Einführung in die Syntax und Morphologie

Vorlesung und Übung

Prof. Dr. phil. habil. Tania Avgustinova

FR Sprachwissenschaft und Sprachtechnologie

Universität des Saarlandes

Organisatorisches

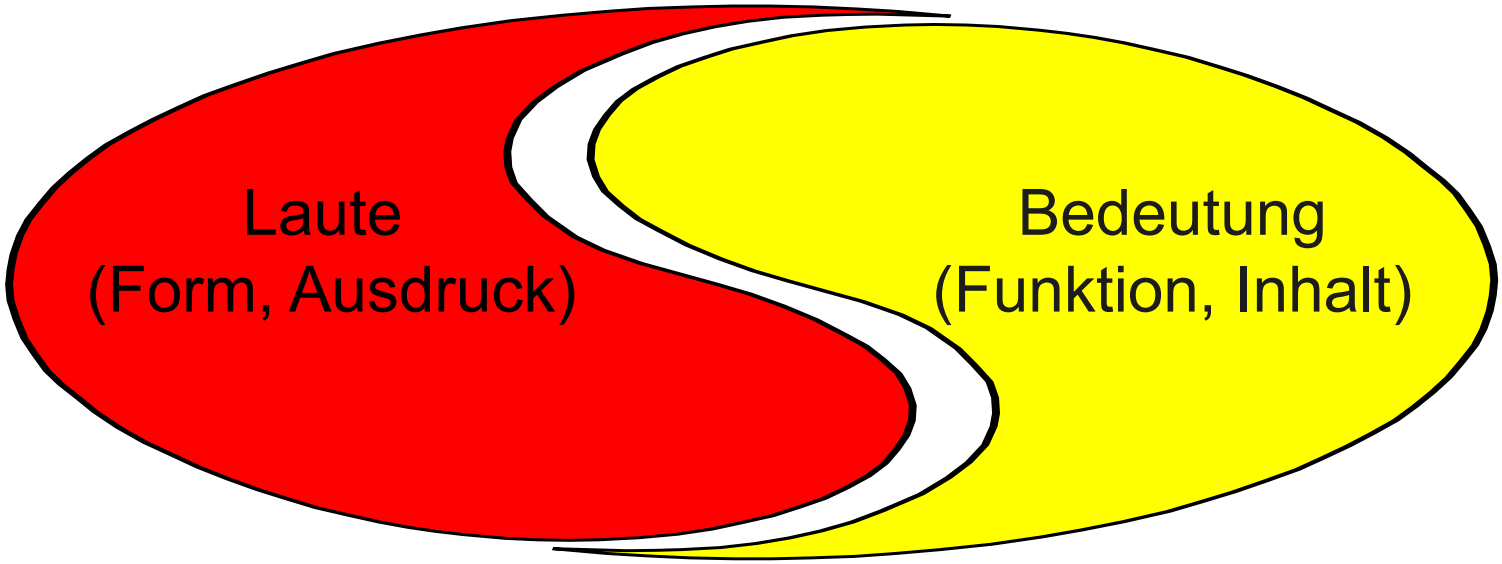
- **Übungsblätter**
  - → nach der jeweiligen Vorlesung ausgeteilt (MSTeams)
  - → regelmäßige Bearbeitung sehr empfohlen

- **Lösungen**
  - → Abgabe bis Do 12:00 (MSTeams)
  - → unbenotet, dienen der Orientierung bzw. als Feedback

- **Übungsgruppen**
  - → Diskussionsforum, evtl. Fragen vorbereiten
  - → Besprechung der Aufgaben inkl. Musterlösungen

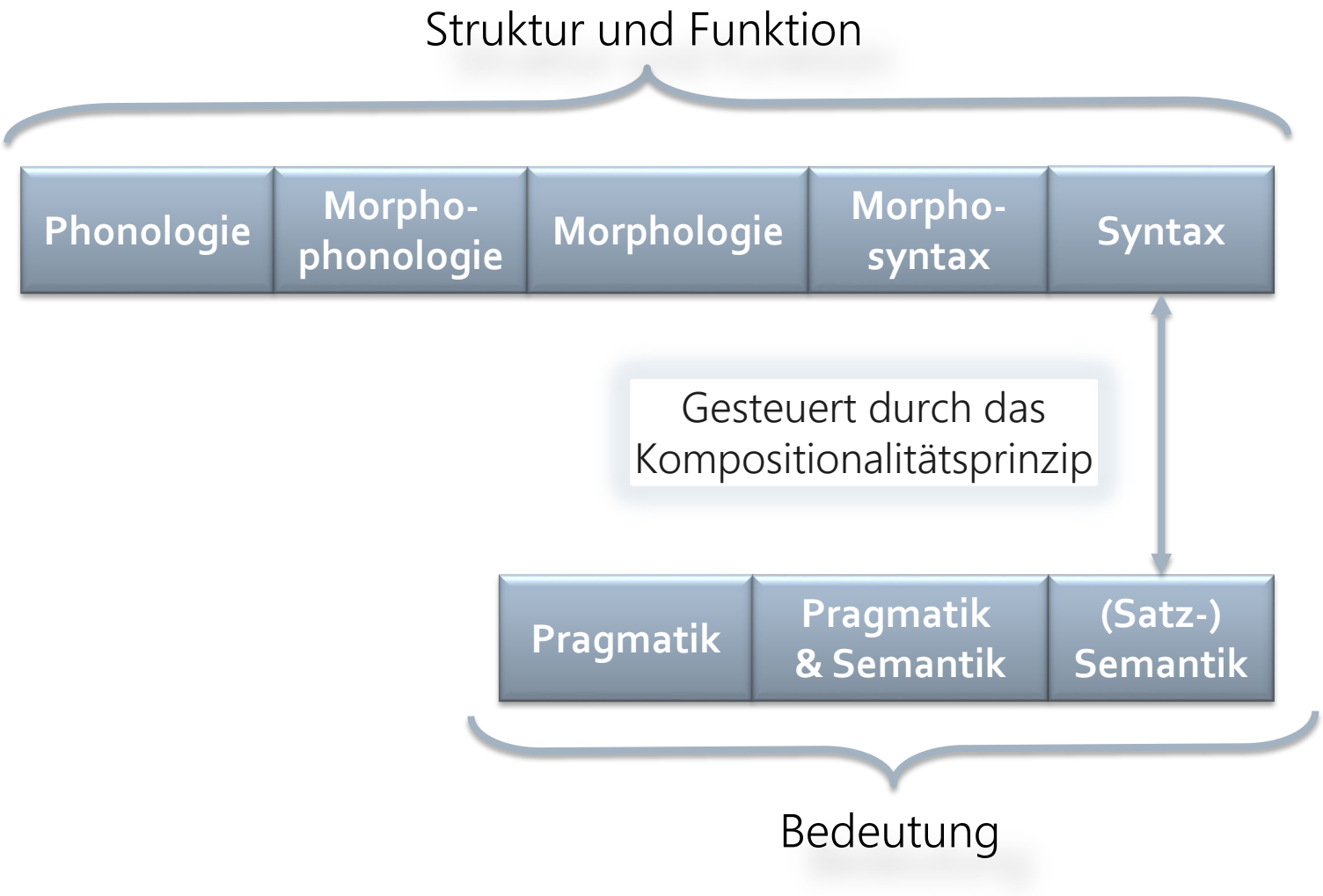
Beschreibung / Modellierung der Sprache

- Empirisch zu beobachten sind:



- Oberflächenorientierung: sprachliche Einheiten werden so beschrieben, wie sie nach einem herkömmlichen Verständnis „tatsächlich“ auftreten.

Beschreibung / Modellierung der Sprache



Beschreibung / Modellierung der Sprache

- Zu sprachlichen Einheiten
  - Laut (Phon/Phonem): nicht weiter zerlegbar, enthält "Bündel" von Merkmalen
  - Morphem: besteht aus (keinem,) einem oder mehreren Lauten
  - Wort: besteht aus einem oder mehreren Morphemen
  - Syntagma: besteht aus einem oder mehreren Wörtern
  - Satz: besteht aus einem oder mehreren Satzgliedern (Syntagmen)
  - Text: besteht aus einem oder mehreren Sätzen
  - Diskurs: besteht aus einem oder mehreren Texten
- … und deren Beschreibung
  1. Mikroebene:
     die kleineren Einheiten erhält man durch Analyse von größeren Einheiten
  2. Makroebene:
     die größeren Einheiten erhält man durch Kombination von kleineren Einheiten

Teilgebiete der Sprachwissenschaft

- **Phonetik**: Sprachlaute, Erzeugung und physikalische Eigenschaften
- **Phonologie**: Funktion der Laute im System der Sprache, d.h. welche Lautunterschiede können Bedeutungsunterschiede nach sich ziehen
- **Morphologie**: **Wort**bestandteilen und ihre Funktion; Form- und Wortbildung
- **Syntax**: **Wort**verbindungen bis zur Ebene des Satzes
- **Semantik**: Bedeutung sprachlicher Ausdrücke
- **Pragmatik**: Beziehung der Sprache zum außersprachlichen Kontext

Morphologie und **Syntax** liefern Hinweise zur Bedeutung durch die ausgedrückten strukturellen **Regelmäßigkeiten**.

Wie viele Wörter?

1. Der Nachrichtensprecher versprach sich.
2. New York ist nicht die Hauptstadt der Vereinigten Staaten.
3. Er kauft gerne am Samstag ein.
4. Sie konnten weder vor- noch zurückgehen.
5. Hans war ganz aus dem Häuschen.

Wie viele Wörter? Ungewiss?

1. Der Nachrichtensprecher versprach sich.
2. New York ist nicht die Hauptstadt der Vereinigten Staaten.
3. Er kauft gerne am Samstag ein.
4. Sie konnten weder vor- noch zurückgehen.
5. Hans war ganz aus dem Häuschen.

Auch ein Wort?

Inuit-Sprachen gelten als typische Vertreter eines **polysynthetischen** Sprachbaus – vgl. Inuktitut nach [Beesley & Karttunen 2003:376]

Paris + mut + nngau + juma + niraq + lauq + sima + nngit + junga

Paris Terminalis-Kasus Weg-nach wollen sagen-dass Vergangenheit Perfektiv Zustand Negativ 1sg-intransitiv

Parimunngaujumaniralauqsimanngittunga

"Ich sagte niemals, dass ich nach Paris gehen will."

Grundbegriffe

- **lexikalisches Wort** (**Lexem**):
  abstrakte Bedeutungseinheit, die einer bestimmten Wortart angehört

- **syntaktisches Wort** (**Wortform**):
  konkrete Realisierung im Verlauf, im Kontext vorkommende Flexionsform

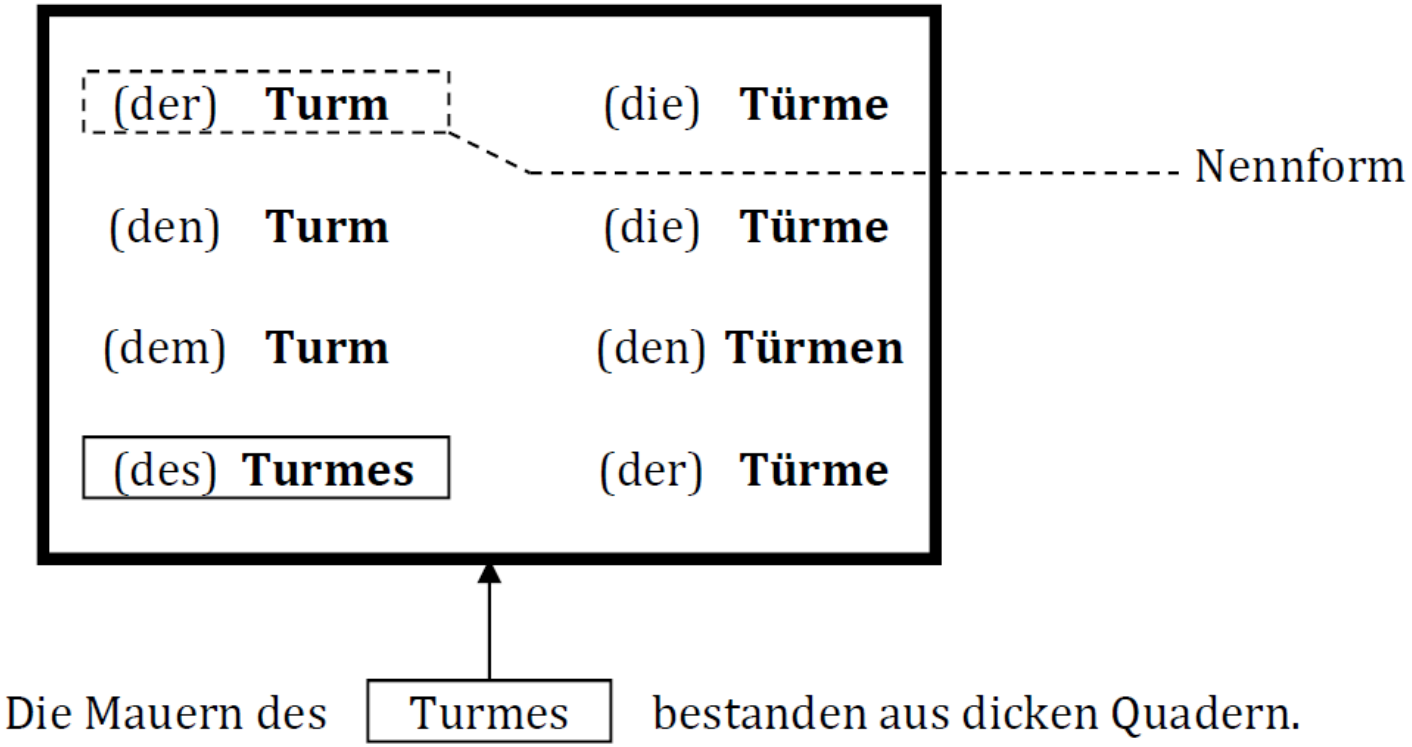
  bzw. **grammatisches Wort** (Wortform + grammatische Funktion)

  - Eine Wortform kann mehrere grammatische Wörter repräsentieren.
  - Die grammatischen Wörter eines Lexems bilden ein Paradigma.

- **Lemma** (Nennform, Grundform, Zitierform): per Konvention aus einem Paradigma zur Repräsentation ausgewählte Wortform

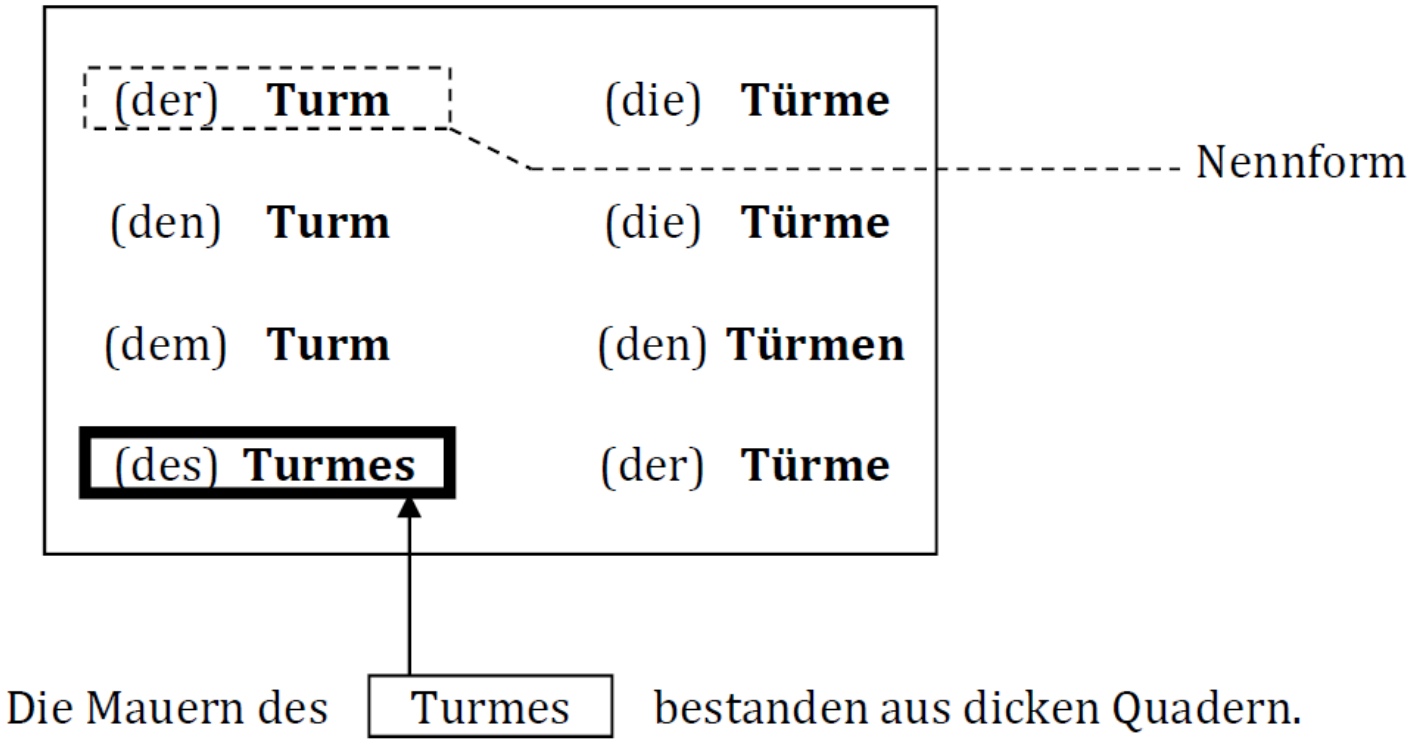
In der Sprachwissenschaft spielen vor allem zwei Typen von »Wort« bzw. zwei Wortbegriffe eine Rolle: das syntaktische Wort und das lexikalische Wort (= Lexem).

Lexem (lexikalisches Wort) mit seinen Flexionsformen (→ Formenreihe, Paradigma)



In der Sprachwissenschaft spielen vor allem zwei Typen von »Wort« bzw. zwei Wortbegriffe eine Rolle: das syntaktische Wort und das lexikalische Wort (= Lexem).

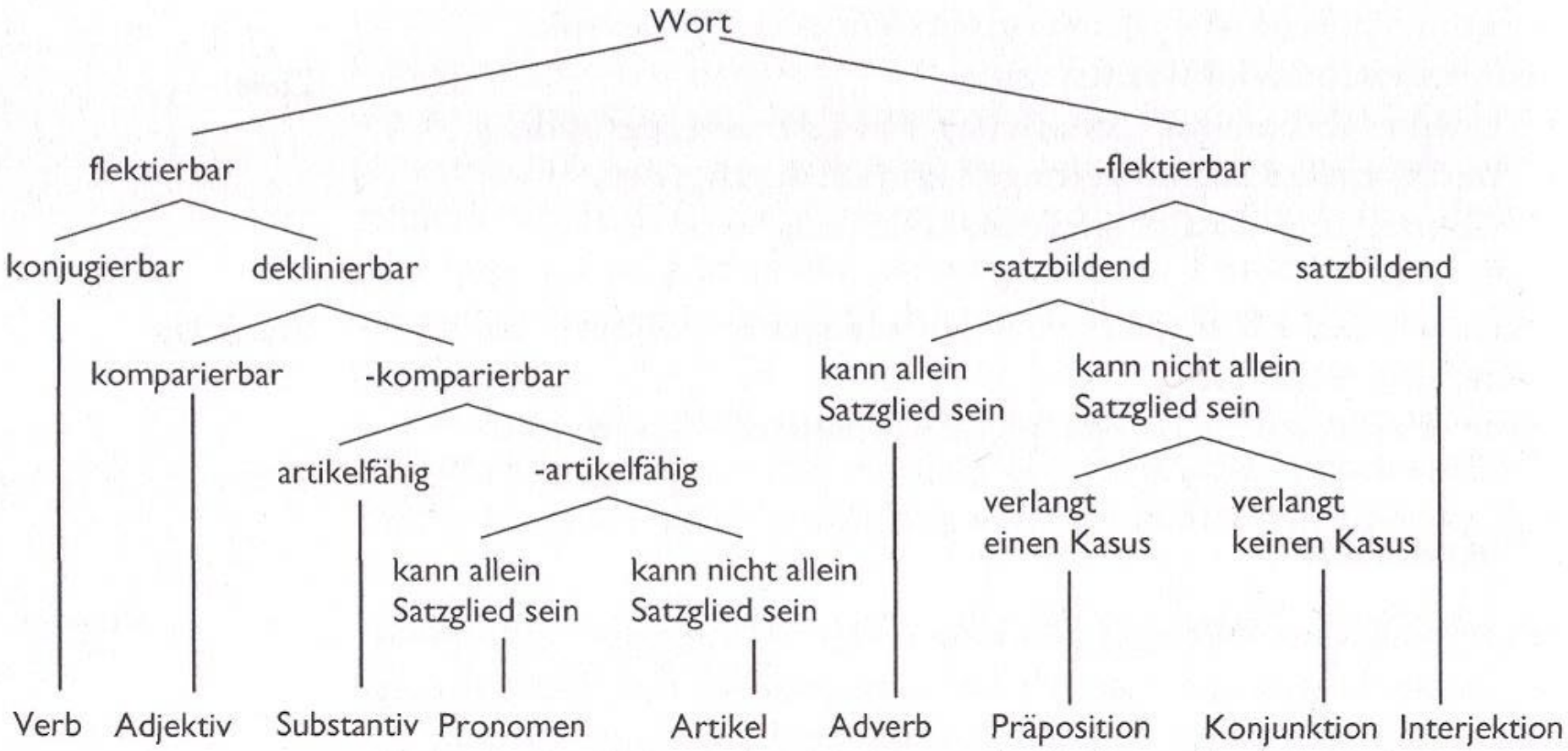
Flexionsform (syntaktisches Wort) mit zugehörigem Lexem:



Wortarten

- Ergebnis der Wortklassifikation nach Form- und Bedeutungsmerkmalen
- Motivation
  - Viele Eigenschaften sind identisch für (große) Klassen von Wörtern
  - Regeln gelten nur für bestimmte Kategorien von Lexemen
  - Kategorisierung der Lexeme nötig ! Generalisierungen werden möglich
  - Wichtig für Anwendungen in CL
- Zahl schwankt je nach Gliederungsaspekten und Zweck (z.B. Tagging)
- Mögliche Gliederungsaspekte:
  - morphologisch: flektierbar?
  - syntaktisch: satzgliedwertig? mit Kasusforderung? artikelfähig? etc.
  - semantisch: Dinglichkeit? Eigenschaft? Prozess? Relation?

Wortarten: eine mögliche Klassifizierung



z.B. Wortarten des Deutschen

Nach Helbig & Buscha (1984)

1. Hauptwortklassen:
   a) Verben
   b) Substantivwörter
   c) Adjektive
   d) Adverbien
2. Funktionswörter:
   a) Artikelwörter
   b) Präpositionen
   c) Konjunktionen
   d) Partikeln
   e) Modalwörter
   f) Negationswörter
   g) Satzäquivalente
   h) Pronomen „es“

Schwierigkeiten bei der Zuordnung

- Wortartwechsel
  - Leid vgl. Das tut mir leid.
  - Klasse vgl. ein klasse Buch
  - ja vgl. Das war ein klares Ja.

- Ambiguität (Zugehörigkeit zu mehreren Wortarten), z.B.
  - aber vgl. Er las, aber er war sehr unkonzentriert (Konjunktion)
    Das kann man aber so nicht sagen (Partikel)
  - Zahlwörter
    - eins / zwei (Kardinalzahl, deklinierbar, z.B. der Bund zweier Kaiser)
    - hundert / tausend (Kardinalzahl, aber vgl. das Hundert vollmachen)
    - Million (eher wie Nomen)

Part-of-Speech (POS) Tagging

- Wörter eines Textes mit dazugehörigen Wortarten kennzeichnen
  - eine Art der Annotierung
  - manuell oder durch Algorithmen

- Inventar an Wortarten → Tagset
  - Welche Granularität braucht man?
  - Was ist universal bzw. sprachspezifisch?

XEROX Tagset für Deutsch (1/2)

Tag	Description	Example
+NOUN	common noun, nominalized adjective, nominalized infinitive, or proper name	Hut, Leute, [das] Gute [das] Wollen Peter, [die] Schweiz
+VVFIN	finite verb form	[er] sagt
+VVINF	infinitive	[er will] sagen, einkaufen
+VVIZU	infinitive with incorporated "zu"	[um] einzukaufen
+VVPP	past participle	[er hat] gesagt
+VAFIN	finite auxiliary	[er] ist, [sie] haben
+VAINF	auxiliary infinitive	[er will groß] sein
+VAPP	auxiliary past participle	[er ist groß] geworden
+VMFIN	finite modal	[er] kann, [er] mochte
+VMINF	modal infinitive	[er wird kommen] können, [er hat kommen]wollen
+VMPP	modal past participle	[er hat es] gekonnt
+ART	article	der [Mann], eine [Frau]

XEROX Tagset für Deutsch (2/2)

+PERSPRO	personal pronoun	ich, du, ihm, mich, uns
+REFLPRO	reflexive "sich"	sich
+REZPRO	reciprocal "einander"	einander
+POSPRO	possessive pronoun	[das ist] meins
+POSDET	possessive determiner	mein [Haus]
+INDDET	indefinite determiner	kein [Mensch]
+RELPRO	relative pronoun	[der Mann,] der [lacht]
+WPRO	interrogative pronoun	wer [ist da?]
+WDET	interrogative determiner	welche [Nummer?]
+ADV	non-adjectival adverb	oft, heute, bald, vielleicht
+CARD	cardinal	1, eins, 1/8, 205
+ORD	ordinal	2., dritter, 1.2.
+COORD	coordinating conjunction	und, oder
+SENT	sentence final punctuation	. ; ? !
+CM	comma	,
+PUNCT	other punctuation, bracket	: ( ) [ ] - "K "

Was ist ein Wort?

- Bessere Formulierung der Frage: Welche **syntagmatische** und **paradigmatische** Beziehungen sind charakteristisch für Wörter?

- Wörter sind typischerweise **komplexe sprachliche Zeichen**, die aus kleineren Einheiten (den Morphemen → Mikroebene) aufgebaut sind und die ihrerseits Bestandteile noch größerer Zeichenkomplexe (z.B. Sätze, Phrasen → Makroebene) sein können.

- Es ist die sprachliche Einheit, die morphologischen Prozessen **Wortbildung** und **Flexion** unterworfen ist.

Identifizierung und Klassifizierung

- Abfolge im **Verlauf** vs. Gegenüberstellung im **System**

  Die Beschreibung einer Sprache geht vom Verlauf aus und zunächst ermittelt durch Segmentierung **syntagmatische** Beziehungen, um dann die so gewonnenen Segmente nach gemeinsamen Eigenschaften (also **paradigmatischen** Beziehungen) zu ordnen und auf diese Weise zur Darstellung des Systems zu gelangen.

Es liegen im **System** und im **Verlauf** dieselben Elemente vor, aber sie werden im **System** von einem anderen Gesichtspunkt betrachtet als im **Verlauf**.

- Sinnvolle terminologische Unterscheidung:
  - Vorkommen im **Verlauf** (syntagmatisch) → Token
  - Element des **Systems** (paradigmatisch) → Type

Morphem als Minimalzeichen

- Beobachtung: Wörter können in minimale Einheiten zerlegt werden, die ihrerseits **einfache sprachliche Zeichen** sind.

  → Er-leb-nis(s)-e Be-schein-ig-en Be-leb-ung

  → „Mehrzahl“: -er „wie ein“: -lich „Infinitivbildung“: -en

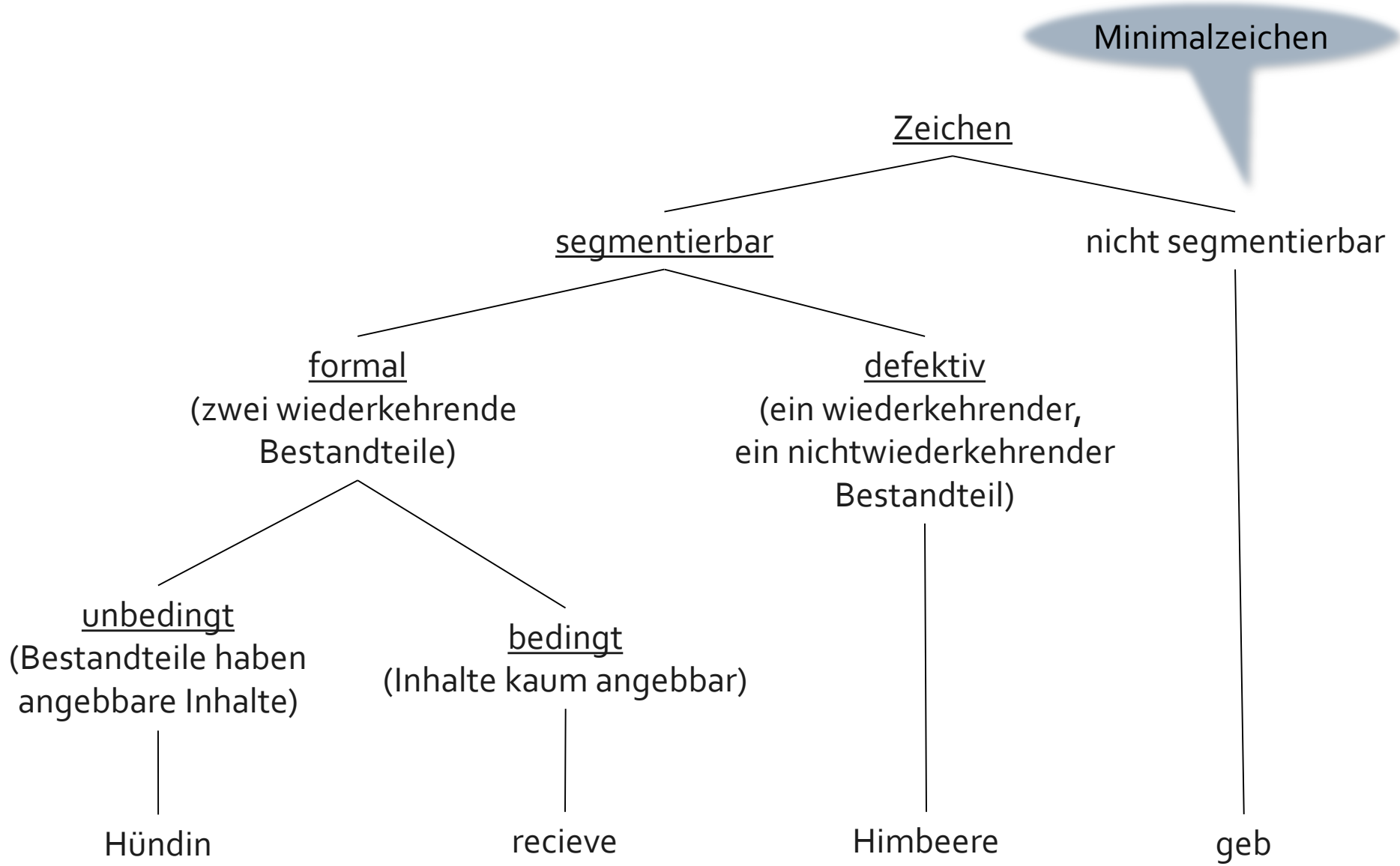
- klassische Definition:

  Ein Morphem ist **die kleinste bedeutungstragende Einheit** einer Sprache.

- revidiert (z.B. Wurzel 1984:38):

  Ein Morphem ist die kleinste, in ihren verschiedenen Vorkommen als formal einheitlich identifizierbare Folge von Segmenten, der (wenigstens) eine als einheitlich identifizierbare **außerphonologische Eigenschaft** zugeordnet ist.

Segmentierbarkeit



morphologisch relevante Einheiten

- **Segmentieren**: K-i-n-d-e-r ◆ l-a-c-h-t-e-n → Minimalzeichen (vs. Silben!)
- **Klassifizieren**:

  ess-
iß-
-gess-
-aß

  ↔ "essen"

  -e
-n
-en
-s
-er
<leer>

  ↔ "plural"

- **Varianz**:
  - lexikalisch gesteuert — Bett-en, Brett-er; knife – knives; index – indces
  - phonologisch bedingt — fütter-t, arbeit-et; cats [s] – dogs [z] – horses [iz]

morphologisch relevante Einheiten

- abstrakte Morpheme (**Grammeme**) stehen oft in Opposition
  - Sg vs. Pl..... Numerus
  - Masc vs. Fem vs. Neut..... Genus
  - Praes vs. Perf vs. Imperf vs. Pluperf vs. Fut..... Tempus
  - Act vs. Pass..... Genus verbi

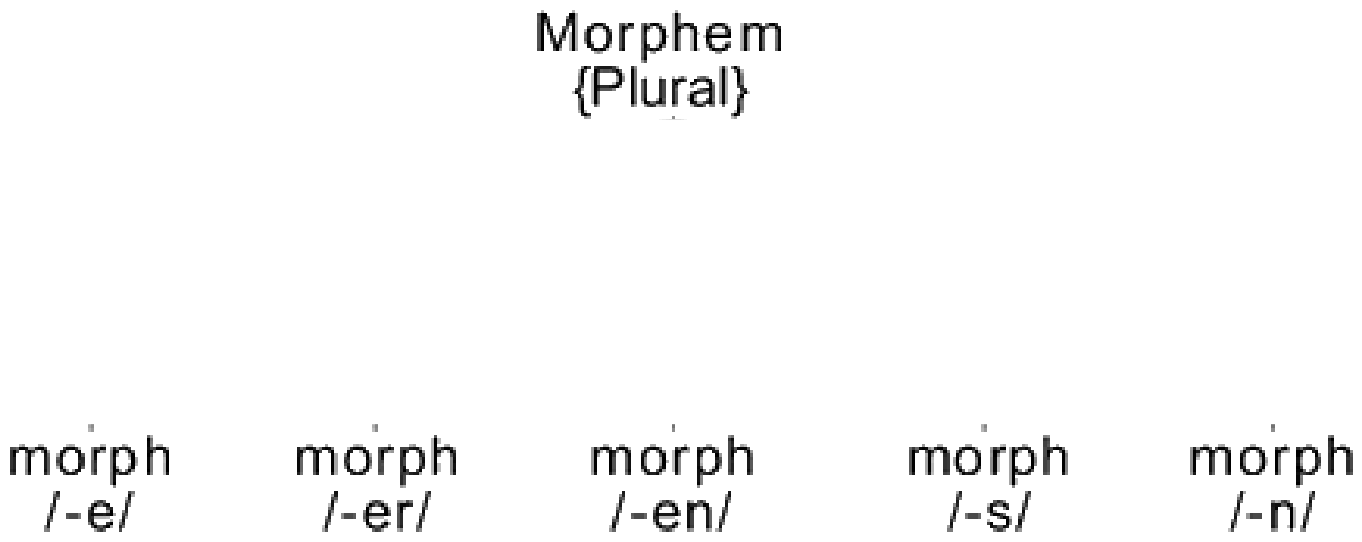
morphologisch relevante Einheiten

- als Ergebnisse des Segmentierens sind dies zunächst die **Morphe**

morph /-e/ morph /-er/ morph /-en/ morph /-s/ morph /-n/

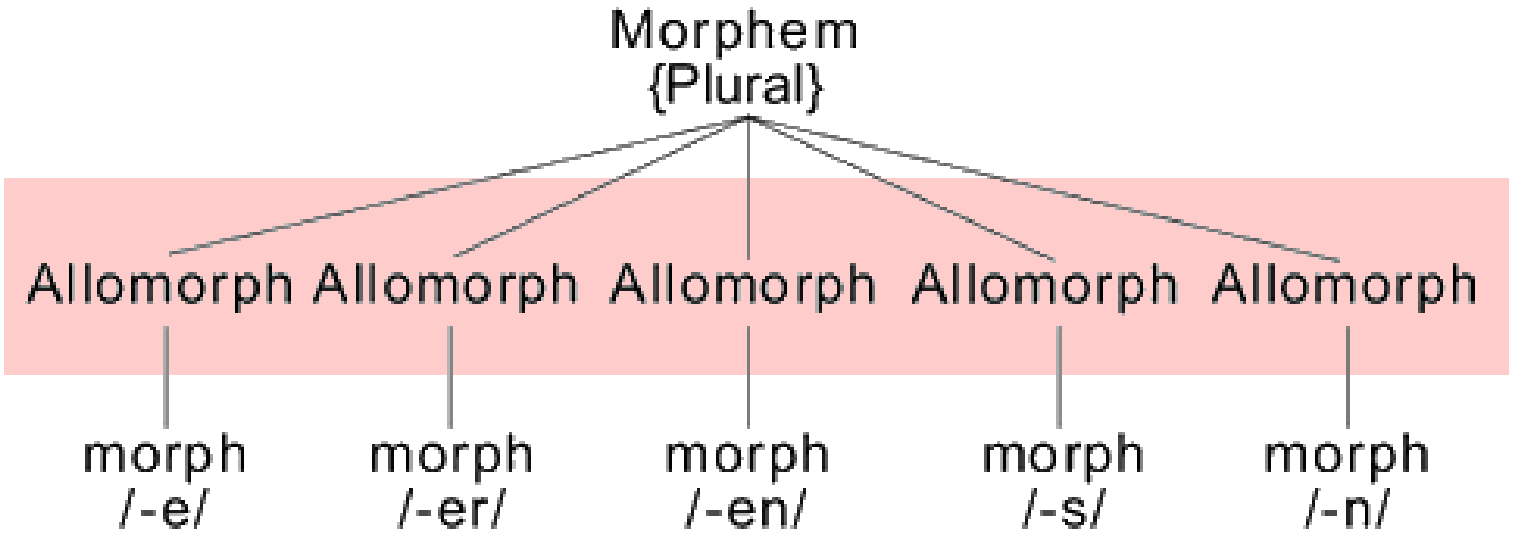
morphologisch relevante Einheiten

- als Ergebnisse des Segmentierens sind dies zunächst die **Morphe**
- diese können abstrakten Einheiten, den **Morphemen**, zugeordnet werden

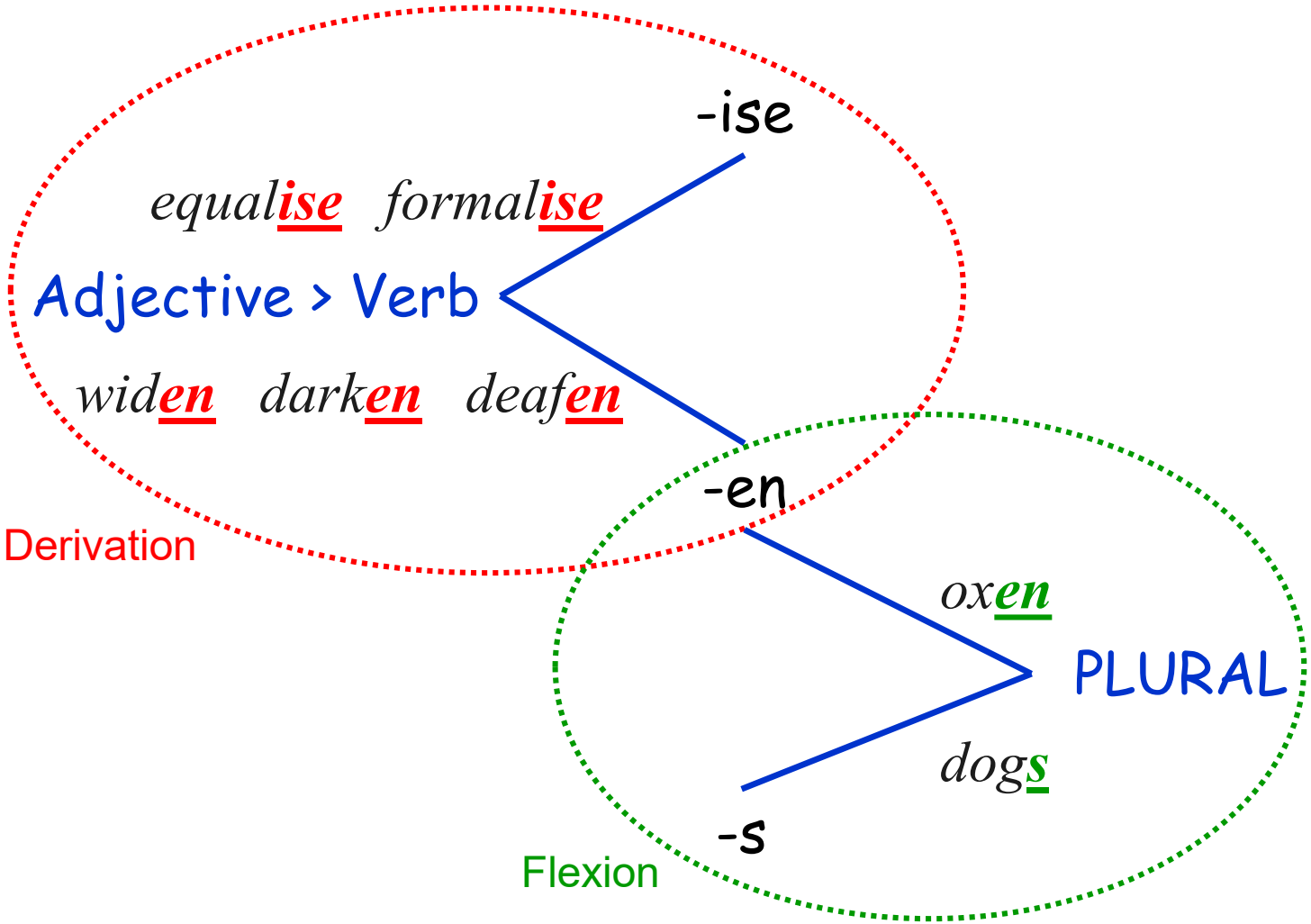


morphologisch relevante Einheiten

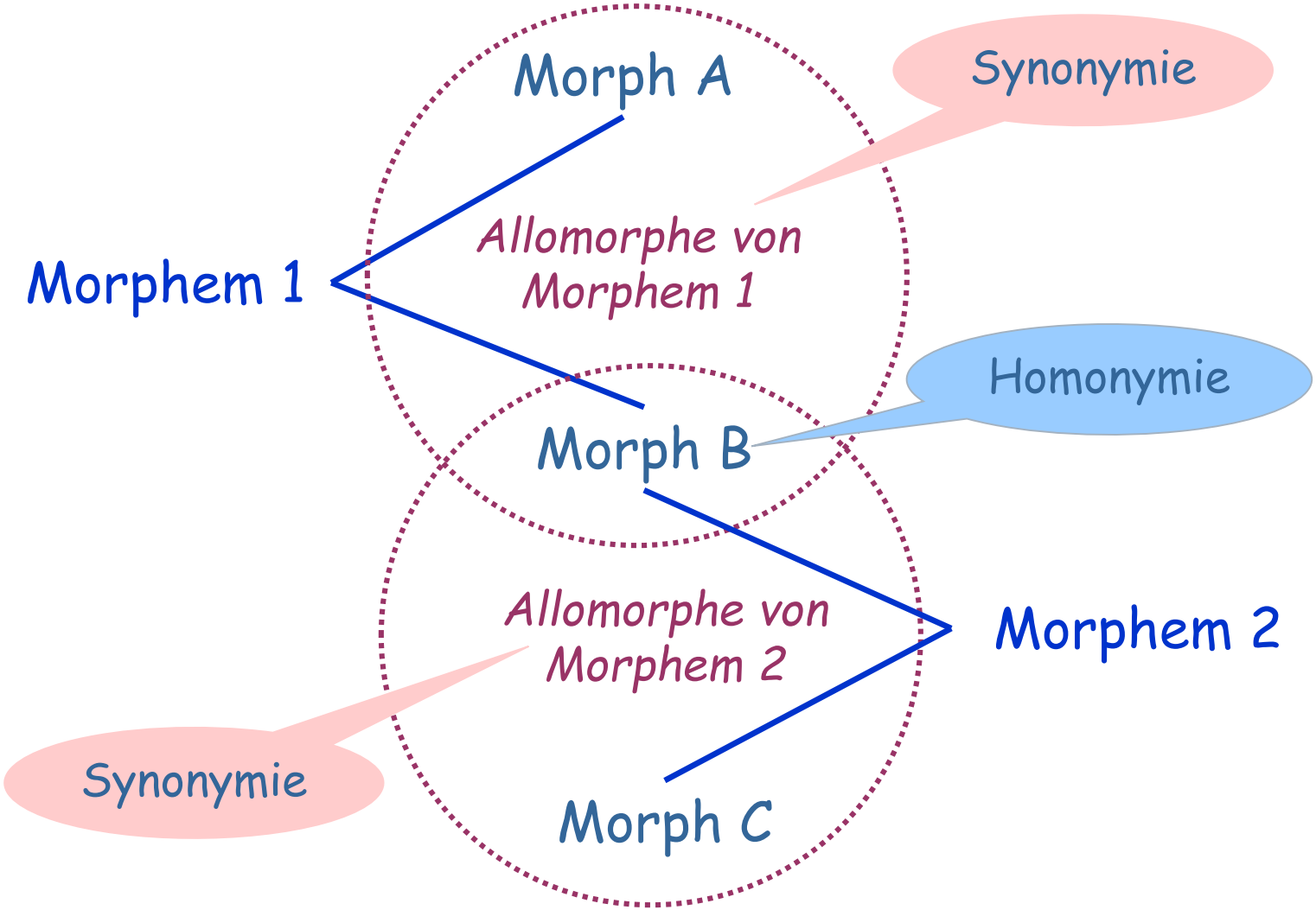
- als Ergebnisse des Segmentierens sind dies zunächst die **Morphe**
- diese können abstrakten Einheiten, den **Morphemen**, zugeordnet werden
- die Klassifizierung von Morphen als **Allomorphe** (Varianten) eines Morphems beruht auf gleicher Bedeutung und komplementärer Verteilung



Identifizierung und Klassifizierung



Identifizierung und Klassifizierung



Morphe – Morphem – Allomorphe

- **Morph**: eine Menge homonymer Minimalzeichen

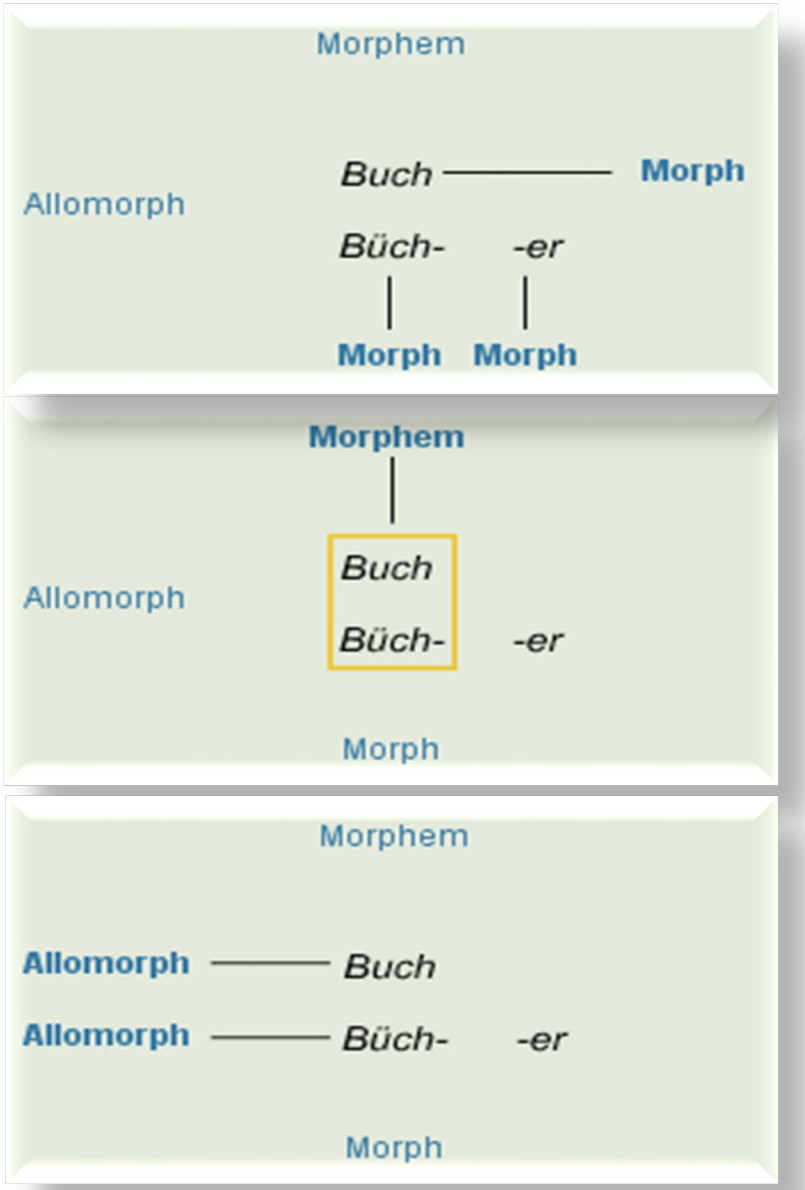
  Minimalzeichen mit derselben Form gehören zu einem Morph

- **Morphem**: eine Menge synonymer Minimalzeichen

  Minimalzeichen mit derselben Bedeutung bilden ein Morphem

- **Allomorphe**: Minimalzeichen, die zu einem Morphem gehören

  Sind synonyme Minimalzeichen zum Morphem klassifiziert worden, sind sie Allomorphe dieses Morphems.



Verfahren zur Ermittlung der Morphe

- Minimalpaaranalyse:
  - /laitete/ (ich) leitete
  - /laite/ (ich) leite
  - /laitetest/ (du) leitetest
  - /leitest/ (du) leitest
- Ergebnis: 4 Morphe
  1. /lait-/ — „leiten, führen“
  2. /-et/ — „Präteritum“
  3. /-e/ — „1.Prs.Sg“
  4. /-est/ — „2.Prs.Sg“

Ein **Morphem** ist eine **maximale Menge** von bedeutungsgleichen **Morphen** in komplementärer Verteilung. Elemente der Menge heißen **Allomorphe.**

→ komplementäre Verteilung:

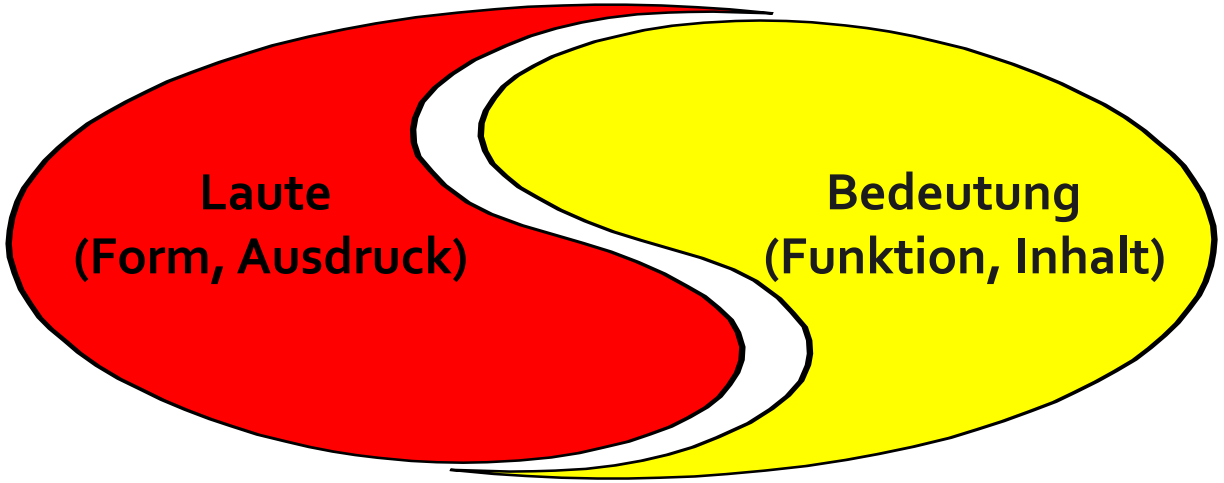
d.h. wenn es keinen Kontext gibt, in dem wahlweise das eine oder das andere Morph auftauchen kann

vgl. -er und -e in Bedeutung Plural:

- Wind – Winde – *Winder
- Kind – *Kinde – Kinder

Gegenstand der Morphologie

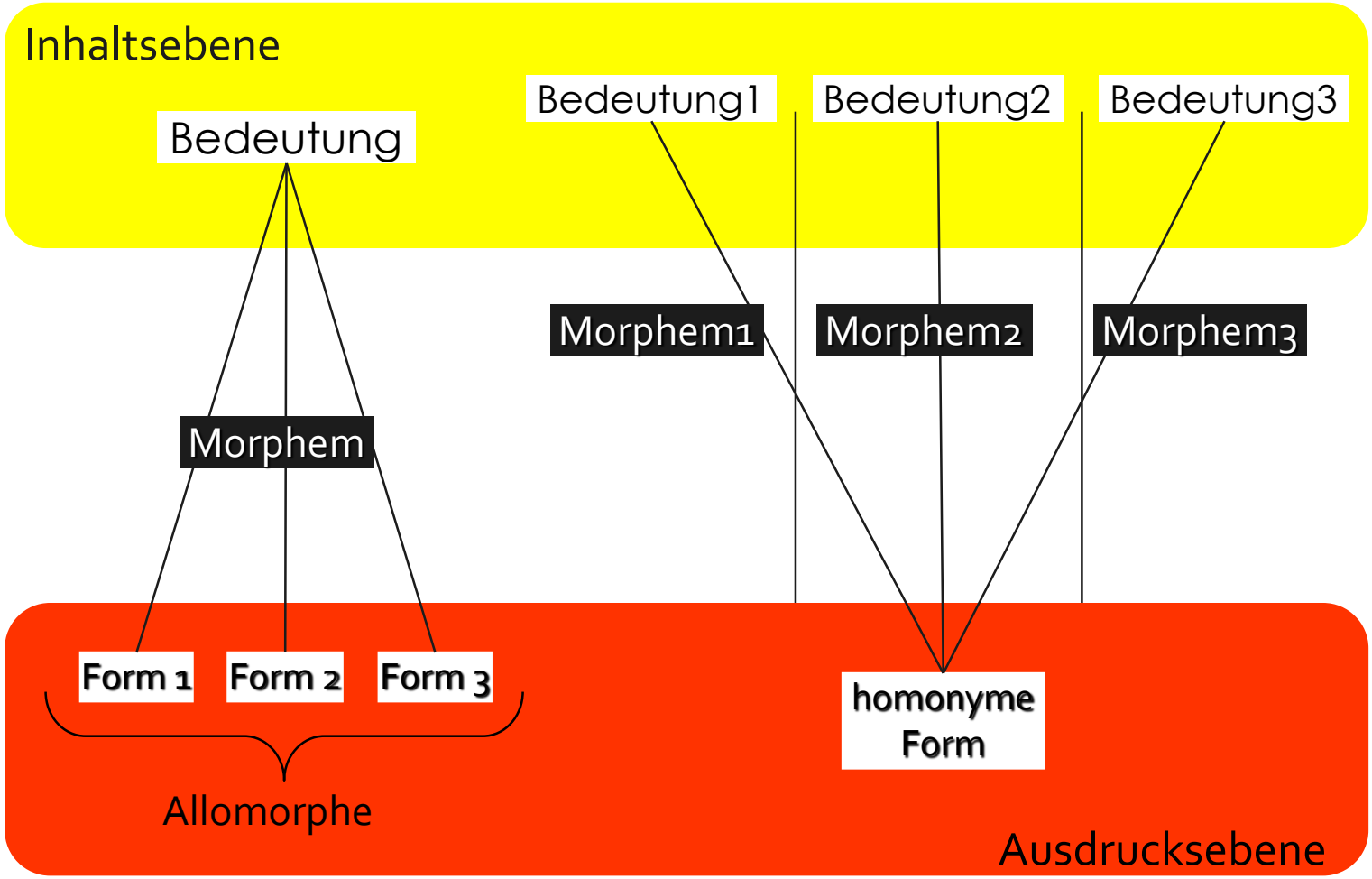
Mengen homonymer und synonymer Minimalzeichen sowie deren Eigenschaften



Zur Erinnerung:

- Jedes sprachliche Zeichen ist **bilateral**, also durch Form und Bedeutung definiert.
- Die Form ist immer **materiell**, also bei der Sprache akustisch bzw. graphisch gegeben.
- Die **Bedeutung** lässt sich nur **über die Form** erschließen.
- Elementare sprachliche Zeichen heißen **Minimalzeichen**.

Die Beziehung zwischen Form und Bedeutung bei Allomorphen und homonymen Morphen



Morphemtypen

- **freies** Morphem
  - → Morphem mit mindestens einem Allomorph, das ohne zusätzliche Morpheme frei auftreten kann
    - wortfähig – vgl. {haus} , {tür}

- **gebundenes** Morphem
  - → Morphem, von dem kein Allomorph ohne zusätzliche Morpheme auftreten kann
    - nicht allein wortfähig – vgl. {ig} , Tür{en}

- **diskontinuierliches** Morphem
  - → Morphem, dessen Allomorphe aus mindestens zwei separaten Teilen bestehen können
    - die Teile werden durch nicht zu diesem Morphem gehörende Segmente linear von einander getrennt – vgl. {ge}sag{t}

Morphemtypen

- **lexikalisches** Morphem → Wurzel
  - entspricht einem Lexem
  - {les} wie in „hat gelesen", {haus} wie in „das schöne Haus"

- **grammatisches** Morphem:
  - hat rein grammatische Funktion
  - in „hat gesagt" ist {hat} ein freies grammatisches Morphem, und {ge- -t} ein gebundenes grammatisches Morphem, das dazu diskontinuierlich ist

- **Flexionsmorphem**: gebundenes grammatisches Morphem.

- **Derivationsmorphem**: gebundenes Morphem, das zur Bildung neuer Lexeme dient
  - {ung} und {heit} dienen zur Bildung der Lexeme „Achtung" und „Menschheit" (aus den Wurzeln {acht} und {mensch}).

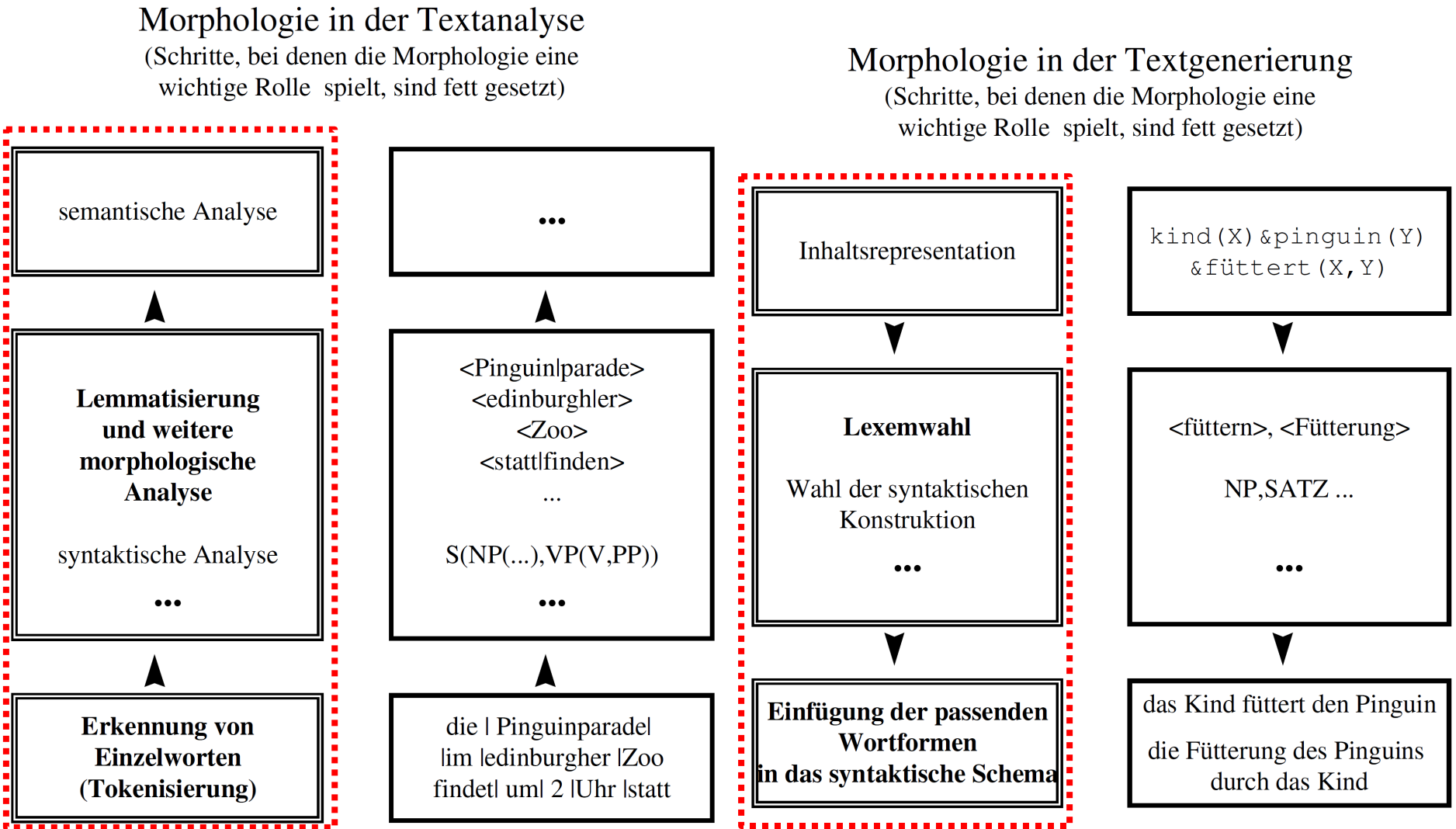
Morphemverbindungen

- **Konkatenation** (von Morphen)
  - geh + st → gehst
  - ab + ge + frag + t + e → abgefragte
  - An den Verbindungsstellen finden oft phonologische Prozesse statt.
    - **Elision**: handel + ung → Handlung; ras + st → rast
    - **Epenthese**: hat + t → hattet; bad + st → badest
    - meistens als Resultat von sprachökonomischen Vereinfachungen, die die Aussprache des zusammengesetzten Wortes erleichtern
- **Nichtkonkatenative Phänomene**: Veränderung des Stammvokals
  - **Umlaut** (z.B. Pluralbildung) Mutter – Mütter
  - **Ablaut** (z.B. Tempusmarkierung) geb – gab

Fragen?

Rolle der Morphologie in Computerlinguistik

Die morphologische Analyse bzw. korrekte Generierung von abgeleiteten Wörtern und von Wortformen ist für fast alle Anwendungen eine notwendige Voraussetzung.



Problemstellungen für Computerlinguistik (1/2)

- Regularitäten erkennen und kodieren
  - Formenregularitäten – Wörter, die derselben morphologischen Klasse angehören bilden Formen in derselben Weise;
  - Ausnahmen müssen berücksichtigt werden, um nicht falsche Wortformen zu tolerieren, unnötige Ambiguitäten zu vermeiden und um in bei der Generierung nicht falsche Formen zu erzeugen.

- Kompositasegmentierung
  - Korrekte Kodierung von Regularitäten der Kompositabildung kann verwendet werden, um Segmentierungsambiguitäten aufzulösen.
  - Allerdings ist dies nicht immer trivial, manchmal ohne Kontextanalyse auch ganz unmöglich (vgl. *Wachstube* – *Wach*|*stube*, *Wachs*|*tube*)

Problemstellungen für Computerlinguistik (2/2)

- Bedeutungsregularitäten in der Wortbildung
  - Bedeutungsregularitäten können z.B. verwendet werden, um semantische Merkmale automatisch zur vergeben.

König <HERRSCHER> - Königin <HERRSCHER&FEM>
Metzger <BERUF> - Metzgerin <BERUF&FEM>

Das Suffix *-in* tritt in systematischer Weise zu maskulinen Menschenbezeichnern - das resultierende Nomen referiert auf weibliche Menschen.

- Ambige Wortformen
  - Zahlreiche Wortformen sind morphologisch ambig z.B. bezüglich Kasus (*Frau*), Numerus (*Treffen*), Wortart (*Treffen*).
  - Die Disambiguierung kann durch syntaktische, semantische oder pragmatische Methoden vorgenommen werden.