**BESCHREIBUNG DES STUDIENFACHS «COMPUTERLINGUISTIK»**

Computerlinguistik ist das Studium der maschinellen Verarbeitung natürlicher Sprachen. Die Computerlinguistik ist eine interdisziplinäre Wissenschaft, die einerseits durch eine Fülle aktueller Anwendungen im Bereich Sprachtechnologie, Informationssuche und Data Mining, andererseits durch enge inhaltliche und methodische Verbindungen zu traditionellen Fächern wie Linguistik, Informatik, Logik und Phonetik gekennzeichnet ist. Ziel des Studiengangs ist es, eine Ausbildung nahe am modernen Berufsbild des Computerlinguisten anzubieten, die gleichzeitig aber auch ein Verständnis der interdisziplinären Rolle und der daraus resultierenden Forschungsfragen vermittelt.

Die Ausbildung soll es den Studierenden ermöglichen, sprachliche Phänomene, Regeln und Zusammenhänge in formaler Weise zu erfassen. Hierzu sind sowohl linguistische wie auch mathematische Kenntnisse notwendig. Zugleich werden die programmiertechnischen Voraussetzungen vermittelt, die es erlauben, Sprachdaten für unterschiedliche Anwendungen in passender Weise zu erfassen und maschinell zu verarbeiten. Schließlich vermittelt der Studiengang einen Eindruck von konkreten Anwendungen in unterschiedlichen Bereichen wie z.B. Websuche und natürlichsprachliche Auskunftssysteme.

Seit dem Wintersemester 2009/2010 wird im Rahmen der neuen Bachelor-Programme ein Hauptfach Computerlinguistik angeboten. Computerlinguistik im Hauptfach ist ein auf sechs Semester ausgelegter Studiengang mit einem Gesamtumfang von 120 ECTS (кредиты)-Punkten. Er besteht hauptsächlich aus den Elementen Sprachwissenschaft, Programmierung sowie Mathematik und Logik.

# ANWENDUNGSGEBIETE

[Sprache-zu-Text](http://www.harald-oertle.de/compling.html%22%20%5Cl%20%22erkennen)
[Text-zu-Sprache](http://www.harald-oertle.de/compling.html#synthese)
[Telefondialogsysteme](http://www.harald-oertle.de/compling.html#dialog)
[Sprachsteuerung](http://www.harald-oertle.de/compling.html#steuern)
[Textgenerierung](http://www.harald-oertle.de/compling.html#generieren)
[Übersetzungstechnologie](http://www.harald-oertle.de/compling.html#uebersetzen)
[Information-Retrieval](http://www.harald-oertle.de/compling.html#suchen)
[Internet-Suchagenten mit natürlich-sprachlicher Schnittstelle](http://www.harald-oertle.de/compling.html#abfragen)
[Informationsextraktion](http://www.harald-oertle.de/compling.html#extrahieren)
[Frage-Antwort-Systeme](http://www.harald-oertle.de/compling.html#frageantwort)
[Automatische Textzusammenfassung](http://www.harald-oertle.de/compling.html#zusammenfassen)
[Rechtschreibkorrektur](http://www.harald-oertle.de/compling.html#rechtschreibung)

 **Sprache-zu-Text:**
 Sprache-zu-Text-Technologie kommt im Rahmen von Diktiersystemen zum Einsatz. Diktiersysteme sind Systeme, die gesprochene Eingaben über Mikrofon in entsprechenden schriftlichen Text umwandeln und den schriftlichen Text auf einem (Computer-)Bildschirm ausgeben.
 Die Spracheingabe wird dabei von einem Spracherkennungssystem analysiert. Ein Spracherkennungssystem ist ein System, das eine gesprochene Eingabe (Äußerung) erkennt, indem es die Schallwelle der gemachten Äußerung in ein entsprechendes elektronisches Sprachsignal umwandelt und das Sprachsignal in Echtzeit analysiert. Aus dem Sprachsignal werden dabei bestimmte signalspezifische Informationen herausgelesen. Mit Hilfe der herausgelesenen signalspezifischen Informationen und mit Hilfe wahrscheinlichkeitstheoretischer Verfahren wird versucht, dem Sprachsignal entsprechende Wörter aus einem Lexikon zuzuordnen. Die Wörter im Lexikon sind deswegen mit einer speziellen Lautumschrift versehen.

Bei der Entwicklung von Spracherkennungssystemen muß berücksichtigt werden, daß identische Äußerungen von unterschiedlichen Sprechern zu nicht ganz identischen Sprachsignalen führen, da Menschen unterschiedliche Stimmen haben. Gründe hierfür sind z.B.:
- Männerstimmen vs. Frauenstimmen,
- Akzent aufgrund einer bestimmten regionalen Herkunft,
- altersabhängige Stimmlage.
Dennoch muß ein Spracherkennungssystem bei identischen Äußerungen von unterschiedlichen Sprechern immer zum gleichen Analyseergebnis kommen.

 **Text-zu-Sprache:**
 Mittels Text-zu-Sprache-Systemen (auch Text-to-Speech-Systeme genannt) können auf einem Computer festgehaltene schriftliche Texte automatisch vorgelesen werden. Das Berechnen und Vorlesen der akustischen Sprachausgabe wird vom Text-zu-Sprache-System übernommen. Ausgabemedium ist ein Lautsprecher. Bei Text-zu-Sprache-Systemen können auch verschiedene Parameter eingestellt werden, wie z.B. Sprachenauswahl (z.B. Deutsch, Englisch, usw.) oder Männer- vs. Frauenstimme.
 Bei Text-zu-Sprache-Systemen werden die schriftlichen Texte mit Hilfe künstlicher "Stimmen" vorgelesen. Dabei kommt es auf die richtige Aussprache und Betonung der einzelnen Wörter und Sätze an. Ein einzelnes vorgelesenes Wort besteht aus einzelnen Lauten. Jeder Laut besteht wiederum aus einzelnen lautlichen Bausteinen. In einem Text-zu-Sprache-System sind alle notwendigen lautlichen Bausteine enthalten. Jeder lautliche Baustein ist ein Stück Sprachsignal. Um ein Laut bzw. ein akustisches Wort zu erzeugen, werden die notwendigen lautlichen Bausteine zusammengefügt, indem die einzelnen Sprachsignalsücke hintereinandergesetzt werden. Welche lautlichen Bausteine dabei auszuwählen sind, hängt auch von den benachbarten Lauten ab (auch wortübergreifend).

 **Telefondialogsysteme:**
 Natürlich-sprachliche Telefondialogsysteme ermöglichen einen wechselseitigen Dialog zwischen einem Menschen und einem Computer. Ziel solcher Systeme ist es, Callcenter-Mitarbeiter von der Beantwortung von Routinefragen zu entlasten oder vollautomatische telefonbasierte Informationsplattformen anbieten zu können, die nur standardisierte Fragen beantworten müssen.
 Beispielsweise können im Rahmen einer telefonbasierten Flugplanauskunft Flugverbindungen abgefragt werden. Nachdem der Benutzer im Rahmen eines solchen Systems seinen Wunsch in Form einer natürlich-sprachlichen Eingabe geäußert hat, wertet das Dialogsystem die Eingabe inhaltlich aus und überprüft dabei, welche Informationen eingegeben worden sind und welche noch fehlen, um eine Flugverbindung ausgeben zu können. Enthält die eingegebene Benutzeranfrage schon alle relevanten Informationen, kann das Dialogsystem gleich eine Antwort präsentieren. Fehlen im eingegebenen Satz noch relevante Informationen, muß das Dialogsystem in separaten Dialogschritten die fehlenden Informationen erfragen.
 Zwei Beispiele im Rahmen einer automatischen Flugplanauskunft soll die Funktionsweise verdeutlichen: Bei der Flugplanauskunft sind Zielflughafen, Ankunftstag und Ankunftsuhrzeit, sowie Abflugsort die relevanten Informationen. (Der Abflugsort soll aber separat eingegeben werden). Die Benutzeranfrage bzgl. Zielort im ersten Dialogbeispiel enthält schon alle relevanten, damit das System in der Lage ist, eine Antwort zu präsentieren.

System>    Wann wollen Sie an Ihrem Zielflughafen sein?
Benutzer>  Ich will nächsten Samstag um 10 Uhr in Paris sein.
System>    Wo wollen Sie abfliegen?
Benutzer>  In Frankfurt.
 System>    Ich kann Ihnen folgende Verbindungen anbieten:
Abflug in Frankfurt am [Datum] um [Uhrzeit]
Abflug in Frankfurt am [Datum] um [Uhrzeit]

 Das nächste Dialogbeispiel soll zeigen, wie sich das Dialogsystem verhält, wenn die Benutzeranfrage noch nicht alle relevanten Informationen enthält. In diesem Fall fehlt zunächst die Ankunftsuhrzeit.
System>    Wann wollen Sie an Ihrem Zielflughafen sein?
Benutzer>  Ich will nächsten Samstag in Paris sein.
System>    Um wieviel Uhr wollen Sie in Paris sein?
Benutzer>  Um 10 Uhr.
System>    Wo wollen Sie abfliegen?
Benutzer>  In Frankfurt.
System>    Ich kann Ihnen folgende Verbindungen anbieten:
                  Abflug in Frankfurt am [Datum] um [Uhrzeit]
                  Abflug in Frankfurt am [Datum] um [Uhrzeit]

 Da im vorletzten Dialogschritt die noch fehlende Ankunftsuhrzeit eingegeben worden ist, sind nun alle Informationen vorhanden, damit die automatische Flugauskunft entsprechende Flugverbindungen ausgeben kann.
 Bei einem natürlich-sprachlichen Telefondialogsystem werden die gesprochenen Äußerungen des Anrufers in Echtzeit von einem Spracherkennungssystem analysiert und von einem Sprachverarbeitungssystem inhaltlich ausgewertet. Das Sprachverarbeitungssystem ermittelt eine formale Bedeutung der getätigten Äußerung, damit das Dialogsystem entsprechend darauf reagieren kann. Das Dialogsystem kann allerdings nur dann die Äußerungen des Benutzers verstehen, wenn sie sich auf die Anwendungsdomäne des Dialogsystems beziehen, welche hier die Flugplanauskunft ist. Aufgrund der Spracheingabe wird eine sprachliche Reaktion seitens des Dialogsystems ermittelt, welche von einem Textgenerierungssystem berechnet und von einem Text-zu-Sprache-System akustisch ausgegeben wird.

 **Sprachsteuerung:**
 Sprachsteuerung elektronischer Geräte bedeutet, dass die Bedienfunktionen eines elektronischen Geräts (wie zum Beispiel eines Mobiltelefons) nicht mittels drücken von Knöpfen und auswählen von Menüpunkten auf dem Bildschirm erfolgt, sondern mittels Sprachbefehlen. Die Sprachbefehle ersetzen hier also die haptische Bedienung.
 Sprachsteuerungen gibt es beispielsweise für:
- Autos (um Infotainmentanwendungen, wie zum Beispiel den Mediaplayer, während der Fahrt
   per Sprache bedienen zu können),
- Smartphones, sowie für
- das "intelligente Haus" (in dem man auf Zuruf Aktionen im und um das Haus) auslösen kann.
 Die Sprachsteuerung besteht typischerweise aus einem Sprachdialogsystem als Mensch-Maschine-Schnittstelle (HMI - Human Machine Interface). Hierbei folgt auf einen Sprachbefehl eine sprachliche Reaktion seitens des Sprachdialogsystems. Die sprachliche Reaktion kann eine akustische Bestätigung für die technische Ausführung einer Aktion sein (wie zum Beispiel, dass der Anruf zum gewünschten Kontakt erfolgt) oder sie kann eine Aufforderung sein, den nächsten Aktionsschritt zu sprechen.
 Ein Sprachbefehl kann ein Wort, eine Wortfolge oder ein Satz sein:
- Bei einer Worteingabe kann es sich beispielsweise um eine Bestätigung mit "ja" oder eine Ablehnung  mit "nein" handeln.
- Wortfolgen als Spracheingabe erlauben das Bedienen komplexerer elektronischer Geräte.
   Sie ermöglichen die Formulierung eines Aktionsschritts im Rahmen der Bedienung.

Typischerweise müssen Sprachbefehle zuvor erlernt werden (außer sie sind intuitiv klar).
Gegebenenfalls müssen zwei (oder mehr) Aktionsschritte per Sprache durchgeführt werden

bis die gewünschte technische Aktion ausgeführt werden kann.

Ein Beispiel im Rahmen der Handy-Sprachbedienung wäre:
Benutzer> "Kontakt anrufen".
System> "Bitte nennen Sie den Namen des Kontakts".
Benutzer> "Peter Fischer".
System> "Peter Fischer wird angerufen" => der Anruf wird daraufhin ausgeführt.

Das nächste Beispiel vereint zwei Benutzer-Dialogschritte zu einem einzigen:
Benutzer> "Kontakt Peter Fischer anrufen".
System> "Peter Fischer wird angerufen" => der Anruf wird daraufhin ausgeführt.

 Sätze als Spracheingabe ermöglichen eine relativ freie Formulierung von Aktionen.
Grundsätzlich soll in einer einzigen sprachlichen Information die Menge an Information "untergebracht" werden können, die sonst über mehrere Dialogschritte hinweg mitgeteilt werden müsste. Diese Interaktionsform wird auch als "natürlich-sprachlich" bezeichnet.

 Ein Beispiel mit natürlich-sprachlicher Eingabe im Rahmen einer Smartphone-Sprachsteuerung wäre:
Benutzer> "Bitte Peter Fischer aus dem Adressbuch anrufen".
System> "Peter Fischer wird angerufen" => der Anruf wird daraufhin ausgeführt.

 Damit ein Sprachbefehl ausgeführt werden kann, bedarf es eines so genannten Spracherkenners. Der Spracherkenner hat die Aufgabe, in der akustischen Äußerung Wörter zu identifizieren und zu extrahieren, damit die Wörter bzw. die Äußerung selbst weiterverarbeitet werden können.
 Bei der technischen Umsetzung gibt es zwei Ansätze:
a) Die Sprachsteuerung läuft auf lokaler Ebene ab. Das heißt, sie ist im elektronischen Gerät selbst integriert. Ein Sprachbefehl wird dabei im Gerät verarbeitet, in eine ausführbare Aktion umgewandelt und ausgeführt.
b) Die Sprachsteuerung läuft über einen Server. Hierbei wird der Sprachbefehl per Mobilfunk an einen Server geschickt, dort verarbeitet, in eine ausführbare Aktion umgewandelt und auf dem Gerät ausgeführt.

 **Textgenerierung:**
 Textgenerierungssysteme ermöglichen das automatische Erzeugen eines ganz neuen Textes. Ausgangsbasis eines solchen Systems ist normalerweise eine nicht-textuelle Datenquelle, wie z.B. eine Tabelle. Textgenierungssysteme werden speziell für eine bestimmte Anwendungsdomäne entwickelt, wie z.B. für die automatische Erstellung von Wetterberichten.
 Ein Textgenerierungsprogramm besitzt eine Textplanungs- und Satzplanungskomponente, sowie eine Wissensbasis, in der das anwendungsspezifische Wissen abgelegt und in die die Daten aus der Datenquelle hineingeladen werden. Die Wissensbasis besitzt hierfür sog. Anwendungskonzepte, denen entsprechende Daten aus der Datenquelle zugeordnet werden. Anwendungskonzepte im Rahmen der Anwendungsdomäne Wetterbericht wären z.B. Windgeschwindigkeit, Temperatur und Luftdruck. Die Textplanungskomponente hat die Aufgabe, zu beschreiben, was der Text aussagen soll (Festlegung der Inhalte) und wie der Text es ausdrücken soll (Festlegung der Mittel). Dabei greift die Textplanungskomponente auf die Wissensbasis zu. Die Satzplanungskonzepte formt aus der inhaltsorientierten Textbeschreibung zunächst entsprechende inhaltsorientierte Satzbeschreibungen und formuliert dann für diese Satzbeschreibungen (unter Zuhilfenahme sprachlichen Wissens und eines domänenspezifischen Lexikons) entsprechende Sätze.
Motivation für die Entwicklung solcher Textgenierungsprogramme ist, daß Informationen in Textform in vielen Fällen leichter zu verstehen sind als in anderen Darstellungsformen.

 **Übersetzungstechnologie:**
 Im Bereich der Übersetzungstechnologie gibt es zwei Ansätze:
- Maschinelle Übersetzung und
- computergestützte Übersetzung.
 Maschinelle Übersetzungssysteme sind Systeme, die einen vorgegebenen Text selbständig in eine andere Sprache übersetzen. Hier gibt es zwei Ansätze:
- Der regelbasierte Ansatz und
- der statistische Ansatz.

 Maschinelle Übersetzungssysteme, die auf dem regelbasierten Ansatz -- wie dem Transfer-Ansatz -- basieren, analysieren nacheinander die einzelnen Sätze des Quelltextes, indem sie auf die Satzstrukturen der quellsprachlichen Sätze eingehen. Für die quellsprachlichen Satzstrukturen werden dann entsprechende zielsprachliche Satzstrukturen berechnet und für diese dann zielsprachliche Sätze erzeugt. Maschinelle Übersetzungssysteme, die auf dem statistischen Ansatz basieren, wählen für jeden Satz des Originalsatzes die wahrscheinlichste Übersetzung aus. Um dies zu ermöglichen, müssen die statistischen Übersetzungssysteme anhand großer, in elektronischer Form vorliegender, Referenzübersetzungen erst einmal trainiert werden. Erst nach erfolgtem Training ist das statistische Übersetzungsprogramm prinzipiell in der Lage, auch Sätze zu übersetzen, mit denen es zuvor noch nicht konfrontiert worden ist.
 Beim maschinellen Übersetzen muß beispielsweise berücksichtigt werden, daß es mehrdeutige Wörter gibt. Ein Beispiel für ein mehrdeutiges Wort ist das deutsche Wort "Bank", welches man im Englischen je nach Kontext entweder mit "bank" (Geldinstitut) oder mit "bench" (Sitzgelegenheit) übersetzen kann. Mit bestimmten Verfahren läßt sich der richtige Kontext (und somit die richtige Übersetzung) bestimmen. Je länger und komplexer die einzelnen zu übersetzenden Quellsätze sind, desto schwerer tut sich das maschinelle Übersetzungssystem mit dem Übersetzen. Deshalb ist oft eine Vorbereitung eines zu übersetzenden Textes oder eine Nachbearbeitung eines maschinell übersetzten Textes notwendig. Im Falle von Anwendungsdomänen, die mit einfachem Satzbau und beschränktem Vokabular auskommen (wie z.B. bei einer technischen Dokumentation) wird auch sog. "kontrollierte Sprache" eingesetzt. Dadurch ist gewährleistet, daß die Texte später reibungslos maschinell übersetzt werden können.
 Im Gegensatz zur maschinellen Übersetzung kommen bei der computergestützten Übersetzung Werkzeuge zum Einsatz, die einen menschlichen Übersetzer bei dessen Übersetzungstätigkeit unterstützen. Hierzu gehören sog. Übersetzungsspeicher (Translation-Memory-Systeme). Übersetzungsspeicher enthalten schon einmal "per Hand" erstellte Übersetzungen. Dabei wird sowohl der Ausgangssatz als auch die dazugehörige fertige Übersetzung gespeichert. Will nun ein menschlicher Übersetzer einen bestimmten Satz übersetzen, so kann er automatisch überprüfen lassen, ob für genau diesen Satz im Übersetzungsspeicher schon eine Übersetzung vorliegt. Findet der Überseztungsspeicher eine Entsprechung für den zu übersetzenden Satz, so zeigt es die gefundene Übersetzung an. Es ist dabei nicht notwendig, den exakt gleichen Satz im Übersetzungsspeicher zu finden. Es reicht aus, wenn auch ein ähnlicher Satz gefunden wird. Der Grad der Ähnlichkeit wird z.B. durch eine Prozentzahl angegeben. Übersetzungsbüros greifen deshalb gerne auf Übersetzungsspeicher zurück, damit Sätze, die schon einmal übersetzt worden sind, immer gleich übersetzt werden (Qualitätssicherung).

 **Information-Retrieval:**
 Information-Retrieval-Systeme haben die Aufgabe, Dokumente (z.B. im Internet) aufzufinden, die den Kriterien einer gemachten Suchabfrage entsprechen. Dabei werden die Suchbegriffe mit den Indizes der Dokumente verglichen. Indizes sind in diesem Zusammenhang Schlagwörter (Begriffe), die ein Dokument möglichst zutreffend beschreiben. Wenn die in der Suchabfrage verwendeten Begriffe auch als Dokumentenindizes vorkommen, wird das Dokument als "Treffer" ausgewählt. Damit man nicht nur solche Dokumente auffindet, die exakt diejenigen Indizes wie in der Suchabfrage enthalten, sondern auch Dokumente mit abweichenden Indizes, welche aber inhaltlich gesehen ebenfalls zutreffend wären, können Retrievalsysteme mit "linguistischer Intelligenz" weiterhelfen.
 Ein Beispiel soll erläutern, wie ein solches Retrievalsystem mit linguistischer Intelligenz funktioniert: Für den Suchbegriff "blaues Fahrzeug" findet ein solches System (auch) Dokumente mit den Indizes {"blau", "Fahrzeug"} bzw. {"blaues", "Fahrzeug"}, da es die Suchabfrage morphosyntaktisch zerlegen und den Wortstamm ermittlen kann. D.h., aus der Wortform "blaues" kann die Grundwortform "blau" automatisch abgeleitet werden.
 Eine weitere Retrieval-Möglichkeit besteht im Einsatz semantischer Methoden. Hier geht es darum, für einen Suchbegriff Textdokumente mit inhaltlich (also semantisch) äquivalenten Indizes aufzufinden. Um eine solche semantische Suche zu ermöglichen, sind sog. Thesauri notwendig, in denen für bestimmte Begriffe entsprechende Alternativen aufgeführt sind. Somit ist es möglich, daß bei dem Suchbegriff "Fahrzeug" auch Dokumente ausgewählt werden, denen die Indizes "Auto" oder Eigennamen wie "Mercedes" oder "Toyota" zugeordnet sind. Auch eine Kombination der semantischen mit der syntaktischen Suche ist möglich. Wenn man als Suchbegriff z.B. "blaues Fahrzeug" verwendet, so würden auch Dokumente mit den folgenden Indizes gefunden: {"blauer", "Mercedes"} oder {"blau", "Auto"}.

 **Internet-Suchagenten mit natürlich-sprachlicher Schnittstelle:**
 Ein Suchagent ist ein Computerprogramm, das selbständig im Internet nach vorgegebenen Informationen sucht. Dabei nutzt es Suchmaschinen und Internet-Datenbanken. Der Suchagent wertet die gefundenen Informationen aus, bereitet sie auf und speichert sie entweder ab oder verschickt die gefundenen Suchergebnisse per E-Mail an den Benutzer. Die zu suchende Information kann dem Suchagenten auch in Form eines (Frage-)Satzes über eine natürlich-sprachliche Schnittstelle mitgegeben werden.
 Besonders gut für die Arbeit von Suchagenten eignet sich das sog. Semantic Web ("Web 3.0"). Im Rahmen des Semantic Web werden Informationen auf Webseiten mittels spezieller Beschreibungssprachen auch inhaltlich (semantisch) beschrieben. Das Semantic Web dient dazu, daß nicht nur Menschen, sondern auch Computerprogramme Webseiten lesen und auswerten können.
 Im Rahmen des Semantic-Web werden den Daten im Internet semantische Umschreibungen hinzugefügt (als Annotationen bezeichnet). Zusätzlich können im Semantic-Web über die Annotationen automatisch logische Schlussfolgerungen gezogen werden. Dies wird dadurch erreicht, indem die Annotationen und deren logische Beziehungen in einer so genannten Ontologie abgelegt sind.
 Ein Anwendungsfall für einen solchen Internet-Suchagenten wäre ein System, das nach bestimmten Studienangeboten sucht. Eine natürlich-sprachliche Abfrage könnte folgendermaßen aussehen:
> Bachelor in Chemie, maximal 100 Kilometer von Stuttgart entfernt.
 Damit entsprechende Angebote gefunden werden können, muß es Internetseiten im Rahmen des Semantic Web geben, die Informationen zum Studium und zur geografischen Lage beinhalten, also: Studiengang, Abschluss, Hochschule, Ort.
 Hierfür müssen betreffende Internetseiten, z.B. einer Hochschule, als Semantic-Web vorliegen und daher auch Annotationen enthalten. Der Annotation [Studiengang] wäre dann der Text "Chemie" zugewiesen, der Annotation [Abschluss] der Text "Bachelor" und den Annotationen [Hochschule], [Ort] entsprechend der Hochschulname und der Hochschulort. Wenn der Suchagent in diesem Zusammenhang auf eine Ontologie zugreift, so könnte er auch aus der Chemie abgeleitete Studiengänge finden, wie zum Beispiel Biochemie oder Lebensmittelchemie.
Hinweis: Die Entfernungen können durch eine separate webbasierte Informationsquelle ermittelt werden, welcher der aktuelle Standort und maximale Radius übergeben wird.
 Bei einem derartigen Suchagent mit natürlich sprachlicher Schnittstelle wird jeder eingegebene Satz von der natürlich-sprachlichen Schnittstelle ausgelesen und einem Sprachverarbeitungsprogramm übergeben. Das Sprachverarbeitungsprogramm analysiert den Satz syntaktisch und berechnet eine formale Satzbedeutung, die für das Semantic-Web geeignet ist. Danach wird der Suchprozess gestartet.

 **Informationsextraktion:**
 Informationsextraktionsysteme sind domänenspezifische Systeme, die Texte inhaltlich auswerten können. Die Aufgabe eines solchen Systems besteht darin, die domänenspezifischen Informationen in einem Text aufzuspüren, zu extrahieren und in einer Datenbank für weitere Recherchen abzuspeichern. Das Informationsextaktionssystem muß hierfür domänenspezifische Konzepte bereitstellen, denen entsprechende Informationen aus dem zu analysierenden Text zugeordnet werden könen.
 Ein Informationsextraktionssystem muß die einzelnen Sätze eines Textes nacheinander analysieren, um die Satzbedeutungen ermitteln zu können und um den Konzepten entsprechende Informationen aus dem zugrundeliegenden Text zuordnen zu können. Erst aufgrund der ermittelten Satzbedeutungen kann das System entscheiden, welche Sätze relevante Informationen enthalten.
 Ein Beispiel soll verdeutlichen, wie ein solches Informationsextraktionssystem arbeitet:
 Ein Informationsextraktionssystem für die Domäne "Unternehmensspezifische Informationen" soll Texte auswerten, in denen Informationen über Unternehmensverkäufe vorkomen. Dem Informationsextraktionssystem stehen hierfür folgende Konzepte zur Verfügung:
- [Verkauftes Unternehmen]:
- [Kaufendes Unternehmen]:
- [Kaufspreis]:
Für einen vorhandenen Text müssen diesen obigen Konzepten entsprechende Informationen aus einem Text zugeordnet werden. In einem fiktiven Text wäre folgender Satz enthalten:
- "Die Schmidt GmH wurde für 5 Millionen Euro an die Müller AG verkauft."
Aufgrund des vorliegenden Satzes könnte das Informationsextraktionssystem nach Ermittlung der Satzbedeutung folgende Zuordnung machen:
- [Verkauftes Unternehmen]: "Schmidt GmbH"
- [Kaufendes Unternehmen]: "Müller AG"
- [Kaufpreis]: "5.000.000 Euro"
 Die aus dem Text extrahierten Informationen und ihre Konzepte können nun in einer Datenbank abgespeichert werden. Hierfür muß natürlich schon eine entsprechende Tabellenstruktur in der Datenbank vorhanden sein. Durch die Abspeicherung in einer Datenank können die extrahierten Informationen auch weiteren Personen zugänglich gemacht werden.

 **Frage-Antwort-Systeme:**
 Frage-Antwort-Systeme sind Systeme, die für einen natürlich-sprachlichen Fragesatz als Eingabe einen entsprechenden Antworttext als Ausgabe erzeugen. Die Eingabe erfolgt in der Regel schriftlich (über Tastatur). Der Antworttext kann ebenfalls schriftlich (über Bildschirm) oder akustisch (über Lautsprecher) ausgegeben werden.
 Der Antworttext soll ausführlich und informativ sein. Erfolgt die Ausgabe des Antworttextes über einen Bildschirm, so kann die Ausgabe auch multimedial angereichert werden, d.h., mit graphischen und akustischen Elementen gekoppelt werden. Frage-Antwort-Systeme sind normalerweise auf eine vorgegebene Domäne, wie z.B. medizinisches oder technisches Wissen, beschränkt.
 Bei Frage-Antwort-Systemen gibt es zwei Ansätze:
- Erzeugung eines neuen Antworttextes,
- Antwortextraktion.
 Im ersten Fall besteht das Frage-Antwort-System aus Analyse und Generierung. Ein eingegebener Fragesatz wird einem Sprachverarbeitungssystem übergeben, das den Satz inhaltlich auswertet und das Ergebnis mit einer Wissensbasis abgleicht. Die Wissensbasis enthält das Domänenwissen (z.B. technisches Knowhow). Für die zum Fragesatz passenden Einträge in der Wissensbasis kann dann ein Textgenerierungssystem einen entsprechenden Text als Antwort (neu) generieren. Im zweiten Fall wird für einen eingegebenen Fragesatz eine Textpassage im zugrundeliegenden Text gesucht, der eine Antwort auf die gestellte Frage darstellt. Hierfür werden auf den zugrundliegenden Text Methoden aus der Informationsextraktion angewandt. D.h., es werden Konzepte bereitgestellt, die die im Text vorkommenden Sätze konzeptionell beschreiben können. Wenn ein natürlich-sprachlicher Fragesatz auf bestimmte Konzepte und den zugeordneten Textinformationen paßt, so wird die entsprechende zugrundeliegende Textpassage als Antworttext präsentiert.
 **Automatische Textzusammenfassung:**

 Systeme zur automatischen Zusammenfassung von Texten sollen für einen vorgegebenen Text eine Zusammenfassung erstellen. Systeme zur automatischen Textzusammenfassung spielen bei der Recherche im Internet (und Intranet) eine wichtige Rolle. Da man nicht alle aufgefundenen Textdokumente komplett durchlesen möchte und kann, bietet es sich an, daß dem Benutzer für jedes gefundene Textdokument eine automatische Zusammenfassung generiert wird. Dadurch kann sich der Benutzer viel leichter einen Überblick verschaffen. Den Grad der Zusammenfassung sollte vom Benutzer selbst eingestellt werden können (d.h., von sehr knapper Zusammenfassung bis hin zu einer umfangreicheren Zusammenfassung).
 Im Rahmen der automatischen Textzusammenfassung gibt es zwei Ansätze:
- Den wissensbasierten Ansatz und
- den statistischen Ansatz.
 Eine Möglichkeit im Rahmen des wissensbasierten Ansatzes ist es, die wichtigen von den unwichtigen Textteilen zu trennen. Hierbei wird für jeden Textabschnitt (eines Textes) die Kernaussage ausfindig gemacht. Die restlichen Aussagen des Textausschnittes bilden nur ergänzende Informationen zur Kernaussage des Textabschnittes. Um zu einer sinnvollen Textzusammenfassung zu gelangen, muß eine auf obigem Verfahren beruhende automatisch erstellte Textzusammenfassung mindestens alle Kernaussagen des ursprünglichen Textes enthalten, damit der Informationsgehalt des ursprünglichen Textes erhalten bleibt. Sie darf aber auch weitere Aussagen enthalten -- je nachdem, welchen Grad der Zusammenfassung der Benutzer gewählt hat. Für domänenspezifische Texte (also Fachtexte) gibt es noch eine weitere Möglichkeit, im Rahmen des wissensbasierten Ansatzes eine Textzusammenfassung automatisch zu erstellen: Unter Einsatz eines Informationsextraktionssystems können die relavanten Informationseinheiten des zugrundeliegenden Textes den domänenspezifischen Konzepten des Informationsextraktionssystems zugeordnet werden. Auf Basis der Konzepte und den zugeordneten Informationseinheiten kann dann mit Hilfe eines Textgenerierungssystems eine Textzusammenfassung erstellt werden.
 Beim statistischen Ansatz werden solche Sätze aus dem zugrundeliegenden Text für die automatisch zu erstellende Textzusammenfassung ausgewählt, die eine möglichst hohe Zahl "wichtiger" Wörter enthalten. Welche Wörter wichtig sind, muß zuvor ermittelt werden, indem schon bei der Entwicklung eines solchen Textzusammenfassungssystems riesige elektronische Textsammlungen (sog. Korpora) ausgewertet werden. Je häufiger ein Wort in einem solchen Korpus vorkommt, desto wichtiger ist es. Ist das Textzusammenfassungssystem für domänenspezifische Texte gedacht, so muß auch der auszuwertende Korpus aus entsprechenden domänenspezifischen Texten bestehen. Dadurch bezieht sich die Wichtigkeit von Wörtern auf die jeweilige Domäne. (Grund: Ein bestimmtes Wort kann in der einen Domäne sehr wichtig sein, in einer anderen Domäne spielt es kaum eine Rolle).
 **Rechtschreibkorrektur:**
 Rechtschreibkorrekturprogramme sind Systeme, die in einem gegebenen Text eventuell vorhandene Rechtschreibfehler aufspüren, markieren und die gefundenen Rechtschreibfehler selbständig korrigieren können. Rechtschreibkorrekturprogramme sind meistens Bestandteil bestehender Textverarbeitungsprogramme, können aber auch separate Anwendungen sein.
 Ein Rechtschreibkorrekturprogramm besitzt ein elektronisches (System-)Lexikon, indem alle gängigen Wörter einer Sprache (z.B. Deutsch) enthalten sind. Ein einfaches Rechtschreibkorrekturprogramm durchläuft einen Text und vergleicht die in dem Text vorkommenden Wörter mit den Wörtern seines Lexikons. Wenn es beim Durchlaufen des Textes auf ein "Wort" (eine Buchstabenfolge) stößt, das nicht im Lexikon enthalten ist, versucht es, in seinem Lexikon ein ähnliches Wort finden. Welches Wort als ähnlich gilt, wird mit Hilfe bestimmter theoretischer Verfahren festgelegt. Die reine wortweise Vorgehensweise ist aber nicht immer ausreichend. Fortschrittlichere Rechtschreibkorrekturprogramme beziehen daher bei ihrer Arbeit auch den Kontext mit ein, d.h., sie betrachten auch die Wortumgebung eines aufgefundenen Rechtschreibfehlers.