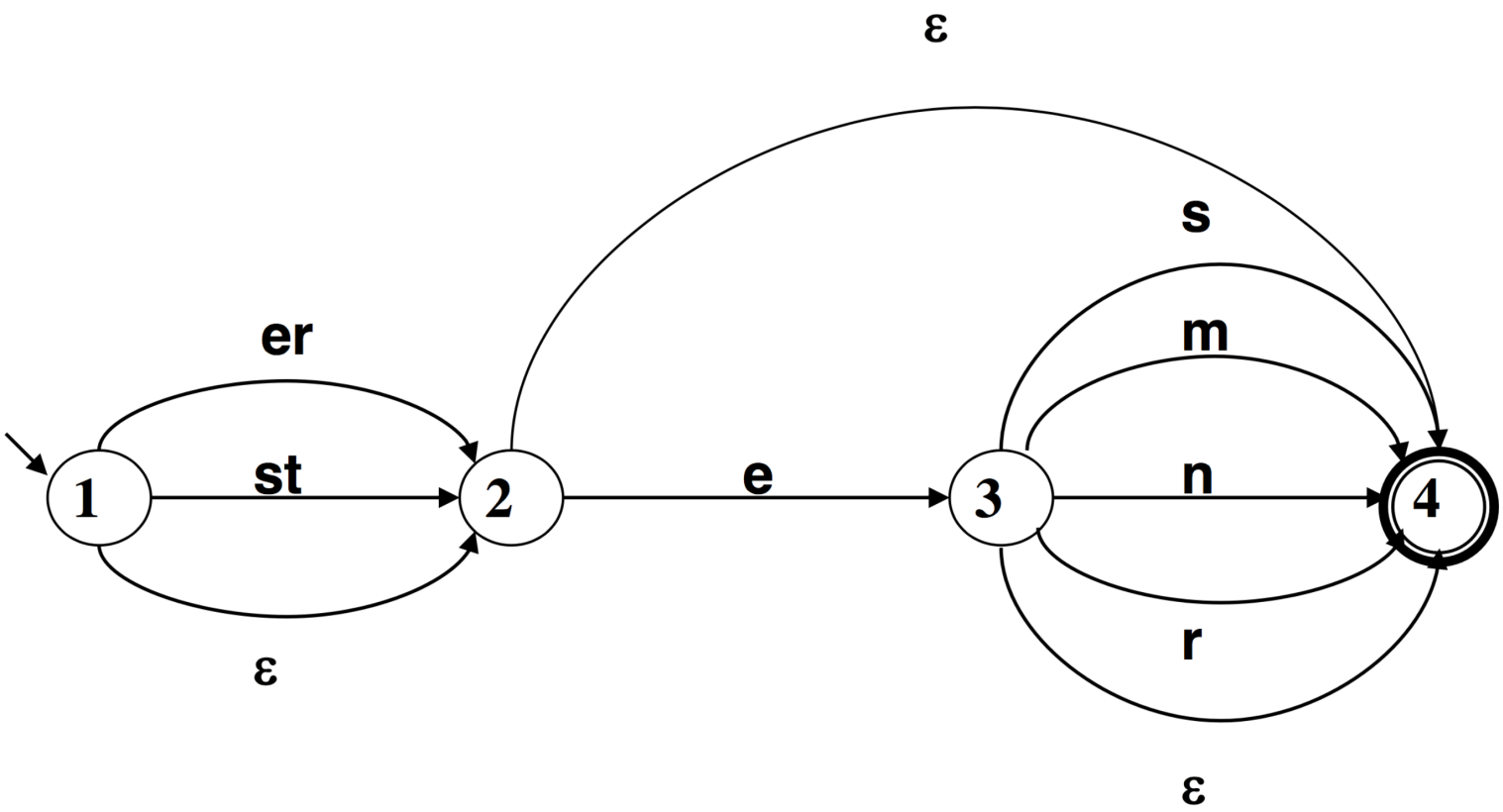


Transducer



smor

Transducer



Zum Schluss: Besonders klausurrelevant

- Morphem
- Flexion
- Derivation
- Komposition
- Morphologische Baumanalyse
- Wurzel, Basis, Affix, Fugenelement, Wortform
- Flexionsparadigma
- Synkretismus

Übung

Tokenisieren und lemmatisieren Sie den folgenden Satz. Bestimmen Sie die Anzahl der Tokens. Bestimmen Sie die Anzahl der Types fuer die zwei in der letzten Woche eingeführten Definitionen von Type.

“Matthias Rose sah den Rosengarten und viele Rosen, die er noch nicht gesehen hatte”

Flowchart

Übung

Zeichnen Sie die morphologischen Analysebäume. Geben Sie bei jeder Basis (das schließt jede Wurzel ein) die Wortart an: V, N, A.

- furchtlos
- Sauberkeit
- lesbar
- Tarnung

Übung

Zeichnen Sie die morphologischen Analysebäume. Geben Sie bei jeder Basis (das schließt jede Wurzel ein) die Wortart an: V, N, A.

- creating
- unhealthy
- seaward
- reconsider
- incompletion
- (das) himbeerigste (Eis, das ich je gegessen habe)
- Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz