

# Language Technology I

## Introduction

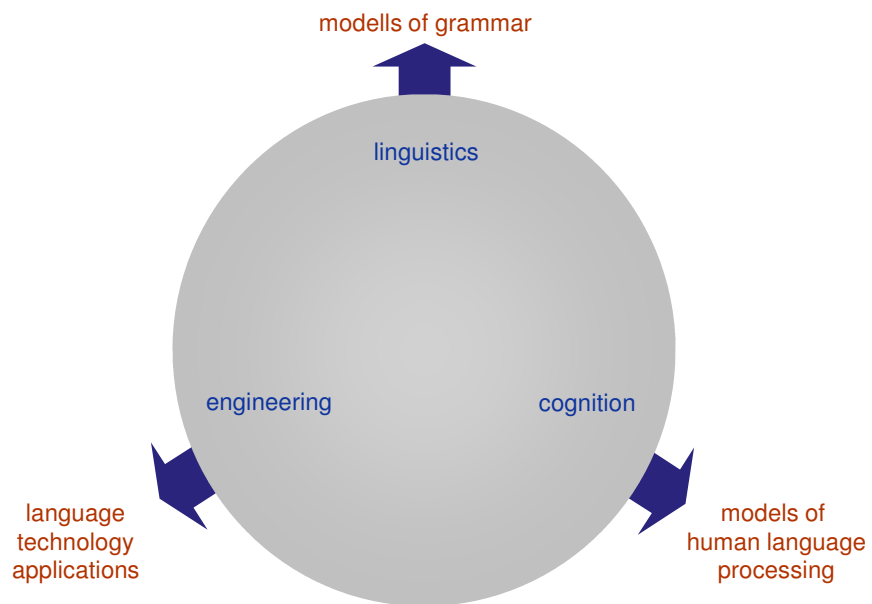
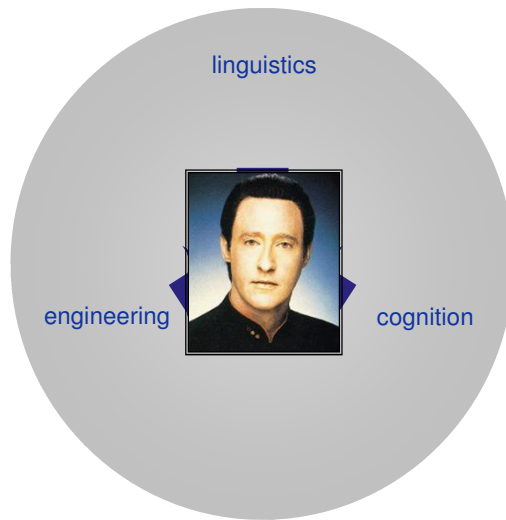
Stephan Busemann

(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



## 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 infrastructure

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

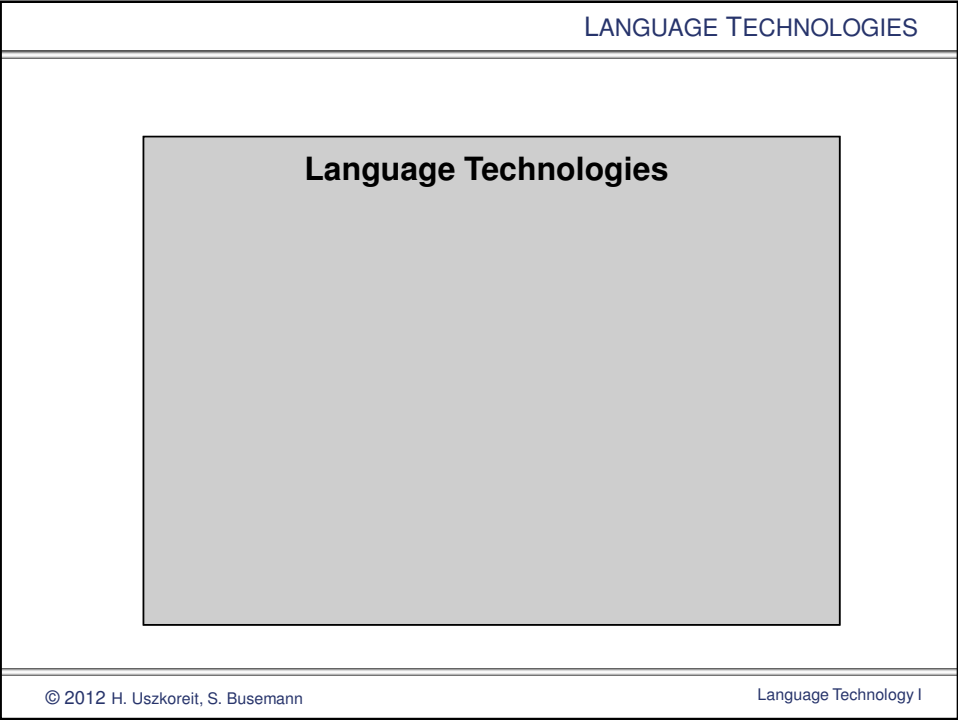
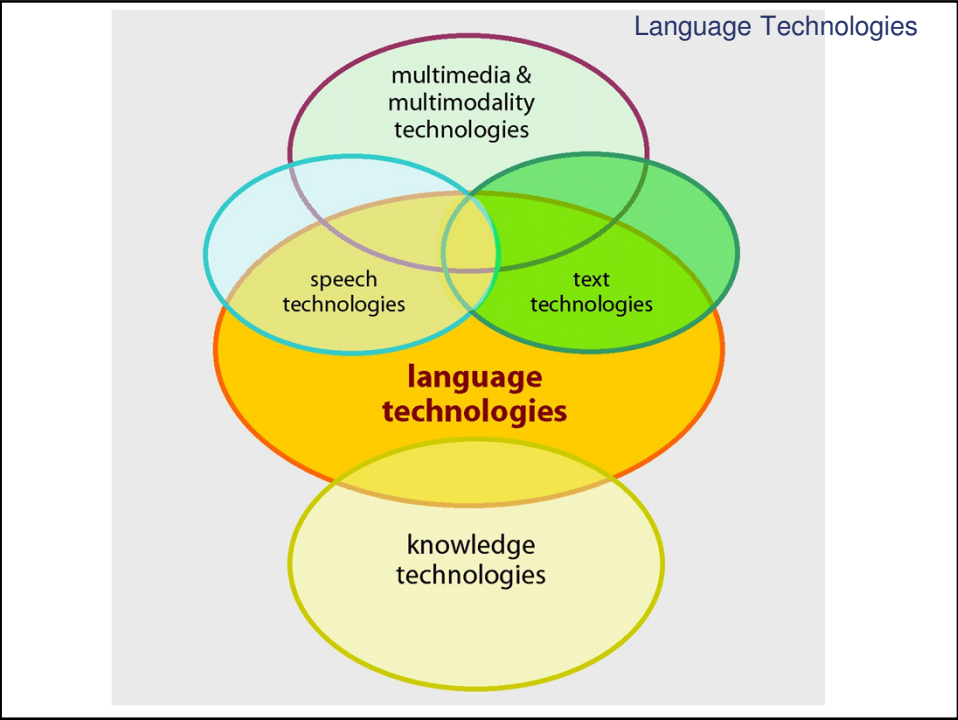
**Synchronicity:** synchronous vs. asynchronous

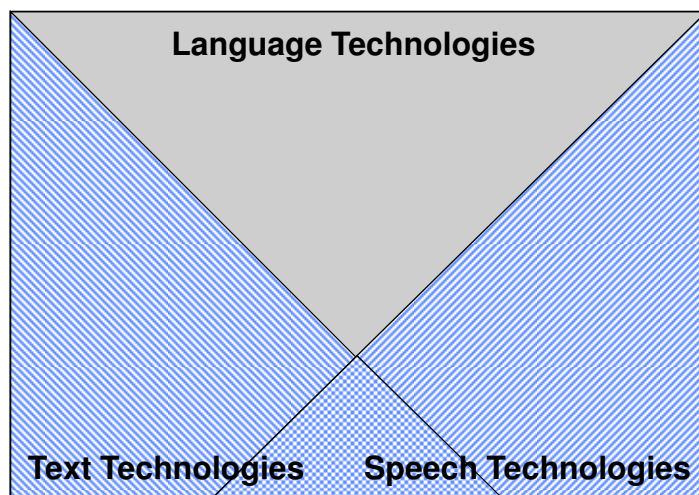
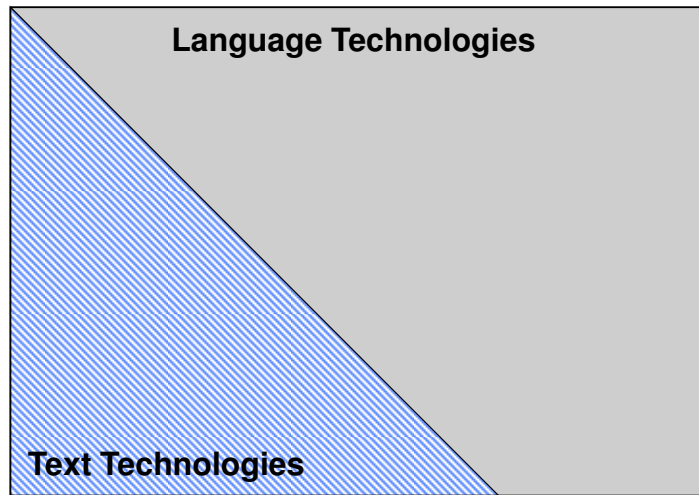
**Situatedness:** sensitivity to context, location, time, plans

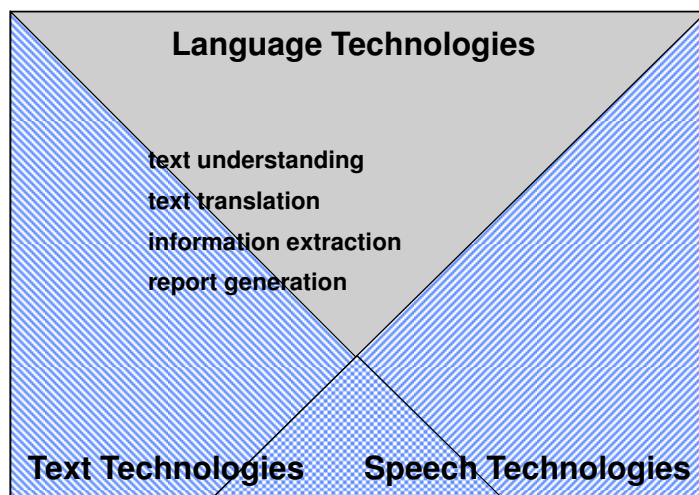
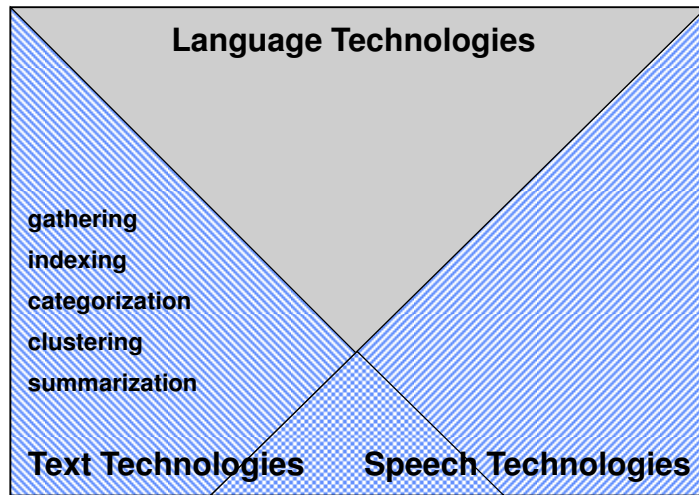
**Type of linguality:** monolingual, multilingual, translingual

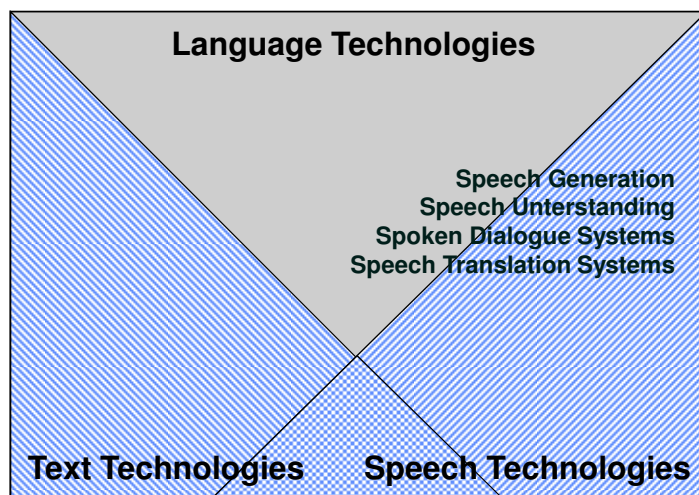
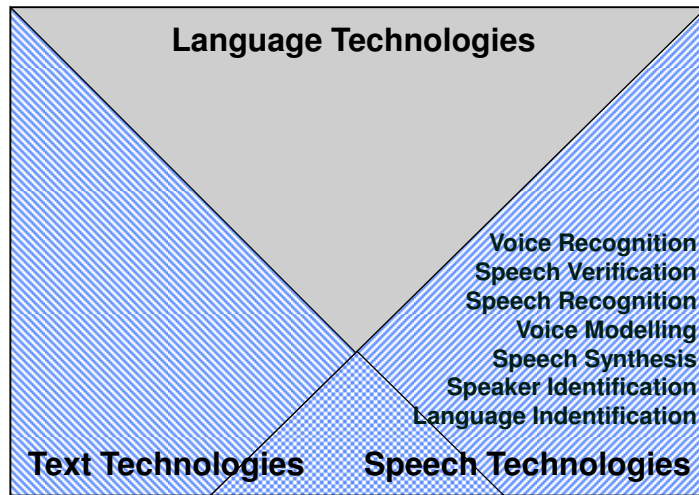
**Type of processing:** Categorization, summarization, extraction, understanding, translating, responding

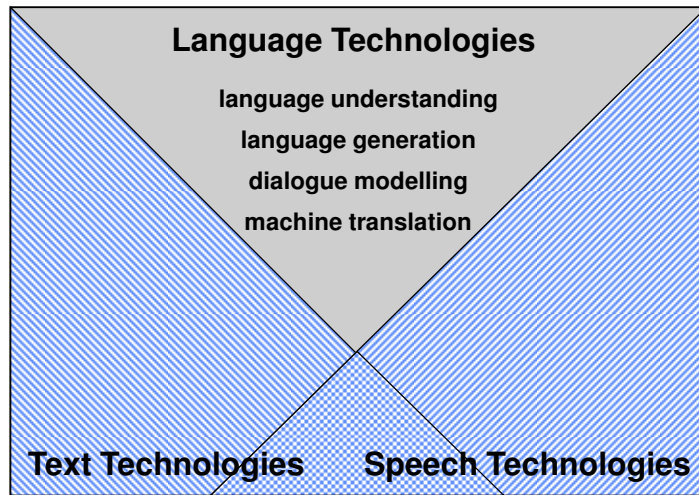
**Level of linguistic description:** phonology, morphology, syntax, semantics, pragmatics





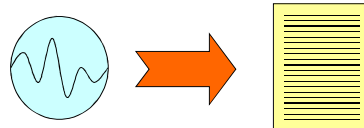






## Speech recognition

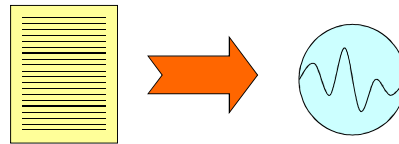
Spoken language is recognized and transformed:  
into text as in dictation systems, into commands as  
in robot control systems, or into some other internal  
representation.



## Speech Synthesis

(also Speech Generation)

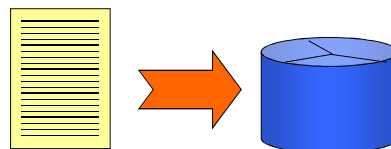
Utterances in spoken language are produced from text (text-to-speech systems) or from internal representations of words or sentences (concept-to-speech systems)



## Text Categorization

(also Text Classification)

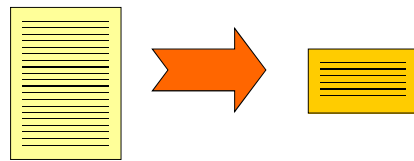
Texts are assigned to given categories. Texts may belong to more than one category, categories may contain other categories. *Filtering* is a special case of categorization with just two categories.



## Text Summarization

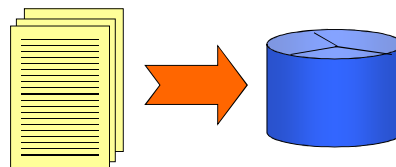
The most relevant portions of a text are extracted as a summary. Summaries may be limited to the needed length. Summarization may be specific to a certain query and/or a user's interests; summaries may be in the same or a different language.

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



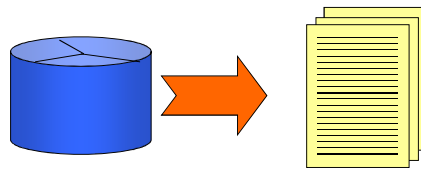
## Text Indexing

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.



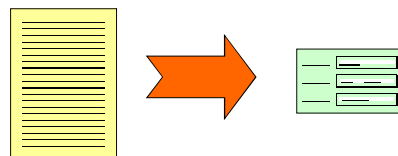
## Text Retrieval

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*.



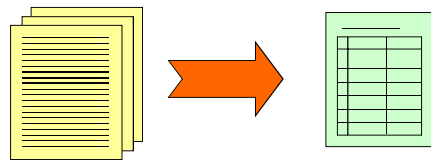
## Information Extraction

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

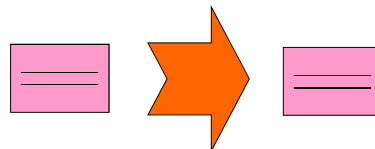
Extracted pieces of information from several sources are combined into one database. Previously undetected relationships may be discovered.



## 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.

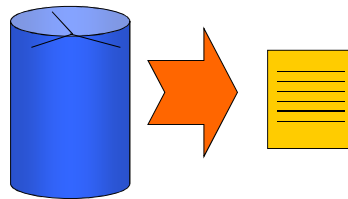
QA on the WWW triggers search engines and exploits their results.



## Report Generation

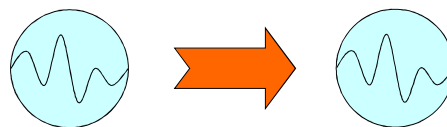
A report in natural language is produced that describes the requested contents or changes of a database. The report can contain accumulated numbers, maxima, minima and the most drastic changes.

Abstract generation reports on contents of a text.



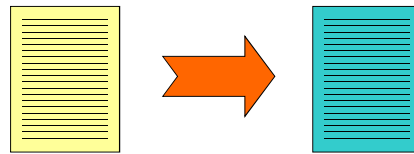
## Spoken Dialogue Systems

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.



## Translation Technologies

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.

### 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.

### 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).

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)

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)**

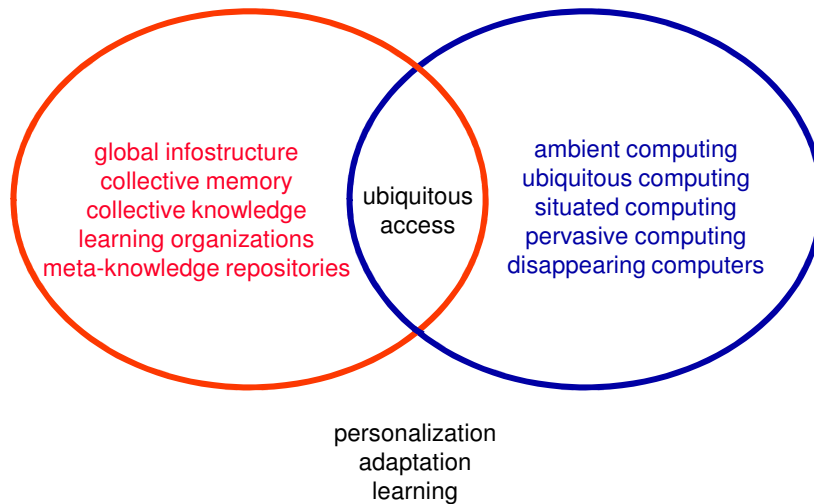
**Complex Information Extraction (Ternary+ Relations)**

**Dense Associative Hyperlinking**

**Concept-Based Information Retrieval**

**Text Understanding**

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



## Overview of Course

22.10.2012	Introduction	Stephan Busemann
29.10.2012	Text Classification	Jörg Steffen
31.10.2012	Exercise	Jörg Steffen
05.11.2012	Information Retrieval	Günter Neumann
07.11.2012	Exercise	Günter Neumann
12.11.2012	Crosslingual Information Retrieval	Günter Neumann
14.11.2012	Exercise	Günter Neumann
19.11.2012	Shallow Analysis and NER	Stephan Busemann
21.11.2012	Exercise	Stephan Busemann
26.11.2012	Dependency parsing	Alexander Volokh
28.11.2012	Exercise	Alexander Volokh
03.12.2012	Relation Extraction	Feiyu Xu
05.12.2012	Exercise	Feiyu Xu
10.12.2012	Opinion Mining	Feiyu Xu
12.12.2012	Exercise	Feiyu Xu
17.12.2012	Textual Inference	Günter Neumann
19.12.2012	Exercise	Günter Neumann
Christmas and New Year's Break		all
07.01.2012	Semantic Web and LT	tbd
08.01.2012	Exercise	tbd
14.01.2012	Science Information Applications	Ulrich Schäfer
16.01.2012	Exercise	Ulrich Schäfer
21.01.2012	Text Generation	Stephan Busemann
23.01.2012	Exercise	Stephan Busemann
28.01.2012	Text Generation	Stephan Busemann
30.01.2012	Exercise	Stephan Busemann
04.02.2012	Text summarization	Stephan Busemann
06.02.2012	Exercise	Stephan Busemann
11.02.2012	EXAM	tbd