Language Technology I

Introduction

Stephan Busemann

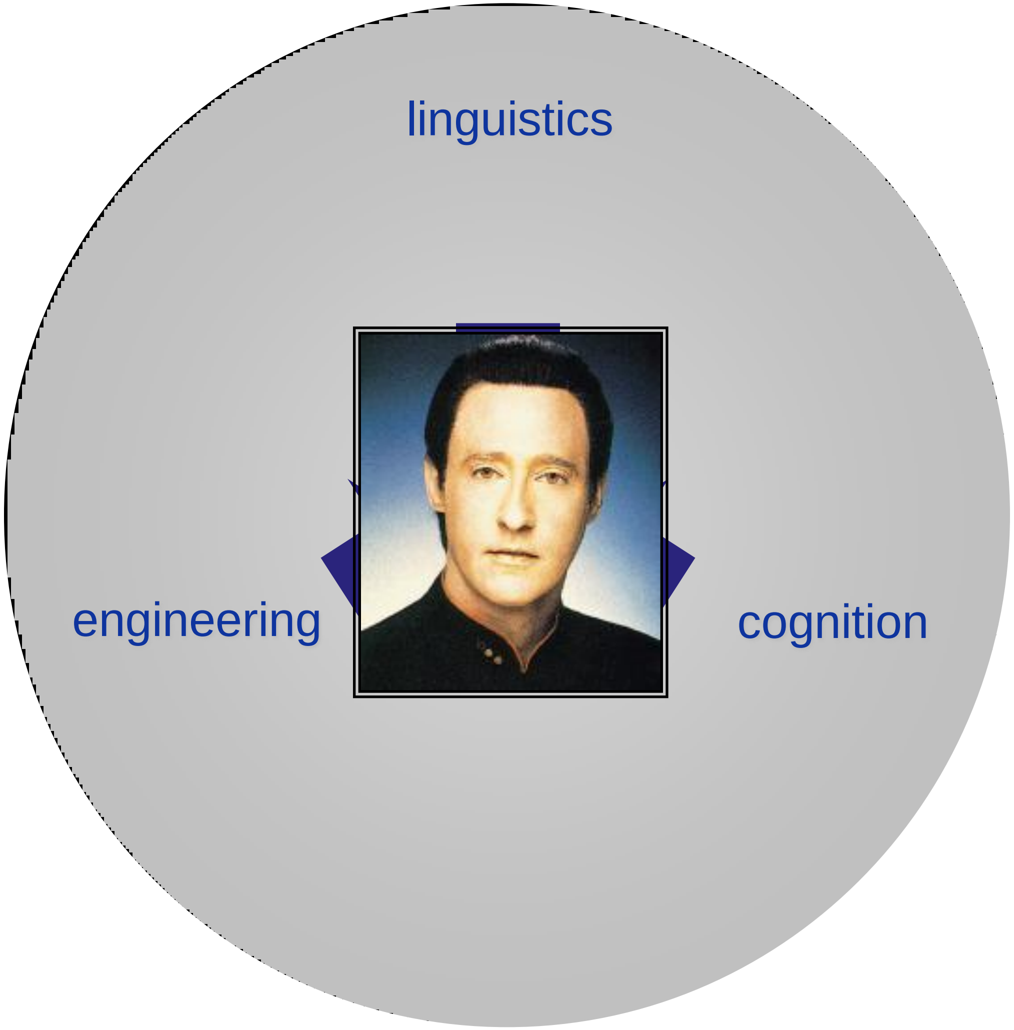
(Slides based on a set by Hans Uszkoreit)

German Research Center for Artificial Intelligence
(DFKI GmbH)

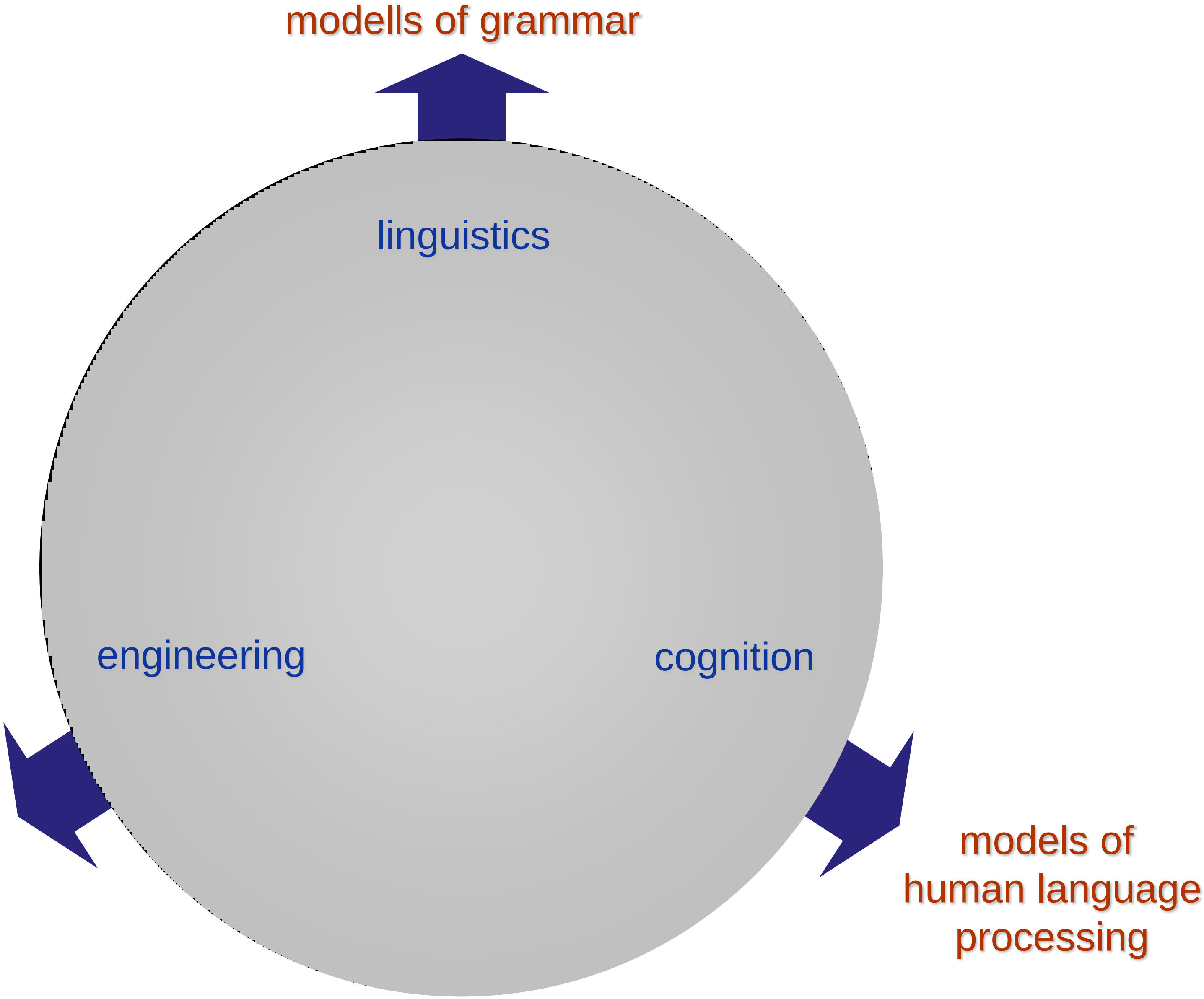
Overview

- What is Language Technology?
- Some Selected Technologies
- Methods
- State of the Art
- Maturity of Technologies
- Megatrends

Motivations



Motivations


modells of grammar
linguistics
engineering
cognition
language technology applications
models of human language processing

What is a Technology?

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

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

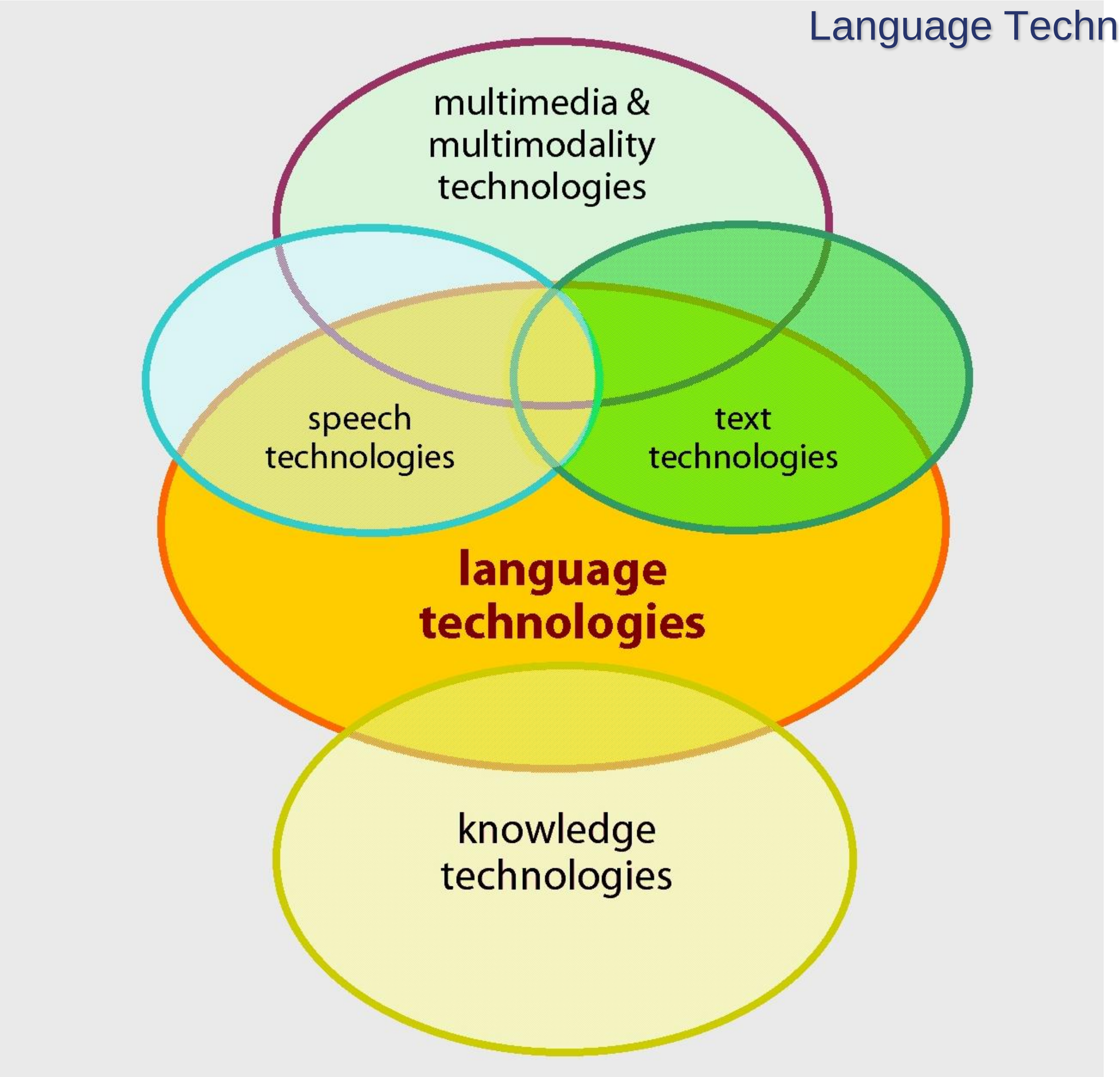
Situatedness: sensitivity to context, location, time, plans

Type of linguality: monolingual, multilingual, translingual

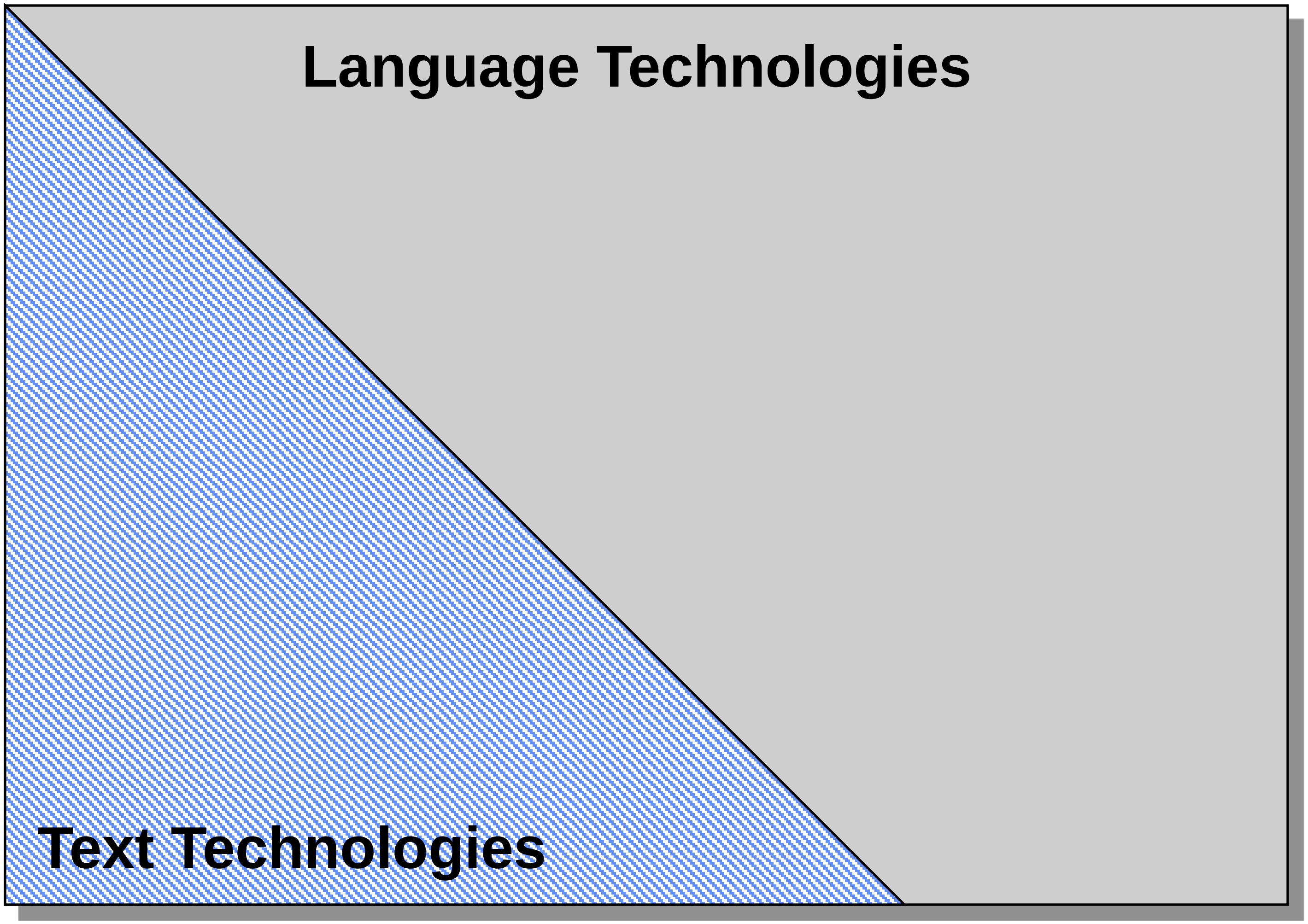
Type of processing: Categorization, summarization, extraction, understanding, translating, responding

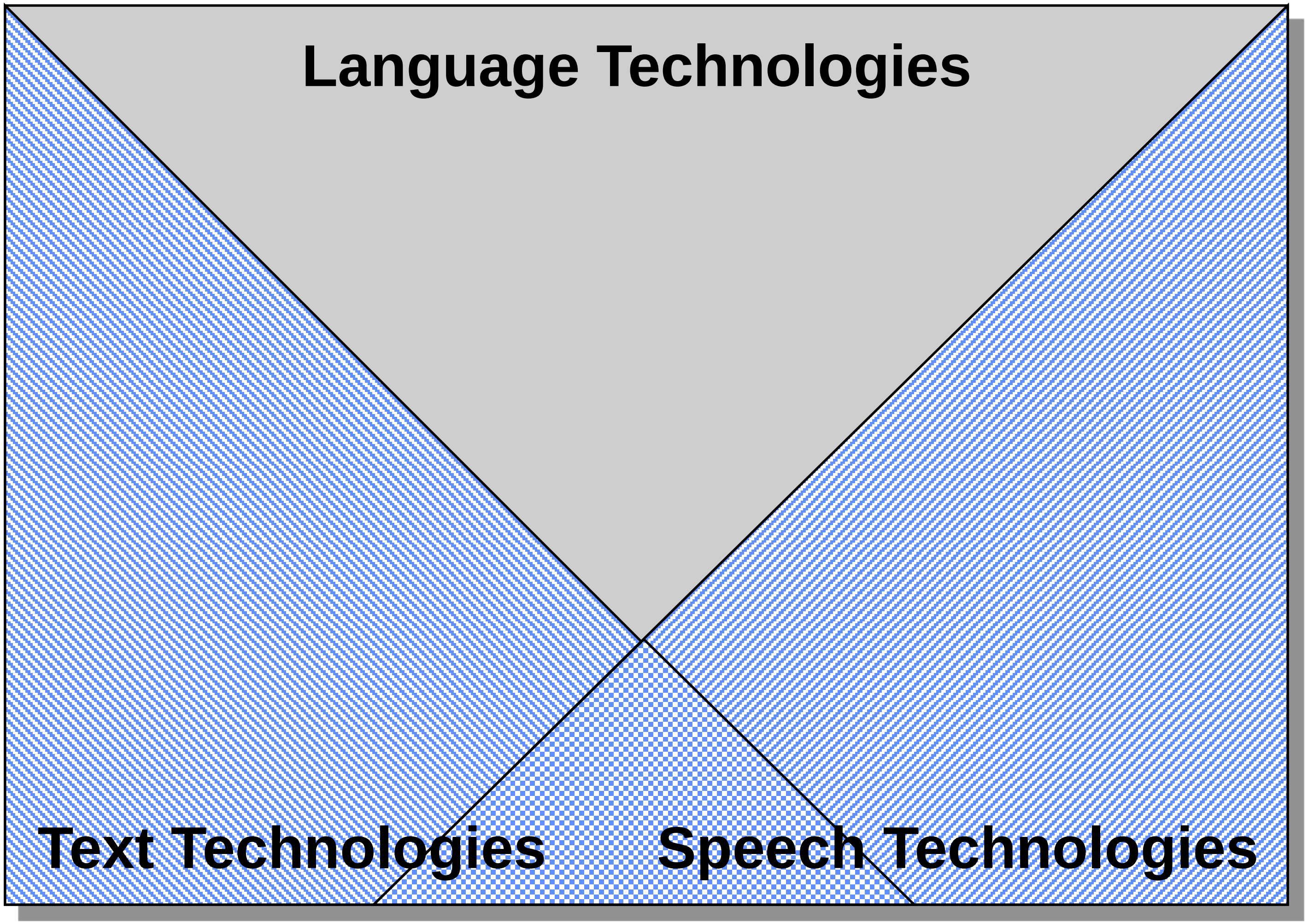
Level of linguistic description: phonology, morphology, syntax, semantics, pragmatics

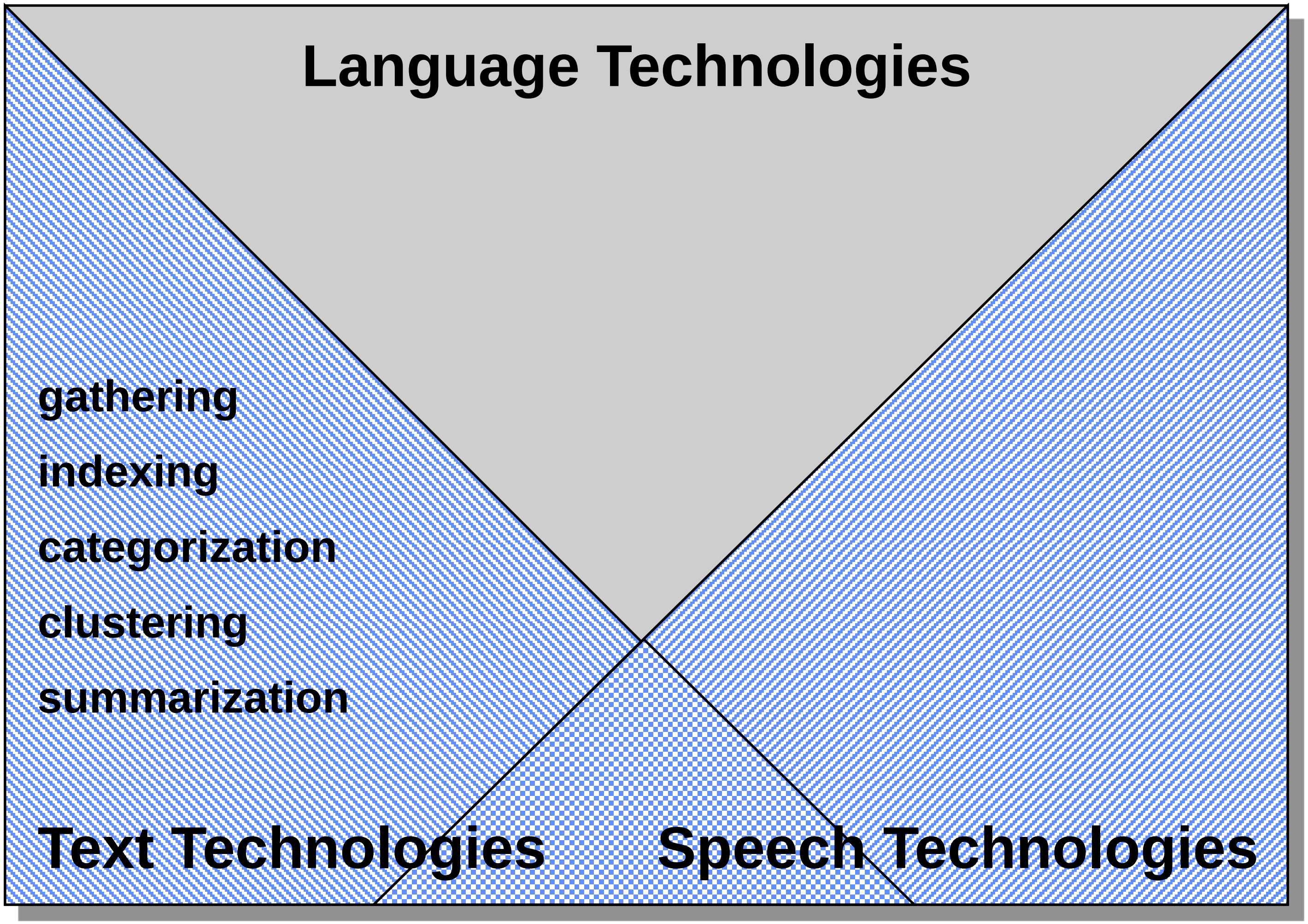
Language Technologies

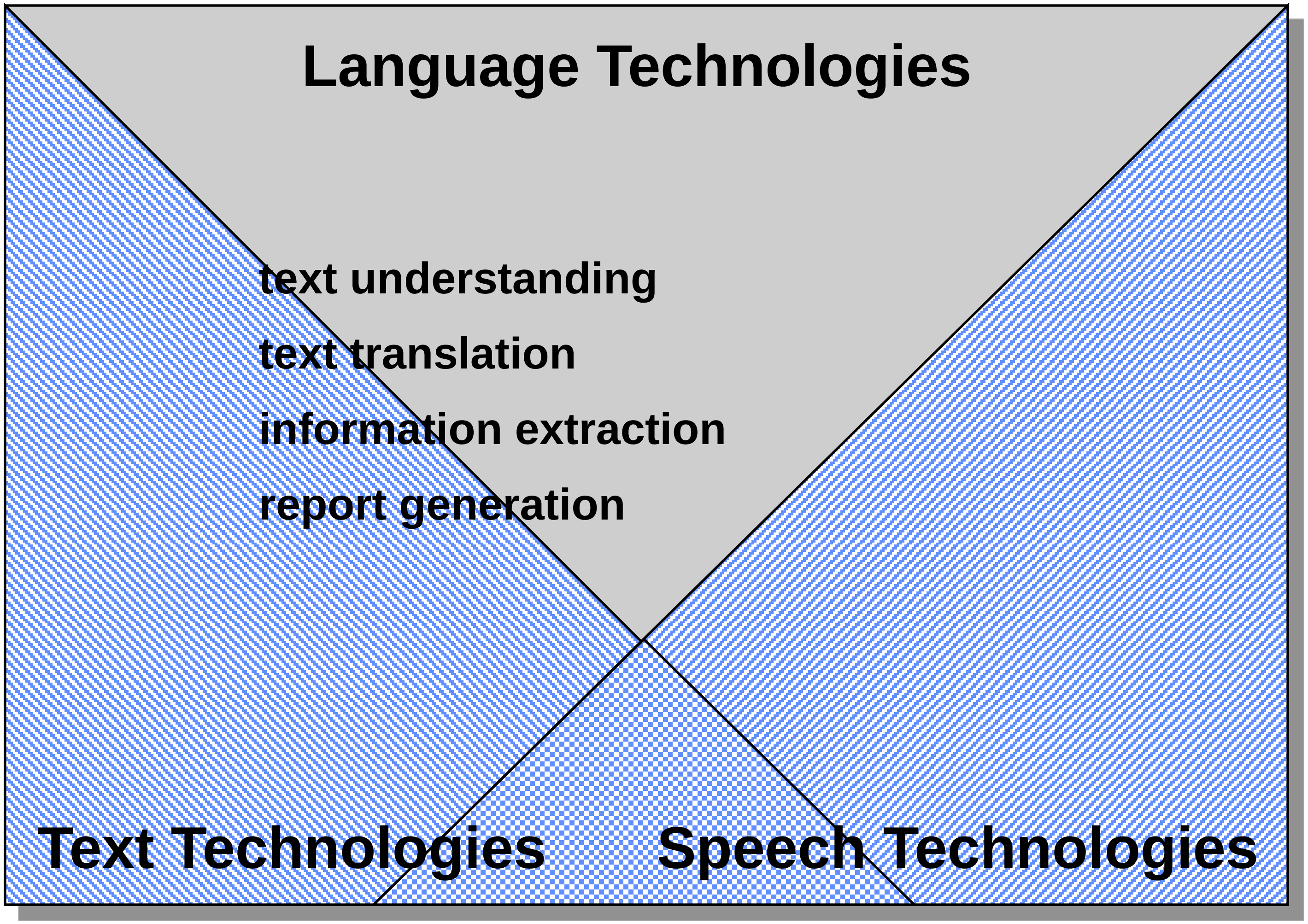


Language Technologies


Language Technologies
Text Technologies


Language Technologies
Text Technologies
Speech Technologies


Language Technologies
gathering
indexing
categorization
clustering
summarization
Text Technologies
Speech Technologies



Language Technologies

text understanding

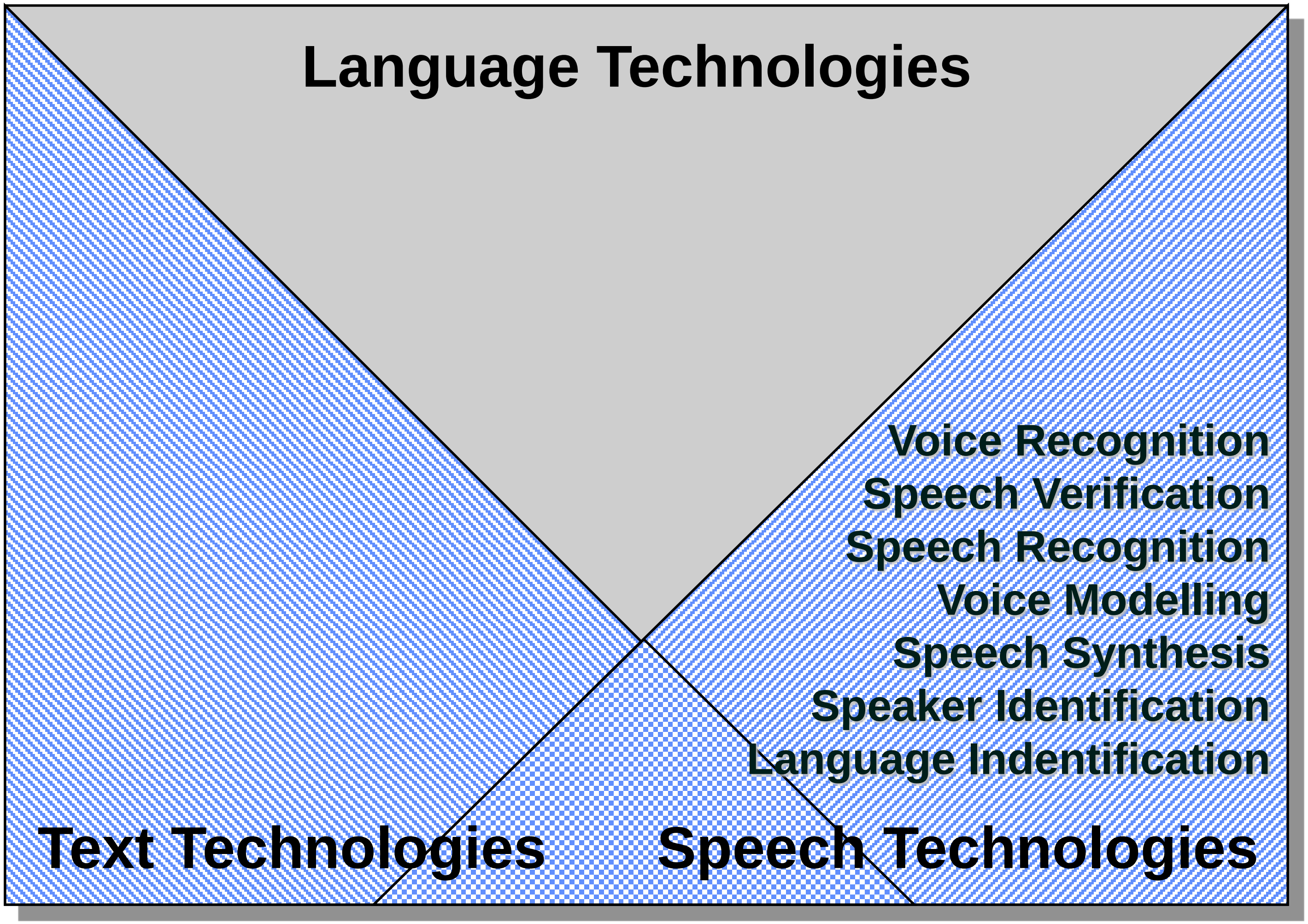
text translation

information extraction

report generation

Text Technologies

Speech Technologies

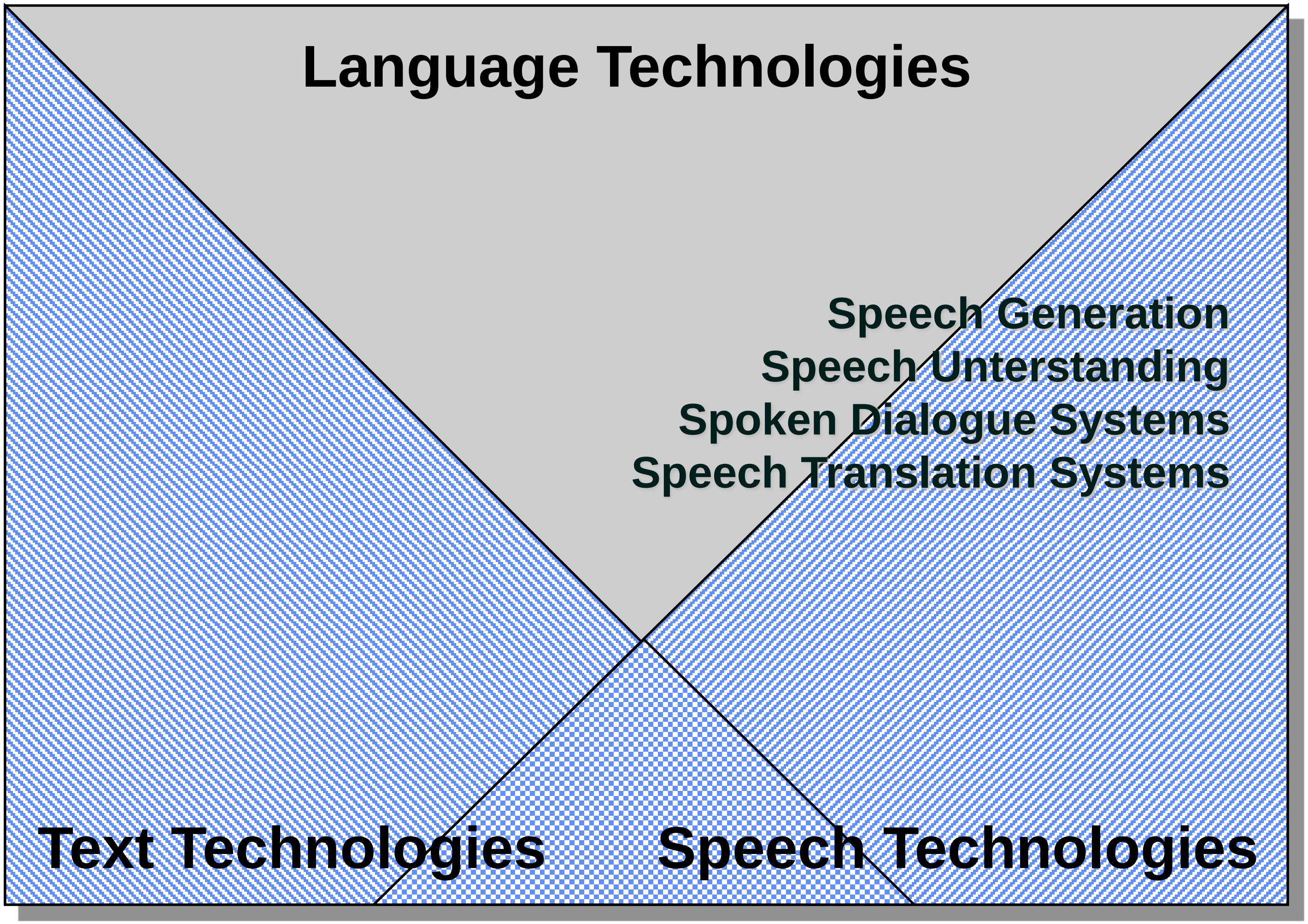


Language Technologies

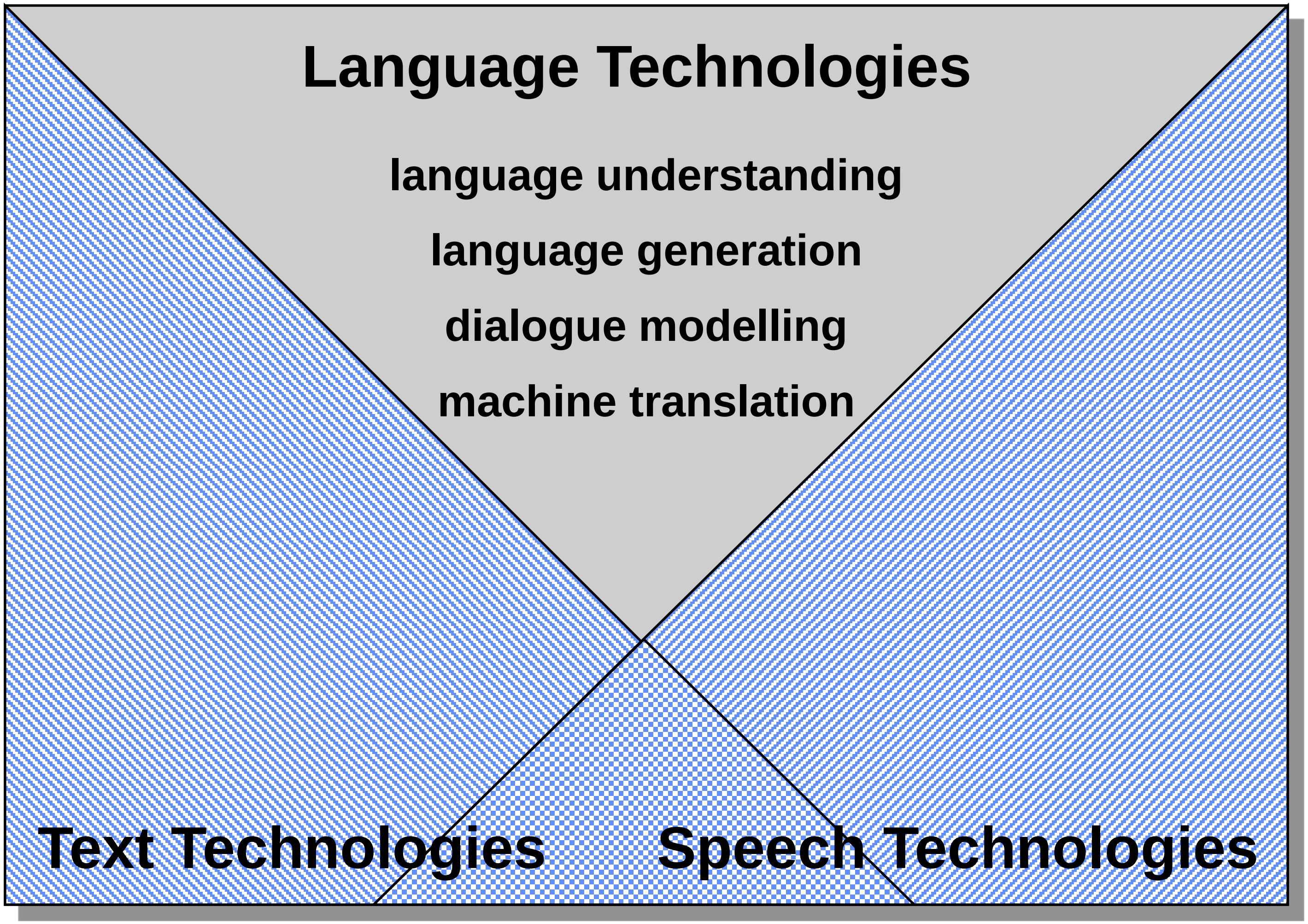
Voice Recognition
Speech Verification
Speech Recognition
Voice Modelling
Speech Synthesis
Speaker Identification
Language Indentification

Text Technologies

Speech Technologies



Speech Generation
Speech Unterstanding
Spoken Dialogue Systems
Speech Translation Systems



Language Technologies

language understanding

language generation

dialogue modelling

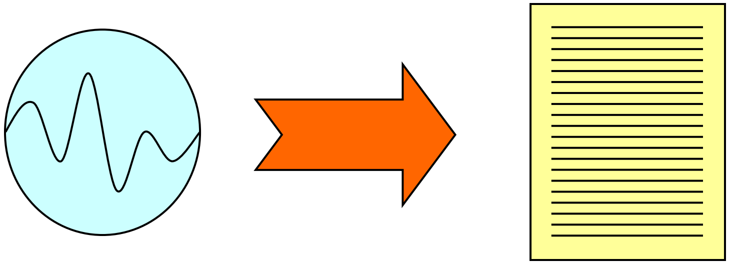
machine translation

Text Technologies

Speech Technologies

Speech recognition

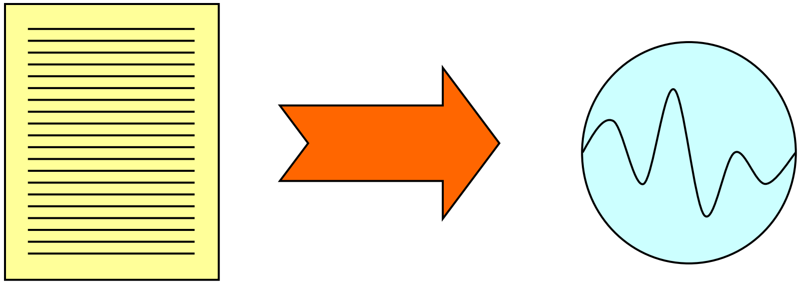
Spoken language is recognized and transformed: into text as in dictation systems, into commands as in robot control systems, or into some other internal representation.



Speech Synthesis

(also Speech Generation)

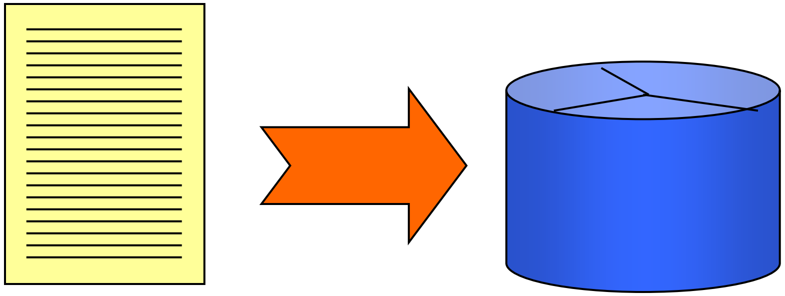
Utterances in spoken language are produced from text (text-to-speech systems) or from internal representations of words or sentences (concept-to-speech systems)



Text Categorization

(also Text Classification)

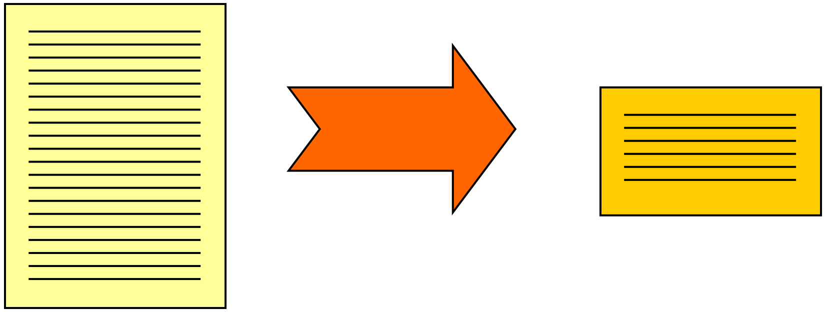
Texts are assigned to given categories. Texts may belong to more than one category, categories may contain other categories. *Filtering* is a special case of categorization with just two categories.



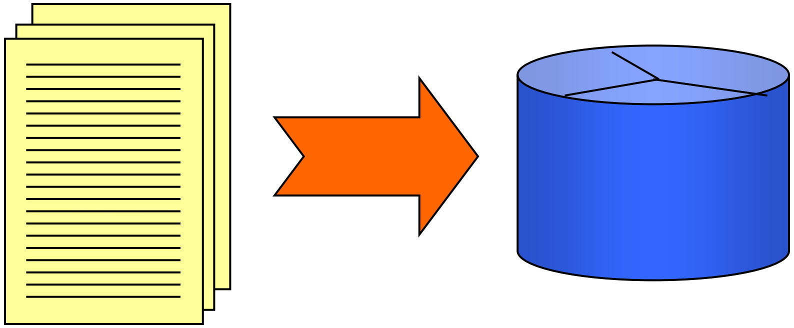
Text Summarization

The most relevant portions of a text are extracted as a summary. Summaries may be limited to the needed length. Summarization may be specific to a certain query and/or a user's interests; summaries may be in the same or a different language.

(Summarization differs from *abstract generation*, which is subsumed under *language generation*)

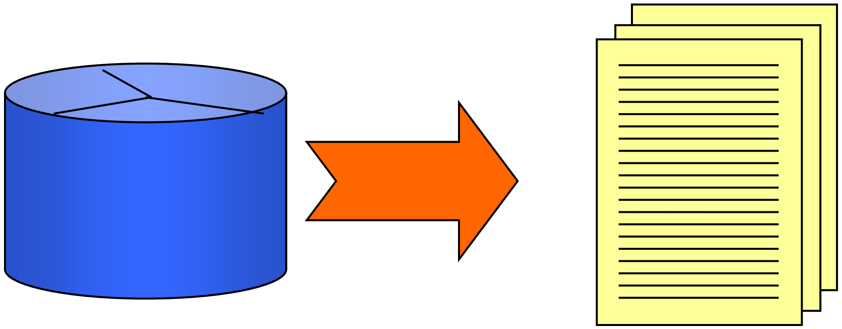


As a precondition for document retrieval, texts are stored in an indexed database. Usually a text is indexed for all word forms or – after lemmatization – for all lemmas. Sometimes indexing is combined with categorization and summarization.

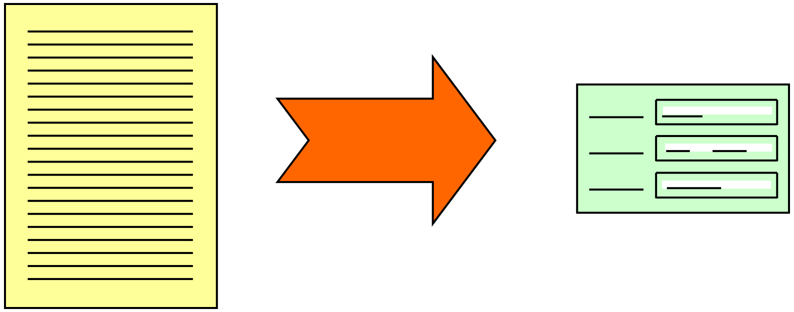


Text Retrieval

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*.

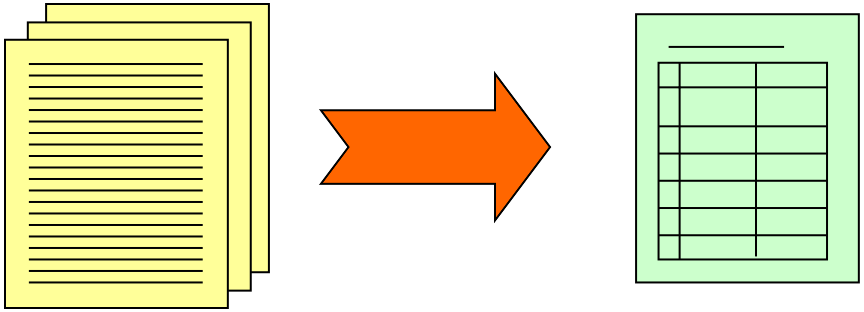


Relevant pieces of information are discovered and marked for extraction. The extracted pieces can be: the topic, named entities such as company, location or person names, simple relations such as prices, destinations, functions etc. or complex relations describing accidents, company mergers or football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



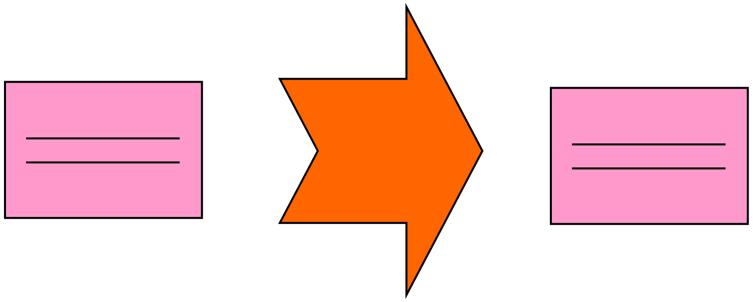
Data Fusion and Text Data Mining

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



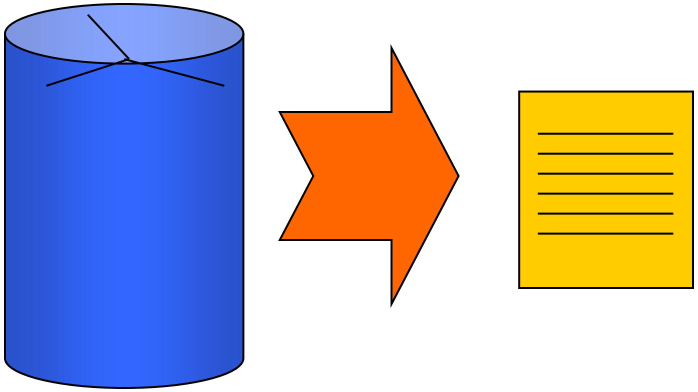
Natural language queries are used to access information in a database. The database may be a base of structured data or a repository of digital texts in which certain parts have been marked as potential answers.

QA on the WWW triggers search engines and exploits their results.

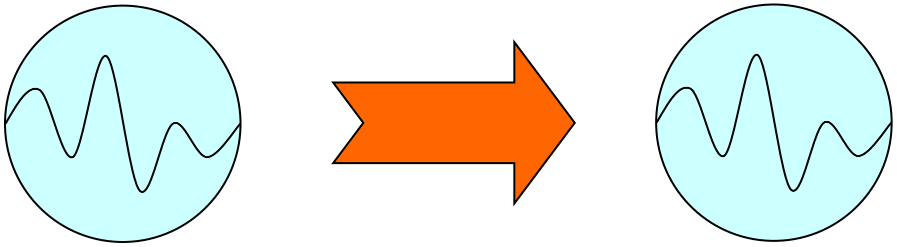


A report in natural language is produced that describes the requested contents or changes of a database. The report can contain accumulated numbers, maxima, minima and the most drastic changes.

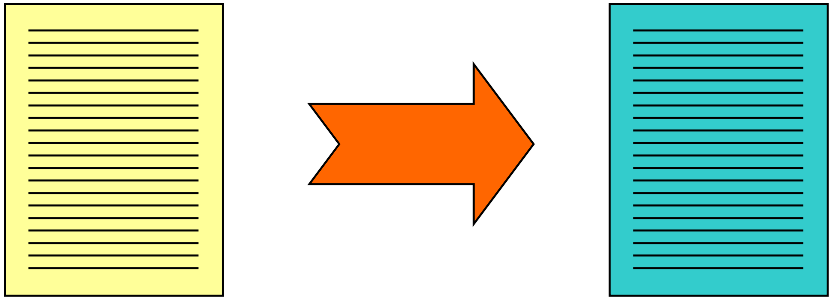
Abstract generation reports on contents of a text.



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.



Texts are translated automatically, or the system assists 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.



Formal and Computational Methods

Generic Computer Science 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.

Non-discrete 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.

Linguistic Methods and Resources

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 and pragmatic interpretation, pronunciation and intonation.

Corpora and Corpus Tools

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

Methods from Cognitive Science (Psychology)

Models of Cognitive Systems and their Components

The interaction of perception, knowledge, reasoning and action including communication is modeled 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 from 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).

State of the Art

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 (probabilistic context-free parsing)**
40%-80%	**Deep parsing of newspaper texts (HPSG or LFG parsing with large lexicon)**

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

Report Generation

Grammar Checkers

Information Extraction

Human Assisted Machine Translation

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)

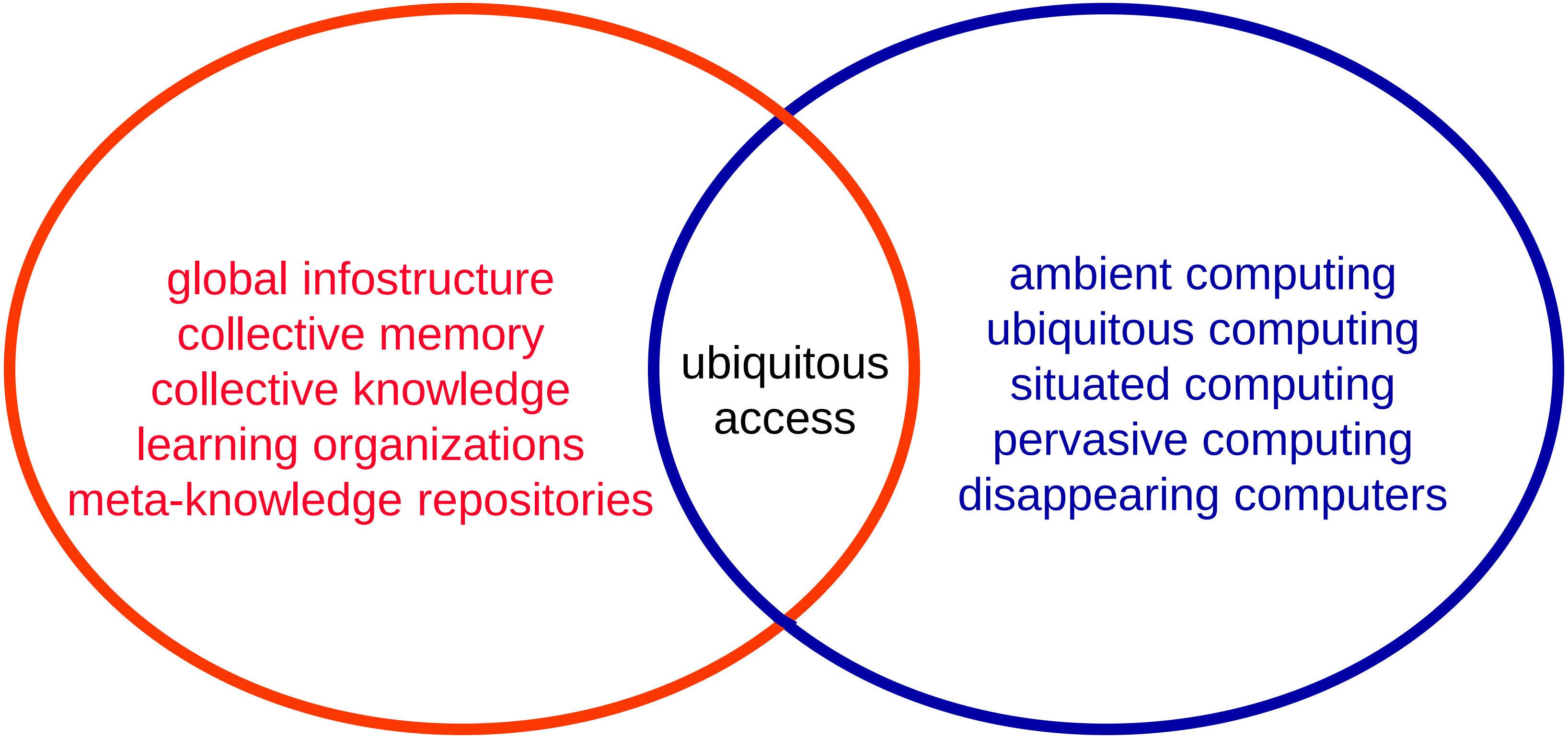
Dense Associative Hyperlinking

Concept-Based Information Retrieval

Text Understanding

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

MEGATRENDS



global infostructure
collective memory
collective knowledge
learning organizations
meta-knowledge repositories

ubiquitous
access

ambient computing
ubiquitous computing
situated computing
pervasive computing
disappearing computers

personalization
adaptation
learning

Further Introduction Online

http://lt-world.org

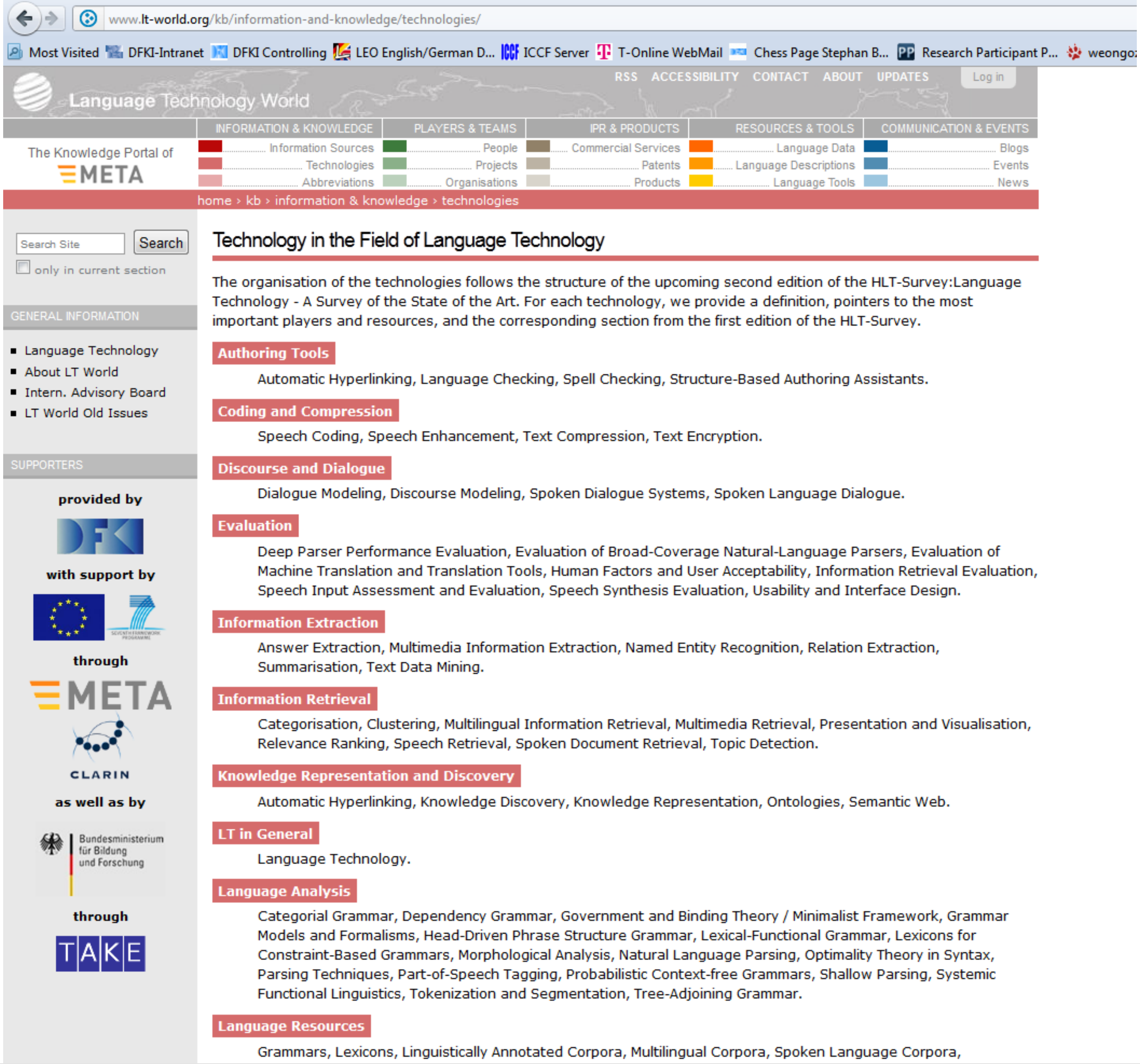
The biggest portal on language technology on the Web

Look around and visit the Technologies section as a complement to today‘s lecture

- List the technologies needed for a system that analyzes Web documents to compile information about people and companies. The user shall be able to learn what affiliations John Doe had between 1995 and 2005, or who was CEO of Dummy Inc. Between 1980 and 2000.
- Use the list to suggest a workflow that fulfills the task by using the technologies (you‘ll need some basic computer science technologies as well)

- Report errors, or inconsistencies, or outdated information you may encounter. Your help is much appreciated!

- Stephan.Busemann@dfki.de



Technology in the Field of Language Technology

The organisation of the technologies follows the structure of the upcoming second edition of the HLT-Survey:Language Technology - A Survey of the State of the Art. For each technology, we provide a definition, pointers to the most important players and resources, and the corresponding section from the first edition of the HLT-Survey.

Authoring Tools

Automatic Hyperlinking, Language Checking, Spell Checking, Structure-Based Authoring Assistants.

Coding and Compression

Speech Coding, Speech Enhancement, Text Compression, Text Encryption.

Discourse and Dialogue

Dialogue Modeling, Discourse Modeling, Spoken Dialogue Systems, Spoken Language Dialogue.

Evaluation

Deep Parser Performance Evaluation, Evaluation of Broad-Coverage Natural-Language Parsers, Evaluation of Machine Translation and Translation Tools, Human Factors and User Acceptability, Information Retrieval Evaluation, Speech Input Assessment and Evaluation, Speech Synthesis Evaluation, Usability and Interface Design.

Information Extraction

Answer Extraction, Multimedia Information Extraction, Named Entity Recognition, Relation Extraction, Summarisation, Text Data Mining.

Information Retrieval

Categorisation, Clustering, Multilingual Information Retrieval, Multimedia Retrieval, Presentation and Visualisation, Relevance Ranking, Speech Retrieval, Spoken Document Retrieval, Topic Detection.

Knowledge Representation and Discovery

Automatic Hyperlinking, Knowledge Discovery, Knowledge Representation, Ontologies, Semantic Web.

LT in General

Language Technology.

Language Analysis

Categorial Grammar, Dependency Grammar, Government and Binding Theory / Minimalist Framework, Grammar Models and Formalisms, Head-Driven Phrase Structure Grammar, Lexical-Functional Grammar, Lexicons for Constraint-Based Grammars, Morphological Analysis, Natural Language Parsing, Optimality Theory in Syntax, Parsing Techniques, Part-of-Speech Tagging, Probabilistic Context-free Grammars, Shallow Parsing, Systemic Functional Linguistics, Tokenization and Segmentation, Tree-Adjoining Grammar.

Language Resources

Grammars, Lexicons, Linguistically Annotated Corpora, Multilingual Corpora, Spoken Language Corpora,