

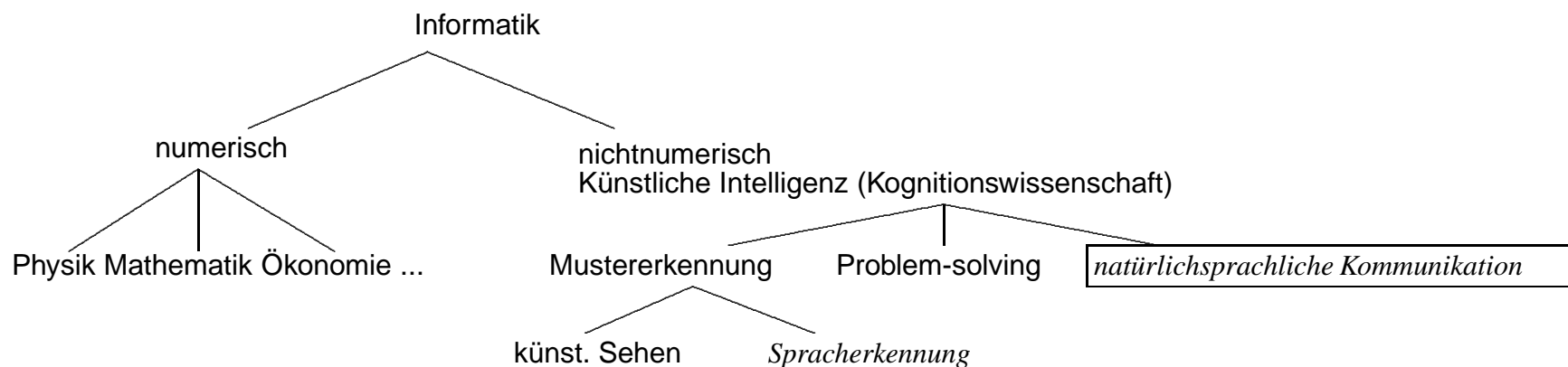
1. Computerbasierte Sprachanalyse

1.1 Mensch-Maschine-Kommunikation

1.1.1 Was ist Computerlinguistik?

Zwei Antworten

- Linguistische Analyse (Philologie) mit Computern
 - Electronische Speicherung großer Mengen realer Daten, z. B. Korpora und Lexika, mit leistungsfähigen Möglichkeiten der Suche und der Reorganisation (verbesserte Datensammlung).
 - Automatisches Testen linguistischer Analysen in Morphologie, Syntax und Semantik an großen Mengen natürlicher Daten (neue Verifikationsmethode)
- Teil der künstlichen Intelligenz, der sich mit natürlichsprachlicher Kommunikation befaßt



Theoretisches Ziel: ein Computermodell der natürlichsprachlichen Kommunikation (wissenschaftliche Erkenntnis).

Praktisches Ziel: Mensch-Maschinenkommunikation in natürlicher Sprachen (maximale Benutzerfreundlichkeit)

1.1.2 Restringsierte vs. nichtrestrictierte Kommunikation

1.1.3 Beispiel restringierter Kommunikation: eine Datensatzstruktur

	Nachname	Vorname	Ort	...
A1	Schmidt	Peter	Bamberg	...
A2	Meyer	Susanne	Nürnberg	...
A3	Sanders	Reinhard	Schwabach	...
	:	:	:	

1.1.4 Anfrage an die Datenbank

Query:

```
select A#
where city = 'Schwabach'
```

Result:

A3 Sanders Reinhard

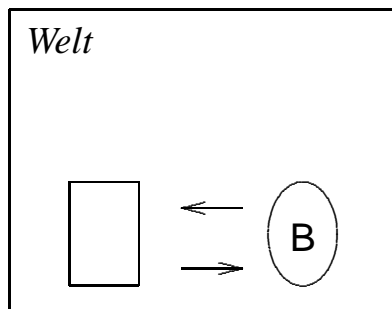
1.1.5 Klassische vs nouvelle KI

Die klassische KI analysiert intelligentes Verhalten als die Manipulation abstrakter Symbole. Ein typisches Beispiel ist ein Schachprogramm. Es arbeitet in Isolation vom Rest der Welt mit fest definierten Figuren auf einem vorgegebenen Brett. Der Suchraum für eine dynamische Gewinnstrategie ist beim Schachspiel zwar astronomisch. Da es sich jedoch um eine abgeschlossene Welt handelt, reichen die technologischen Voraussetzungen eines Standardcomputers aus.

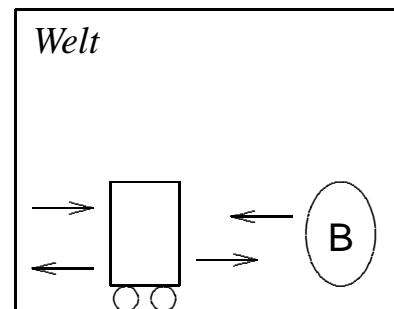
Das Ziel der nouvelle KI ist dagegen die Entwicklung selbständig agierender Roboter (*autonomous agents*). Da sich die Umwelt ständig in unvorhersehbarer Weise ändern kann, muß das System sie mit Hilfe von Sensoren kontinuierlich beobachten. Die Strategie der *task level decomposition* zerlegt den Komplex von Kognition und Verhalten in viele kleinere, problemgerechte Teilaufgaben, wobei die Folgerungsmethoden direkt auf den lokalen Wahrnehmungsdaten aufsetzen.

1.1.6 Drei Typen der Mensch-Maschine-Kommunikation

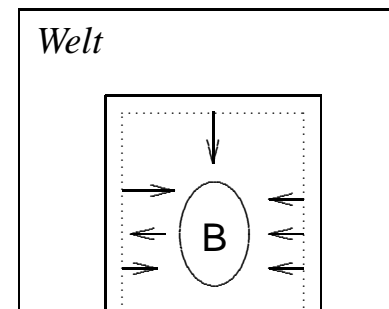
Standardcomputer



autonomer Roboter

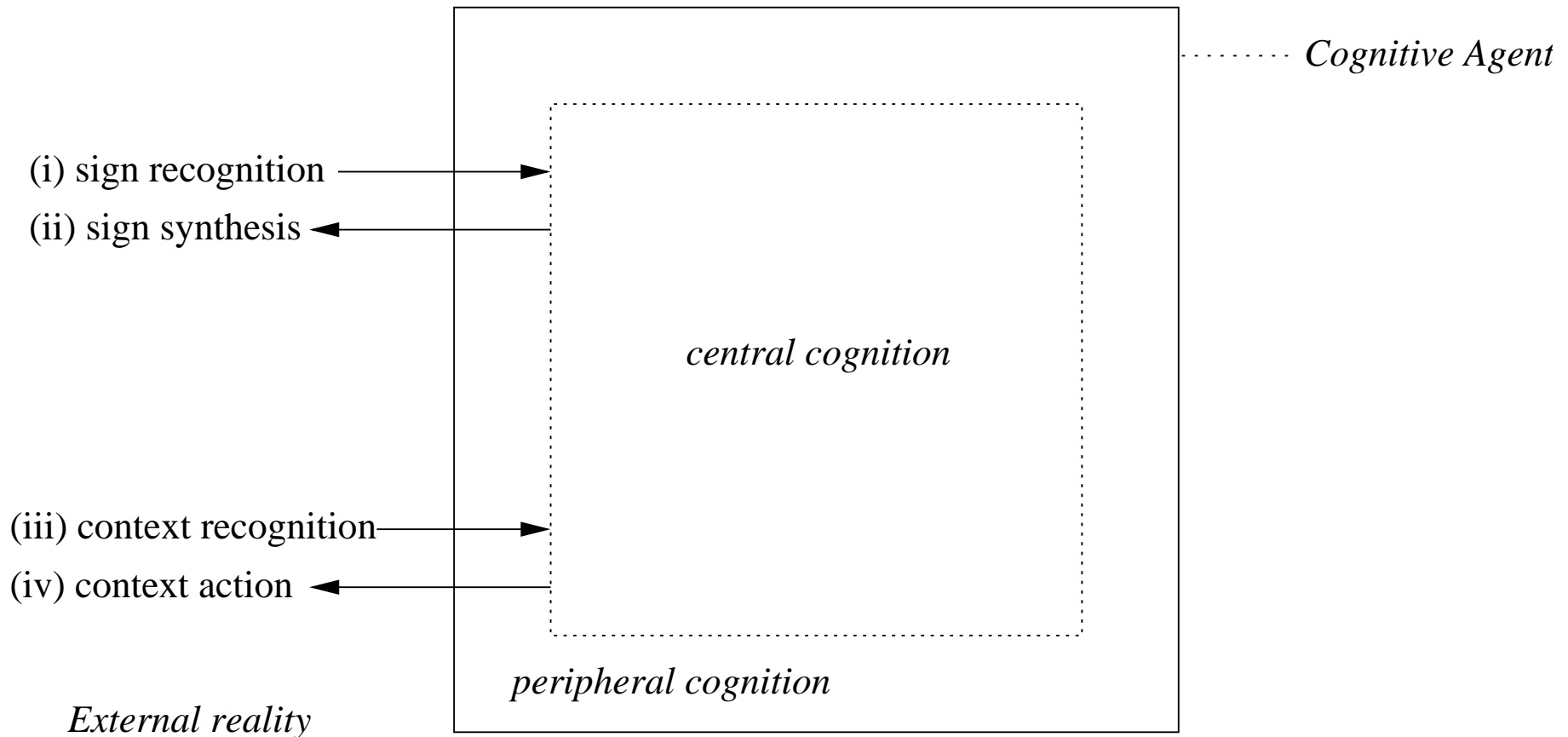


virtuelle Realität

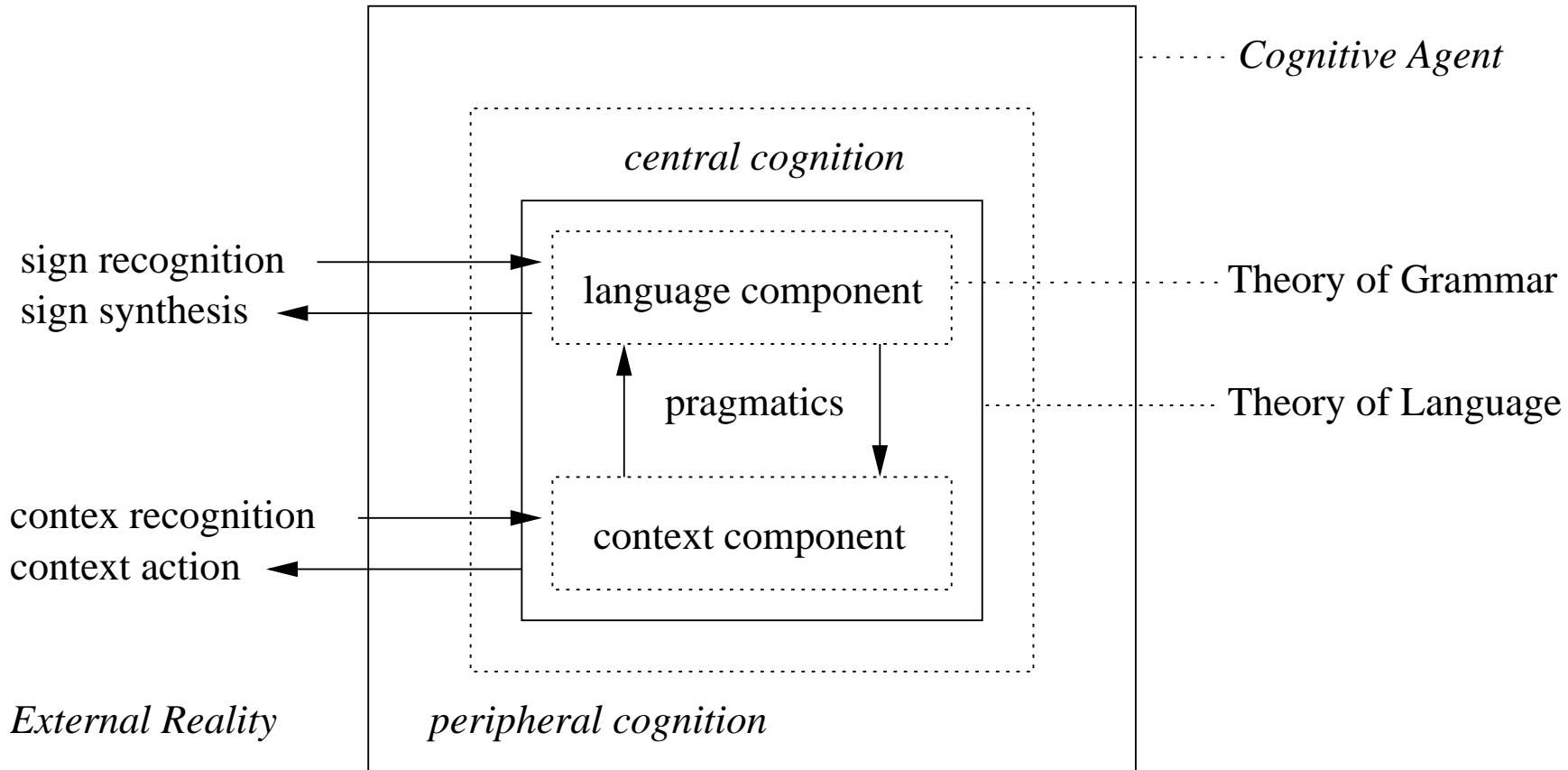


1.2 Schnittstellenbasierte Entwicklung einer Sprachtheorie

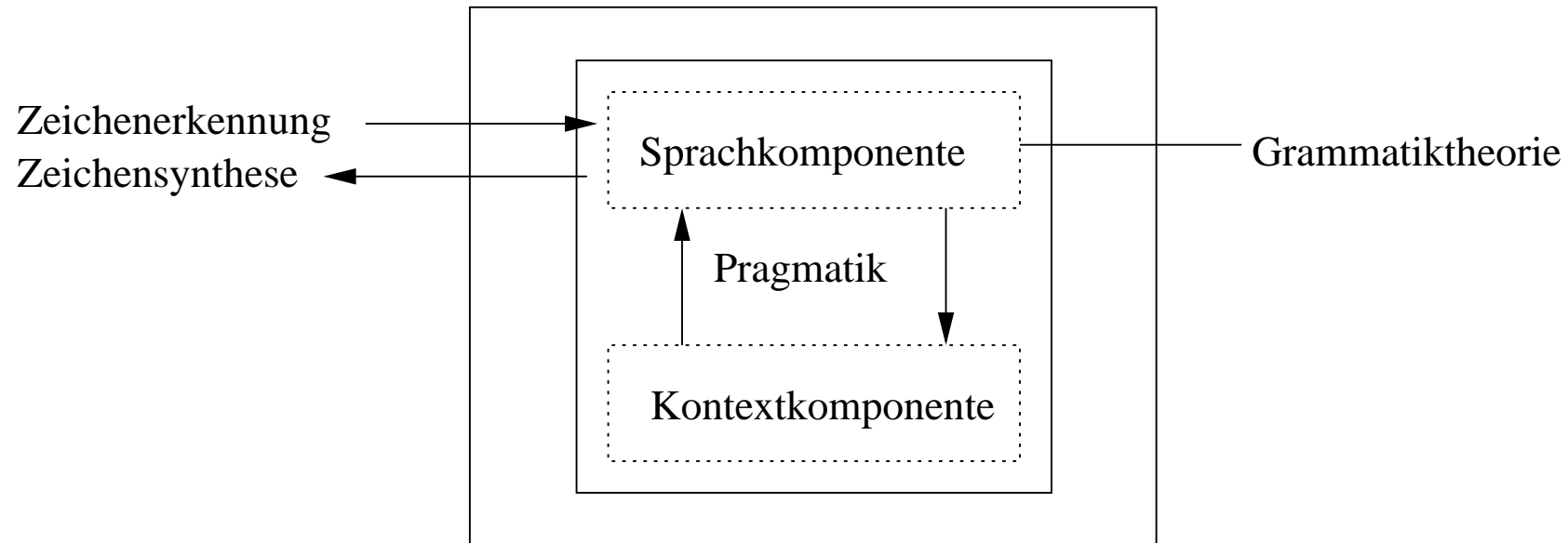
1.2.1 Schnittstellen (interfaces) eines kognitiven agenten



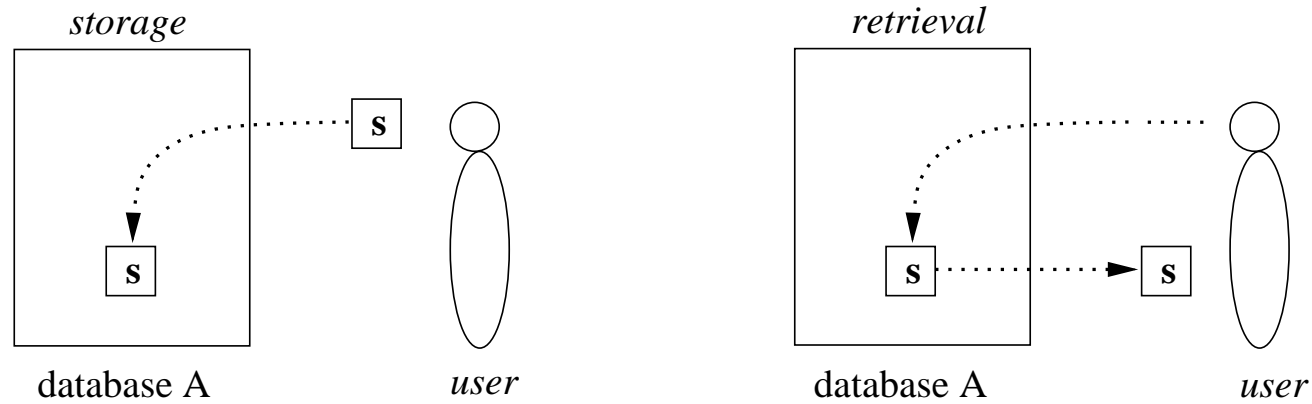
1.2.2 Sprachtheorie als Teil der zentralen Kognition



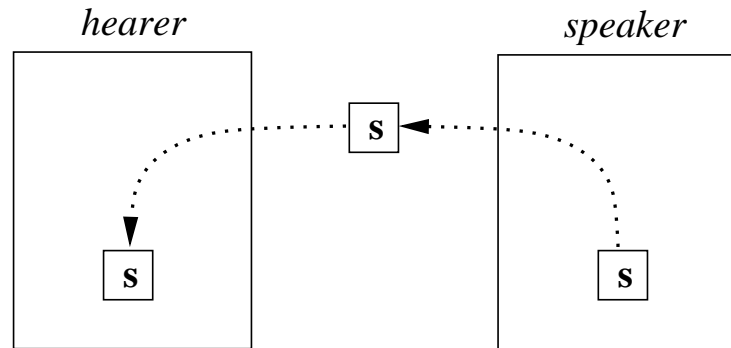
1.2.3 Komponenten einer vereinfachten Sprachtheorie



1.2.4 Interaktion mit einer konventionellen Datenbank



1.2.5 Interaction between speaker and hearer



Datenbankmetapher erfolgreicher Kommunikation

1.3 Sprachwissenschaft und ihre Komponenten

1.3.1 Varianten der Sprachwissenschaft

- *Traditionelle Grammatik*
- *Theoretische Linguistik*
- *Computerlinguistik*

1.3.2 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

1.4 Methoden und Anwendungen der Computerlinguistik

1.4.1 Methodologie des Parsing

- *Zerlegung* eines komplexen Zeichens in seine elementaren Bestandteile,
- *Klassifikation* der gefundenen Teile mit Hilfe des Lexikons, und
- *Zusammensetzen* der klassifizierten Teile zur Ableitung einer grammatischen Gesamtanalyse des komplexen Zeichens.

1.4.2 Praktische Aufgaben der Computerlinguistik

- Indizieren von und Abruf aus textuellen Datenbanken
- Maschinelle Übersetzung
- Automatische Textproduktion
- Automatische Textüberprüfung
- Automatische Inhaltsanalyse
- Automatisierter Unterricht
- Dialogsysteme und automatische Auskunft

1.5 Modalitäten und Medium in Spracherkennung und -synthese

1.5.1 Modalitäten der Zeichenerkennung

Zeichenerkennung basiert auf folgenden Eingabegeräten

- Ohren – gesprochene Sprache
- Augen – handgeschriebene, gedruckte, und Gehörlosensprache
- Tastsinn – Blindenschrift (Braille)

1.5.2 Modalitäten der Zeichensynthese

Zeichensynthese basiert auf folgenden Ausgabegeräten

- Stimmbänder in Kombination mit dem Mund – gesprochene Sprache
- Hand – geschriebene Sprache, einschließlich Braille
- Hand-Arm-Gesichtsgesten – Gehörlosensprache

1.5.3 Medien der Sprache

Nichtelektronische Medien:

- *Laute* der gesprochenen Sprache
- *Buchstaben* der handgeschriebenen oder gedruckten Sprache
- *Gebärden* einer Gehörlosensprache

Elektronisches Medium:

- Modalitätsabhängige Repräsentationen:
 - Tonbandaufnahmen gesprochener Sprache
 - Bitmap von geschriebener Sprache
 - Videoaufnahme einer Gebärdensprache
- Modalitätsunabhängige Repräsentation:
 - digital kodierte elektronische Zeichensequenzen, z. B. ASCII

1.5.4 Korrelation von Modalität und Medium

Monomedial

Handschrift, Druck, Gebärdensprache
gesprochene Sprache, Tonband
Braille

Eingabemodalität

Augen
Ohren
Tastsinn

Multimedial

Film

Augen, Ohren

Monomedial

Handschrift, Druck, Braille
gesprochene Sprache, Tonband

Ausgabemodalität

Hand
Stimmbänder mit Mund

Multimedial

Gebärdensprache

Hand-, Arm-, Gesichtsgebärden

1.5.5 Transfer zwischen modalitätsabhängigen und modalitätsunabhängigen Repräsentationen

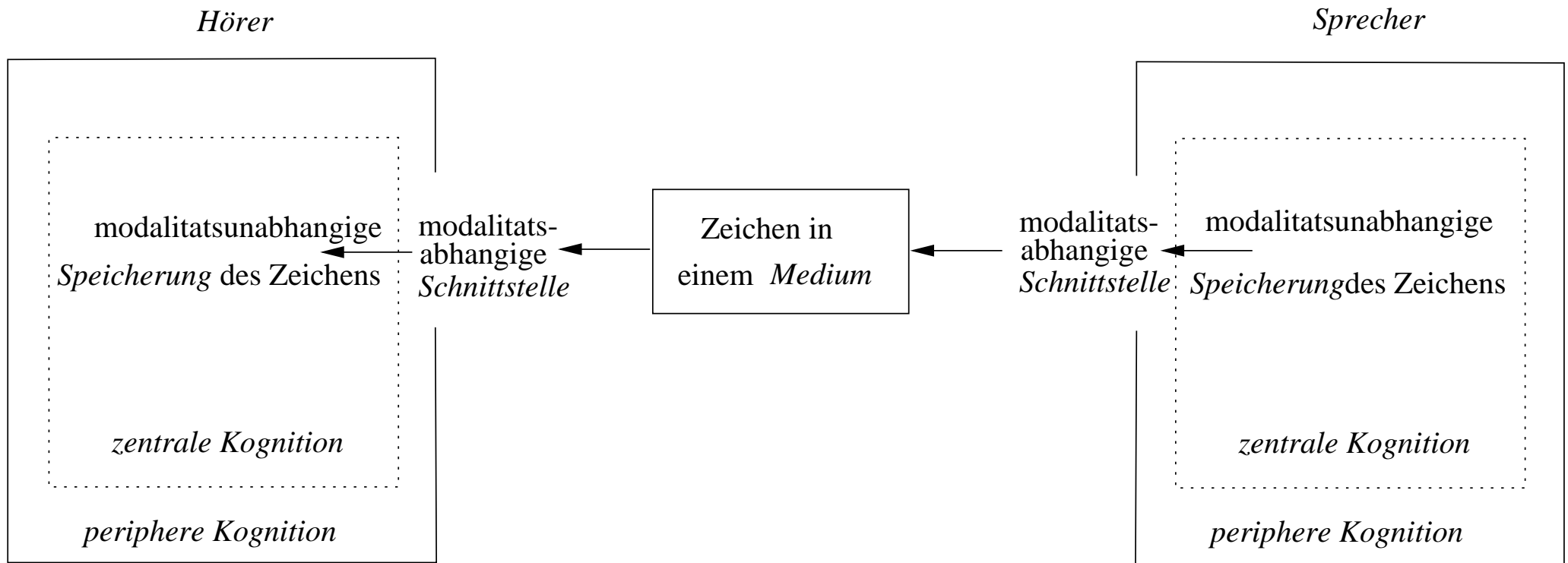
Spracherkennung: $ma \Rightarrow mu$ Transfer

modalitäts-abhängige Repräsentationen werden in modalitäts-unabhängige Repräsentationen abgebildet.

Sprachsynthese: $mu \Rightarrow ma$ transfer

modalitäts-unabhängige Repräsentationen werden in modalitäts-abhängige Repräsentationen abgebildet.

1.5.6 Speicherung, Modalität und Medium in Spracherkennung und Sprachsynthese



Vergleich: eine sprachgesteuerte Schreibmaschine

1.5.7 Methoden des $ma \Rightarrow mu$ Transfer

Nicht-automatischer $ma \Rightarrow mu$ Transfer: überläßt die Spracherkennung dem Menschen, der gesprochene oder geschriebene Sprache in den Computer eintippt.

Automatischer $ma \Rightarrow mu$ Transfer: akustische oder optische Mustererkennung.

1.5.8 Desiderata der automatischen Spracherkennung

- *Sprecherunabhängigkeit*

Das System soll verschiedene Sprecher mit verschiedenen Tonhöhen, Dialekten etc. spontan verstehen, – ohne daß eine anfängliche Lernphase erforderlich ist, in der das System an einen bestimmten Sprecher angepaßt werden muß.

- *Kontinuierlichkeit*

Das System soll kontinuierliche Sprache in unterschiedlicher Geschwindigkeit bewältigen, – ohne daß dabei unnatürliche Pausen zwischen den einzelnen Wörtern erforderlich sind.

- *Domänenunabhängigkeit*

Das System soll in der Lage sein, gesprochene Sprachzeichen unabhängig vom Inhalt zu erkennen – ohne daß ihm vorher eingegeben werden muß, welches Vokabular es zu erwarten bzw. nicht zu erwarten hat.

- *Realistischer Wortschatz*

Das System soll in der Lage sein, mindestens ebenso viele Wortformen zu erkennen wie ein durchschnittlicher Sprecher.

- *Robustheit*

Auch bei Abbrüchen, Kontraktionen und Verschleifungen soll das System in der Lage sein, die intendierten Wortformen zu erschließen.

1.5.9 Wie gut ist die automatische Spracherkennung heute?

It will be many years before unlimited vocabulary, speaker-independent continuous dictation capability is realized.

[Es wird noch viele Jahre dauern bis eine sprecherunabhängige, kontinuierliche Diktatfähigkeit realisiert sein wird.]

Zue, Cole & Ward 1998

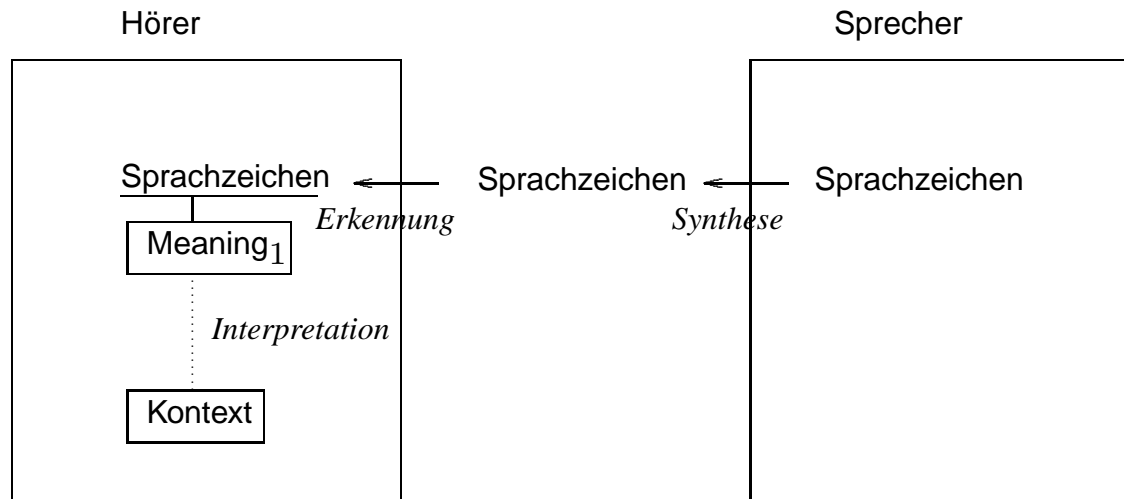
1.5.10 Die entscheidende Frage für die Konstruktion einer wirklich leistungsfähigen Spracherkennung:

Wie sollte das Grammatik- und Domänenwissen am besten organisiert werden?

Die Antwort ist offensichtlich:

Im Rahmen einer funktionalen, mathematisch effizienten und computergerechten Sprachtheorie.

1.5.11 Verkleinerung des Suchraums bei der automatischen Spracherkennung



- automatische Wortformerkennung: entspricht die Hypothese einer möglichen Wortform?
- automatische Syntaxanalyse: entspricht die Hypothese einem wohlgeformten Ausdruck?
- automatische semantisch-pragmatische Analyse: macht die Hypothese einen Sinn in bezug auf den aktuellen Verwendungskontext?

1.6 Zweite Gutenbergsche Revolution

1.6.1 Die Erste Gutenberg Revolution

Basierend auf der technischen Innovation des Druckens mit beweglichen Lettern, stellte sie der breiteren Öffentlichkeit eine Fülle von Wissen zur Verfügung.

1.6.2 Die Zweite Gutenberg Revolution

Basierend auf der automatischen Sprachverarbeitung im elektronischen Medium hat sie zum Ziel, dem Benutzer die gewünschte Information präzise, schnell und komfortabel zu *finden*.

1.6.3 SGML: *standard generalized markup language*.

A family of ISO standards for labeling electronic versions of text, enabling both sender and receiver of the text to identify its structure (e.g. title, author, header, paragraph, etc.)

Dictionary of Computing, p. 416 (ed. Illingworth et al. 1990)

1.6.4 Zeitungstext mit SGML-Steuerzeichen

```
<HTML>
<HEAD>
<TITLE>9/4/95 COVER: Siberia, the Tortured Land</TITLE>
</HEAD>
<BODY>
<!-- #include "header.html" -->
<P>TIME Magazine</P>
<P>September 4, 1995 Volume 146, No. 10</P>
<HR>
Return to <A href="../../../time/magazine/domestic/toc/
950904.toc.html">Contents page</A>
<HR>
<BR>
<!-- end include -->
<H3>COVER STORY</H3>
<H2>THE TORTURED LAND</H2>
<H3>An epic landscape steeped in tragedy, Siberia suffered
grievously under communism. Now the world's capitalists covet
its vast riches </H3>
<P><EM>BY <A href="../../../time/bios/eugenelinden.html">
EUGENE LINDEN</A>/YAKUTSK</EM>
<P>Siberia has come to mean a land of exile, and the place
easily fulfills its reputation as a metaphor for death and
```

1.6.5 Verschiedene Textsorten

- Zeitungsartikel
- Buch
- Theaterstück
- Drehbuch
- Lexikon

1.6.6 TEI

Text encoding initiative: defines a DTD (*document type definition*) for the markup of different types of text in SGML.

1.6.7 Unterschiedliche Ziele der Textauszeichnung

- Funktionsorientiert: SGML und TEI
- Druckbildorientiert: T_EX und L^AT_EX
- Benutzerorientiert: Winword, WordPerfect, etc.

1.6.8 Alphabetische Liste von Wortformen

10	in	STORY
146	in	suffered
1995	in	sun
20	its	than
4	its	that
a	LAND	The
a	land	the
a	landscape	the
a	like	the
a	LINDEN	the
above	Magazine	the
across	markers	the
and	mean	the
and	metaphor	the
and	midnight	the
and	midsummer	the
Arctic	million	through
as	mist	Throughout
as	more	to
barracks	mossy	to
bits	muting	TORTURED

2. Technologie und Grammatik

2.1 Indizieren und Finden in textuellen Datenbanken

2.1.1 Indizieren

Die Indizierung einer textuellen Datenbank ist eine Tabelle, die für jeden Buchstaben sämtliche Positionen (Adressen) im elektronischen Speichermedium der Datenbank auflistet, in denen dieser Buchstabe steht.

2.1.2 Vorteile einer elektronischen Indizierung

- Feinheit der Suche
- Flexibilität
 - Allgemeine Spezifikation von Mustern
 - Kombination von Mustern
- Automatischer Aufbau der Indexstruktur
- Bequemlichkeit, Geschwindigkeit, Zuverlässigkeit
 - Eingabe
 - Ausgabe