Sprachtechnologie

Hans Uszkoreit

Universität des Saarlandes

und

Deutsches Forschungszentrum für Künstliche Intelligenz

Eigentlich eine Klasse von Technologien in der Informations-technologie, die Wissen über die Struktur des menschlichen Sprachen verwenden, um die maschinelle Verarbeitung der Sprache zu ermöglichen bzw. zu verbessern.

Beispiel: Microsoft Word verarbeitet zwar Sprache, enthält aber im Kern nur sehr wenig Sprachtechnologie.

Sprachtechnologie steckt aber in der Erkennung von Satzgrenzen für die Formatierung, in der automatischen Silbentrennung, in der Rechtschreibkontrolle und in der Grammatikkontrolle.

Nach Meinung der führenden Experten in der Computerindustrie ist die Sprachtechnologie eine Schlüsseltechnologie für den weiteren Fortschritt in der Computertechnik.

Akzeptanz

Das Hauptproblem für die Akzeptanz des Computers ist das Sprachproblem

- Der Standardanwender beherrscht keine Computersprachen.
- Der Standardanwender mag Computersprachen nicht.
- Der Standardanwender will auch keine Computersprachen lernen.
- Die Sprache, die der Mensch bestens beherrscht, ist seine Muttersprache.
- Das natürlichste Medium für die unmittelbare Übermittlung von Information ist die gesprochene Sprache.
- Die wichtigste Klasse von Daten sind Texte.
- Der Standardanwender verwendet die Maschine zur Produktion von Texten in menschlicher Sprache.
- Computer tun sich schwer in der Verarbeitung und Verwaltung von Texten

Aber der Computer beherrscht die menschliche Sprache nicht!

Types of Technologies

Communication partners: humans and machines (technology),
humans and humans
humans and infostructure

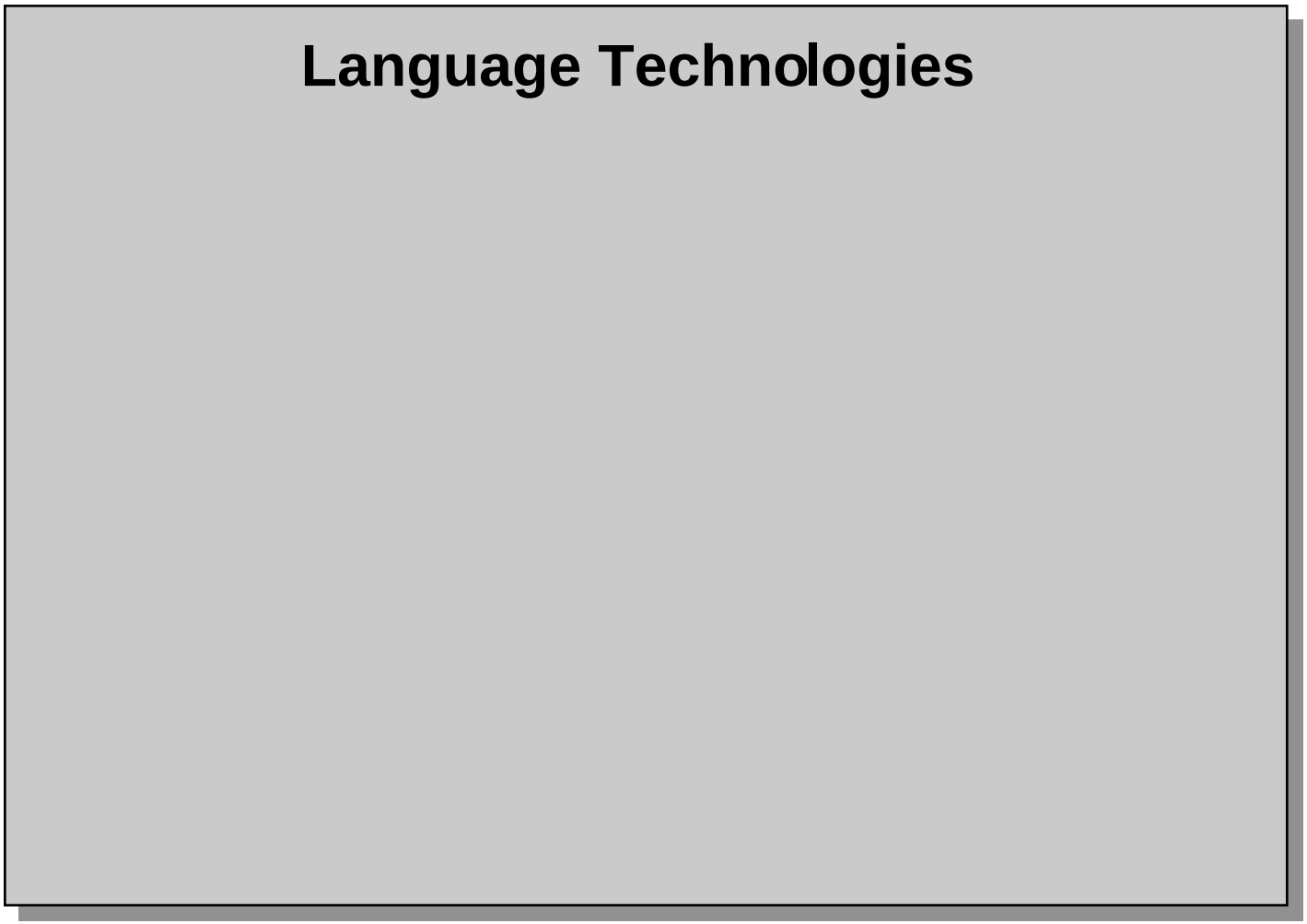
Modes and media for input and output: text, speech, pictures, gestures

Synchronicity: synchronous vs. asynchronous

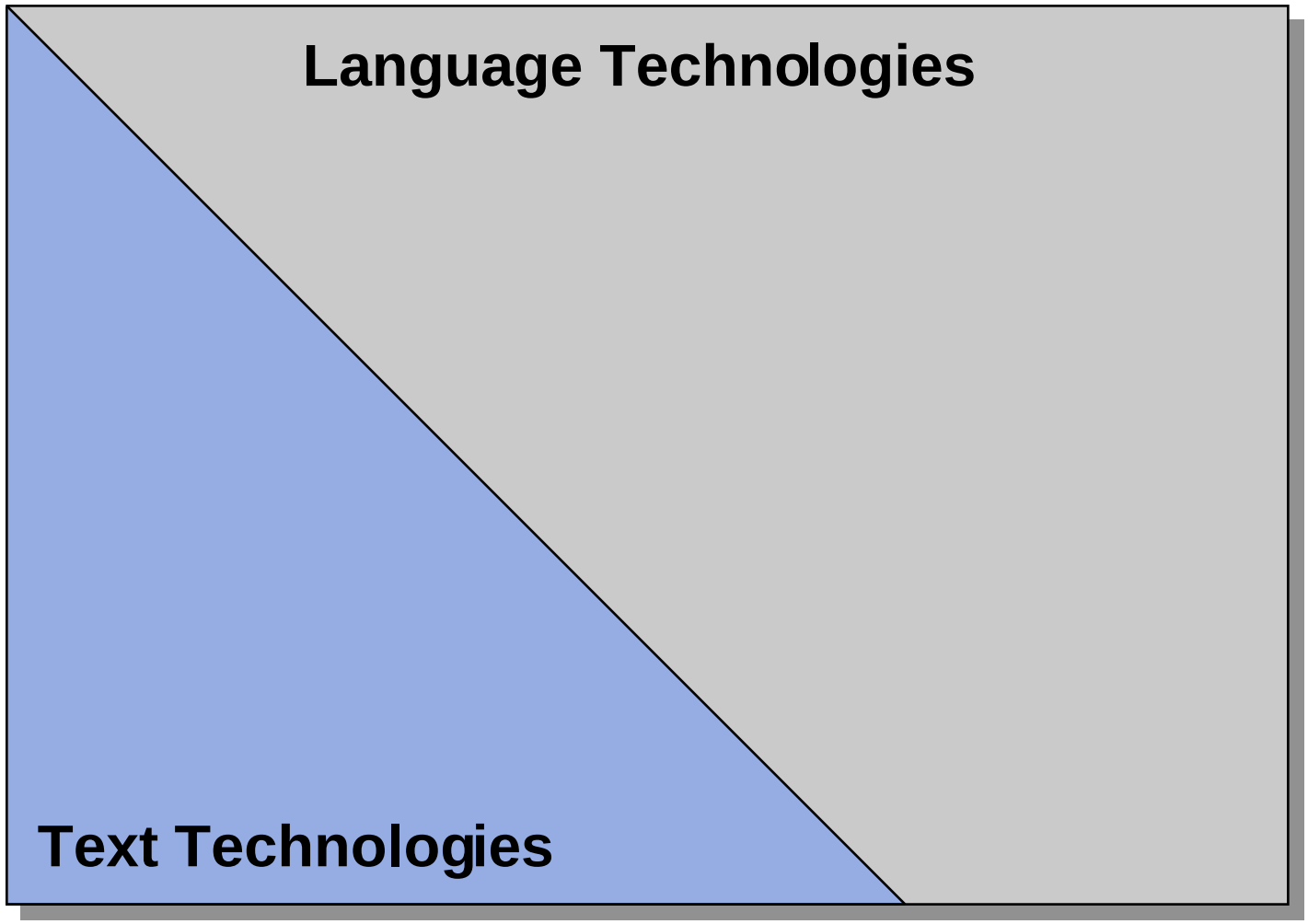
Situatedness: sensitivity to context, location, time, plans

Type of linguality: monolingual, multilingual, translingual

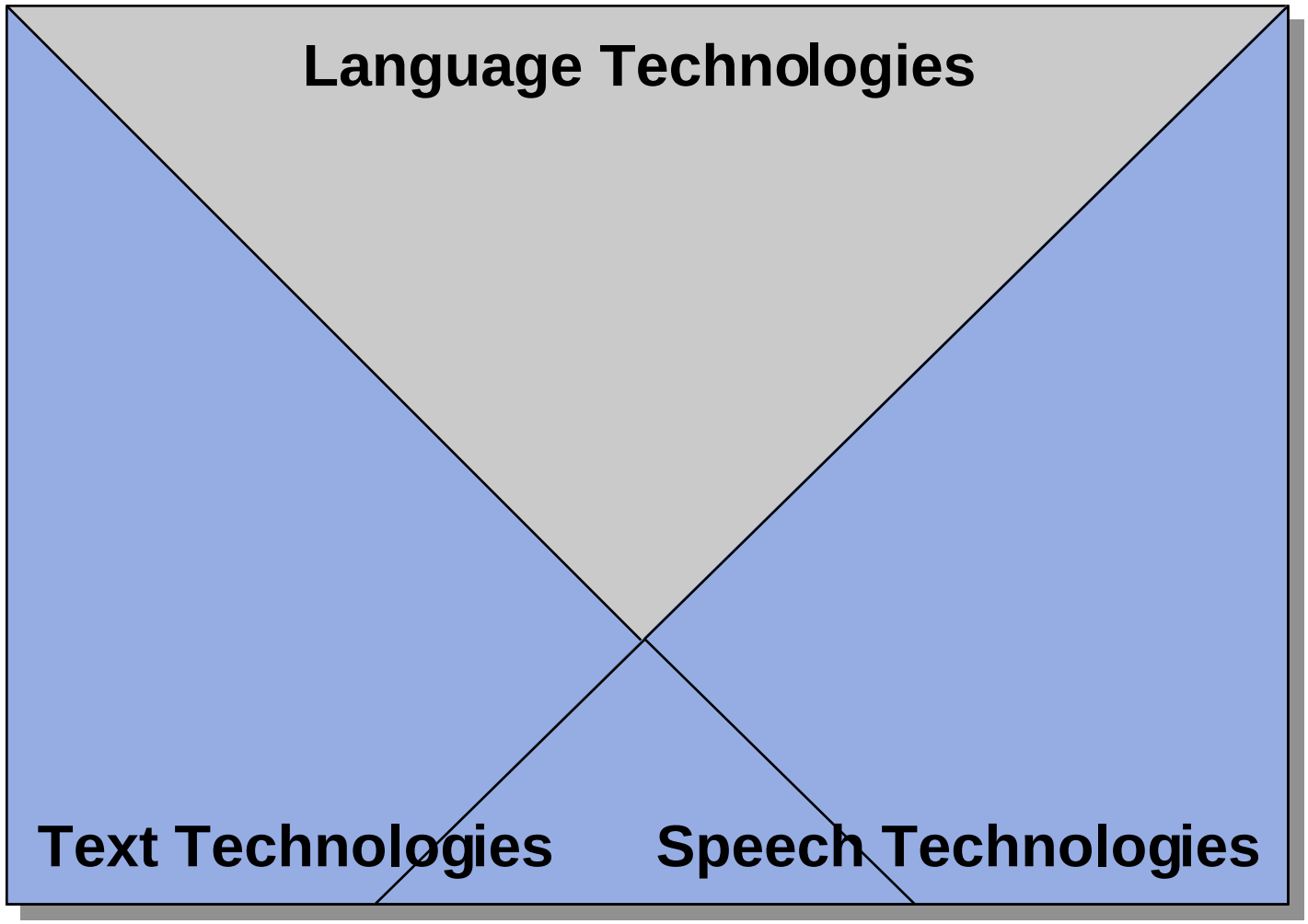
Type of processing: Categorization, summarization, extraction,
understanding, translating, responding



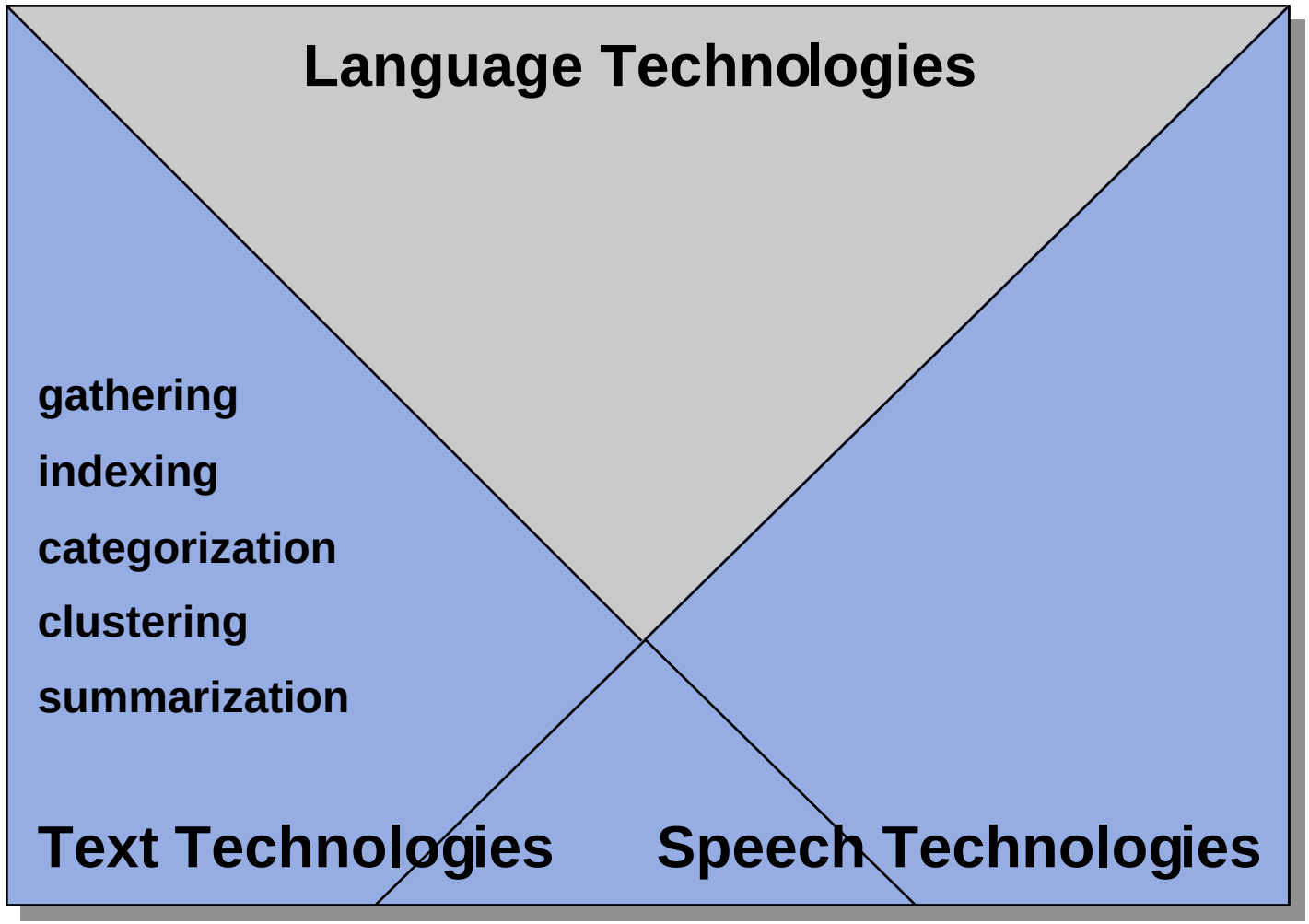
LANGUAGE TECHNOLOGIES



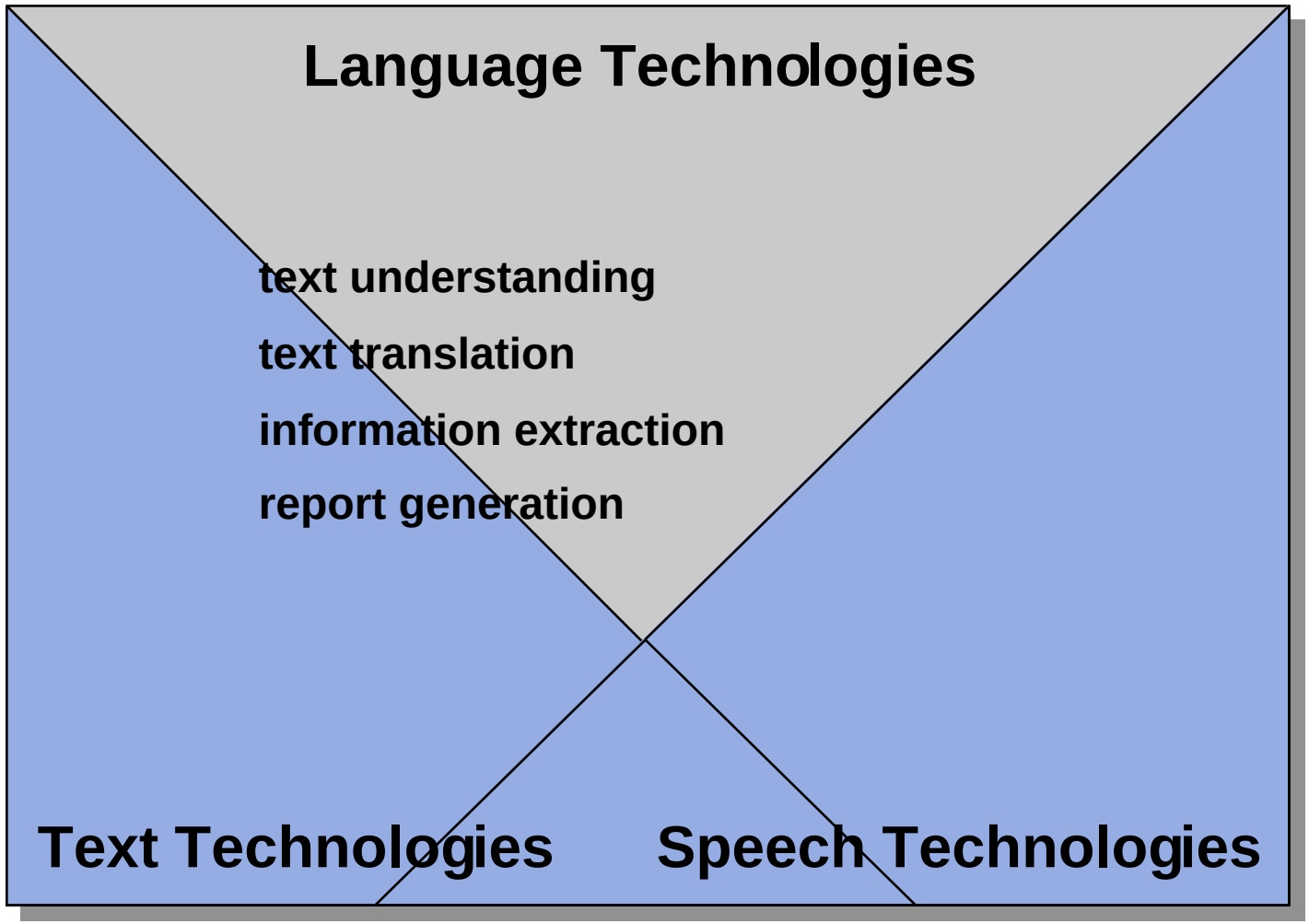
LANGUAGE TECHNOLOGIES



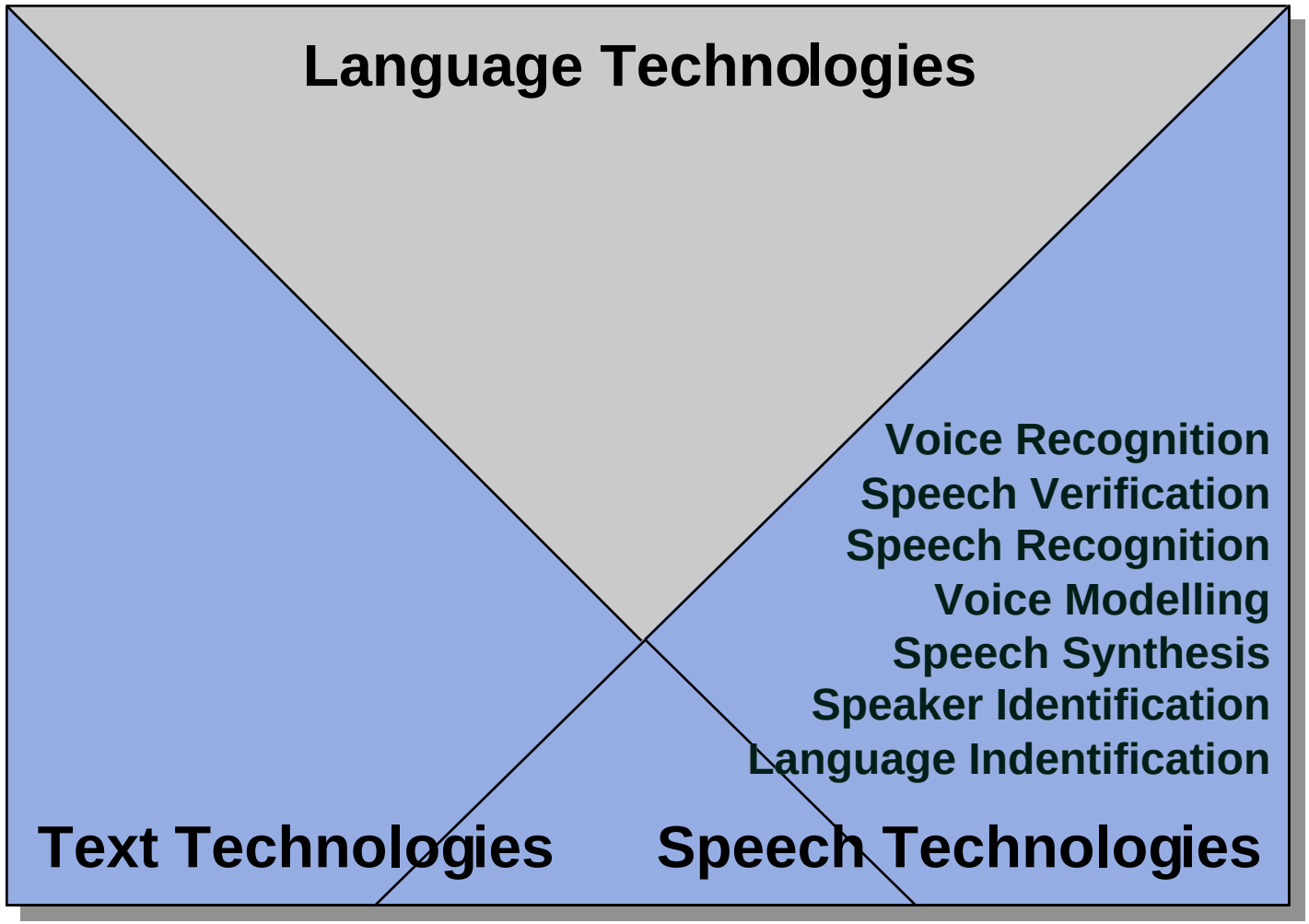
LANGUAGE TECHNOLOGIES



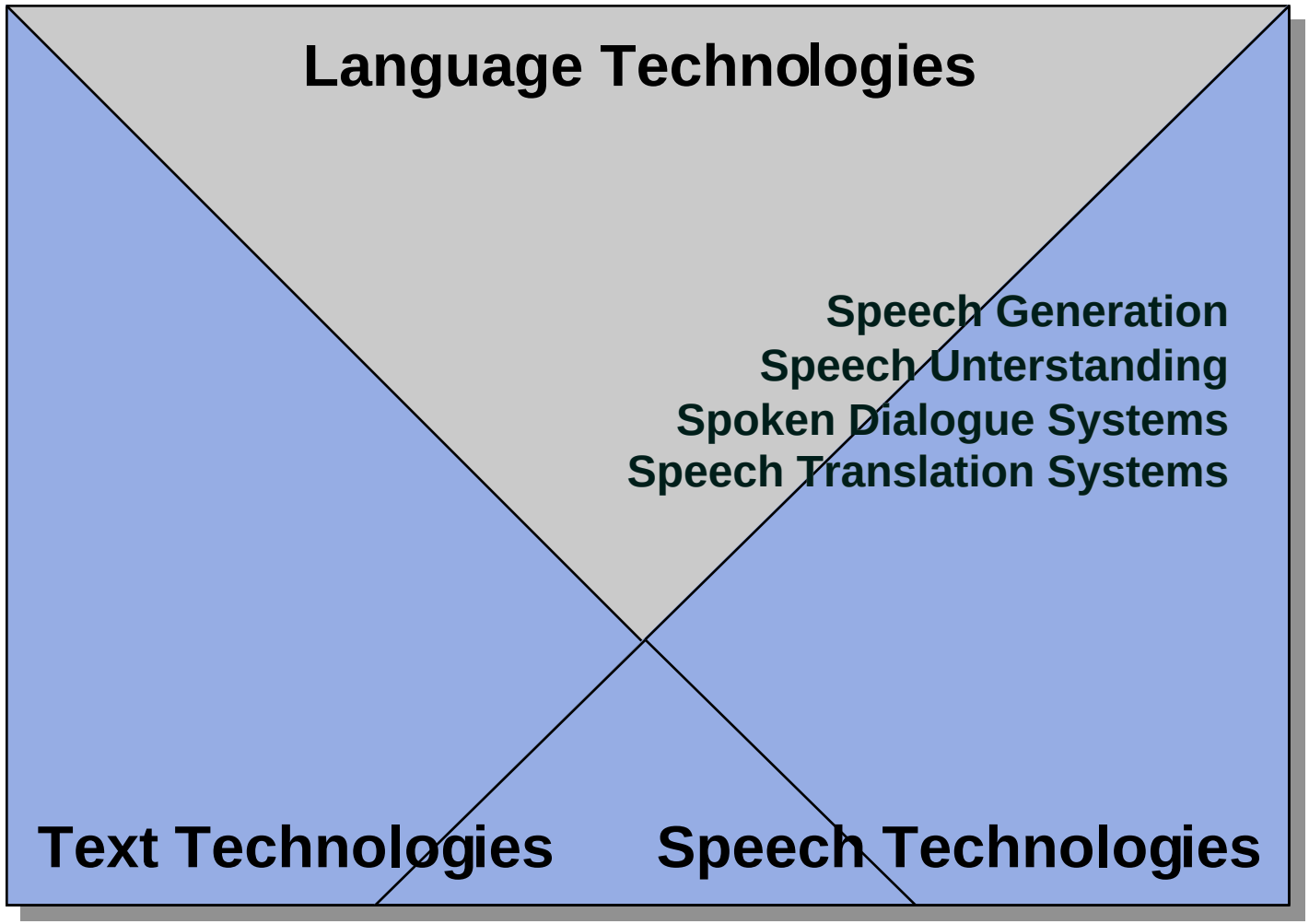
LANGUAGE TECHNOLOGIES



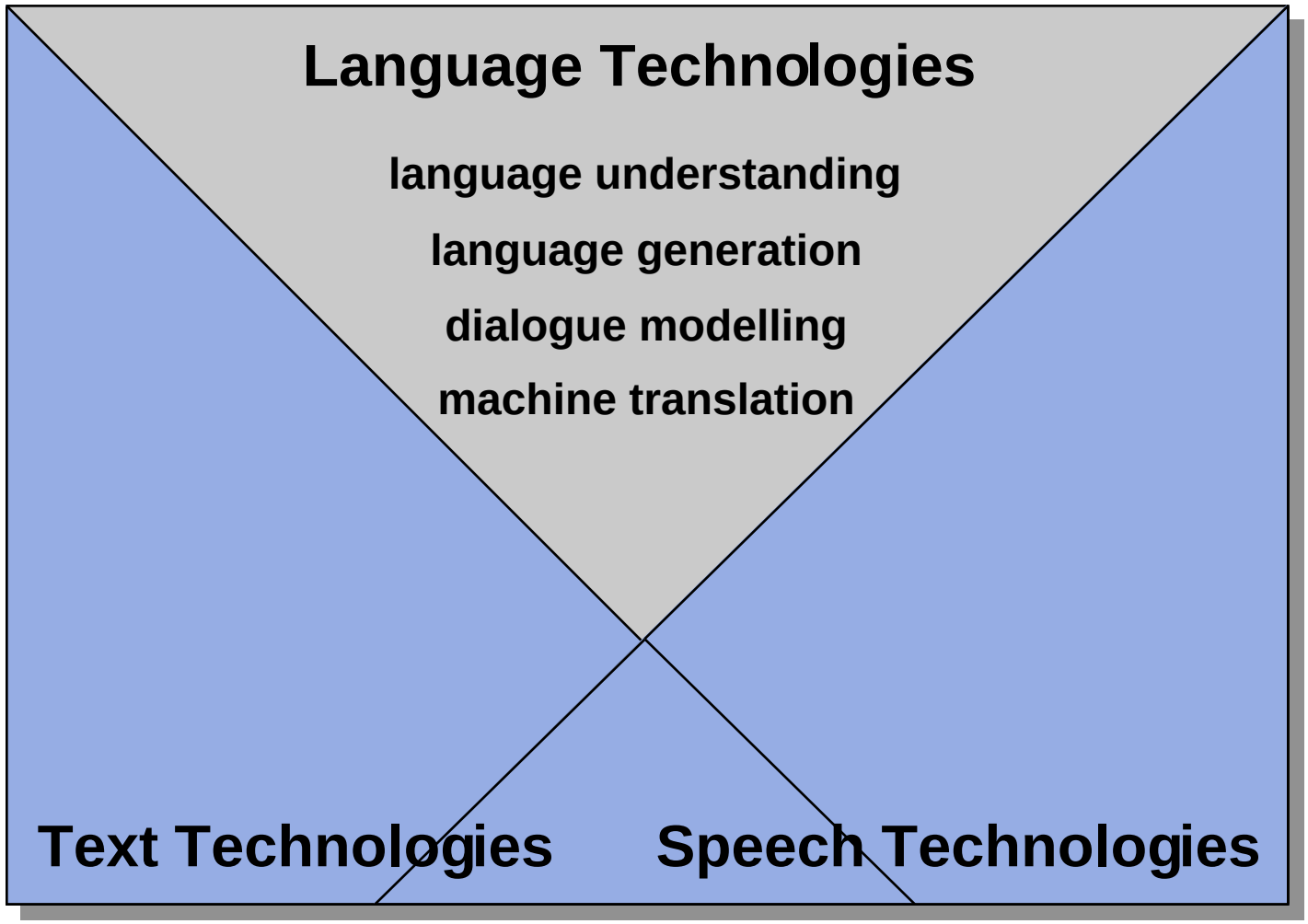
LANGUAGE TECHNOLOGIES



LANGUAGE TECHNOLOGIES



LANGUAGE TECHNOLOGIES



Maturity of Speech Technologies

Voice Control Systems

Dictation Systems

Text-to-Speech Systems

Machine Initiative Spoken Dialogue Systems

Identification and Verification Systems

Spoken Information Access

Mixed Initiative Spoken Dialogue Systems

Speech Translation Systems

Deployed. On the market
Mature or close to maturity
Research prototypes in R&D

Maturity of Text Technologies

Spell Checkers

Machine-Assisted Human Translation

Translation Memories

Indicative Machine Translation

Grammar Checkers

Information Extraction

Human Assisted Machine Translation

Report Generation

High Quality Text Translation

Text Generation Systems

Deployed. On the market
Mature or close to maturity
Research prototypes in R&D

Maturity of IM Technologies

Word-Based Information Retrieval

Summarization by Simple Condensation

Simple Statistical Categorization

Simple Automatic Hyperlinking

Cross-Lingual Information Retrieval

Automatic Hyperlinking With Disambiguation

Simple Information Extraction (Unary, Binary Relations)

Complex Information Extraction (Ternary+ Relations)

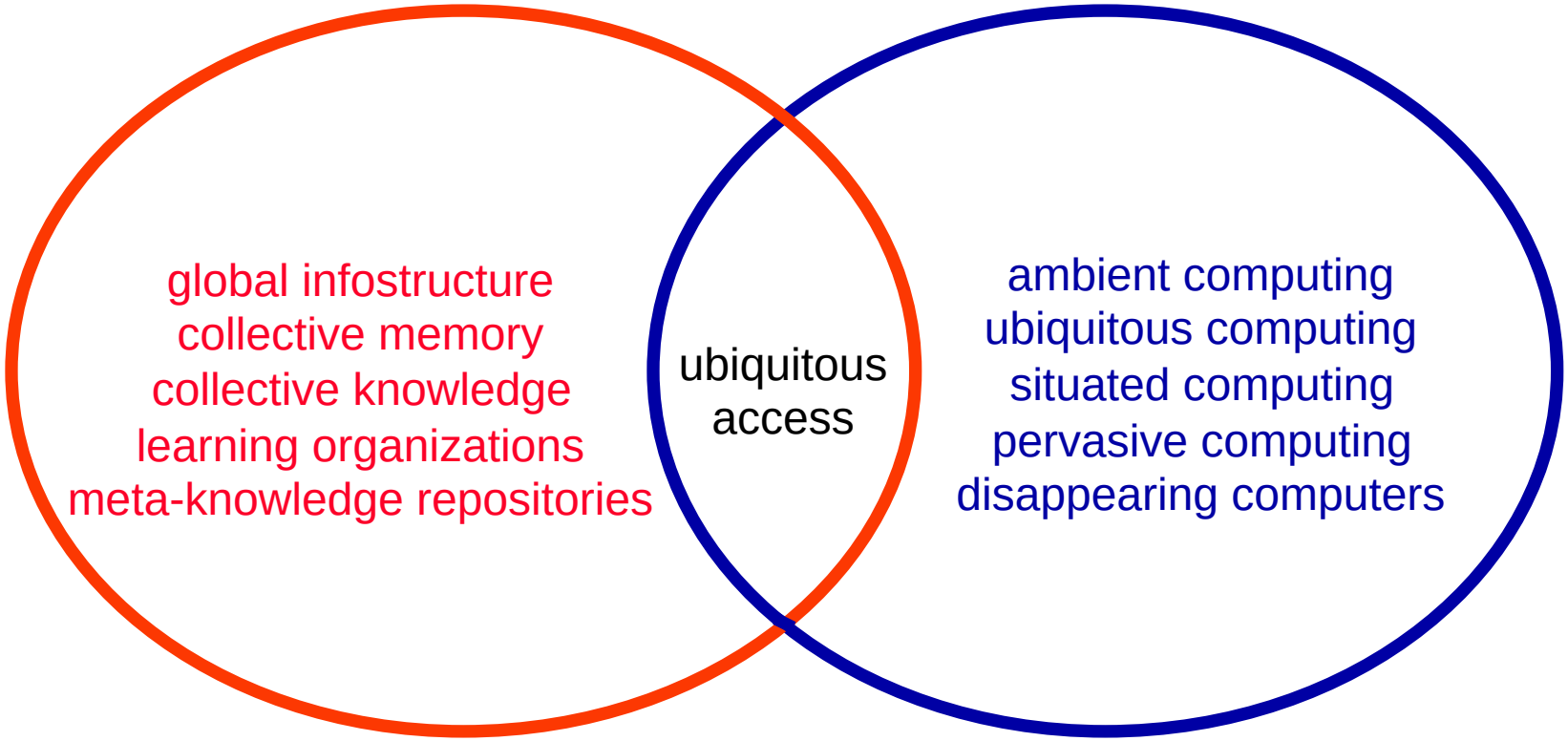
Dense Associative Hyperlinking

Concept-Based Information Retrieval

Text Understanding

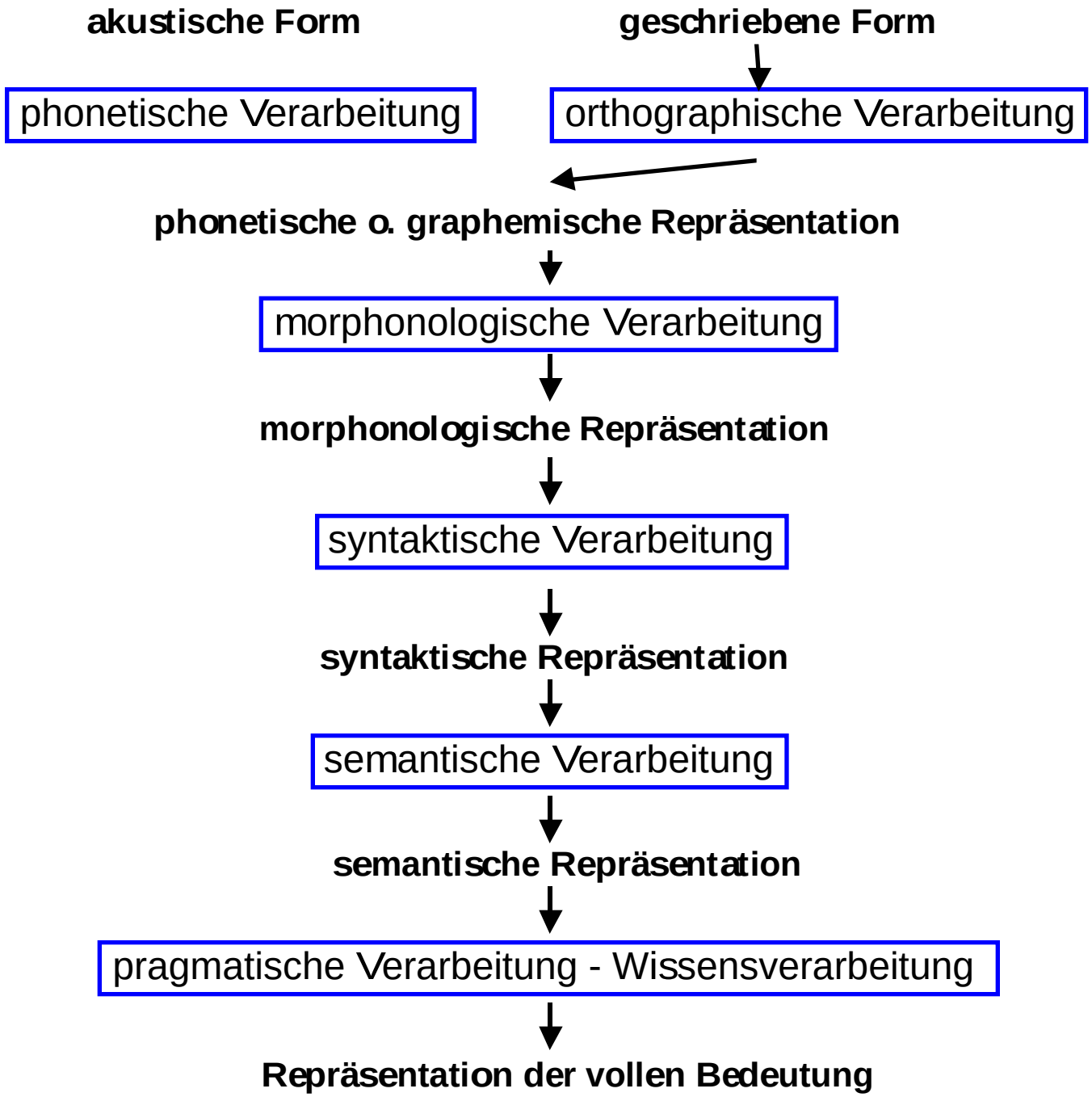
Deployed. On the market
Mature or close to maturity
Research prototypes in R&D

GENERAL TRENDS

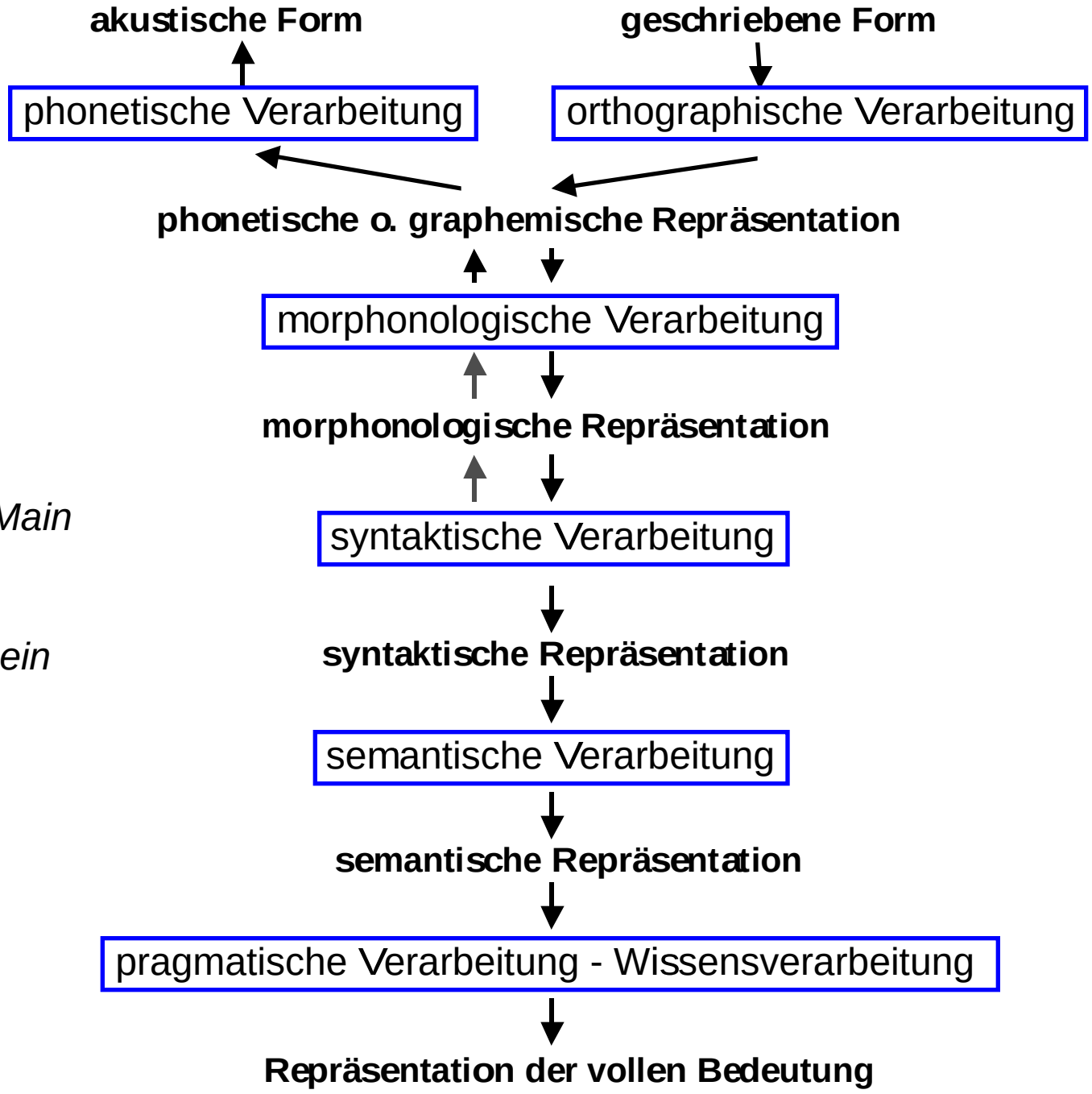


personalization
adaptation
learning

Textverstehen



Diktat

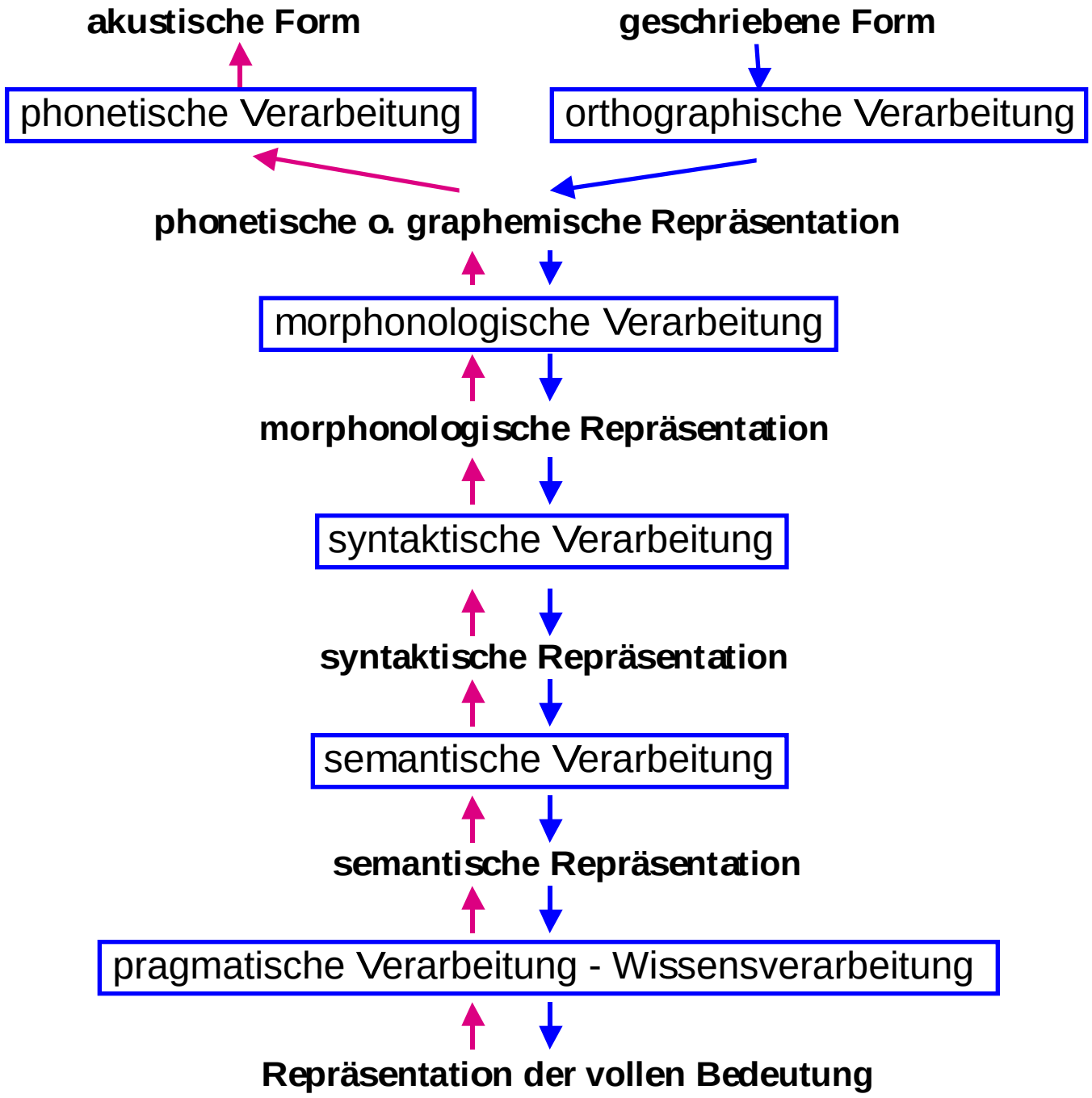


das Boot auf dem Main

oder

daß bot auf dem mein

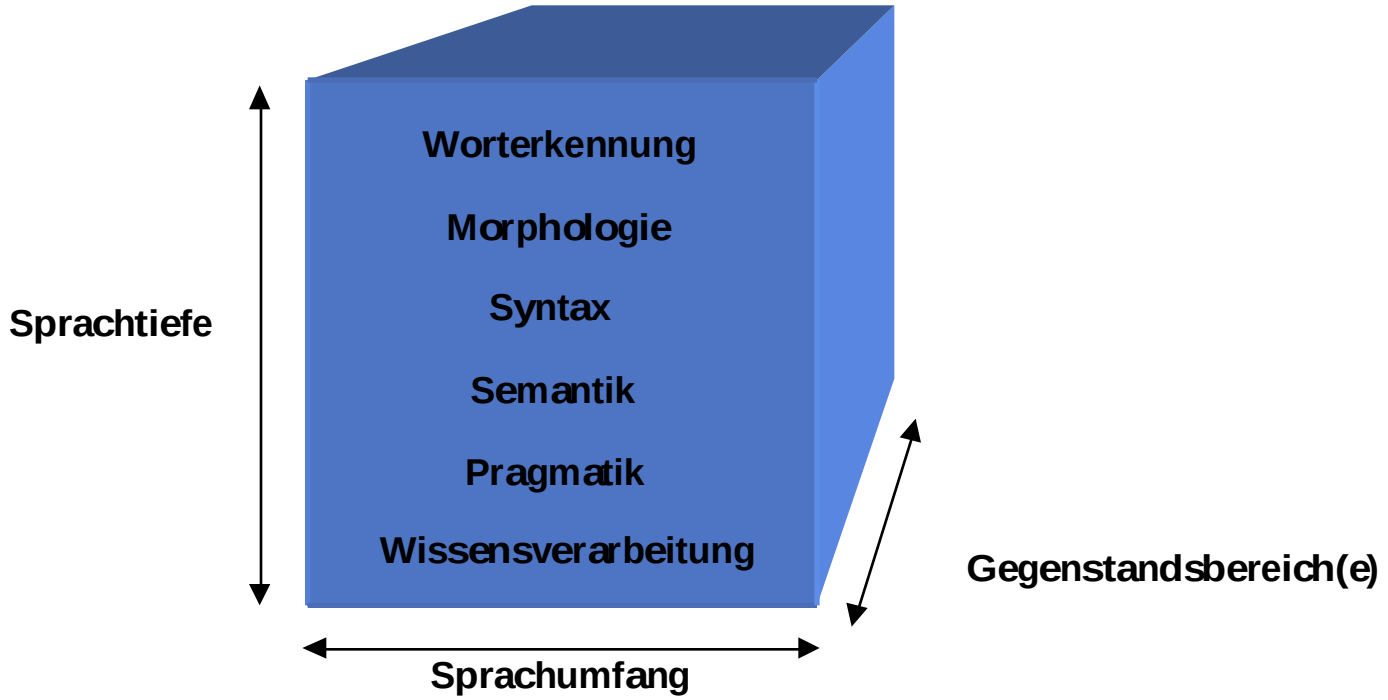
Maschinelle Übersetzung



Performanzkriterien

efficiency	**geringer Zeit- und Speicherbedarf**
accuracy	**Fähigkeit, linguistisch korrekte Lösungen zu liefern**
robustness	**Fähigkeit, mit allen möglichen Eingaben fertigzuwerden**
coverage	**größtmögliche Abdeckung der Sprache**
specificity	**Fähigkeit, die richtige Analyse zu selegieren**

Das Problem der Sprachbeherrschung ist zu komplex



Es gibt viele Anwendungen, die nur begrenzte Sprachbeherrschung benötigen!

Grenzen der Technologie

Der Computer wird so bald nicht sprechen und schreiben wie wir

KEINE SPRACHBEHERRSCHUNG OHNE ALLGEMEINES WISSEN

Wörterbücher und Grammatiken können wir formalisieren

Bei der Semantik wird es schon schwerer

Dialog kann nur recht einfach modelliert werden

Begrenztes Domänenwissen ist machbar

Beim Allgemeinwissen und großem Fachwissen hört es auf

Wir müssen das Problem begrenzen, um zu vernünftigen Anwendungen zu gelangen

Weder Sprachverstehen noch Sprachproduktion sind bisher gelöst.

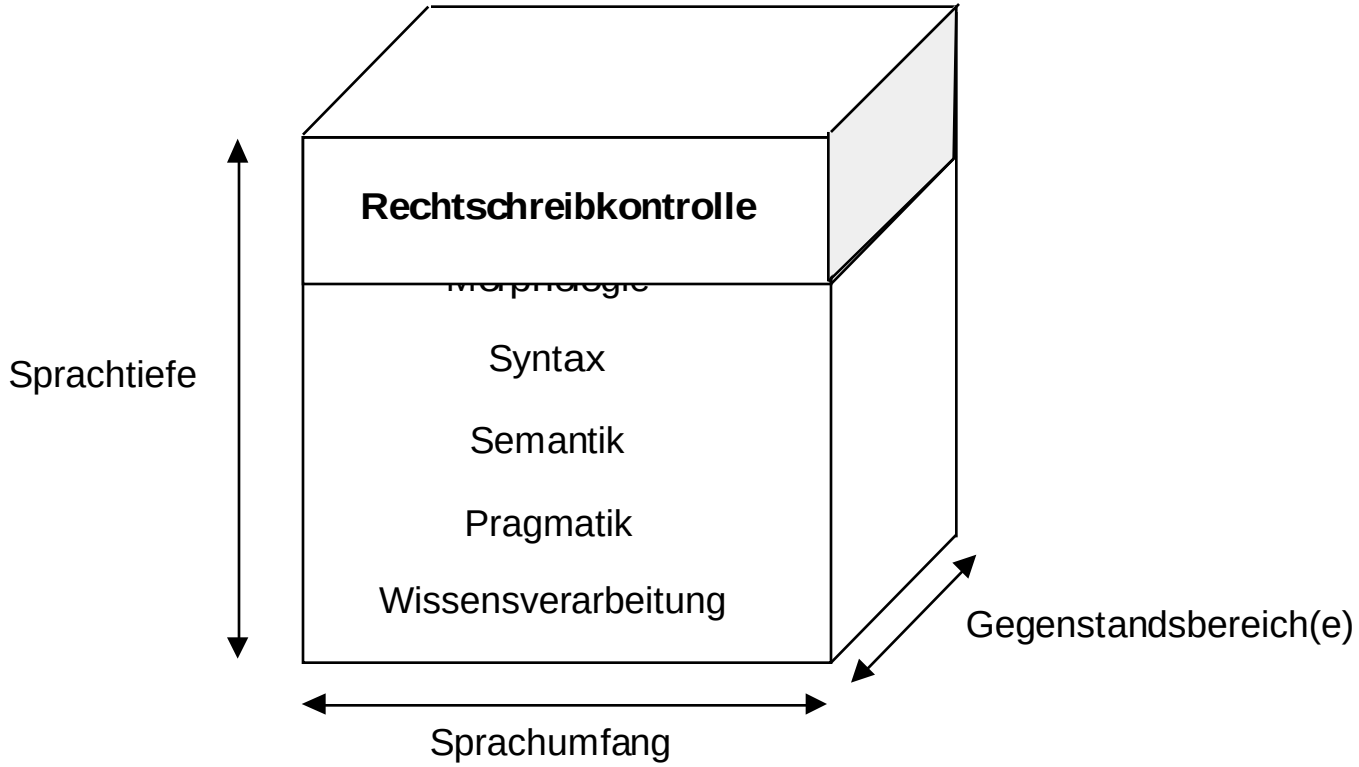
Aber: Wir besitzen heute sogenannte flache Verfahren, die zwar kein Verstehen ermöglichen, aber für viele Anwendungen oft völlig ausreichen.

flache Ansätze (effizient und robust)

- **statistische Methoden, Mustergrammatiken**

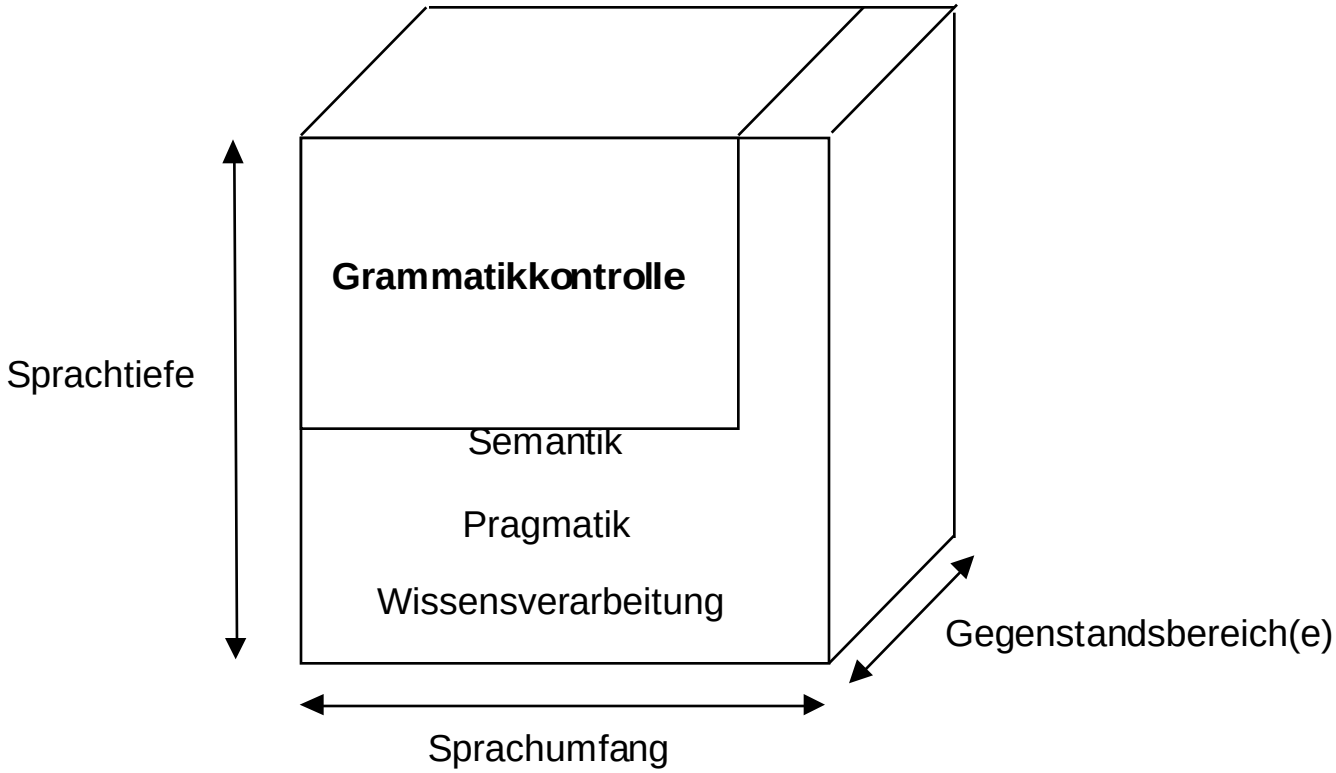
tiefe Ansätze (präzise und korrekt)

- **linguistische Prinzipien, Constraints oder komplexe Regelwerke**



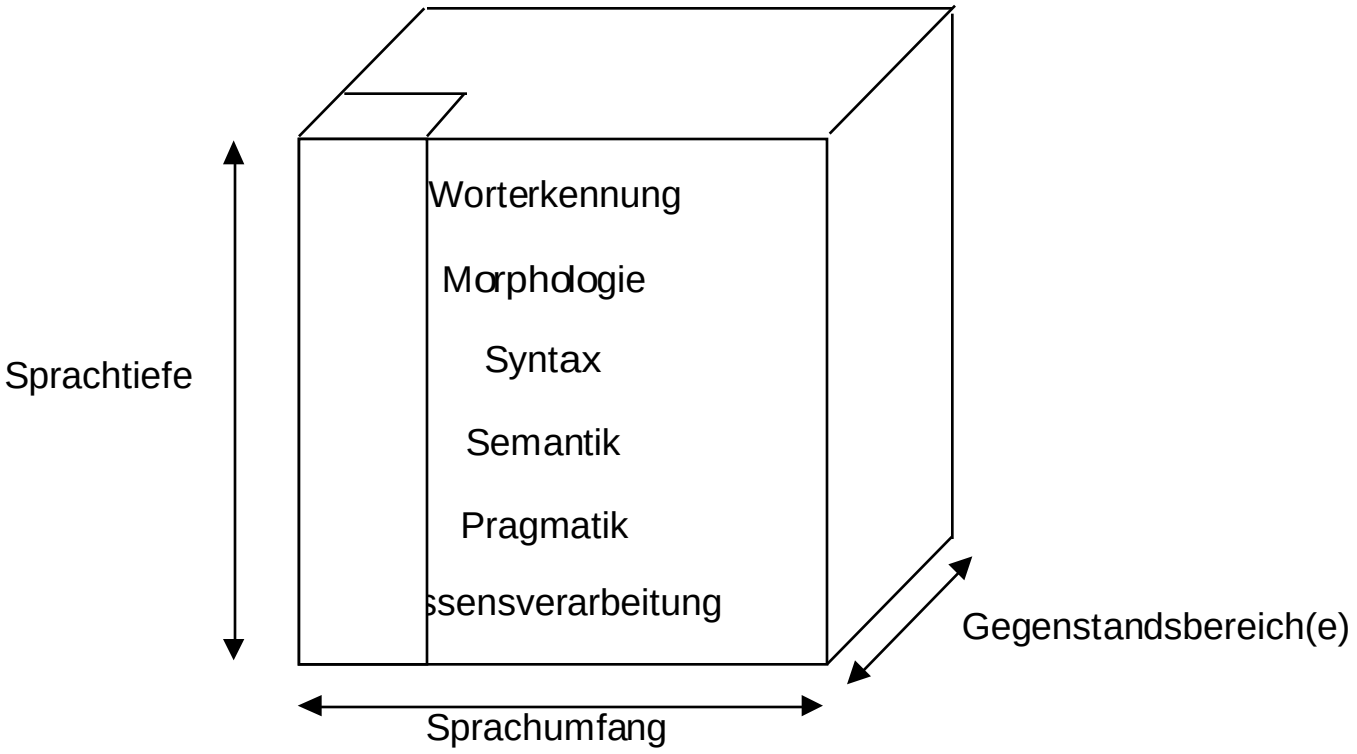
Diese Anwendung ist die bisher meistverkaufte Sprachanwendung

GRAMMATIKKONTROLLE

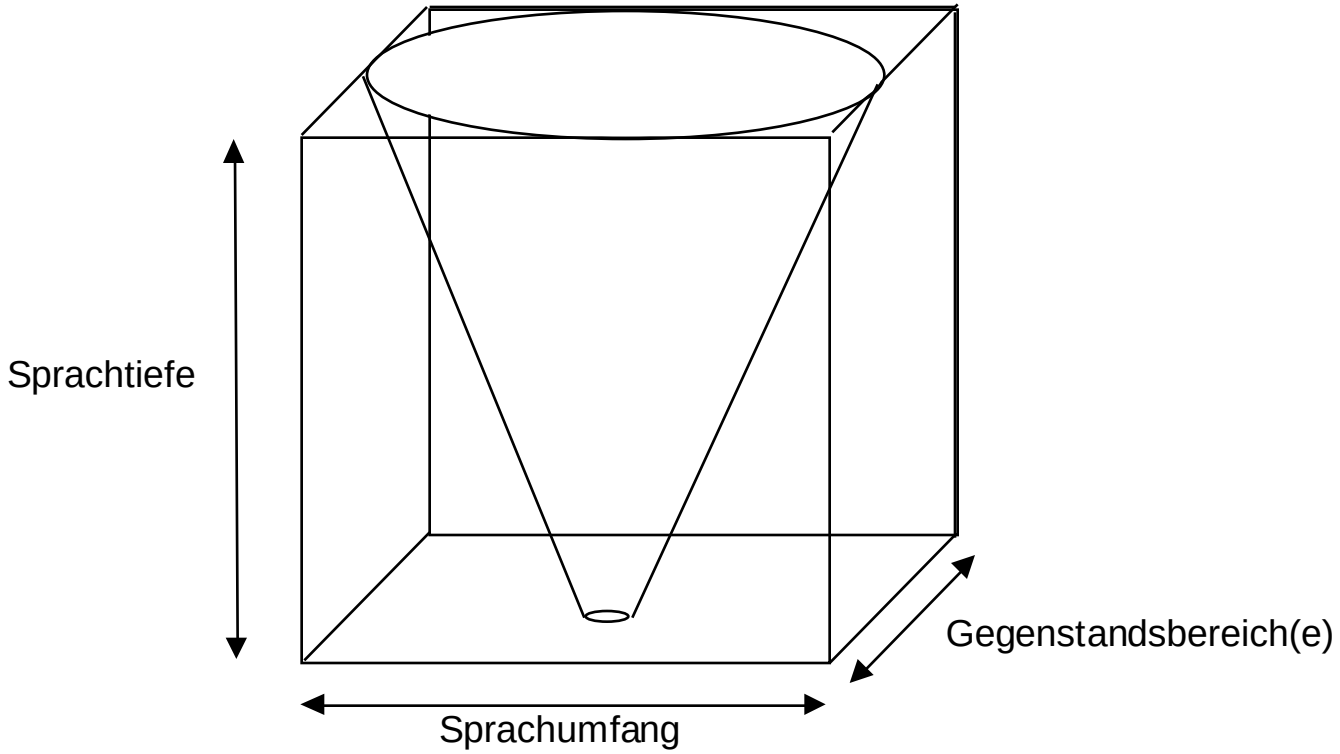


Hier beginnt das Geschäft erst gerade

EINFACHE ABFRAGESYSTEME



Der Bedarf kommt mit der akustischen Spracherkennung


Sprachtiefe
Gegenstandsbereich(e)
Sprachumfang

Heutige Suchtechnologie

Wort-Index

Boolsche Kombinationen

verschiedene Indexierungsverfahren

eingeschränkte Morphologie

Sortierung nach Relevanz

Suche in mehreren Sprachen

Sie finden nicht genug!

Andere Wortformen

der Herzog, des Herzogs, die Herzöge

Unter- und Überbegriffe

Alfa Romeo Zagato → roadster → sports car → car → motor vehicle → vehicle

Paraphrasen

steuerliche Gründe, Steuergründe, steuerliche Erwägungen,
steuerliche Überlegungen, fiskalische Erwägungen, um Steuern zu sparen, ...

Nehmen wir an, Sie suchten nach Automobilfirmen

und gäben daher der Suchmaschine (z.B. HOTBOT) den Suchbegriff

“Automobilfirmen”

Im Englischen suchten Sie nach:

“automobile companies”

Suche

automobile companies	704

Automobilfirmen	55

Suche

automobile companies	704
car builders	233
car makers	1846
auto makers	2307
automobile makers	181
car companies	3046
cars companies	14
motor companies	194
auto companies	1345
car manufacturers	3056
motor manufacturers	582
automobile manufacturers	4263
manufacturers of cars	151
manufacturers of autos	15
manufacturers of automobiles	165
manufacturers of motor vehicles	55

Automobilfirmen	55
Autohersteller	320
Autobauer	131
Autoproduzenten	26
Autofabrikant	89
Autofirmen	86
Pkw Hersteller	15
Automobilunternehmen	57
Automobilhersteller	602
Kfz-Hersteller	42
Autounternehmen	9
Automobilkonzerne	83
Unternehmen der Automobilbranche	4
Hersteller von Autos	4
Hersteller von Automobilen	13
Hersteller von Kraftfahrzeugen	3

Sie finden zu viel!

Ambiguität

- deutsch: *Zug, Bahn, Leitung, Schalter*
- englisch: *terminal, line, engine*

Polysemie

- *Buch, Schule, printer*

Eigennamen

- Personennamen: *Maurer, Washington, Chase*
- Ortsbezeichnungen: *Essen, Halle, Bismarck*

AUFGABEN DES IM (1)

- **Die Aufgabe des Informationsmanagements ist die Verwaltung und Nutzbarmachung von sehr großen Informationsmengen, wie wie sie heute bereits auf dem WWW, in Intranets und in großen Text-Datenbanken finden.**

- **Das Netz macht sie erst einmal nur verfügbar.**

- **Im Gegensatz zu herkömmlichen Datenbanken ist die Information viel weniger vorstrukturiert (in Sinne der Strukturierung von Computerdaten). Auf der anderen Seiten sind die relevanten inhaltlichen Strukturen natürlich weitaus komplexer. Durch die Digitalisierung von großen Teilen des menschlichen Wissen (z.B. digitale Bibiliotheken, Filmarchive etc.) wird dieses Problem noch zunehmen.**

Distributivität

Die Information liegt auf verschiedenen Maschinen

Heterogenität

Vielzahl von Dokumentformaten
Multilingualität
Multimedialität (z.B. Sprache, Bilder, Klänge),
Multimodalität (z.B. geschr. u. gesprochene Sprache, Filmdateien o. Realzeitübertragungen)

Unstrukturiertheit

keine einheitliche Klassifikation,
keine einheitliche interne Strukturierung.
keine einheitliche u. verläßliche Hypertextverknüpfung

Redundanz

Viele Informationen sind mehrfach vorhanden.

Information wird

- **gewonnen**
- **kategorisiert**
- **gefiltert**
- **zusammengeführt**
- **strukturiert**
- **dem Benutzer zugeführt**
- **adäquat präsentiert**

- **Sammeln (gathering)**
- **Indizieren (indexing)**
- **Kategorisierung (categorization)**
- **Gruppierung (clustering)**
- **Zusammenfassung (summarization)**
- **Informationsextraktion (information extraction)**
- **Automatische Verknüpfung (automatic hyperlinking)**
- **Datenschürfen (text data mining)**
- **Informationsfusion (information fusion)**
- **Berichtsgenerierung (report generation)**
- **Textübersetzung (text translation)**

- **Sammeln (gathering)**
- **Data Mining auch Text Mining**
- **Konversion z.B. Einscannen, OCR, Transkription**
- **Agenten z.B. NetBots, WebBots**

- **Indizieren (indexing)**
- **Kategorisierung (categorization)**
- **Gruppierung (clustering)**
- **Zusammenfassung (summarization)**

- **Informationsextraktion (information extraction)**
- **Hyperverknüpfung (hyperlinking)**
- **Informationsfusion (information fusion)**
- **Trendanalyse (trend analysis)**
- **Berichtsgenerierung (report generation)**

- **Suchschlüsselerweiterung (query expansion)**
- **Relevanzsortierung (relevance ranking)**
- **Dublettenerkennung (redundancy check)**
- **thematische Gruppierung (thematic clustering)**
- **Erkennung verwandter Information (information association)**

- **Ergebnispräsentation (result presentation)**
- **Informationsvisualisierung (information visualization)**
- **virtuelle Navigation (virtual navigation)**

Robuste Extraktion von relevanten Begriffen, Phrasen, Aussagen aus Texten.

Erfolgsraten (Vollständigkeit und Präzision) hängen von der Aufgabe und vom Gegenstandsbereich ab.

Bereits eingesetzt in verschiedenen Anwendungen, z.B. für

- Firmennamenerkennung,
- Nachrichtenkategorisierung,
- Übersichten zu Firmenindikatoren (Umsatz, Gewinn, Kurse)
- Nachrichtenübersichten zu speziellen Themenbereichen

In der IE werden gezielt relevante Informationen aus Texten herausgesucht und strukturiert.

Bremen, 14. 10. 1997, wiwo: Lagersoftware weiter im Aufwind

Die Bremer Firma Trade Consult hat auf einer Pressekonferenz in Hannover die Version 2.0 ihrer erfolgreichen Lagerverwal-tungssoftware Store Age vorgestellt..
Die neue Version ermöglicht jetzt auch ...

Auf der Pressekonferenz gab Geschäftsführer Franz Merleback auch die Umsatzzahlen der Softwareschmiede für das 3. Quartal bekannt. Wurden im zweiten Quartal bereits über 30 Millionen Mark umgesetzt, so konnte Merleback jetzt das stolze Ergebnis von 42,5 Millionen verkünden.

...

In der IE werden gezielt relevante Informationen aus Texten herausgesucht und strukturiert.

Bremen, 14. 10. 1997, wiwo: Lagersoftware weiter im Aufwind

Die Bremer Firma Trade Consult hat auf einer Pressekonferenz in Hannover die Version 2.0 ihrer erfolgreichen Lagerverwal-tungssoftware Store Age vorgestellt..
Die neue Version ermöglicht jetzt auch ...

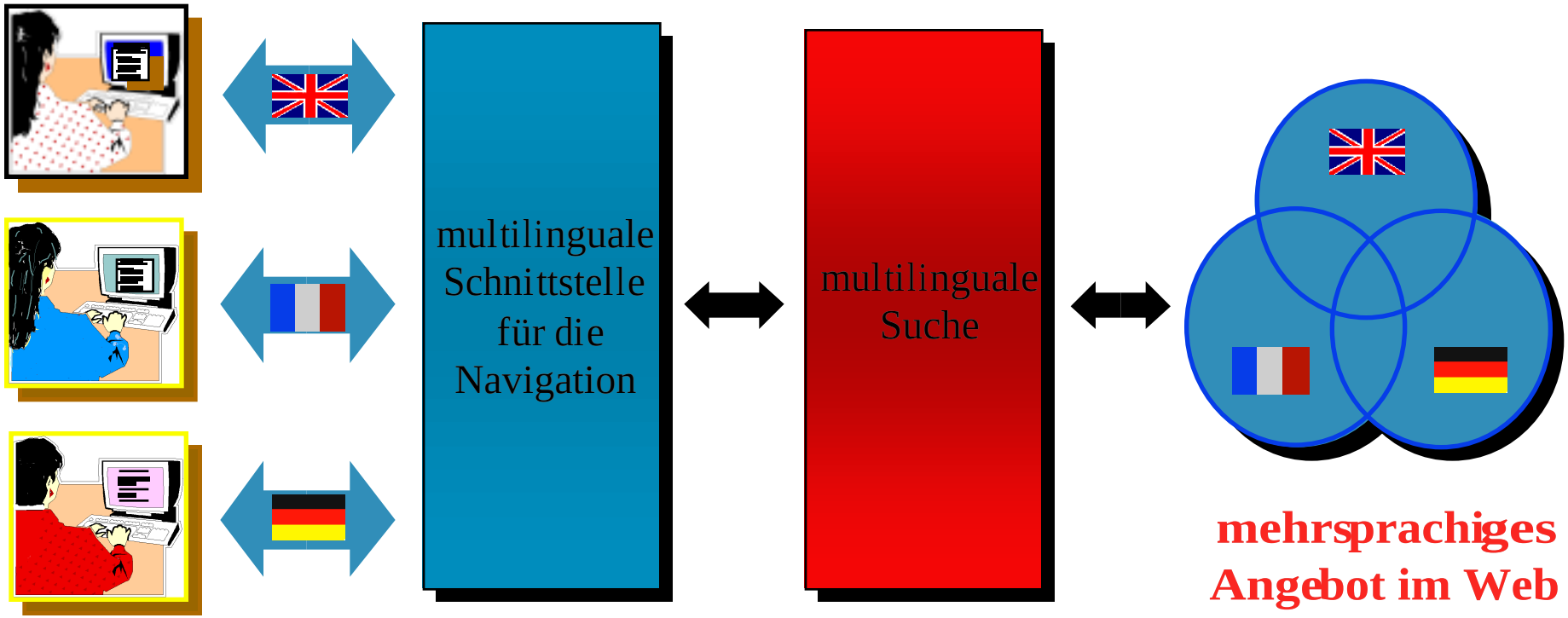
Auf der Pressekonferenz gab Geschäftsführer Franz Merleback auch die Umsatzzahlen der Softwareschmiede für das 3. Quartal bekannt. Wurden im zweiten Quartal bereits über 30 Millionen Mark umgesetzt, so konnte Merleback jetzt das stolze Ergebnis von 42,5 Millionen verkünden.

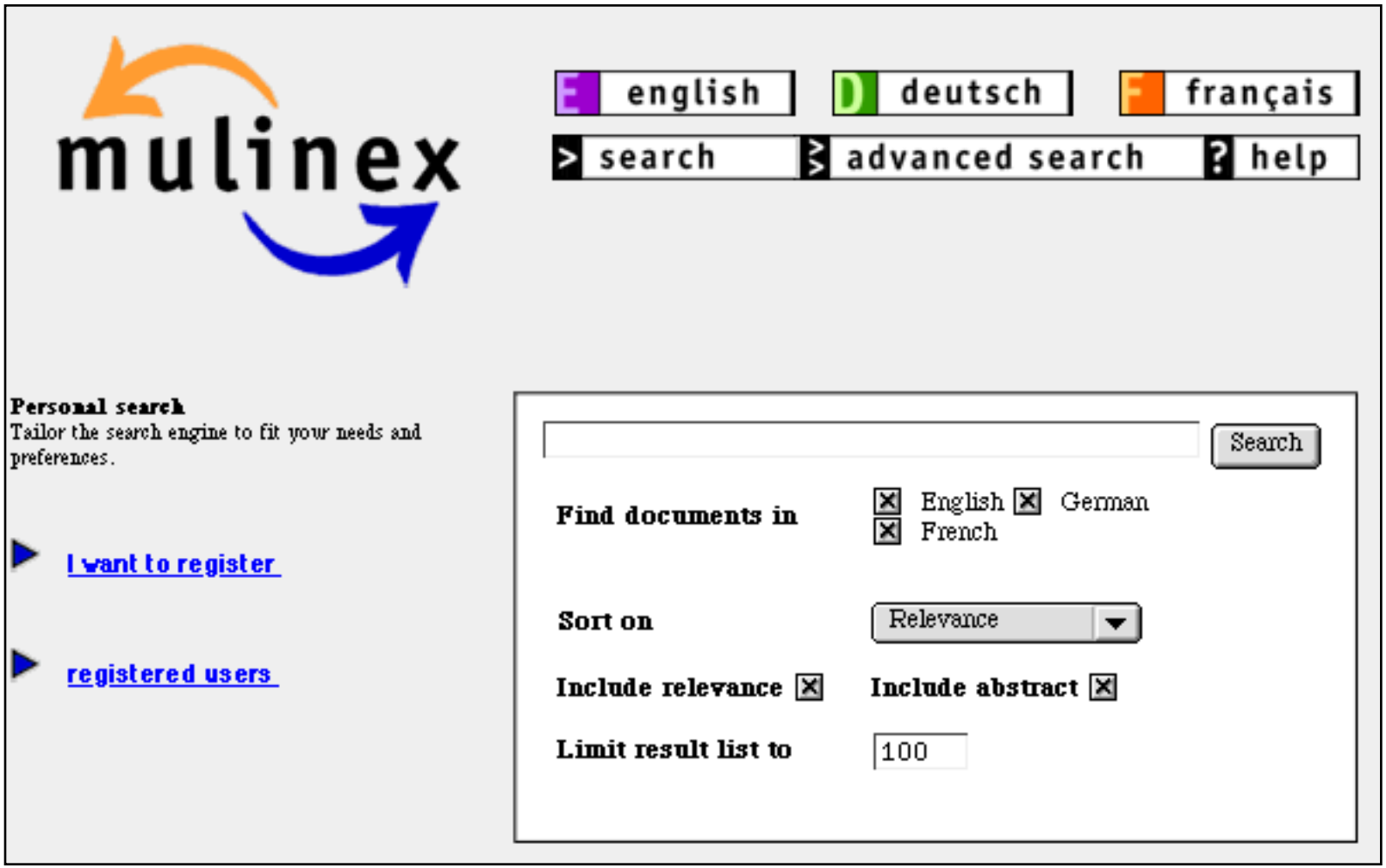
...

BEISPIEL: INFORMATIONSEXTRAKTION (2)

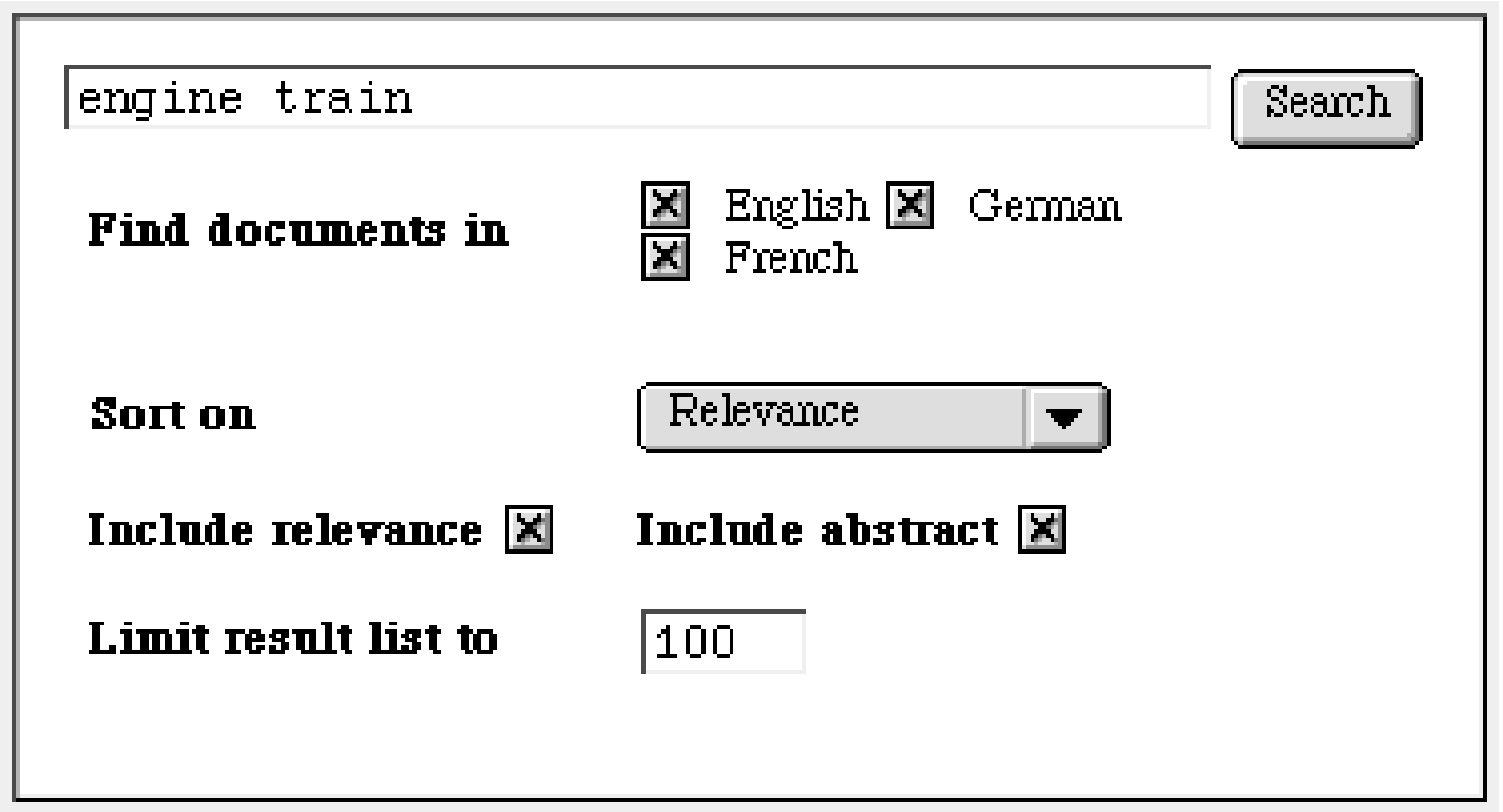
Firma	96Q4	1996	97Q1	97Q2	97Q3	97Q4	1997	Diff
ComSoft		120Mio					110Mio	
Trade Consult				30 Mio	42,5Mio			
Z&M					71,0Mio			

Crosslingual Information Retrieval (CLIR)





Suchschlüssel-Eingabe



Anfrageassistent

English query terms	**German** translations	**French** translations
engine	☒ 1 Motor ☒ 1 Maschine ☒ 1 Triebwerk ☒ 1 Lokomotive	☒ 1 locomotive ☒ 1 machine ☒ 1 moteur
train	☒ 1 dressieren ☒ 1 trainieren ☒ 1 Zug ☒ 1 Eisenbahn	☒ 1 rame ☒ 1 entraîner ☒ 1 train ☒ 1 dresser ☒ 1 suite ☒ 1 escort ☒ 1 cortège ☒ 1 clique