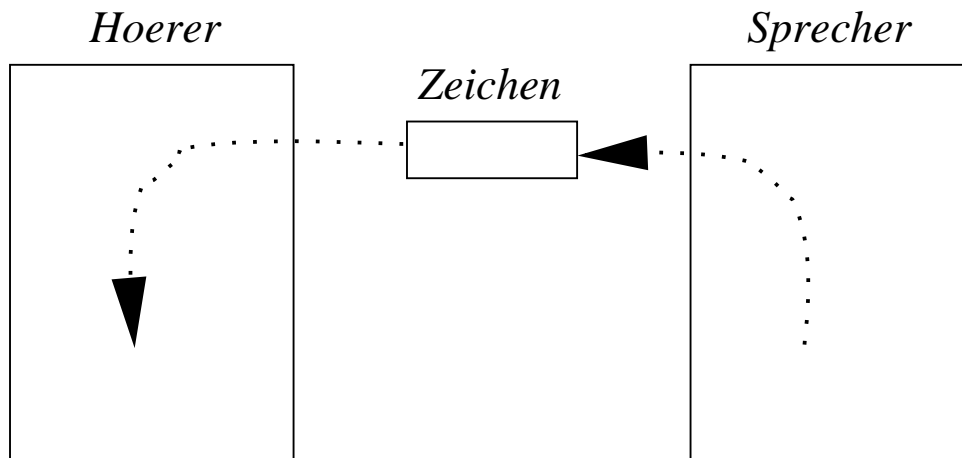


13. Wörter und Morpheme

Platz der Morphologie im Gesamtsystem

Verhältnis von Sprecher, Hörer und Zeichenoberfläche



Input and output devices of sign recognition and production

Sign recognition is based on the input devices of

the *ears* (spoken language),

eyes (written, signed language), and

skin (braille).

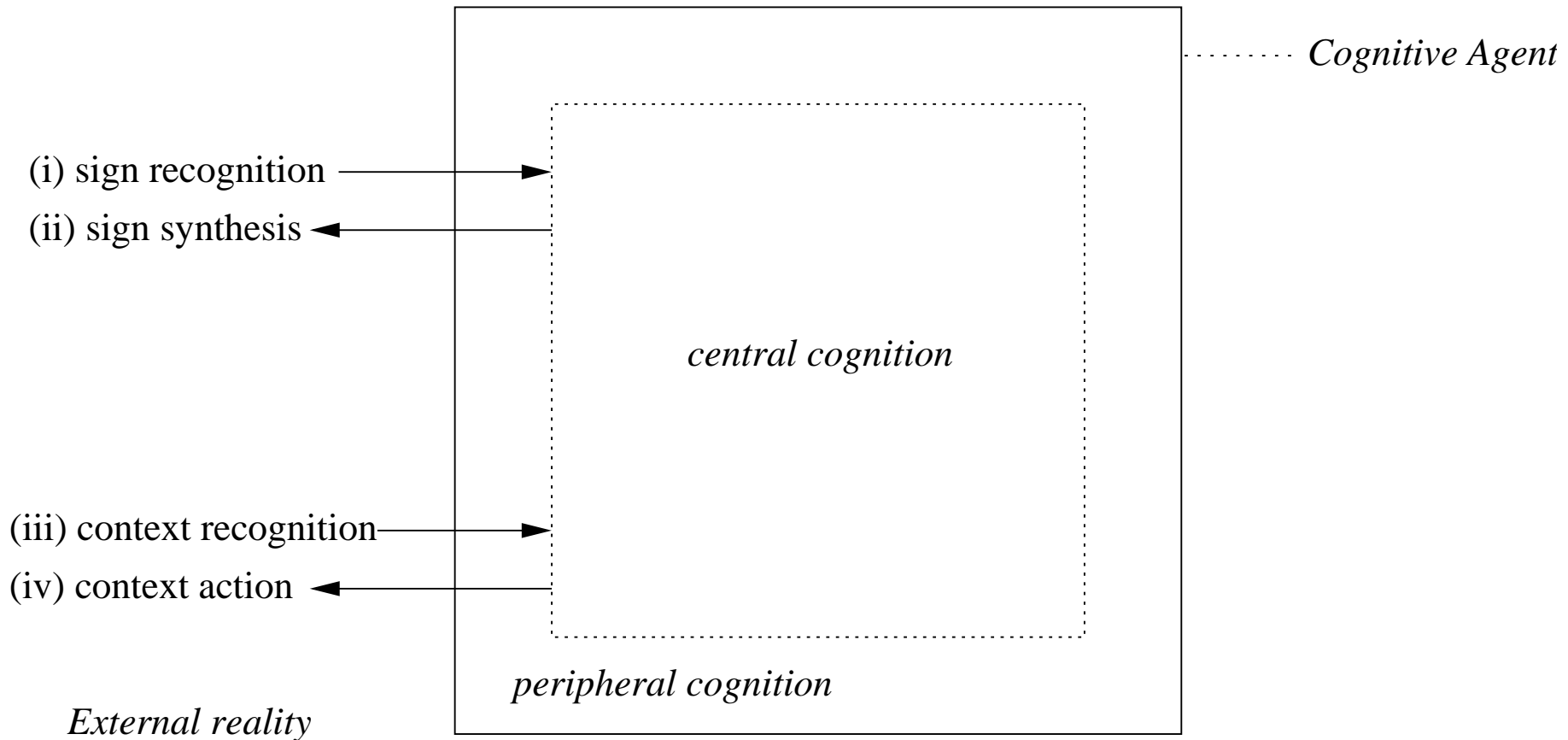
Sign production is based on the output devices of

the *vocal tracts* in combination with the mouth (spoken language),

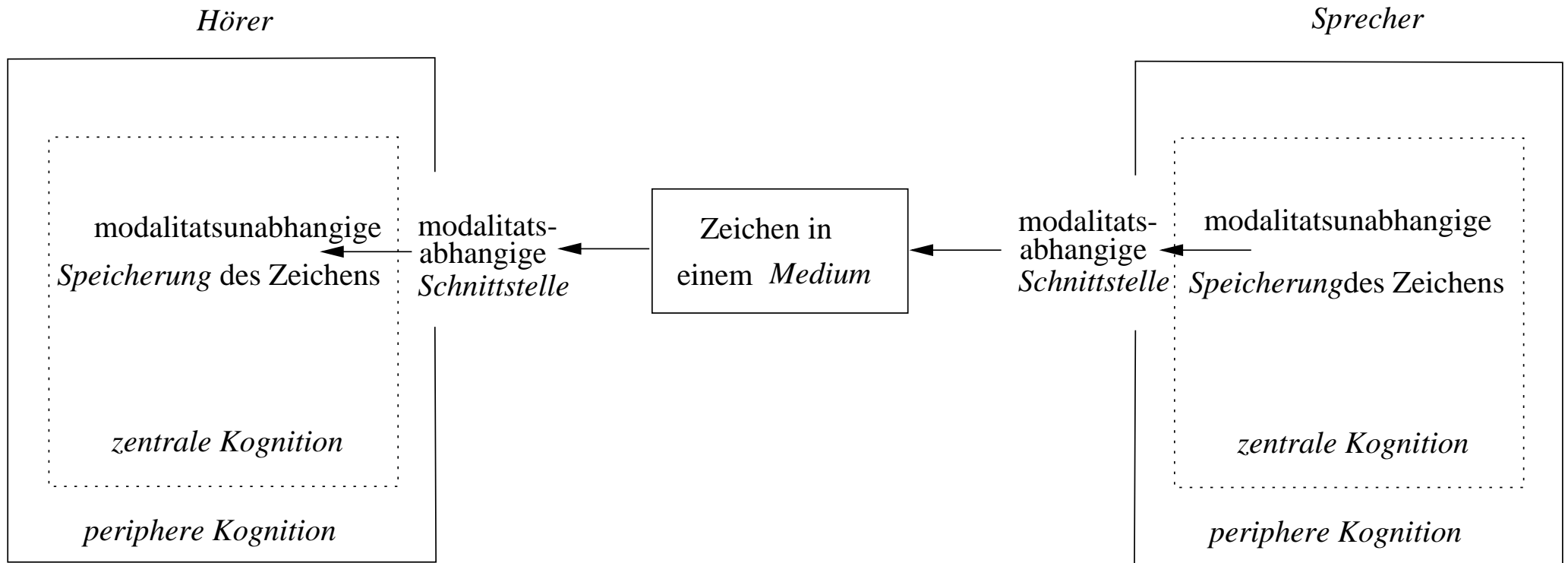
the *hands* (written language, including braille), and

hand-arm-face gestures (signed language).

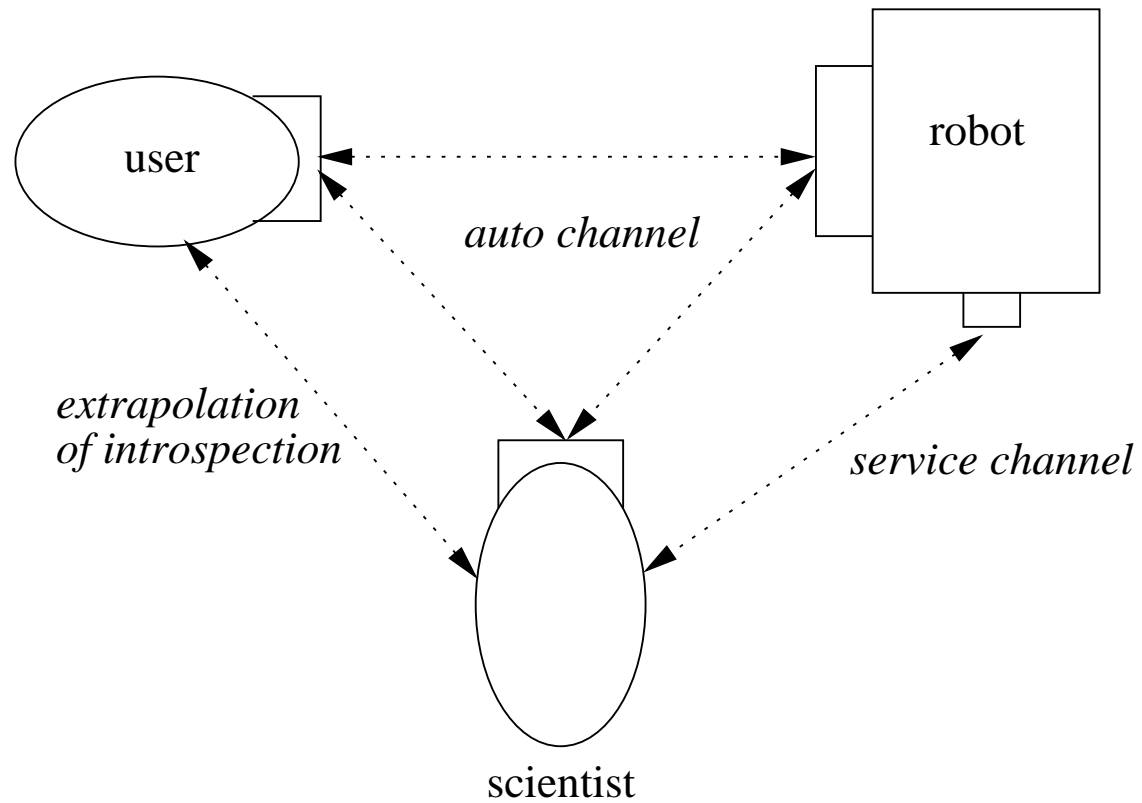
Interfaces of a cognitive agent



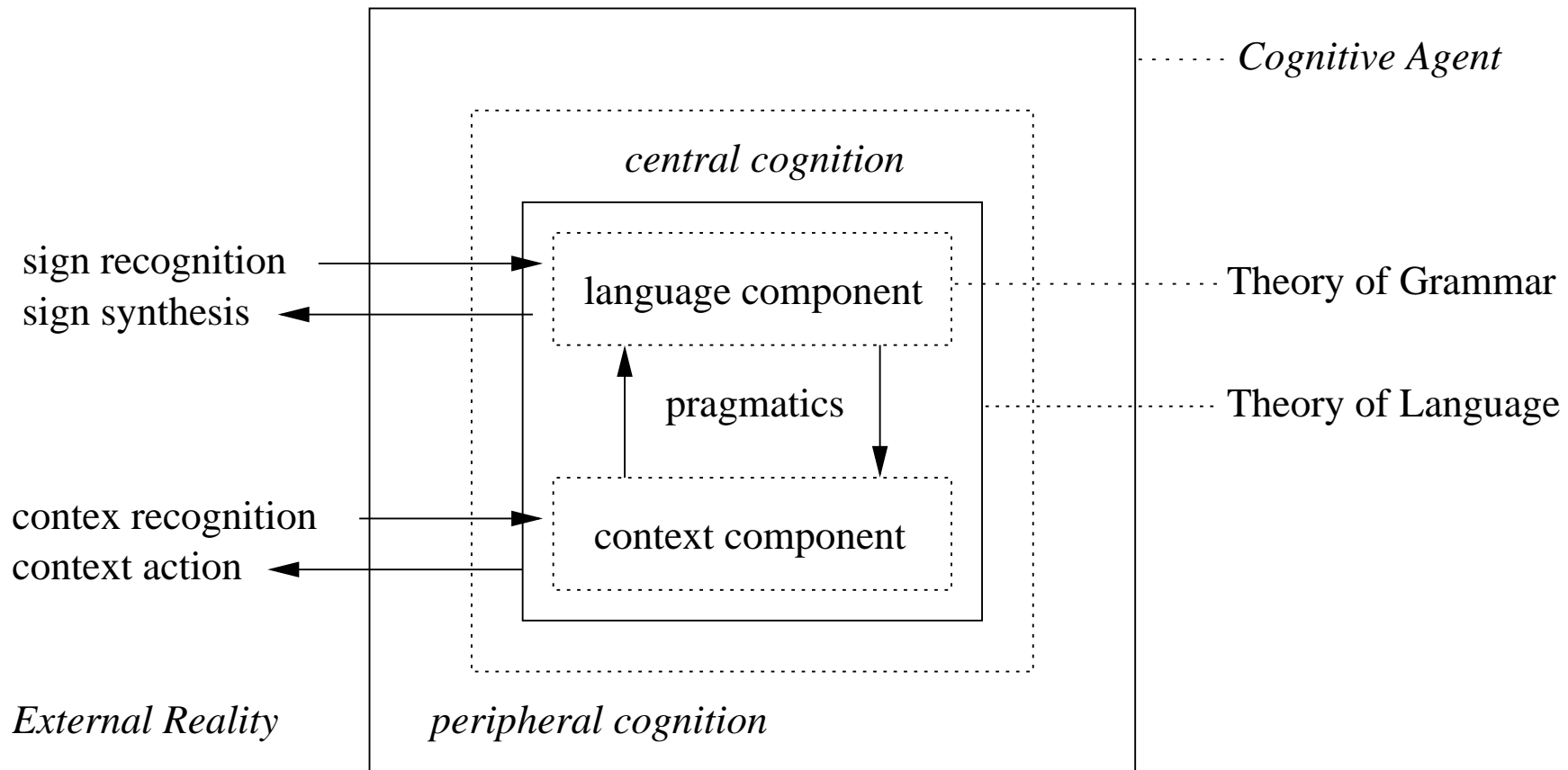
Storage, interface, medium and sign in communication



Constellation for modelling human-machine communication



Structuring central cognition in agents with language



Die Komponenten der Grammatik

- *Phonologie*: Lehre von den Sprachlauten
- *Morphologie*: Lehre von den Wortformen
- *Lexikon*: Auflistung der Wörter
- *Syntax*: Lehre von der Komposition der Wortformen
- *Semantik*: Lehre von den wörtlichen Bedeutungen
- *Pragmatik*: Lehre von den Verwendungen

13.1 Wörter und ihre Wortformen

13.1.1 Unterschiedliche Kompatibilität der Wortformen

*Buch

*Buches

Die Bücher gefallen Bärbel.

*Büchern

13.1.2 Francis' & Kučerás Definition eines graphematischen Wortes (*graphic word*) (1982)

Ein Wort ist eine zusammenhängende alphanumerische Zeichenkette, die auf beiden Seiten von Leerzeichen begrenzt ist. Die Zeichenkette kann Bindestriche und ein Apostroph aber keine anderen Interpunktionszeichen enthalten.

“A word is a string of continuous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks.”

13.1.3 Kombinationsprinzipien der Morphologie

1. *Flexion* die systematische Variation, mit der ein Wort sich an verschiedene syntaktische Umgebungen anpaßt (bzw. verschiedene syntaktische u. semantische Funktionen ausübt)
Beispiele: lern/e, lern/st, lern/t, lern/en, lern/te,
2. *Derivation* Verbindung eines Wortes mit einem Affix (z.B.: -lich, un-) in ein neues Wort; beinhaltet sowohl Suffigierung als auch Präfigierung
Beispiele: un/schön, Schön/heit, er/lernen, klein/lich
3. *Komposition* Verknüpfung zweier oder mehrerer Wörter zu einem neuen Wort.
Beispiele: Stiefel/knecht, Hart/holz, gras/grün, schlitten/fahren, Bet/schwester, spritz/gießen

13.1.4 Definition des Begriffes *Wort*

Wort =_{def} {Zugehörige analysierte Wortformen}

13.1.5 Beispiel einer analysierten Wortform

[Buches (-FG) Buch]

[Bücher (P-D) Buch]

13.1.6 Analyse eines flektierenden Wortes

<i>Wort</i>	<i>Wortformen</i>
Buch = _{def}	{ [Buch (N-G) Buch], [Buches (NG) Buch], [Buche (ND) Buch], [Bücher (P-D) Buch] [Büchern (PD) Buch]}

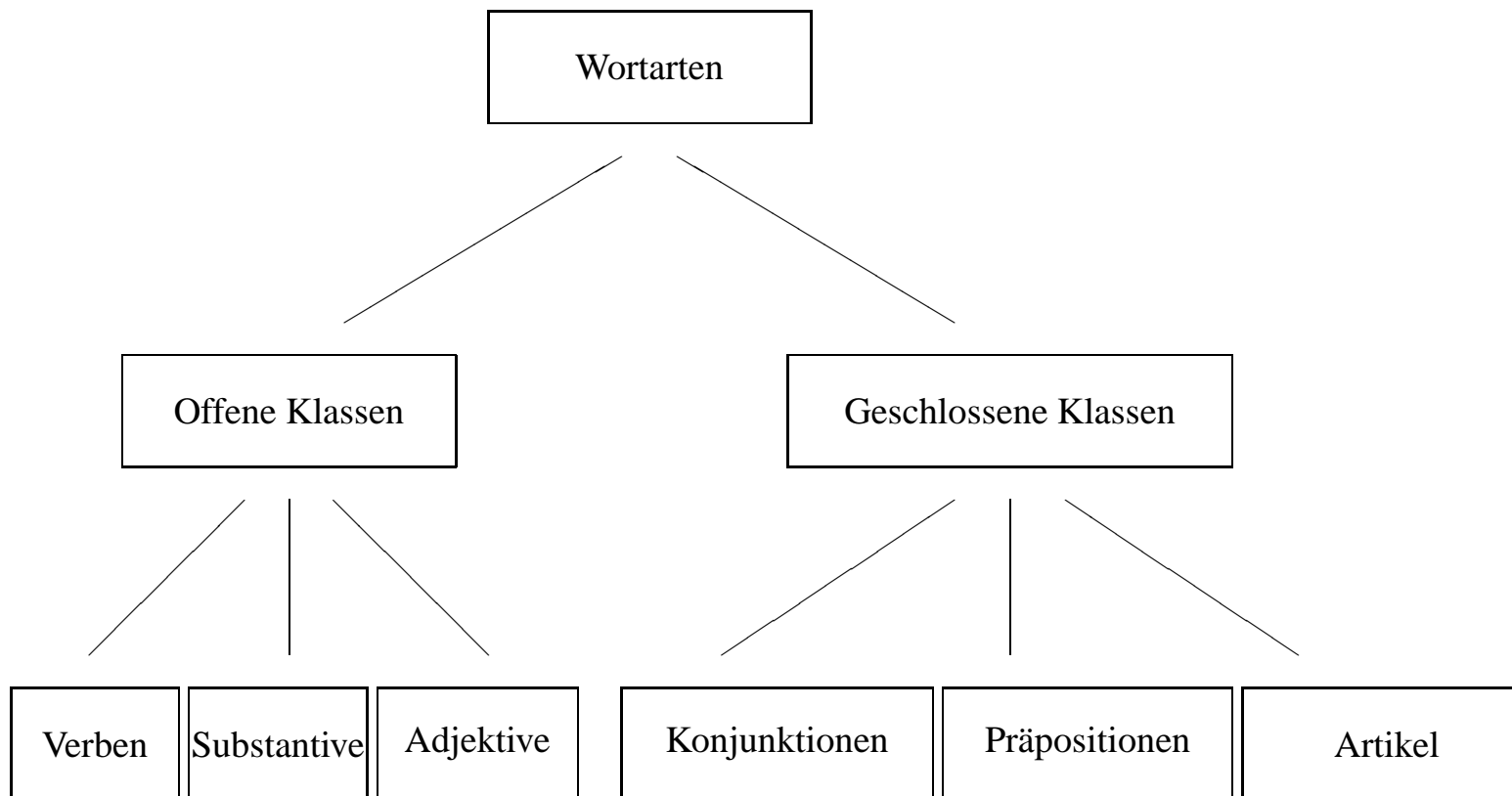
13.1.7 Analyse eines nichtflektierenden Wortes

<i>Wort</i>	<i>Wortformen</i>
und = _{def}	{ [und (knj) und] }

13.1.8 Die verschiedenen Wortarten im Deutschen

- *Verben*, z.B., gehen, lesen, geben, helfen, lehren . . .
- *Substantive*, z.B., Buch, Tisch, Frau, Bote, Boot, Arena . . .
- *Adjektiv-Adverbien*, z.B., schnell, gut, leise . . .
- *Konjunktionen*, z.B., und, oder, weil, nachdem, obwohl . . .
- *Präpositionen*, z.B., in, auf, über, unter, vor . . .
- *Artikel*, z.B., der, die, das, ein, einem . . .
- *Partikel*, z.B., zu, nur, schon, eben. . .

13.1.9 Offene und geschlossene Wortklassen



13.1.10 Vergleich der offenen und geschlossenen Klassen

- Die offenen Klassen umfassen einige 10 000 Elemente, die geschlossenen nur einige hundert.
- Die morphologischen Prozesse der Flexion, Derivation und Komposition sind in den offenen Klassen produktiv.
- In den offenen Klassen verändert sich der Gebrauch der Worte ständig: Neue Worte werden in Gebrauch genommen, alte verschwinden. Die geschlossenen Wortklassen weisen keine vergleichbare Fluktuation auf

13.1.11 Wortklassen und Zeichentypen

Aus semantisch-pragmatischer Sicht werden die Elemente der offenen Wortklassen auch *Inhaltswörter* genannt, während die der geschlossenen Klassen *Funktionswörter* genannt werden.

Bei dieser Definition muß allerdings auch der Zeichentyp berücksichtigt werden. Nur die *Symbole* unter den Substantiven, Verben und Adjektiven sind Inhaltswörter im eigentlichen Sinn. *Indizes* werden dagegen zu den Funktionswörtern gerechnet, auch wenn sie (wie z.B. die Personalpronomina *er*, *sie*, *es* . . .) von der Kategorie Substantiv sind.

Auch der Zeichentyp *Name* nimmt unter den Substantiven eine Sonderstellung ein.

13.2 Segmentierung und Konkatenation

13.2.1 Verhältnis Wörter zu Flexionsformen

	Grundformen	Flexionsformen
Substantive:	23 000	92 000
Verben	6 000	144 000
Adjektiv-Adverbien:	11 000	198 000
<hr/>		
	40 000	434 000

13.2.2 (Theoretische) Zahl von Substantiv-Substantiv-Kompositionen

- Länge 2: n^2
Beispiele Haus/schuh, Schuh/haus, Jäger/jäger.
- 20 000 Substantive \mapsto 400 000 000 mögliche Kompositionen der Länge 2 (Grundformen).
- Länge 3: n^3
Beispiele: Haus/schuh/sohle, Sport/schuh/haus, Jäger/jäger/jäger.
- 20 000 Substantive \mapsto 8 000 000 000 000 (acht Milliarden) mögliche Kompositionen der Länge 3 (Grundformen).

13.2.3 Mögliche Wörter, Gebräuchliche Wörter, Neologismen

- Mögliche Wörter

Es gibt keine grammatische Obergrenze für die Länge von Substantiv-Kompositionen. Das heißt, die Zahl der möglichen Wortformen im Deutschen ist prinzipiell unendlich. Die meisten dieser Wörter existieren nur potentiell aufgrund der Produktivität der morphologischen Wortbildungsprozesse.

- Gebräuchliche Wörter

Die Menge der Wörter, die innerhalb eines bestimmten Zeitabschnitts von der Sprachgemeinschaft verwendet werden, ist endlich.

- Neologismen

Neologismen werden von Sprachteilnehmern auf der Grundlage bekannter Wörter und morphologischer Regeln spontan hervorgebracht.

13.2.4 Beispiel für Neologismen (Spiegel-Artikel v. 18.10.93, S. 256 - 261)

Genitivmetaphern	radikalbiographisch
Tiefenplumpheit	Zwangskontinua
Hyperpolitik	Psychofinalismen
neubürgerlich	Fundamentalpositivismus
Großweltrisiko	psychonautisch

13.2.5 Definition des Begriffs *Morphem*

Morphem =_{def} {zugehörige analysierte Allomorphe}

13.2.6 Das regelmäßige Morphem: lern

<i>Morphem</i>	<i>Allomorphe</i>
lern = _{def}	{[lern (N ... V) lernen]}

13.2.7 Das unregelmäßige Morphem: sprach

<i>Morphem</i>	<i>Allomorphe</i>
sprech = _{def}	{[sprech (N ... V1) sprechen], [sprich (N ... V2) sprechen] [sprach (S13 ... V) sprechen_i] [spräch (S13... V) sprechen_k2] [sproch (N... V) sprechen]}

13.2.8 Vergleich von Morphem und Wort: Buch

<i>Morphem</i>	<i>Allomorphe</i>	<i>Wort</i>	<i>Wortformen</i>
Buch = _{def}	{Buch, Büch}	Buch = _{def}	{Buch, Buch/es} Buch/e, Büch/er, Büch/er/n}

13.2.9 Unterschiedliche Zerlegungen einer Wortform

Allomorphe:	lern/en
Silben:	ler/nen
Phoneme:	l/e/r/n/e/n
Buchstaben:	l/e/r/n/e/n

13.3 Morpheme und Allomorphe

13.3.1 Das regelmäßige Morphem: lern

<i>Morphem</i>	<i>Allomorphe</i>
lern = _{def}	{[lern (N ... V) lernen]}

13.3.2 Das unregelmäßige Morphem: sprach

<i>Morphem</i>	<i>Allomorphe</i>
sprech = _{def}	{[sprech (N ... V1) sprechen], [sprich (N ... V2) sprechen] [sprach (S13 ... V) sprechen_i] [spräch (S13... V) sprechen_k2] [sproch (N... V) sprechen]}

13.3.3 Beispiel für Suppletion: gut

<i>Morphem</i>	<i>Allomorphe</i>
gut = _{def}	{[gut (ADV POS IR) gut], [bess (ADV KOMP IR) gut] [be (ADV SUP IR) gut]}

13.3.4 Beispiel für ein gebundenes Morphem (hypothetisch): -en (Pluralmorphem)

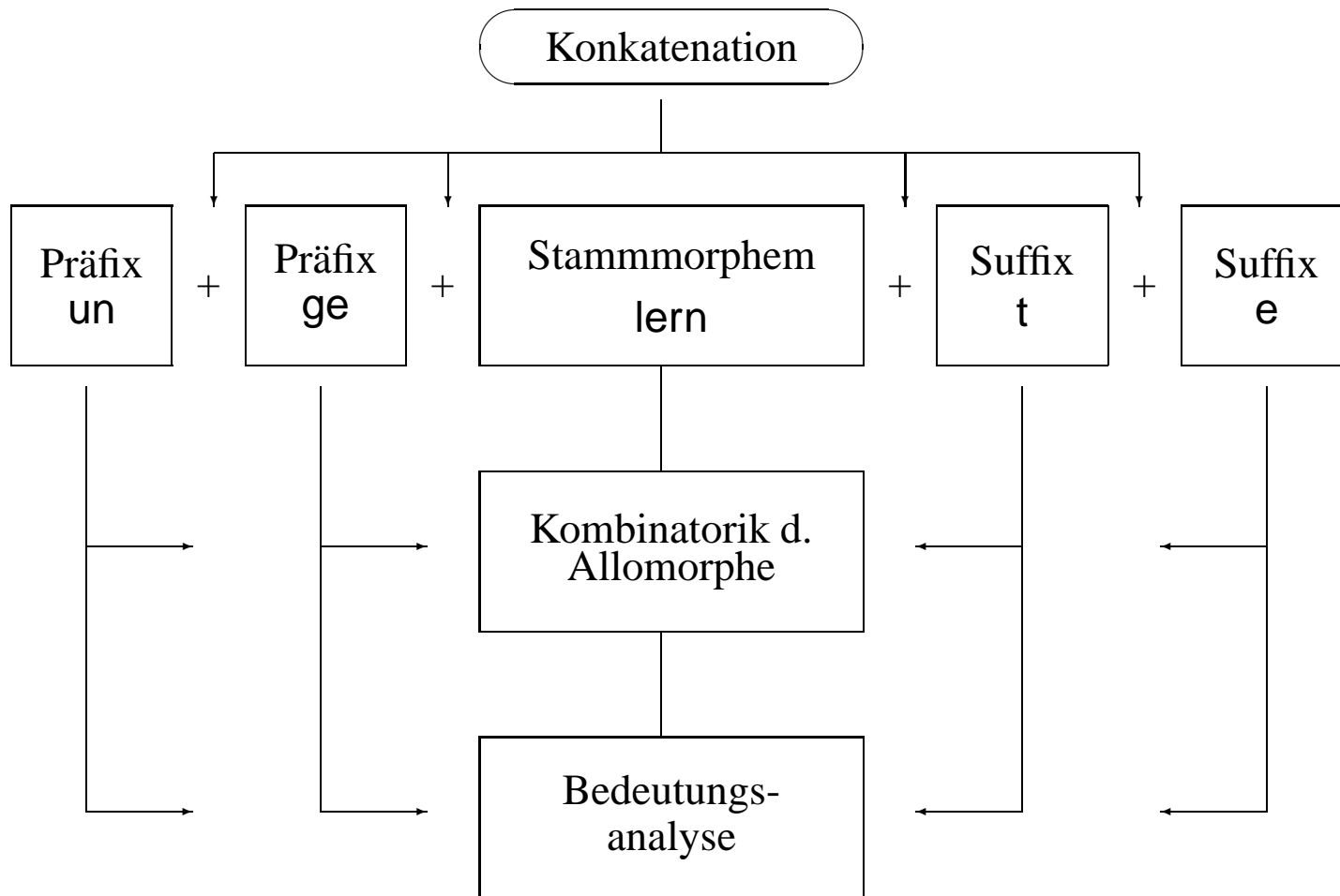
<i>Morphem</i>	<i>Allomorphe</i>
-en = _{def}	{[en (PL1) Plural], [er (PL2) Plural] [e (PL3) Plural] [n (PL4) Plural] [se (PL5) Plural] [s (PL6) Plural] *[# (PL7) Plural]}

13.4 Kategorisierung und Lemmatisierung

13.4.1 Schritte der morphologischen Analyse

- Segmentierung:
Die Oberfläche wird in ihre elementaren Bestandteile zerlegt
- Allomorph-Lexikon:
Abruf der zu den Segmenten gehörenden Definitionen mit den grammatischen Eigenschaften aus dem Lexikon (Lexical Lookup)
- Konkatenation:
Regelbasierte Zusammensetzung der analysierten Bestandteile (betrifft gleichzeitig die Oberfläche, die morphosyntaktische Kategorie und die semantische Repräsentation)

13.4.2 Morphologische Analyse von ungelernete



13.4.3 Schematische Darstellung einer LAG-Ableitung von ungelernete

(“un” (CAT1) MEAN-a) + (“ge” (CAT2) MEAN-b)
 (“un/ge” (CAT3) MEAN-c) + (“lern” (CAT4) MEAN-d)
 (“un/ge/lern” (CAT5) MEAN-e) + (“t” (CAT6) MEAN-f)
 (“un/ge/lern/t” (CAT7) MEAN-g) + (“e” (CAT8) MEAN-h)
 (“un/ge/lern/t/e” (CAT9) MEAN-i)

13.4.4 Komponenten der automatischen Wortformerkennung

- *Online-Lexikon* Für jedes Element (z.B. Morphem) der natürlichen Sprache muß eine lexikalische Analyse definiert und elektronisch gespeichert werden.
- *Erkennungsalgorithmus* Mit Hilfe des Online-Lexikons muß jede *Wortform* (z.B. Büchern) (segmentiert,) kategorisiert und lemmatisiert werden:
 - Die *Kategorisierung*
ist die Zuweisung einer Wortklasse (z.B. Substantiv) und der spezifischen morphosyntaktischen Eigenschaften (z.B. Dativ Plural), die für die syntaktische Analyse benötigt werden.
 - Die *Lemmatisierung*
ist die Zuweisung der Grundform (z.B. Buch), wodurch der Zugriff auf entsprechende Einträge in einem semantischen Lexikon ermöglicht wird.

13.4.5 Grundstruktur eines Lemmas

[Oberfläche (Lexikalische Beschreibung)]

13.4.6 Lemma eines traditionelle Wörterbuchs (*Ausschnitt*)

Buch <n.12 u> **1** <urspr.> zusammengebundene Täfelchen aus Buchenholz zum Schreiben; <dann> zusammengeheftete od. eingebundene, beschriebene od. bedruckte, oft illustrierte Papierbogen; größeres Druckwerk; ein Satz Spielkarten; Teil eines größeren Schriftwerkes, z.B. der Bibel; <Kaufmannsspr.; oft Pl.> Geschäftsbuch ...

(aus: Wahrig, "Wörterbuch der Deutschen Sprache")

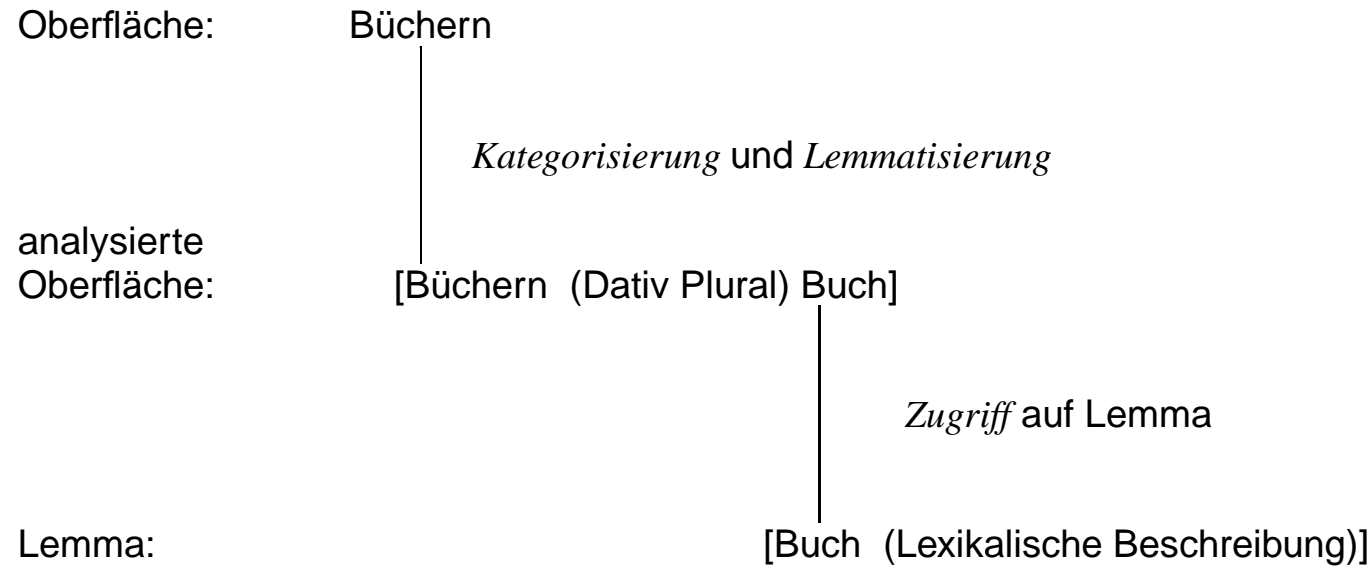
13.4.7 Abgleichung einer Oberfläche auf einen Schlüssel

Oberfläche der Wortform: Buch

| *Abgleichung*

Lemma: [Buch (Lexikalische Beschreibung)]

13.4.8 Zweischnittverfahren bei der Wortformerkennung



13.4.9 Gründe für das Zweischnittverfahren:

In natürlicher Sprache

- ist die Zahl der Wortformen beträchtlich höher, als die Zahl der Worte (zumindest in flektierenden und agglutinierenden Sprachen)
- Lemmata bestehender Lexika definieren in der Regel Worte und keine Wortformen

13.5 Methoden der automatischen Wortformerkennung

13.5.1 Wortform-Methode (Vollformansatz)

Basiert auf einem Lexikon analysierter Wortformen

13.5.2 Analytierte Wortform als lexikalisches Lemma

[Hirtenhunden (WORTART: Substantiv, NUMERUS: Plural, KASUS: Dativ, GRUNDFORM: Hirtenhund)]

Hier wird die Kategorisierung und Lemmatisierung nicht durch Regeln vorgenommen. Sie basiert vielmehr allein auf dem lexikalischen Eintrag.

13.5.3 Vorteile und Nachteile der Wortform-Methode

- Vorteil

Der simpelste Erkennungs-Algorithmus der drei Methoden: Die Oberfläche einer unbekanntem Wortform (z.B. Hirtenhunden) wird als ganzes auf den entsprechenden Schlüssel des Analyse-Lexikons abgepaßt.

- Nachteil

Die Erstellung des Analyse-Lexikons ist extrem aufwendig. Das resultierende Lexikon ist sehr groß und es gibt mit dieser Methode keine Möglichkeit Neologismen zu erkennen.

13.5.4 Die Morphem-Methode

Basiert auf einem Lexikon analysierter Morpheme

13.5.5 Schema der Morphem-Methode

Oberfläche:	Bücher	
		<i>Segmentierung</i>
Allomorphe:	Büch/er	
	↓ ↓	<i>Reduktion</i>
Morpheme:	Buch + er	<i>Grundform-Lookup und Konkatenation</i>

- (1) Segmentierung in Allomorphe,
- (2) Reduktion von Allomorphen in Morpheme,
- (3) *Lookup* – Identifizieren der Morpheme im Analyse-Lexikon und
- (4) regelbasierte Konkatenation der Morpheme zur Ableitung der morphosyntaktischen Kategorie der Wortform.

13.5.6 Vorteile und Nachteile der Morphem-Methode

- Vorteile

Das kleinste Lexikon der drei Methoden. Neologismen können zur Laufzeit erkannt und analysiert werden (durch regelbasierte Segmentierung und Konkatenation der analysierten Morpheme).

- Nachteil

Ein maximal komplexer Erkennungsalgorithmus (\mathcal{NP} -vollständig).

13.5.7 Allomorph-Methode

Basiert auf einem Lexikon elementarer Grundformen, aus dem vor der Laufzeit regelbasiert ein Lexikon analysierter Allomorphe generiert wird (Allomorphlexikon).

13.5.8 Schema der Allomorph-Methode

Oberfläche:	Bücher	
		<i>Segmentierung</i>
Allomorphe:	Büch/er	<i>Allomorph-Lookup und Konkatenation</i>
	↑ ↑	<i>Ableitung der Allomorphe vor der Laufzeit</i>
Morpheme:	Buch er	

Während der Laufzeit stellt das Analyselexikon sämtliche Allomorphe als vorberechnete analysierte Formen zur Verfügung. Sie bilden die Grundlage einer simplen Segmentierung, indem auf die unbekanntes Oberflächen von links nach rechts die passenden Allomorphe abgepaßt werden.

13.5.9 Vergleich der drei Methoden der Wortformerkennung

