Language Technology I

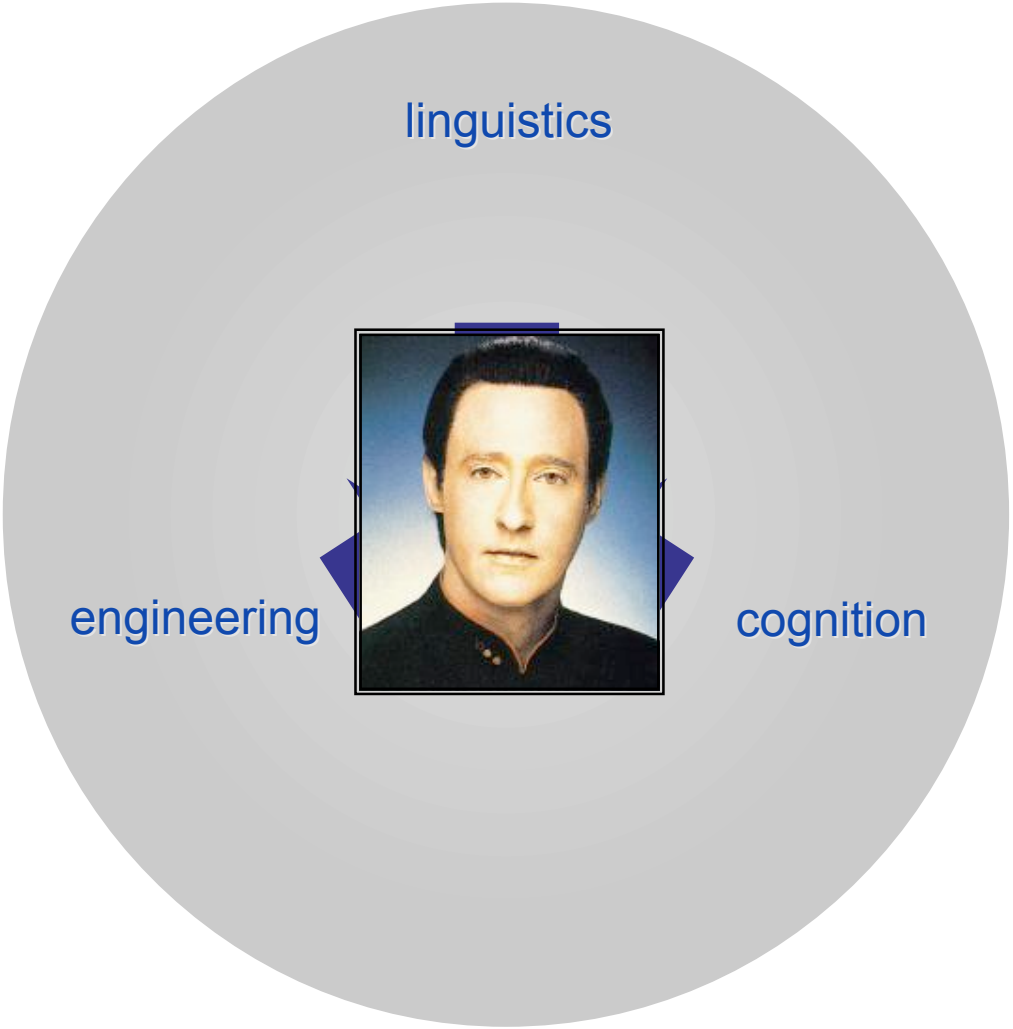
2006/07

Hans Uszkoreit

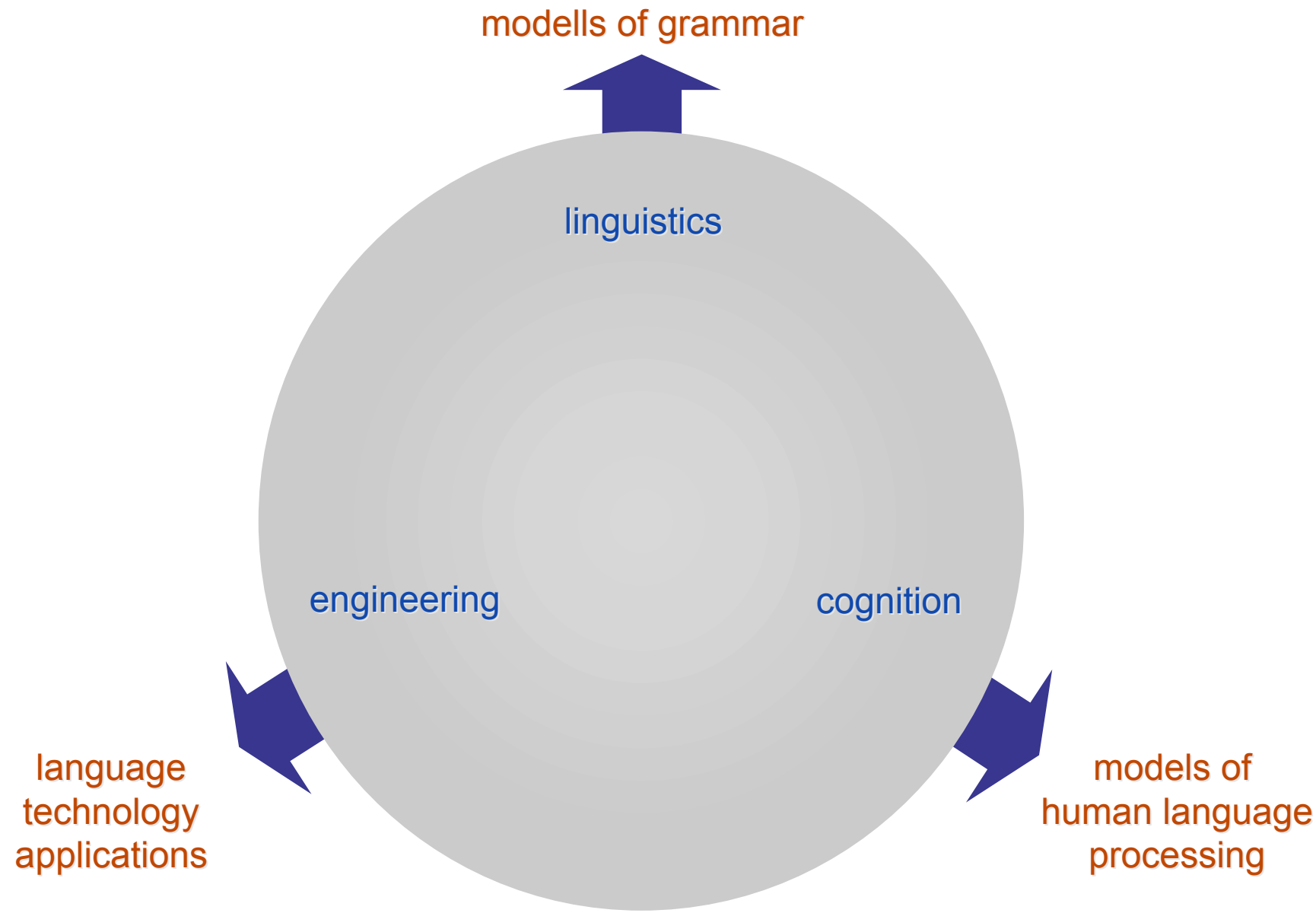
Universität des Saarlandes
and
German Research Center for Artificial Intelligence (DFKI)

- ☆ What is Language Technology
- ☆ Some Selected Technologies
- ☆ Some Selected Applications
- ☆ Information Extraction
- ☆ Cross-Linguistic Information Retrieval
- ☆ Email Management
- ☆ Language Checking

Motivations



Motivationen



Technology: methods and techniques that together enable some application.

In real life usage of the word there is a continuum between methods and applications.

method/technique	finite state transduction
component technology	tokenizer
technology	named entity recognition
	high precision text indexing
application	concept based search engine

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

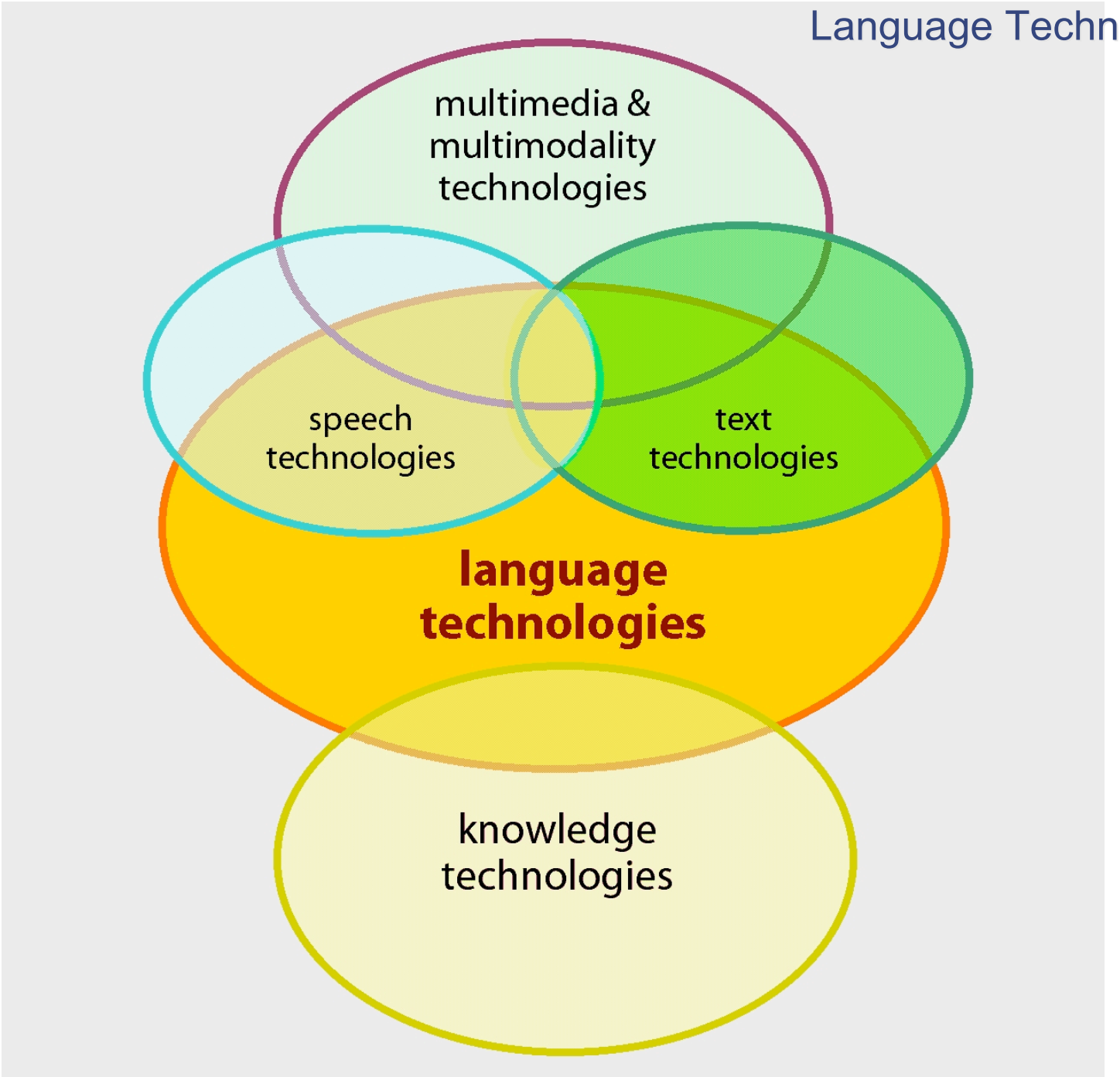
Situatedness: sensitivity to context, location, time, plans

Type of linguality: monolingual, multilingual, translingual

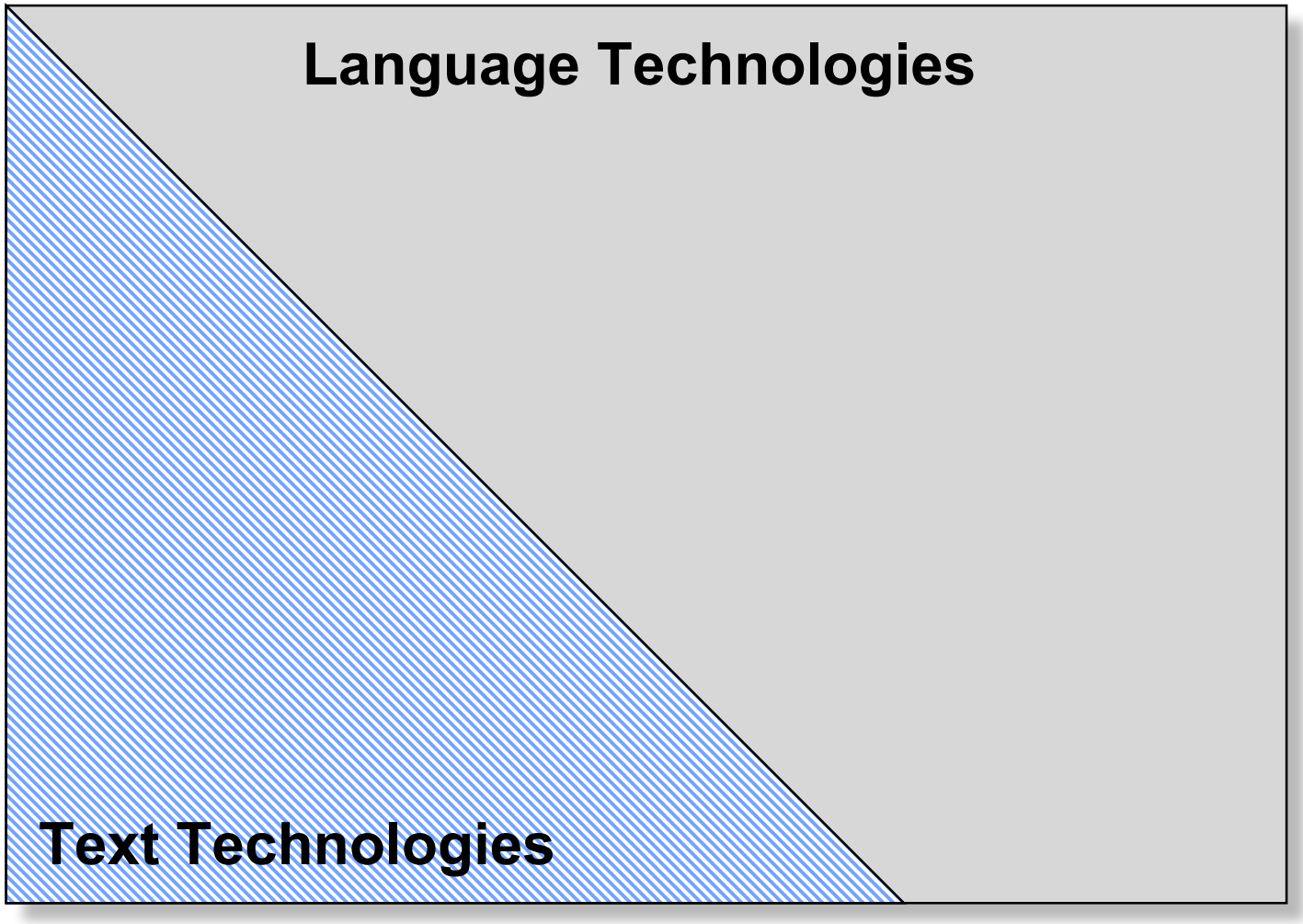
Type of processing: Categorization, summarization, extraction, understanding, translating, responding

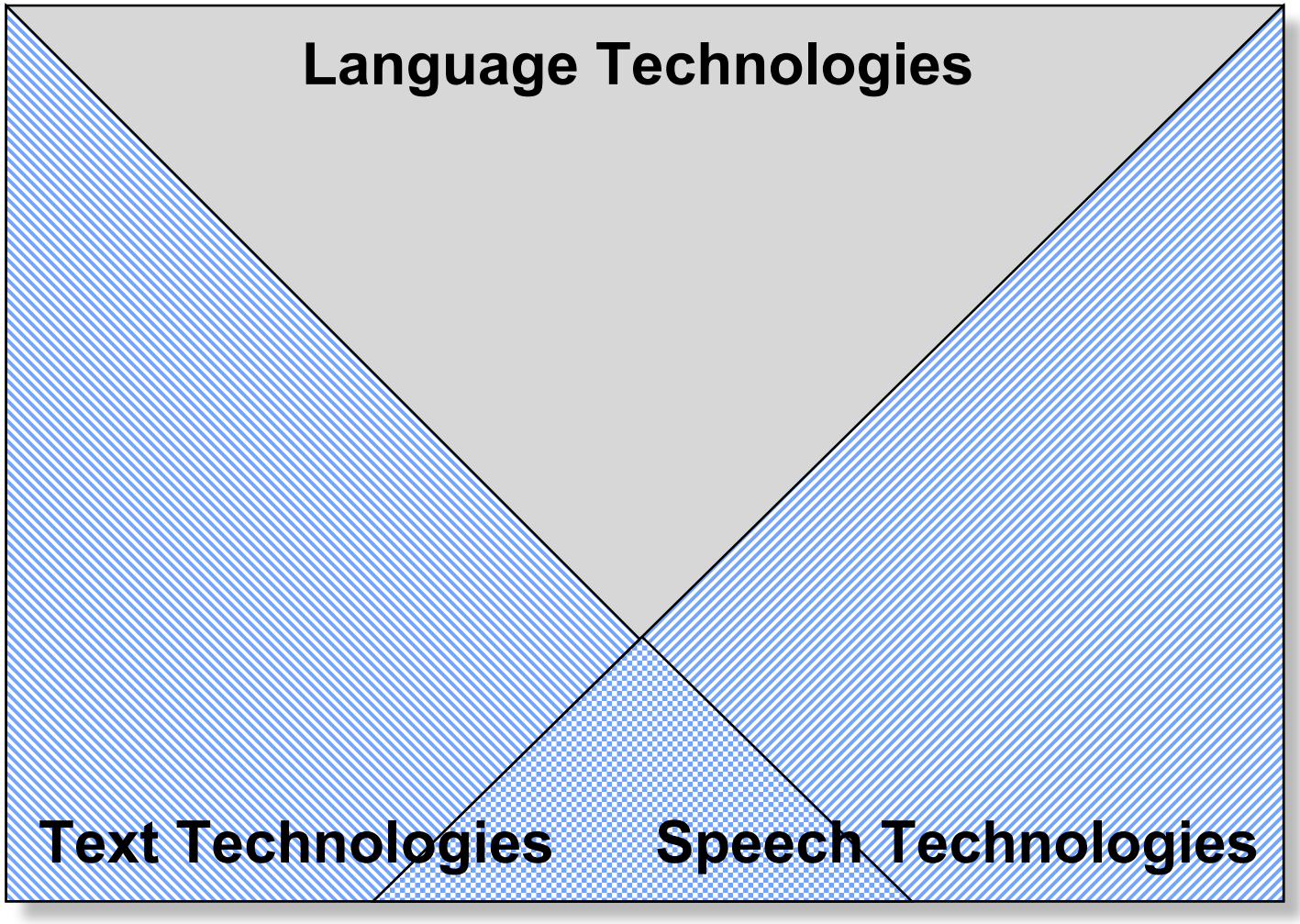
Level of linguistic description: phonology, morphology, syntax, semantics,pragmatics

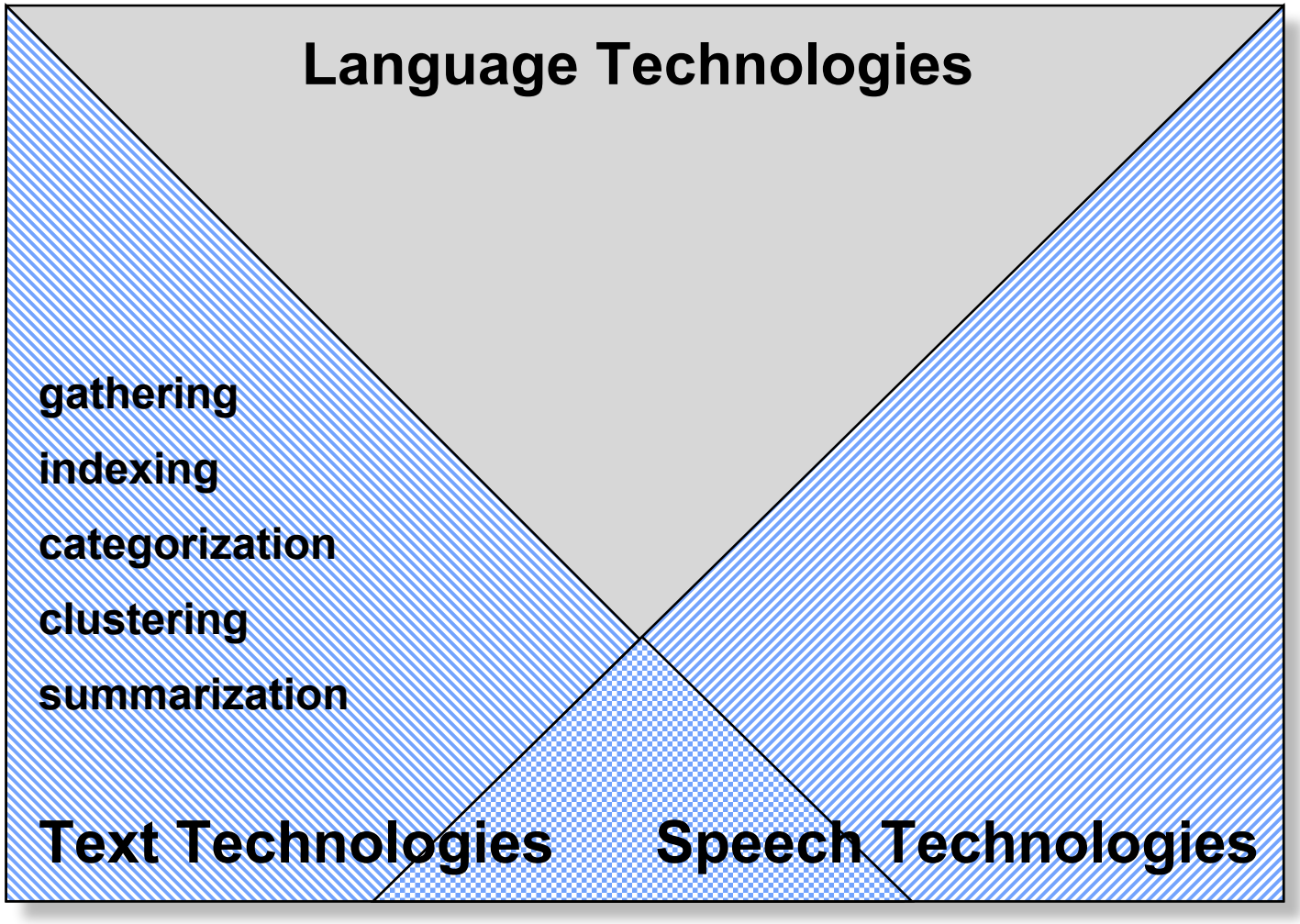
Language Technologies

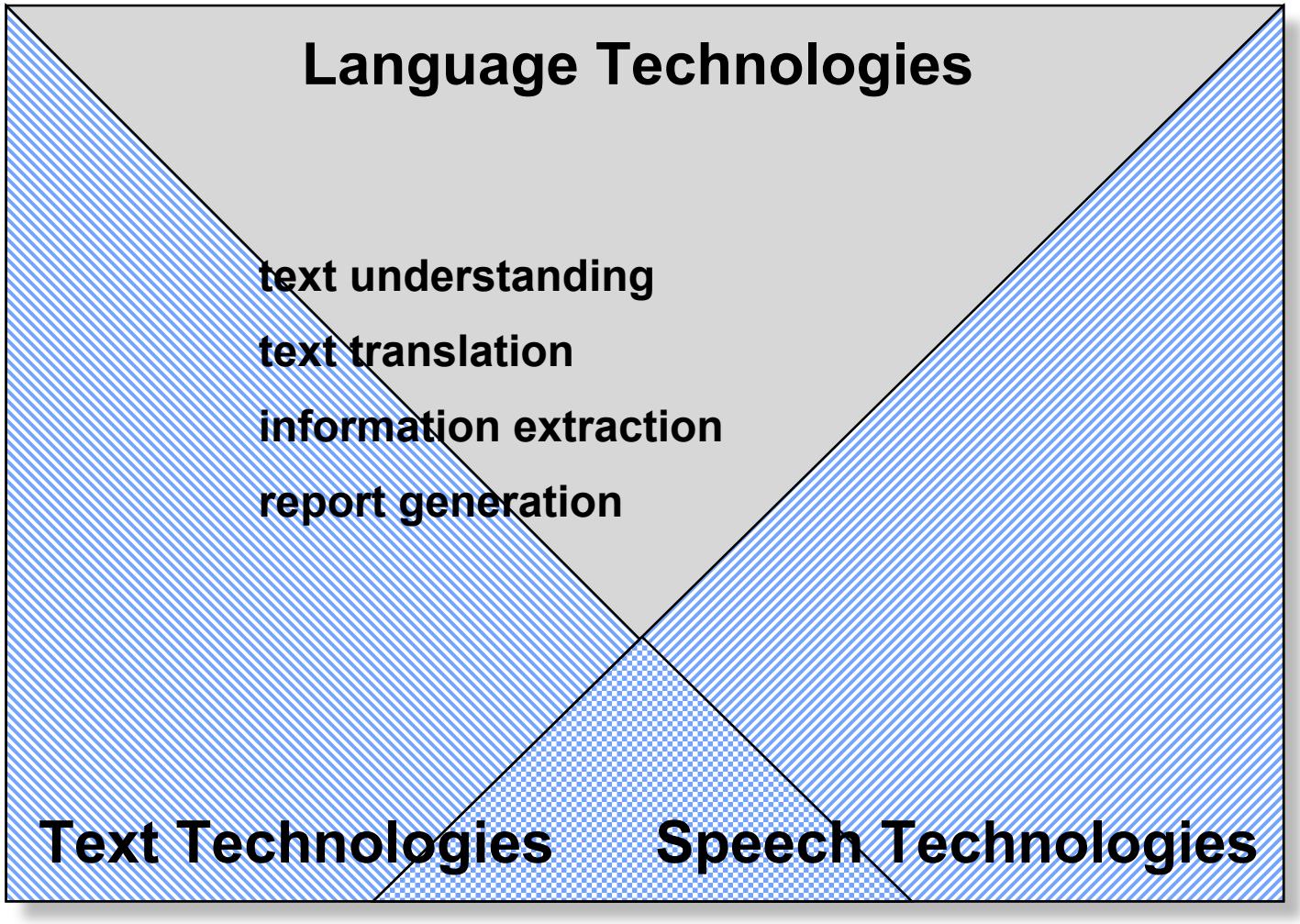


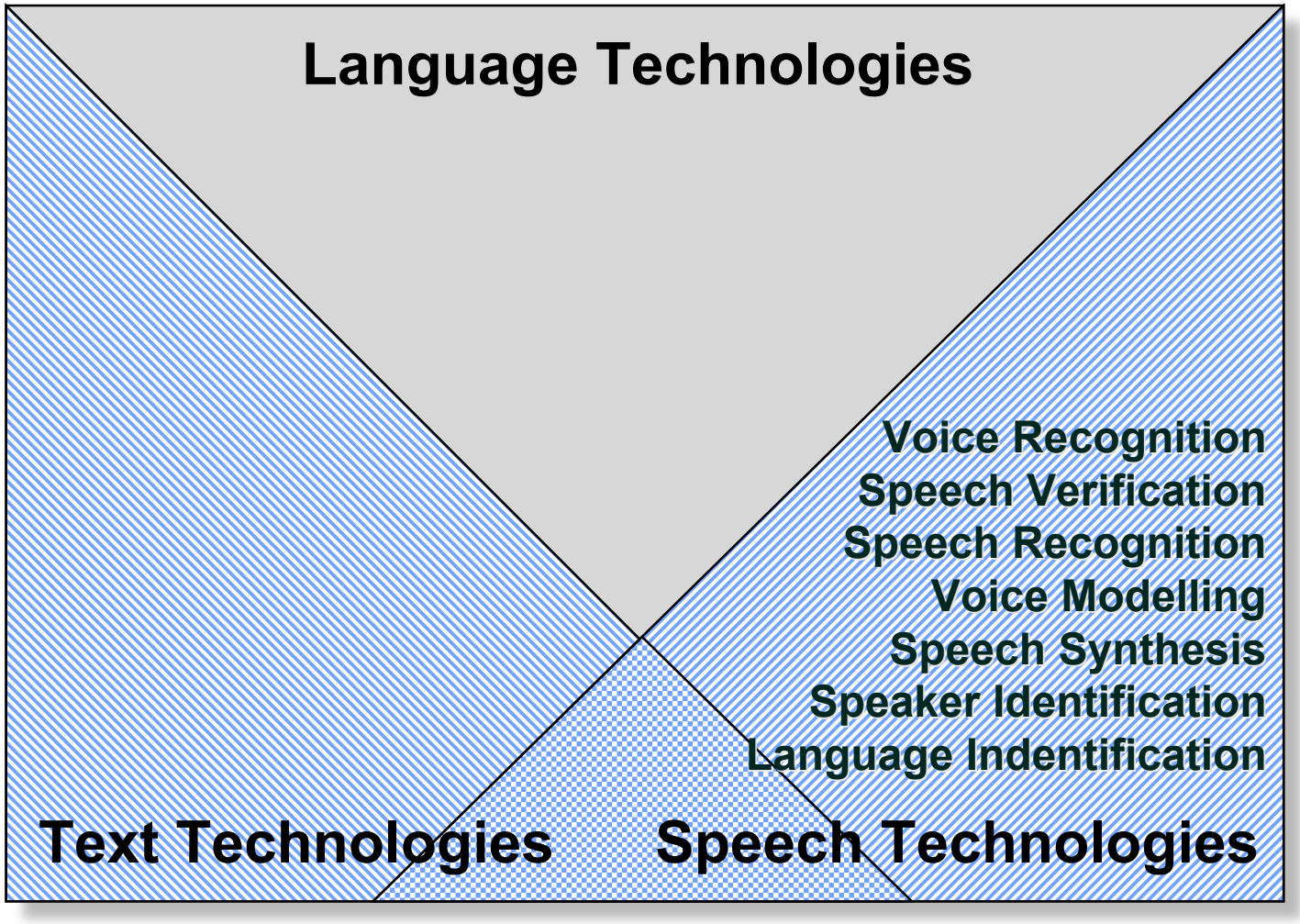
Language Technologies

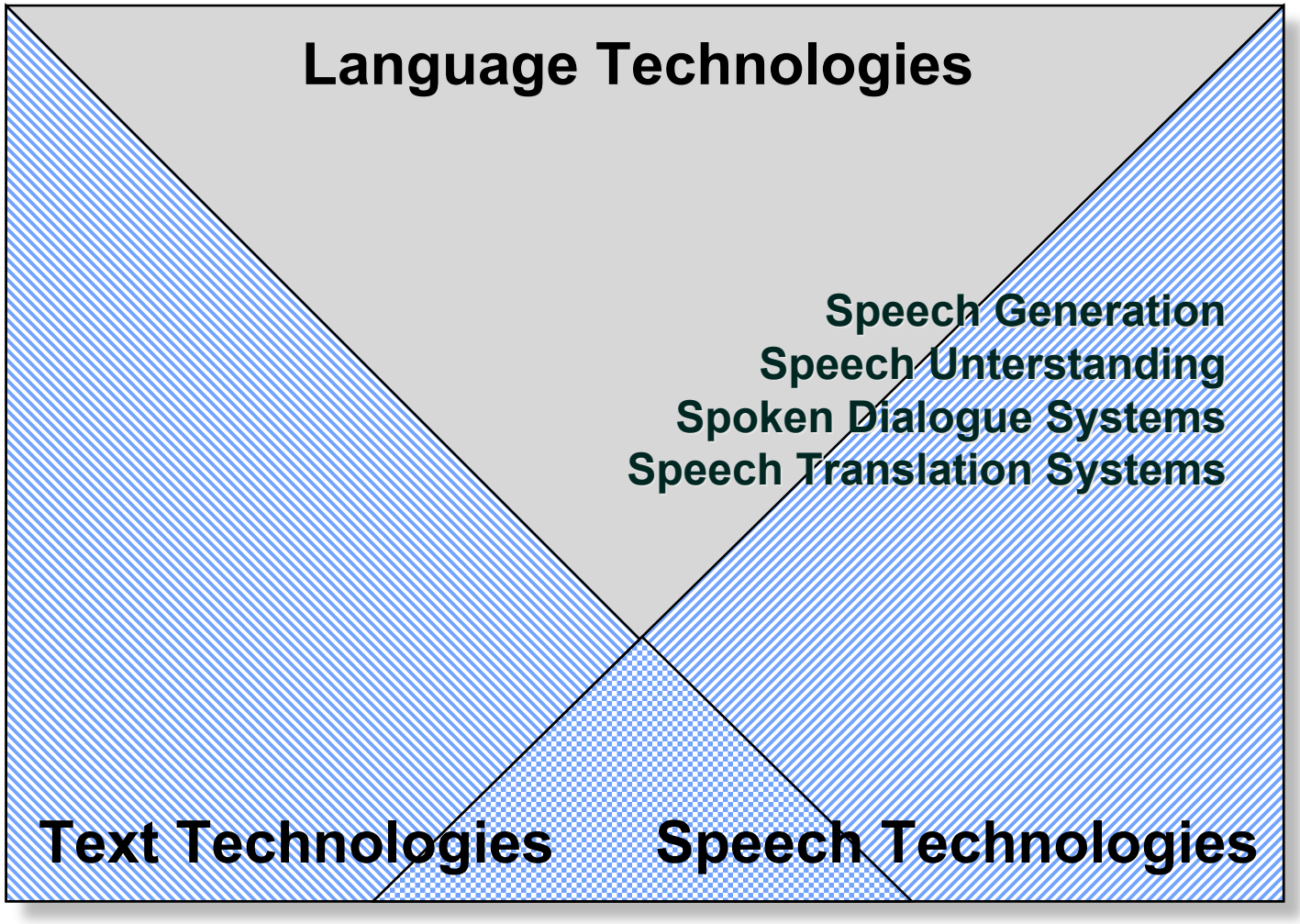


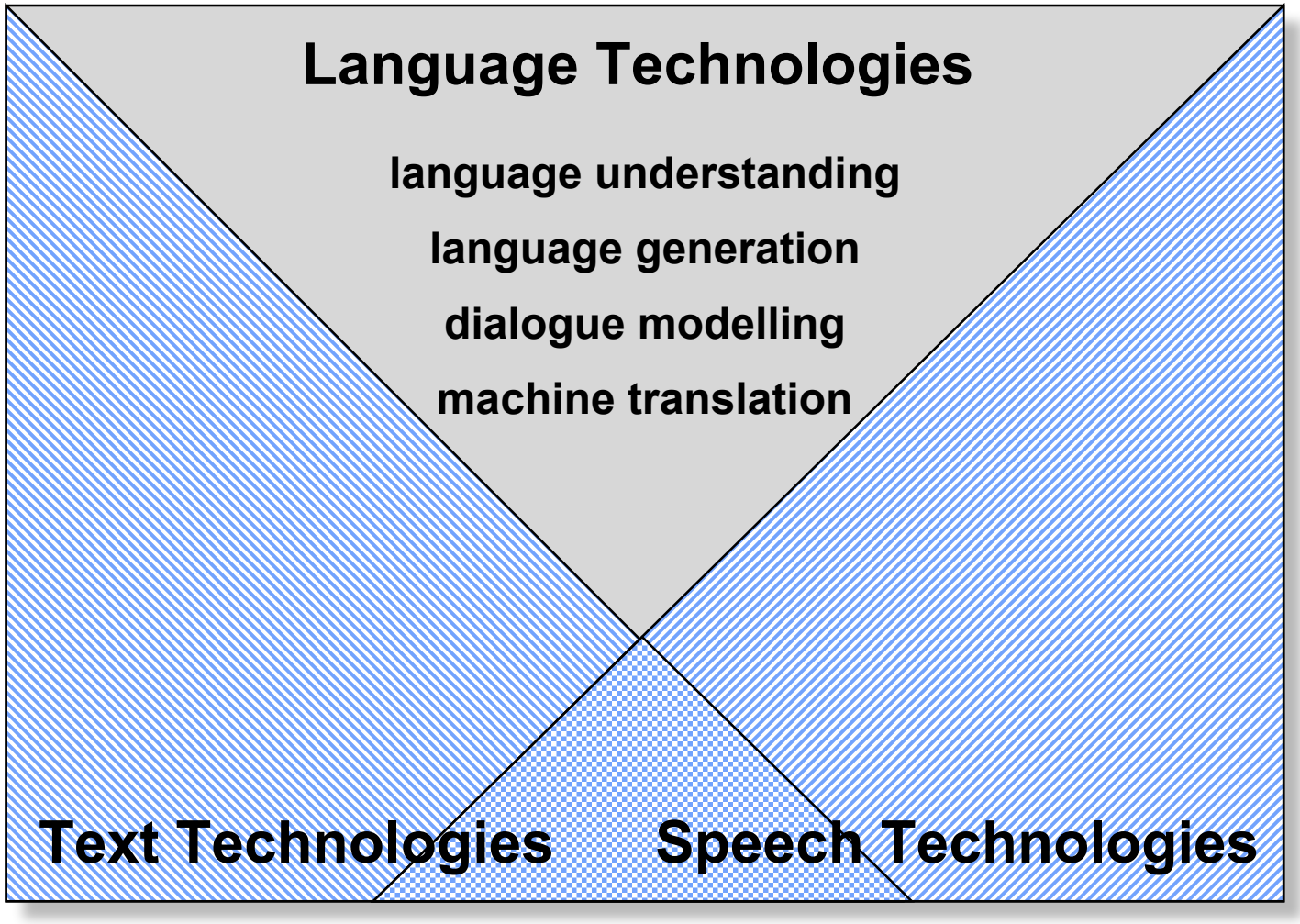




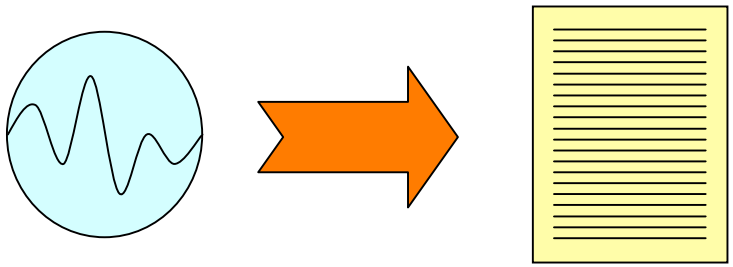





Language Technologies
Speech Generation
Speech Unterstanding
Spoken Dialogue Systems
Speech Translation Systems
Text Technologies
Speech Technologies

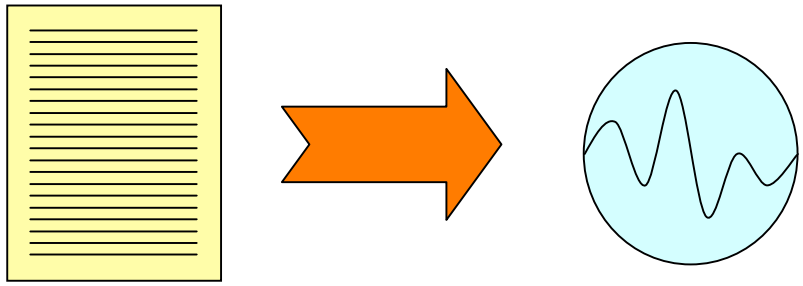


Spoken language is recognized and transformed in into text as in dictation systems, into commands as in robot control systems, or into some other internal representation.

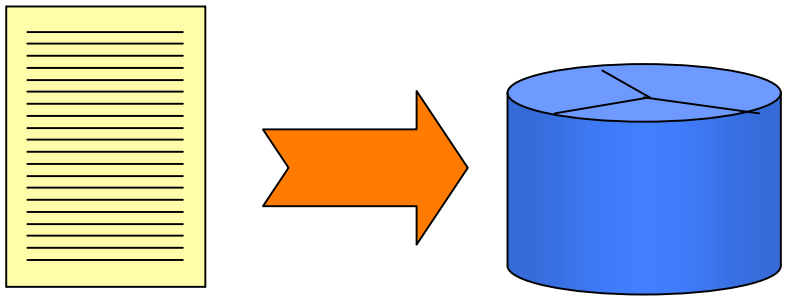


(also Speech Generation)

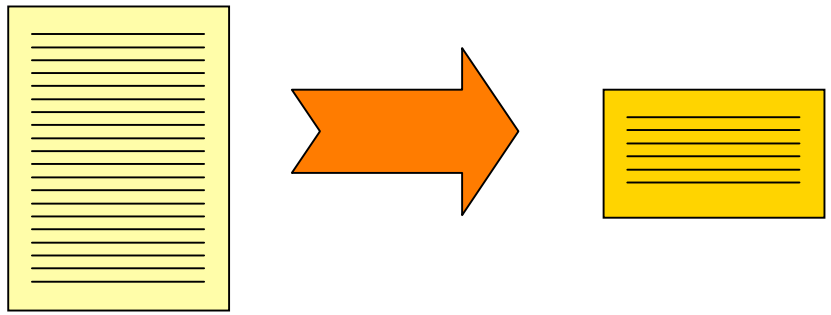
Utterances in spoken language are produced from text (text-to-speech systems) or from internal representations of words or sentences (concept-to-speech systems)



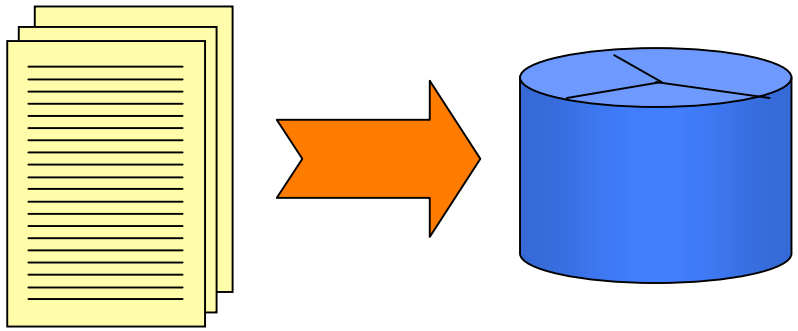
This technology assigns texts to categories. Texts may belong to more than one category, categories may contain other categories. Filtering is a special case of categorization with just two categories.



The most relevant portions of a text are extracted as a summary. The task depends on the needed lengths of the summaries. Summarization is harder if the summary has to be specific to a certain query.

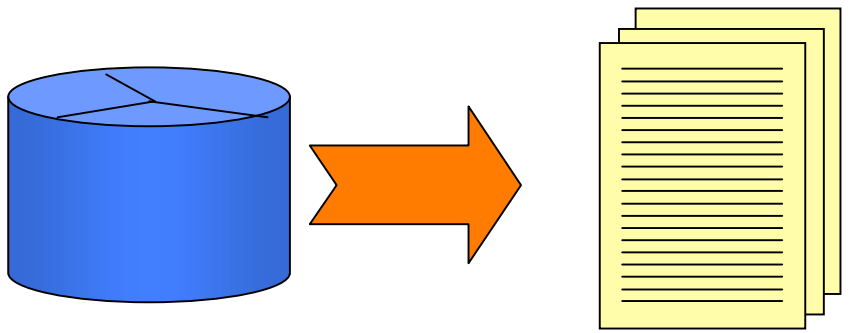


As a precondition for document retrieval, texts are
are stored in an indexed database. Usually a text
is indexed for all word forms or – after lemmatization –
for all lemmas. Sometimes indexing is combined
with categorization and summarization.

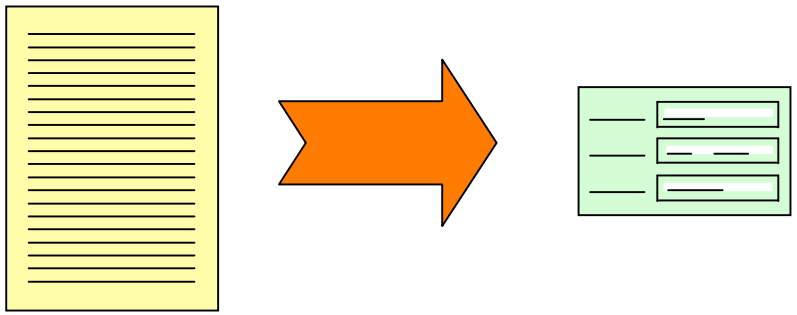


Texts are retrieved from a database that best match a given query or document. The candidate documents are ordered with respect to their expected relevance.

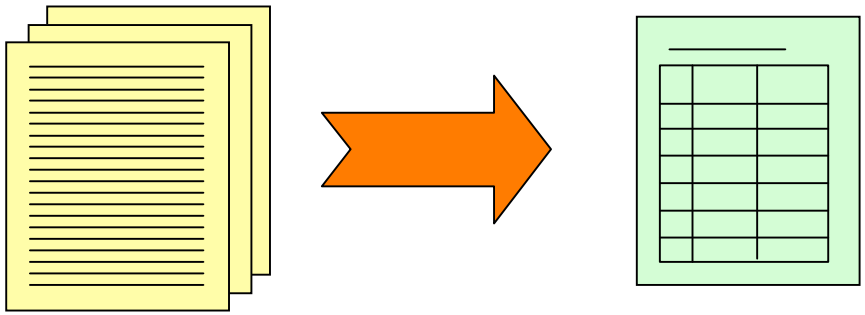
Indexing, categorization, summarization and retrieval are often subsumed under the term information retrieval.



Relevant information pieces of information are discovered and marked for extraction. The extracted pieces can be: the topic, named entities such as company, place or person names, simple relations such as prices, destinations, functions etc. or complex relations describing accidents, company mergers or football matches.

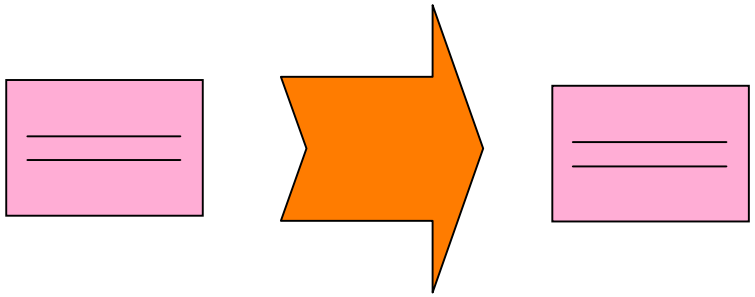


Extracted pieces of information from several sources are combined in one database. Previously undetected relationships may be discovered.

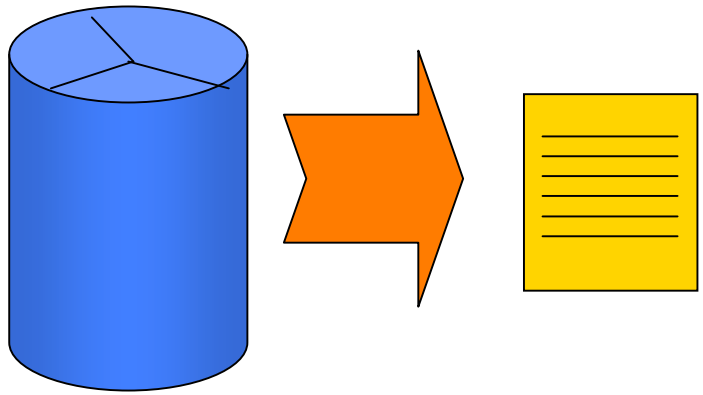


Question Answering

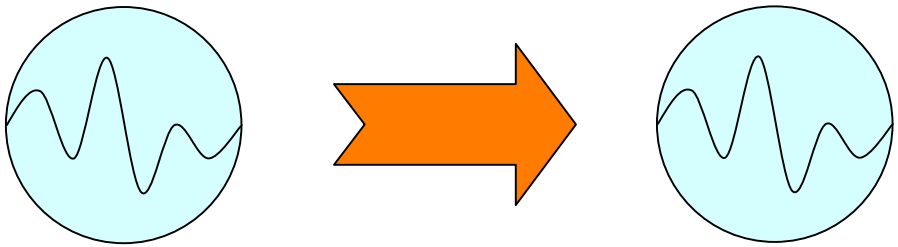
Natural language queries are used to access information in a database. The database may be a base of structured data or a repository of digital texts in which certain parts have been marked as potential answers.



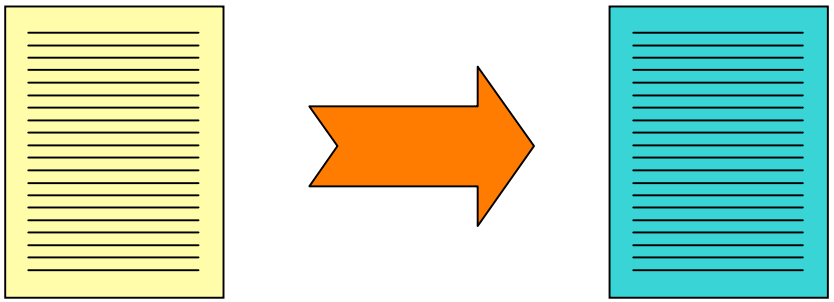
A report in natural language is produced that describes the essential contents or changes of a database. The report can contain accumulated numbers, maxima, minima and the most drastic changes.



The system can carry out a dialogue with a human user in which the user can solicit information or conduct purchases, reservations or other transactions.



Technologies that translate texts or assist human translators. Automatic translation is called machine translation. Translation memories use large amounts of texts together with existing translations for efficient look-up of possible translations for words, phrases and sentences.



Generic CS Methods

Programming languages, algorithms for generic data types, and software engineering methods for structuring and organizing software development and quality assurance.

Specialized Algorithms

Dedicated algorithms have been designed for parsing, generation and translation, for morphological and syntactic processing with finite state automata/transducers and many other tasks.

Nondiscrete Mathematical Methods

Statistical techniques have become especially successful in speech processing, information retrieval, and the automatic acquisition of language models. Other methods in this class are neural networks and powerful techniques for optimization and search.

Logical and Linguistic Formalisms

For deep linguistic processing, constraint based grammar formalisms are employed. Complex formalisms have been developed for the representation of semantic content and knowledge.

Linguistic Knowledge

Linguistic knowledge resources for many languages are utilized: dictionaries, morphological and syntactic grammars, rules for semantic interpretation, pronunciation and intonation.

Corpora and Corpus Tools

Large collections of application-specific or generic collections of spoken and written language are exploited for the acquisition and testing of statistical or rule-based language models.

Models of Cognitive Systems and their Components

The interaction of perception, knowledge, reasoning and action including communication is modelled in cognitive psychology. Such models can be consulted or employed for the design of language processing systems. Formalized models of components such as memory, reasoning and auditive perception are also often utilized for models of language processing.

Empirical methods fromn Experimental Psychology

Since cognitive psychology investigates the intelligent behavior of human organisms, many methods have been developed for the observation and empirical analysis of language production and comprehension. Such methods can be extremely useful for building computer models of human language processing (Examples: "Wizard of Oz Experiments" and measurements of syntactic and semantic processing complexity.

95%-98%	Correct recognition of word categories (part-of-speech-tagging)
85%-98%	recognition of names of people, companies, places, products (named-entity-recognition)
95%	statistical recognition of major phrases (HMM chunk parsing)
91%	parsing of newspaper texts by statistically trained parsers (probibilistic context free parsing)
40%-60%	deep parsing of newspaper texts (HPSG or LFG parsing with large lexicon)

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

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

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)

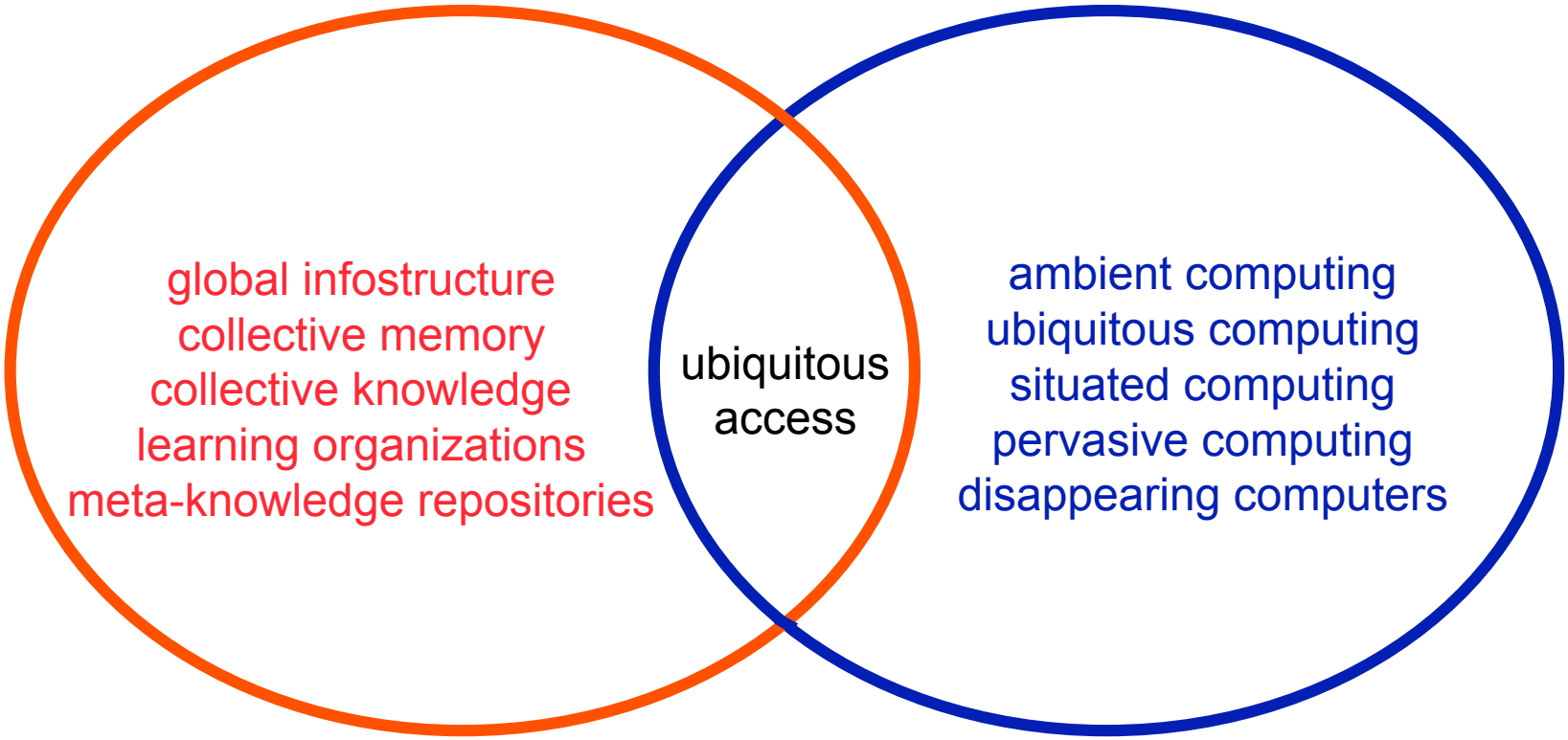
Complex Information Extraction (Ternary+ Relations)

Dense Associative Hyperlinking

Concept-Based Information Retrieval

Text Understanding

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



personalization
adaptation
learning