# 3.2 Sprachbasierte und multilinguale Systeme

Dr. Reinhard Busch

## 3.2.1 Interaktionskonzepte

Im Rahmen von Forschungsarbeiten werden Assistenz- und Agentensysteme entwickelt, die den Benutzer beim Zugriff auf Inhalte und Systemfunktionalität unterstützen. Eines der Ziele ist die Integration der Spracherkennungstechnik und der multilingualen Gesprächsassistenz in die Plattform, zur Unterstützung kollaborativer Tätigkeiten bei Contenterstellung und Beratungsleistungen bei Finanzdienstleitern.

Im Bereich der sprachbasierten Komponenten kommen Natural-Language-Werkzeuge zur Beschreibung der Inhalte, zur multilingualen Kommunikation, zur Anbindung von Sprach-Schnittstellen und zur Ergänzung der Fähigkeiten von virtuellen Beratern zum Einsatz. Der Benutzer kommuniziert über Tools zur sprachlichen und inhaltlichen Aufbereitung komplexer Informationen einschließlich multilingualer Übersetzung (maschinelle Übersetzung/MÜ) in natürlicher Weise mit dem System bzw. das System mit dem Benutzer.

Ein- und Ausgabemedium ist die gesprochene Sprache (Spracherkennung und Sprachsynthese), zusätzlich auf der Systemseite ein Textdisplay und eine graphische Projektion von Text- und Inhaltsrelationen, die komplexe Informationsstrukturen illustrieren und den Benutzer zur Exploration und zum Verständnis dieser Information befähigen. Die Texteingabe des Benutzers ist die klassische Form von Interaktion bei der Benutzung von MÜ. Zusätzlich sind intelligente Assistenzsysteme eingebunden, um z.B. Übersetzungen zu erleichtern, den Wechsel zwischen Sprachen zu unterstützen und automatisch fremdsprachige Phrasen in Texten phonetisch korrekt vorzulesen.

Der Informationstransfer ist durch hoch interaktive Prozesse charakterisiert, die benutzergesteuerte Interaktionsformen ermöglichen. Das System bietet sich dem Benutzer als kooperativer Partner an, die Interaktion ist benutzerinitiativ und intuitiv. Intelligente Assistenten, die die Interaktion mit dem System beim Abfragen und Einbringen von Inhalten unterstützen, erlauben dem Benutzer eine Fokussierung auf die wesentlichen Aufgaben. Zugleich werden Interaktion und Informationstransfer unterstützt durch Systemkomponenten wie mehrsprachige Lexika mit hohem Abdeckungsgrad der in Informationsnetzen verwendeten Terminologie sowie Synonym- und Äquivalenzlisten in Ausgangs- und Zielsprache.

Die Interaktion läuft zwischen Benutzer und Sprachkomponenten in beiden Richtungen ab (System ↔ Benutzer). Die dabei wirksamen Interaktionskonzepte, die einzelnen Interaktionsschritte sowie die während der Interaktion intern angestoßenen System-Reaktionen, die die intuitive und effektive Kooperation von Benutzer und System erst ermöglichen, werden im folgenden kurz skizziert.

## Interaktionskonzepte Benutzer-System

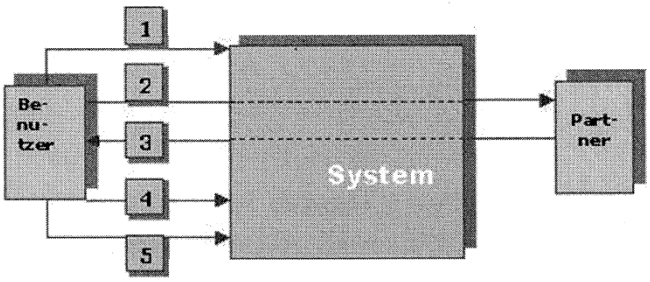


*Abb. 1: Interaktion Benutzer-System*

- Interaktion: Benutzer-System (Abbildung 1, Nummer 1)
  (Benutzer-initiativ)
  Via Sprach-Eingabe teilt der Sprecher dem System unmittelbar seine initiative Aussage und seine Intentionen in Form einer mehrteiligen Äußerung mit.
- Interaktion: Benutzer-Benutzer/Transfer-Medium (Abbildung 1, Nummer 2)
  (Benutzer-initiativ + system-kontrolliert)
  Via Sprach-Eingabe teilt der Sprecher mittelbar über das System weiteren Benutzern seine Intentionen mit („Dialog-“ und „Chat-“ Funktionen; auch mit mehreren Sprechern).
- Interaktion: Benutzer-Benutzer/Transfer-Medium (Abbildung 1, Nummer 3)
  (Partner-getriggert)
  Der Sprecher empfängt mittelbar über das System Reaktionen weiterer Benutzer als natürlichsprachliche Antwort (Sprache und Text; („Dialog-“ und „Chat-“ Funktionen).
- Interaktion: Benutzer-System (Abbildung 1, Nummer 4)
  (Benutzer-initiativ)
  Der Sprecher reagiert auf Aktionen des Systems durch Eingabe neuer Information (gesprochene Sprache; („Dialog-“ und „Chat-“ Funktionen).
- Interaktion: Benutzer-System (Abbildung 1, Nummer 5)
  (Benutzer-initiativ)
  Text-Eingabe des Benutzers
  („klassischer“ MÜ-Benutzer).

## Interaktionskonzepte System-Benutzer

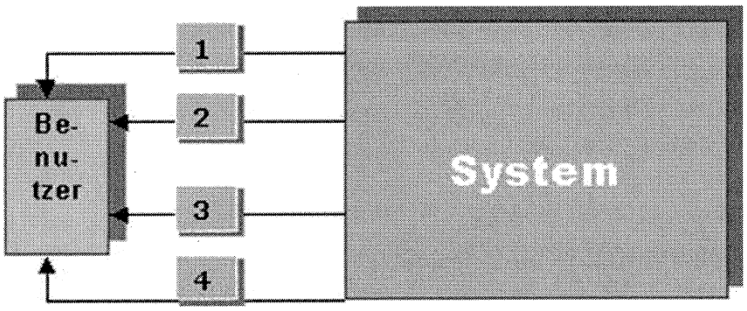


*Abb. 2: Interaktion System-Benutzer*

- Interaktion: System-Benutzer/system-initiativ (Abbildung 1, Nummer 1)
  Der Sprecher empfängt unmittelbar Reaktionen des Systems als natürlich-sprachliche Antwort – gesprochene Sprache.
- Interaktion: System-Benutzer/system-initiativ (Abbildung 1, Nummer 2)
  Der Sprecher empfängt unmittelbar Reaktionen des Systems als natürlich-sprachliche Antwort – Textdisplay.
- Interaktion: System-Benutzer/system-initiativ (Abbildung 1, Nummer 3)
  Der Sprecher empfängt unmittelbar Reaktionen des Systems als Antwort – graphische Projektion von Informationsstrukturen („assoziative Begriffsfelder").
- Interaktion: Benutzer-System/benutzer-initiativ (Abbildung 1, Nummer 4)
  Text-Ausgabe des Systems
  (bei „klassischer" MÜ-Benutzung).

**Interne System-Reaktion**

- Interaktion: System bereitet Ausgabe für weitere Benutzer vor
  (System-initiativ)
  Das System identifiziert den Spracheingabe des Benutzers (→ Spracherkennung).
- Interaktion: System bereitet Ausgabe für weitere Benutzer vor
  (System-initiativ)
  Das System bringt Sprach-Eingabe in eine textuell verarbeitbare Form (→ Sprache zu Text, Speech to Text/STT).
- Interaktion: System bereitet Ausgabe für weitere Benutzer vor
  (system-initiativ)
  Das System interpretiert und bearbeitet die transformierte Eingabe (Bereitstellung für MÜ und/oder graphische Interpretation des Inhalts).
- Interaktion: System bereitet Ausgabe für Benutzer vor
  (system-initiativ)
  Das System beschreibt und visualisiert die Strukturen des Inhalts und assoziiert Begriffe/Begriffsfelder.
- Interaktion: System bereitet Ausgabe für Benutzer vor
  (system-initiativ)
  Das System initiiert und realisiert multilinguale Bearbeitung (MÜ).
- Interaktion: System bereitet Ausgabe für Benutzer vor
  (system-initiativ)
  Das System retransformiert Text zu Sprach-Äußerungen (→ Text zu Sprache, Text to Speech/TTS)

**Hardware und Plattformen : Interaktionsmanager**

- Eingabemedium: Headset, Tastatur
- System-Plattform: PC, PDA

Die Funktionen sind auf Basis der Projekt-Struktur implementiert und in die Gesamtarchitektur der Plattform integriert. Die Voice-Funktionen sind Bestandteil der Benutzeroberfläche und sind auf den jeweiligen Zielplattformen in das Benutzer-Interface integriert. Die Steuerung der Ausprägung der Benutzeroberfläche erfolgt über den Interaktionsmanager.

## 3.2.2 Sprachsysteme: State of the Art

**Gesprochene Sprache**
Spracherkennung hat zum Ziel, dem Endbenutzer mehr Flexibilität zu ermöglichen. Die jahrzehntelange Grundlagenforschung in der Spracherkennung ist in letzter Zeit wesentlich vorangekommen. Für die allgemeine Textverarbeitung, die Telekommunikation und den Zugriff auf Datenbanken, sowie beim Dialog mit wissensbasierten Systemen, wird an besonders leistungsfähigen Spracherkennungssystemen geforscht. In steigendem Maße breitet sich Sprach-Technologie bei stimmaktiven Web-Applikationen, Smart-Phones, PDAs und anderen drahtlosen Einheiten aus. Während bis 1997 die meisten Spracherkennungsprogramme nur diskrete gesprochene Sprache unterstützten (diskret = mit einer Unterbrechung vor und nach jedem Wort), ist nunmehr Reden ohne spezielle Pausen oder Betonung mit normalem Stimmton durch kontinuierliche Spracherkennung allgemein ermöglicht. Dies bedeutet einen wichtigen Schritt zu höherem Benutzerkomfort: Eingabe in fließender Sprache statt Tastatur, besonders auch für den mobilen Einsatz.

Diktat vs. Befehl und Steuerung sind die zwei Aufgaben, die Spracherkennungsprogramme bewältigen können, d.h. Text zu diktieren und über gesprochene Befehle zu steuern. Hochwertige Systeme sind in der Lage, beide Aufgaben auszuführen.

Diktiersoftware verwandelt die Rede des Benutzers in Text. Es ist möglich, den so gewonnenen Text zu editieren, zu drucken und aufzubewahren wie ein normales maschinenlesbares Dokument. Diktatprogramme sind (noch) sprecherabhängig, d.h. sie verlangen, dass der Sprecher das System trainiert, damit es sein spezifisches Stimmenprofil und das Vokabular versteht. In neuester Zeit konnte bei der Spracherkennung die Trainingszeit eines Systems wesentlich reduziert werden. linguatecs VoicePro benötigt als kommerzielles Produkt nur noch eine Einarbeitungszeit von ca. 10 Minuten.

Sprecherunabhängige Erkennungssysteme existieren bisher nur für spezielle Domänen mit beschränktem Vokabular, z.B. Informationsdienste für Zugverbindungen, Aktienpreise, Verkehrsansagen. Während Spracherkennungssysteme mit Universalvokabular eine Installation auf dem Client-PC erfordern, können Spracherkennungssysteme mit einem eingeschränkten, auf spezielle Anwendungen ausgerichteten Vokabular auf einem Server installiert werden; zur Spracherkennung im Detail (vgl. Ihm 1999).

Obwohl Fortschritte bei der maschinellen Transkription von kontinuierlicher gesprochener Sprache mit umfangreichem Vokabular in den letzten Jahren gemacht wurden, ist die neuste Technik nur unter kontrollierten Bedingungen wie niedrigem Lärmpegel, sprecherabhängiger Erkennung, Client-Installation und vorgelesener Sprache wirkungsvoll.

Das schwierigste Problem in der Spracherkennung ergibt sich aus der Tatsache, dass ein Wort nie zweimal auf gleiche, identische Weise ausgesprochen werden kann, so sehr man sich auch bemüht. Die größere Varianz der Aussprache ist durch physische und psychische Verhältnisse bedingt, denen der Sprecher möglicherweise unterliegt, ebenso wie durch den sprachlichen Kontext, die Merkmale des Mikrophons oder andere Begleitumstände. Außer Prosodie und Tonhöhe wird beim Gespräch zusätzliche Information über Gestik und Ge-

sichtsausdruck übermittelt, z.B. über die Stimmung einer Person. Diese Informationskanäle sind der Maschine unzugänglich.

Ein anderes Problem bei der Spracherkennung ergibt sich aus der Tatsache, dass Menschen nicht nur in der Lage sind, physisch zu verstehen, was geäußert wird, sondern vor allem die Bedeutung der Äußerung zu erfassen. Wenn etwa die Qualität der gesprochenen Sprache beträchtlich durch Hintergrundlärm wie Musik und Stimmengewirr beeinträchtigt wird, ist der Mensch immer noch sehr gut in der Lage, die richtige Bedeutung zu erraten, weil durch Erwartungshaltung und Kontext klar wird, was gesagt wird.

Ein besonderes Problem ergibt sich, wenn Hintergrundgeräusche zeitlich stark schwanken oder sich laufend verändern, wie z.B. im fahrenden Auto, wo in entsprechenden Situationen das Gespräch völlig im Fahrgeräusch untergehen kann, im nächsten Augenblick aber wieder verständlich wird. Bisher übliche Methoden arbeiten unflexibel, indem ein durchschnittliches breitbandiges Rauschen generell vom Gesamtsignal abgezogen wird. Dabei wird die Dynamik des sich ständig verändernden Rauschens nur unvollkommen berücksichtigt.

Im Max-Planck-Institut für Physik komplexer Systeme in Dresden wurde ein flexibles Verfahren entwickelt, das nicht-deterministische und nicht-monotone Rauschsignale, die ein Spracherkennungssystem nicht von vornherein kennt, beherrschen kann (vgl. Wahlster 2000). Das Rauschunterdrückungssystem basiert auf der Theorie des deterministischen Chaos, d.h. im scheinbar rein chaotischen Verhalten (bei nicht-deterministischen Rauschsignalen) werden wiederkehrende Strukturen aufgedeckt.

Neueste Forschungs- und Entwicklungstendenzen bestehen darin, Lippenbewegungen und andere Äußerungen der Artikulationswerkzeuge des Menschen in den Spracherkennungsprozess einzubeziehen, um daraus die kontinuierliche Wortfolge abzuleiten und in Text umzusetzen. Nach Versuchen mit ersten Prototypen besteht die Aussicht, Erfolg versprechend bei der Artikulation die Lippenbewegungen des Sprechers zu identifizieren sowie über eine am Headset integrierte Kamera Zunge und Zähne zu beobachten und somit eine verbesserte Spracherkennung bzw. eine Erkennung ohne Stimmeinsatz oder -interpretation zu erreichen (IBM-Prototyp; vgl. Butscher 2003).

Sprachsynthese-Systeme sind mittlerweile in der Lage, gesprochene Sprache in vernünftiger Qualität zu produzieren. Verfügbare Sprachcodierer können Sprache mit unter 10 kb/s übertragen. Dies hat in Verbindung mit der wachsenden Geschwindigkeit von Mikroprozessoren und Signalverarbeitungshardware zu einer steigenden Anzahl von praktischen Anwendungen geführt und weitere Forschung stimuliert. Im Allgemeinverständnis bezieht sich Sprachsynthese auf Systeme, die aus einem unbegrenzten Vokabular von Wörtern einen unbeschränkten Eingabetext synthetisieren können. Derartige Systeme versuchen sich an der schwierigsten Aufgabe der Sprachsynthese und werden z.B. zum Lesen von E-Mails eingesetzt, wo es unmöglich ist, den Inhalt vorherzusagen. Sie unterliegen dabei einer Reihe von technischen Bedingungen.

Sprachsyntheseprogramme konvertieren schriftliche Eingabe in gesprochene Ausgabe durch automatisches Generieren von synthetischer, gesprochener Sprache. Mechanische Sprach-Einheiten wurden rasch durch elektronische Sprachsimulatoren ersetzt. Modelle des mensch-

lichen Stimmtrakts in Form von elektrischen Schaltungen konnten unter Verwendung von Basis-Artikulations-Modellen basierend auf elektrisch-akustischen Theorien erstellt werden und führten zur Realisierung von dedizierten Hardware Sprachsynthesizern.

Fortschritte in der Informatik haben das Forschungsfeld um Sprachsynthese so ausgeweitet, dass nicht nur die menschliche Sprachausgabe, sondern auch die Verarbeitung von Text modelliert ist. Dabei kommt eine Reihe von Regeln zur Anwendung, die von phonetischen Theorien und akustischen Analysen abgeleitet sind. Diese Technik wird regelbasierte Sprachsynthese genannt. Es gibt jedoch keine systematischen Studien um zu bestimmen, wie und wo die akustischen Parameter von Einheiten am besten zu extrahieren sind oder welches Sprachkorpus als optimal betrachtet werden kann.

In Sprachsynthesesystemen werden auch Spracheinheiten verwendet, die normalerweise kleiner als Wörter sind, um gesprochene Sprache von beliebigem Eingabe-Text zu synthetisieren. Da es beispielsweise im Englischen 10.000 verschiedene mögliche Silben gibt, wurden wesentlich kleinere Einheiten wie Phoneme und dyads (Phonempaare) als Typen modelliert. Insgesamt gibt es folgende Einheiten: Diphone, Triphone, Halbphoneme, Halbsilben, Silben, Wörter und sogar verschiedene längere Einheiten. Die spektralen Merkmale eines Sprachsegments variieren je nach phonetischem Kontext, so, wie er durch benachbarte Phoneme, durch Betonung und positionelle Unterschiede definiert ist. Für die Synthese von natürlich-klingender Sprache ist es wesentlich, Prosodie zu steuern und entsprechend Rhythmus, Tempo, Akzent, Intonation und eine korrekte Betonung sicherzustellen.

Sprachsynthesesysteme können an der Fähigkeit zum Text-Processing, der Fähigkeit zur Aussprache inklusive (Eigen-)Namen, der Verständlichkeit und der Natürlichkeit gemessen werden. Es gibt aber keine universell akzeptierten Referenz-Tests oder Standardsysteme. Die informelle vorwissenschaftliche Auswertung eines Systems, indem lediglich ein paar Sätze angehört werden, kann zu ungenauen Schlüssen über Textverarbeitungsfähigkeit oder Aussprachefähigkeiten des Systems führen. Informelles Zuhören könnte zwar einen besseren Einblick in Verständlichkeit und Natürlichkeit eines Systems liefern, jedoch haben manche Systeme bei bestimmten Sätzen eine bessere Leistung als bei anderen. Daher ist für die Beurteilung erst eine weitergehende, evaluierende Auswertung wirksam. Hierbei ist der Unterschied zwischen Verständlichkeit und Natürlichkeit wichtig. Die Verständlichkeit eines Systems ist der Maßstab, wie gut ein Hörer verstehen kann, was gesagt wird. Natürlichkeit ist stärker subjektiv und bezieht sich im Wesentlichen darauf, wie sehr die synthetische Stimme wie die eines Menschen klingt.

Systeme, die mit der akustisch parametrisierten, regelbasierten Formant-Technologie arbeiten, bieten gesprochene Sprache mit geminderter Natürlichkeit an, sind aber in der Lage, Text aus einem unbegrenzten Vokabular mit unbeschränktem Eingabetext zu synthetisieren. In einer virtuellen Situation, die nicht auf einen fixierten Inhalt beschränkt ist, wird der Benutzer eine akzeptable Sprachausgabe für eine offene Domäne mit ausgedehntem Basisvokabular fordern. Systeme mit diphonbasierter Verkettungstechnik dagegen können für eine eingeschränkte Domäne fast natürlich klingende gesprochene Sprache produzieren. Text außerhalb dieser Domäne jedoch wird in vielen Fällen in minderer Sprech-Qualität produziert oder kann im schlimmsten Fall überhaupt nicht gelesen werden. In dieser technologi-

schen Linie sind konkatenierende Systeme entwickelt, die nur in einer eingeschränkten Domäne operieren und Wort- oder Phrasen-Verkettungstechniken anwenden.

Eine weitere konkatenierende Technologie ist die Unit-Selection-Synthese. Die Grundidee dabei ist, dass auf einer Menge natürlichsprachlicher digitalisierter Signale und deren Zerlegung in zeitliche Intervalle mit eindeutigem Index eine Abbildungsvorschrift zwischen den phonetischen und prosodischen Symbolen der Eingabe und den Indizes der Bausteine konstruiert wird. Die Intervalle beschreiben einen sinnvollen Abschnitt der Sprachsignale. Die Konkatenation der durch die Indizes bezeichneten Signalabschnitte ergibt das gesuchte synthetische Signal. Die Indizes werden dabei derart gewählt, dass das resultierende Sprachsignal optimal im Sinne der Sprachwahrnehmung ist. Die erzeugten Referenzsignale sollten ähnliche Ergebnisse liefern wie natürliche Sprache.

Verständlichkeit und Natürlichkeit stehen nicht in Korrelation. Ein System kann äußerst verständlich, aber nicht sehr natürlich sein oder natürlich sein, aber eine schwache Verständlichkeit haben. Natürlichkeit ist wichtig, Verständlichkeit allerdings ist von größerer Tragweite; denn bei den meisten Anwendungen ist es irrelevant, wie angenehm das System klingt, wenn man nicht verstehen kann, was gesagt worden ist. Gute Ergebnisse werden erst mit den besten Features beider Ansätze erzielt. Ein diphonbasierter verkettender Unit-Selection-Algorithmus zur Synthese liefert die derzeit besten Resultate, insbesondere, wenn Prosodiefaktoren selektiv modifizierbar sind und eine Feinkontrolle der für die Synthese zu wählenden Einheiten ausgeübt werden kann (z.B. Wort-Konkatenierung, vgl. Wahlster 2000). Ein derartiger Ansatz kommt auch bei der Sprachsynthese in INVITE zum Einsatz (vgl. Kapitel 3.2.3).

**Maschinelle Übersetzung**

Über den direkten Ansatz für maschinelle Übersetzung (MÜ), bei dem in der Frühphase MÜ mehr als ein kryptographisches Problem gesehen wurde, in dem ein ausgangssprachliches (Text-)Symbol durch ein zielsprachliches ersetzt wird (so genannte „Wort-für-Wort"-Übersetzung), versuchte man infolge der schlechten Ergebnisse mit neuen Ansätzen aus dem Bereich der Computerlinguistik und anderer Disziplinen die Ergebnisse zu verbessern.

Seit Beginn der achtziger Jahre wurden Systeme entwickelt, die sich Forschungsergebnisse aus dem Bereich der künstlichen Intelligenz (KI) zunutze machten: In so genannten wissensbasierten Systemen sollte die Implementierung von Weltwissen semantische Mehrdeutigkeiten auflösen. Der Interlingua-Ansatz benutzt eine semantisch motivierte Repräsentationssprache. Die Interlingua ist eine sprachunabhängige, abstrakte Repräsentation der Bedeutung des ausgangssprachlichen Satzes, die unabhängig von der Ausgangssprache ist und aus der direkt der Satz in entsprechende Zielsprachen generiert werden kann. Um die Bedeutung ermitteln zu können, ist eine tiefe linguistische Analyse nötig.

Theoretisch ist der Ansatz interessant, da er modular ist und neue Sprachpaare jeweils nur zwei neue Module benötigen: die Analyse, die die komplexe Interlingua-Repräsentation aufbaut und die Generierung, die den zielsprachlichen Satz aus der Interlingua generiert. Praktisch ergibt sich jedoch das Problem, dass eine vollständige Interlingua ein Welt-Modell mit allen semantischen Merkmalen enthalten müsste. Eine solche sprachunabhängige Reprä-

sentation ist ein bisher ungelöstes Problem. In jüngster Zeit gibt es wieder vermehrt Forschungsprojekte im Bereich von Interlingua und Semantik (z.B. das internationale Projekt zur Realisierung der Universal Networking Language (UNL); vgl. Uchida 2003).

Die sprachenpaarabhängige, syntaktisch orientierte Transfer-Architektur besitzt neben Grammatiken und Lexika eine sprachenpaarabhängige Transferkomponente zwischen ausgangssprachlicher und zielsprachlicher Text-Repräsentation. Die Übersetzung wird in drei verschiedenen Phasen durchgeführt: in der Analyse wird der Quelltext analysiert und als abstrakte Repräsentation – (meist) in Form eines syntaktischen Baumes – an den Transfer weitergeliefert, in dem sprachenpaarbezogene Transformationen ausgeführt werden, bevor der Text in der Zielsprache generiert wird. Die meisten kommerziellen Systeme, die sich im praktischen Einsatz bewährt haben, basieren auf dem Transferansatz. Linguatec bildet hier keine Ausnahme, benutzt allerdings einen modernen und effektiven State-of-the-Art-Grammatikformalismus (die so genannte „slot grammar"; vgl. McCord 1990).

Seit den 90er Jahren kommen nicht-linguistische statistische Ansätze zur Anwendung, die mit bedingten Wahrscheinlichkeiten und stochastischen Modellen arbeiten. Benutzt wird ein „Übersetzungsmodell", um durch Vergleich aus verschiedenen Übersetzungshypothesen, die von einem Algorithmus generiert werden, den besten zielsprachlichen Satz herauszufiltern. So wird beispielsweise mit Hilfe eines Übersetzungsmodells ein englischer Satz gesucht, der die wahrscheinlichste Übersetzung eines gegebenen französischen Satzes ist und gleichzeitig wird mittels des englischen Sprachmodells überprüft, ob der englische Satz wohlgeformt und korrekt ist. Dies kann zur bedingten Wahrscheinlichkeit berechnet werden.

Das Problem statistischer Ansätze ist, dass sowohl das Übersetzungsmodell als auch das Sprachmodell sehr komplex sind und eine Vielzahl von Parametern benutzen. Die Präzision der Systeme hängt stark von der Menge der Trainingsdaten ab. Umfangreiche statistische Systeme sind insofern meist im Forschungskontext zu finden (vgl. z.B. Höge 2002).

In jüngster Zeit gibt es eine verstärkte Tendenz zu hybriden Systemen. unterschiedliche Ansätze werden in einem MÜ-System integriert, um die Vorteile verschiedener Verfahren zu nutzen und ihre Schwächen zu umgehen.

Eine der hybriden Methoden besteht darin, den zu übersetzenden Satz parallel verschiedenen MÜ-Systemen mit unterschiedlichen technologischen Ansätzen einzugeben und die gesamte Ausgabe zu kombinieren, indem die besten Phrasen zusammengesucht werden und der zielsprachliche Satz aufgebaut wird. Für das beste Ergebnis werden zugewiesene Korrektheitswerte normalisiert und bei der Integration statistisch ausgewertet.

In Zukunft sind sicherlich generell kombinierte Ansätze die Erfolg versprechende Lösung, sei es die Kombination von Übersetzungsarchiv und maschineller Übersetzung, hybride statistisch-linguistische Übersetzungsverfahren oder ein Multi-Engine-Übersetzungssystem, oder eine Kombination von allen. Vor der weitgehenden Verfügbarkeit derartiger System auf Produktniveau ist allerdings derzeit noch Forschungsarbeit nötig.

**Speech-to-Speech-Übersetzung**
Ein bedeutsames aktuelles Entwicklungsfeld sind Übersetzungssysteme für gesprochene Sprache – einer der innovativsten Bereiche der MÜ. Gegenwärtig ist die vollständige Speech-to-Speech-Übersetzung mit Sprach-Ein- und Ausgabe meist ein Thema der Forschung. Labor-Prototypen können nur eingeschränkte Domänen behandeln, und die Forschung untersucht hauptsächlich, wie die verschiedenen Techniken der Verarbeitung von gesprochener und geschriebener Sprache integriert werden können.

Ein erster Schritt bei der Entwicklung von Speech-to-Speech-Übersetzungen ist das Verbmobil Projekt des BMBF (vgl. Wahlster 2000). Verbmobil hat einen Prototyp eines MÜ-Systems, das als Dolmetscher fungiert, entwickelt; die automatische Übersetzung einer gesprochenen sprachlichen Äußerung und ihre zielsprachliche Wiedergabe ebenfalls in gesprochener Sprache. Der Prototyp ist sprecherunabhängig gestaltet, verarbeitet spontane gesprochene Sprache und bietet Hilfe bei mehrsprachigen Dialogsituationen an.

Prinzipiell ist festzustellen, dass bisher die Qualität von Speech-to-Speech-Übersetzungen nicht sehr zufrieden stellend war. Falsch erkannte gesprochene Worte können nämlich dazu führen, dass die gesamte Satzkonstruktion vom MÜ-System falsch analysiert und übersetzt wird, d.h. hier potenzieren sich mögliche Fehler beider Technologien (Spracherkennung/MÜ. Da aber inzwischen auch kommerzielle Spracherkennungstools bessere Ergebnisse liefern, ist die automatische Speech-to-Speech-Übersetzung realistisch geworden. Ein effektiver Anwendungsrahmen für die Technologie bietet sich in INVITE durch den automatischen Fremdsprachen-Assistenten. linguatec begann als erstes Unternehmen, Spracherkennung, MÜ und Sprachsynthese für einen marktfähigen Produktprototyp zu kombinieren und brachte auf Grund der Erfahrungen das Produkt „talk & translate“ (vgl. www.linguatec.de) auf den Markt. Auf der Basis dieser Technologie wurde im Rahmen des INVITE-Projekts das multilinguale Konferenz-/Chat-System entwickelt (siehe Kapitel 3.2).

### 3.2.3 Innovationen und Ergebnisse

Eines der Interaktionsparadigmen in INVITE ist die „Multimodale Interaktion für kreative Team-Sitzungen“. In diesem Rahmen wurde ein produktnaher Prototyp („automatischer Fremdsprachen-Assistent“) entwickelt. Der Fremdsprachen-Assistent hat zum Ziel, Benutzer mit unterschiedlichen Voraussetzungen bei der mehrsprachigen Kommunikation effektiv und intelligent zu unterstützen.

Das intelligente Assistenzsystem ist durch den aufeinander abgestimmten Einsatz aktueller Sprachtechnologien charakterisiert, um möglichst weitgehend mittels innovativer Mensch-Maschine-Technologie eine intuitive und reziproke Interaktion zwischen Benutzer und System zu ermöglichen.

Das Paket „Fremdsprachen-Assistent“ wurde dabei unter mehrfachen Aspekten bearbeitet und so realisiert, dass eine Assistenz in zwei zentralen Anwendungsbereichen angeboten ist:

1. Fremdsprachlicher Beratungsassistent, realisiert als System für multilingualen Chat;
2. Assistent zur Gesprächsunterstützung durch Begriffsassoziationen.

In der Projektlaufzeit gelang es, innovative Entwicklungen für beide Gebiete zu realisieren und positive Ergebnisse zu erzielen, die als an der Spitze der Innovation stehend bezeichnet werden können. Für die Arbeiten am multilingualen Chat wurde linguatec im Herbst 2003 von der Europäischen Kommission mit den European IST Prize ausgezeichnet.

Neben Neu-Entwicklungen wurden innovative und funktionale Verbesserungen am vorhandenen maschinellen Backend-Übersetzungsystem vorgenommen, die positive Auswirkungen auf den Einsatzbereich und die Kommunikationsfähigkeit des Systems gezeigt haben. Prinzipiell wurden darüber hinaus innovative Methoden für die Realisierung des technischen Umfelds, d.h. die progammtechnische Umsetzung, die eigentliche Implementierung und die Konfiguration der Hardware-Umgebung erarbeitet und realisiert.

Im Folgenden werden beide Anwendungsbereiche des maschinellen Gesprächsassistenten ausführlicher vorgestellt.

**Multilingualer Chat**

Mehrsprachige Dialoge

Die Unterstützung mehrsprachiger Dialoge in e-Business-Anwendungen (Chats) ist ein neues Merkmal, das es ermöglicht, Gesprächsteilnehmer unterschiedlicher Muttersprache und unterschiedlicher Sprachfähigkeiten in einen Dialog einzubeziehen (vgl. Abbildung 3):

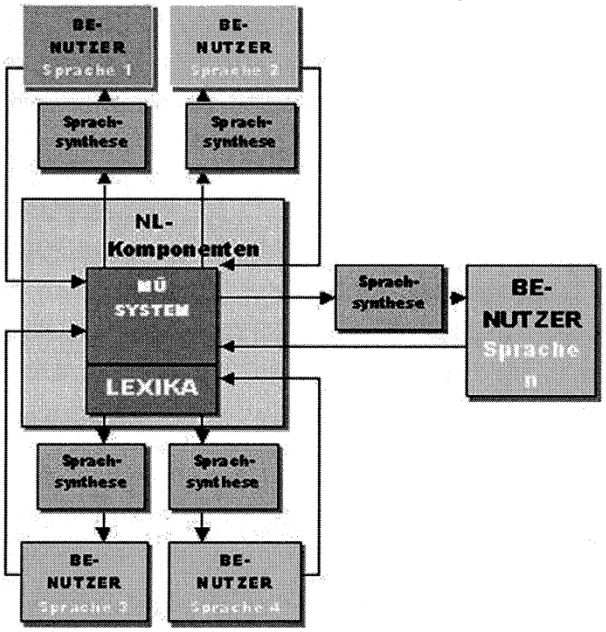


*Abb. 3: Interaktion System-Benutzer*

Ziel für den Ausbau einer solchen kollaborativen Anwendung war im INVITE Projekt die Realisierung einer multilingualen Kommunikationplattform zwischen verschiedensprachigen Partnern mit Szenarien im Bereich e-Commerce, d.h. community-basierte Erstellung und Nutzung komplexer Finanzdienstleistungen und Kundenberatung.

Der für diesen Zweck entwickelte multilinguale Chat-Assistent erlaubt mehrsprachige Sitzungen verschiedensprachiger Benutzer, die ihre Sprachen gegenseitig nicht verstehen, aber eine gemeinsame Verständigung über die für eine Sitzung aktuellen Themen erreichen wollen. Das System vereint neueste Technologien und linguistische Strategien. Zur Systemfunktionalität vgl. Abbildung 4, die die Benutzer-Oberfläche während einer Beratungsitzung mit drei Partnern zeigt – einem Franzosen, einer Engländerin und einem Deutschen:

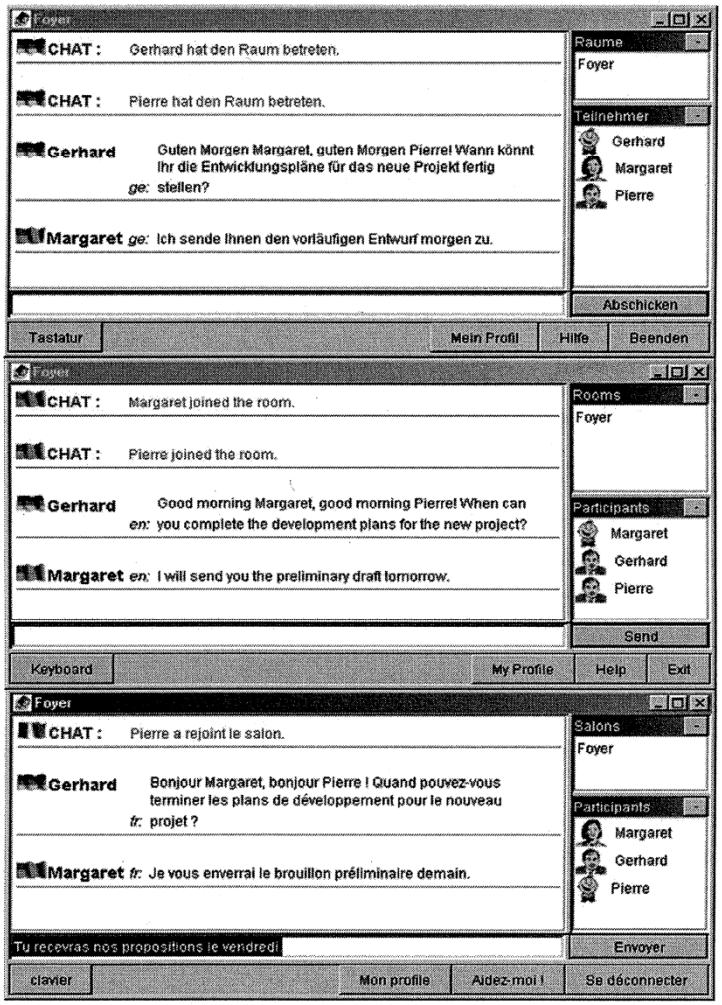


*Abb. 4: Multilingualer Chat; Interface*

Die jeweilige Quellsprache wird maschinell bearbeitet. Eine spezielle Lösung ermöglicht ihre automatische Identifikation. Beim Öffnen eines Dokumentes oder der Eingabe eines Dialogtextes wird auf diesem Weg festgestellt, um welche Sprache es sich handelt („language identifier"). Wenn mehrere Sprachen in einem Text bzw. Dialog vorkommen, kann dies ermittelt und angezeigt werden. Im Dialog wird/werden folgerichtig die entsprechende(n) Übersetzung(en) aktiviert und der identifizierte Eingabe-Text automatisch in die Sprachen der anderen Teilnehmer an der Konferenz übersetzt.

Das System verkörpert somit das die Partner verbindende Kommunikationsglied, das nicht nur die fließende Kommunikation der Chat-Sitzung gewährleistet, sondern auch Sprachbar-

rieren überwindet und Kommunikation zwischen Gesprächspartnern ermöglicht, die nicht die Sprachen ihrer jeweiligen Partner beherrschen.

Integrierte Sprachsynthese
Eine weitere Funktionalität des Fremdsprachen-Assistenten besteht darin, dass die Übersetzung mit Sprachsynthese gekoppelt ist, d.h. alle übersetzten und in der jeweiligen Sprache generierten Texte werden dem Partner in seiner Muttersprache vorgelesen. Somit erfährt jeder Teilnehmer in seiner Sprache vom System die vom Partner gesendeten und über die Tastatur eingetippten Nachrichten nicht nur schriftlich, sondern hört sie auch in gesprochener Sprache. Der multilinguale Chat-Assistent mit integrierter Sprachsynthese hat in dieser Funktionalität ein Alleinstellungsmerkmal im Vergleich zu anderen bekannten Systemen.

Die Sprachsynthese-Funktionalität ist sukzessiv integriert, wobei eine Kombination der modernsten Technologien der Sprachsynthese (diphon-basierte Konkatenationssynthese mit Unit-Selection Synthese) evaluiert und angewendet wurde. Die Anforderungen nach einer möglichst natürlichen Mensch-Maschine-Interaktion werden am besten von einem Verfahren erfüllt, das die Robustheit der Diphonsynthese mit der klanglichen und prosodischen Qualität der Unit-Selection Synthese kombiniert. Dieses Verfahren wurde gewählt und integriert. Das Ergebnis ist eine weitestgehend natürliche Sprachsynthese, die gut verständlich ist. Die Multimodalität ist durch die Integration einer prosodie-optimierten Sprachsynthese erweitert.

Ein besonderes Kennzeichen ist die schnelle Adaptierbarkeit des Sprachsynthesemoduls hinsichtlich unterschiedlicher Domänen- und Terminologieanforderungen. Dabei orientieren sich die Tools für eine phonetische Konvertierung von Eigennamen und Neologismen an internationalen Standards wie z.B. IPA (International Phonetik Alphabet) und SAMPA (Speech Assessment Methods Phonetic Alphabet).

Auswirkung auf die MÜ und technische Realisierung
Im Bezug auf die MÜ als Backend-System wurden die Erkennungs- und Verarbeitungsalgorithmen hinsichtlich der Finanztexte und -terminologie verfeinert und optimiert. Die Lern- und Anpassungsfähigkeit der Plattform ist durch eine automatische Identifizierung von unbekannten Begriffen und eine benutzerinitiierte Erweiterungsoption sowie durch die Visualisierung dieser Information transparent gemacht.

Das Vokabular ist auch hier um Neologismen erweitert und besonders an die Erfordernisse von e-Business angepasst, was sich als ein fortlaufender Prozess herausgestellt hat, da gerade durch die Dynamik neuer Technologien viele neue Begriffe geschaffen werden, die zusammen mit ihren Übersetzungen verfügbar sein müssen. Die Berücksichtigung sprachlicher und übersetzungsrelevanter Merkmale in Benutzerprofilen ist ein innovativer Aspekt bei der Benutzermodellierung, der hier erstmals in diesem Rahmen untersucht und angewendet wird.

Die Integration der Sprachsynthese in den multilingualen Chat stellt besondere Anforderungen an die Synchronisierung der Datenflüsse. Die hohen Speicher- und Prozessoranforderungen für Übersetzungs- und Sprachsynthese-Server bedingen eine skalierbare, dezentrale Architektur. Um sicherzustellen, dass Text und Sprache bei allen Clients (Teilnehmern) gleichzeitig dargestellt werden, übernimmt der Chat-Server auch die erforderliche Synchro-

nisierung. Das Führen eines zwei- oder mehrsprachigen Dialogs wird durch die dezentrale Architektur für die Szenarien des e-Commerce ermöglicht.

Die Übersetzungstechnologie ist auf die Gegebenheiten der jeweiligen Anwendungssituation hin adaptiert. Für das Kundenberatungsszenario ist eine Offline-Übersetzungsfunktion installiert, die auf die Übersetzung von Nachrichten im Hintergrund ausgerichtet ist und jederzeit vom Benutzer angefordert werden kann. Die Übersetzungsfunktion wurde dabei auf die individuellen Fremdsprachenkenntnisse des Benutzers abgestimmt. Die Skala reicht vom situativen Terminologie-Lookup bis hin zur Volltextübersetzung.

Die wissenschaftlich-technischen Ergebnisse tragen dazu bei, mehrsprachige Beratungsdialoge und die Nutzung fremdsprachiger Informationsquellen zu erleichtern. Es ergeben sich vielfältige Anwendungsmöglichkeiten, die sich zeitnah umsetzen lassen.

**Gesprächsunterstützung durch Begriffsassoziationen**

Die zweite Anwendung, bei der Gesprächsassistenz geleistet wird, ist die Unterstützung des Gesprächs durch visuelle, dynamisch die Gesprächsthematik aktualisierende Hilfsmittel. Im Rahmen des Einsatzszenarios lassen die Gesprächspartner den von ihnen gesprochenen Text per Spracherkennung aufzeichnen. Das Gesprächsprotokoll wird ausgewertet, indem aus dem laufenden Textstrom wichtige Begriffe und ihre Zusammenhänge isoliert werden, um

- den Inhalt des Gesprächs auszuwerten und eine Übersicht über die zum Thema gehörenden Fachbegriffe zu erhalten;
- eine graphische Repräsentation des Gesprächgegenstandes zu erzeugen;
- den Gesprächspartnern Informationen über weitere damit zusammenhängende Begriffe zu geben. Diese werden automatisch assoziiert und für neue Zwecke aktualisiert, z.B. der thematischen Ausweitung des Gesprächs (vgl. das Schema in Abbildung 5):

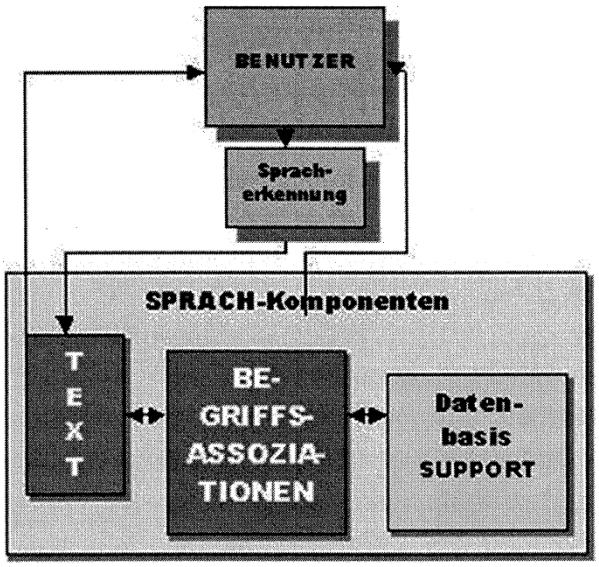


*Abb. 5: Datenfluss – Gesprächsunterstützung durch Begriffsassoziationen*

Darüber hinaus ist es möglich, dass bei einem Gespräch zwischen Experten einem Interessierten typische Fachbegriffe und Relationen zwischen diesen graphisch vermittelt werden. Dieses Wissen kann er nutzen, um ins Gespräch einzugreifen und kompetent Zusammenhänge zu erfragen oder Stichwörter für den weiteren Gesprächsverlauf zu geben. (Vgl. z.B. zur optischen Repräsentation den Assoziationsgraph des Begriffs „Börse“, Abbildung 6).

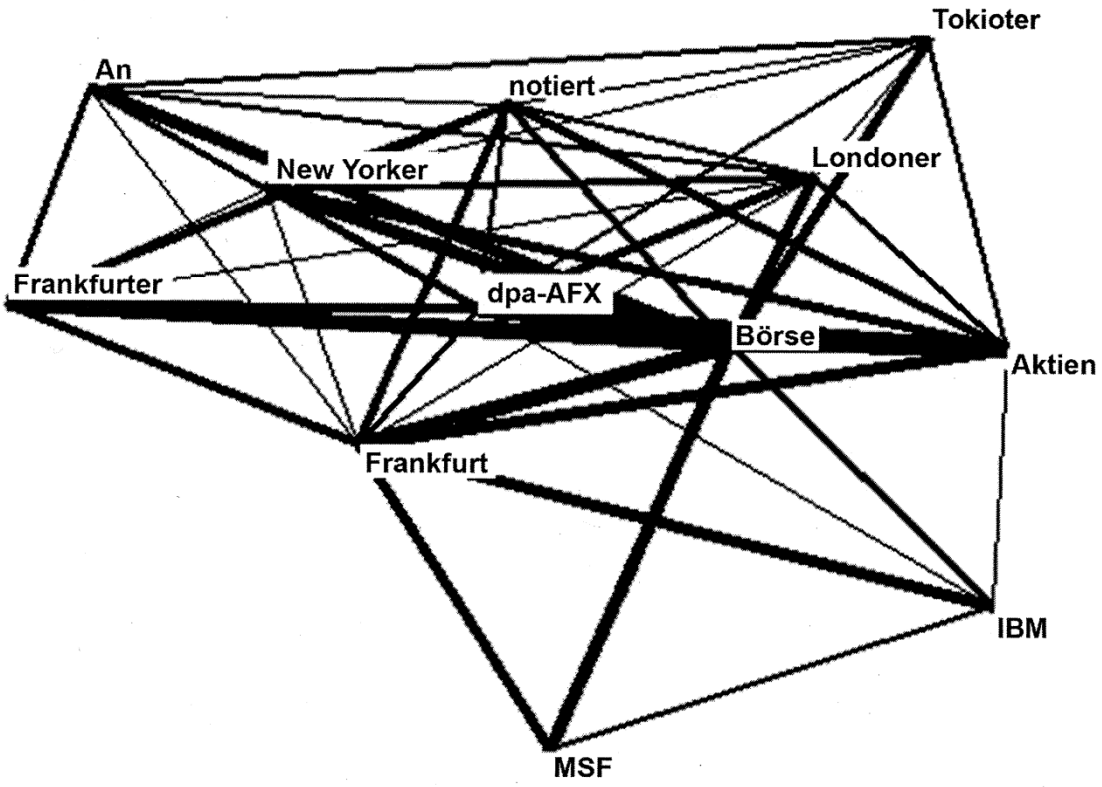


*Abb. 6: Assoziationsgraph für „Börse“*

Diese Art der Gesprächsunterstützung basiert auf gesprochener Spracheingabe. Die gesprochenen Passagen werden identifiziert und in schriftlichen Text umgewandelt. Der Text wird in Echtzeit auf dem Bildschirm gezeigt.

Gleichzeitig wird der Text interpretiert, auf Schlüsselwörter untersucht und über diese inhaltlich erweitert. Dies geschieht, indem identifizierten Schlüsselwörtern verwandte Begriffe assoziiert und graphisch sichtbar gemacht werden. In einem separaten Fenster werden Graphen präsentiert, deren Knoten Schlüsselwörter und ihnen assoziierte Begriffe sind. Dabei sind Schlüsselwort und assoziierter Begriff graphisch über Kanten verbunden und werden im Gesprächsverlauf dynamisch aktualisiert bzw. entsprechend dem Inhalt des Gesprächs sowie den in der Datenbasis gefundenen Kollokationen erweitert (vgl. Abbildung 7):

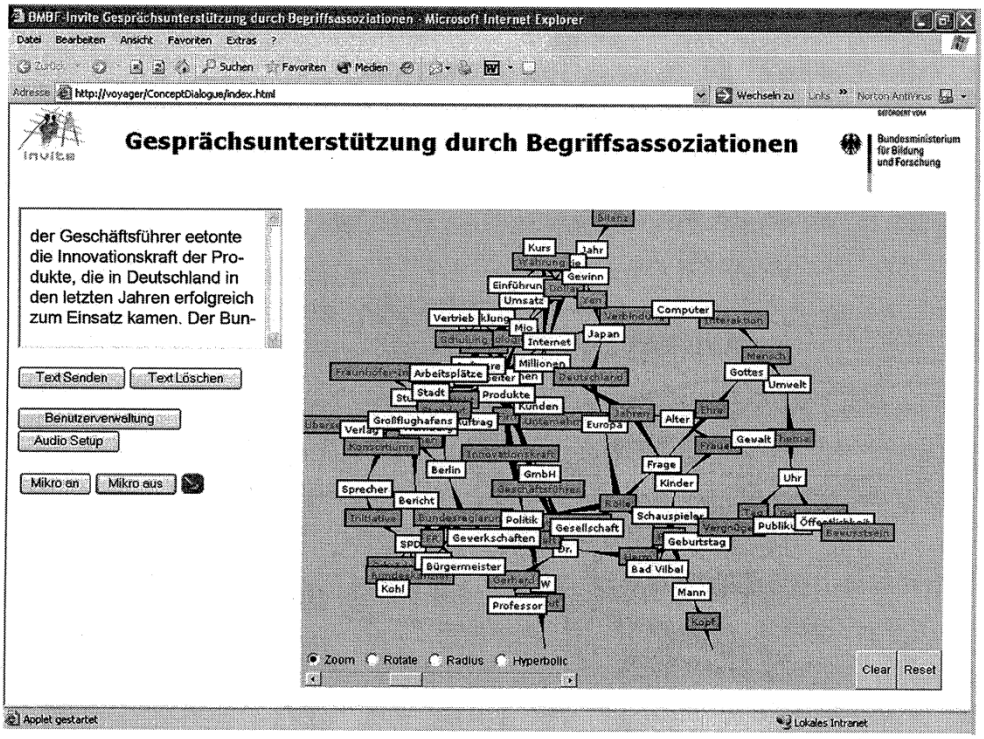


*Abb. 7: Gesprächsunterstützung durch Begriffsassoziation – Interface*

Die Assoziationsmuster basieren auf der Auswertung einer umfangreichen Datenbasis (vgl. Heyer et al. 2003, Quasthoff 1998), in der Begriffsfelder aus Zeitungsberichten über circa 10 Jahre hin gespeichert und nach statistischen Kriterien ausgewertet wurden.

Von besonderer Bedeutung sind die Kollokationen der Begriffe, d.h. Paare von Begriffen, die statistisch auffällig gemeinsam miteinander auftreten und häufig menschlichen Assoziationen entsprechen. Das Wiederauffinden der Assoziationen im Expertengespräch ist der Kernpunkt der Gesprächsunterstützung. Die Begriffsassoziationen werden vollautomatisch ermittelt und können in Echtzeit mit den Begriffen aus dem Gespräch verglichen werden.

Die Auswertung des Gesprächs ist als so genannter Web-Service implementiert, d.h. der auszuwertende Text wird als Ganzes oder in Teilen mittels eines standardisierten Protokolls an den Server geschickt, der das gesamte bzw. partielle Auswertungsergebnis als textuelle Beschreibung und Graphik zurückschickt. Dieses Ergebnis wird auf dem Bildschirm angezeigt.

Bei der Analyse kommen folgende Daten und Methoden zum Einsatz:

- der Aufbau einer applikationsspezifischen Datenbank, die Wissen aus bisherigen Gesprächen und anderen Texten zum Gesprächsthema enthält;
- der Vergleich der Auftretenshäufigkeiten mit einer allgemeinen linguistischen Datenbank, die Daten über den „normalen" Sprachgebrauch enthält und zum Vergleich benötigt wird;
- der Text des Gesprächs, der zunächst mit linguistischen Verfahren (Segmentierung, Grundformenreduktion, Erkennung von Mehrwort-Begriffen) analysiert wird;
- die Extraktion der wichtigen Fachbegriffe mittels der Vergleichsdatenbanken;

- die Bewertung ihrer Assoziationen, die durch statistisch auffällig gemeinsames Auftreten in den gesammelten Texten gemessen wurden;
- die graphische Darstellung der relevanten Assoziationen durch Simulated Annealing (d.h. ein randomisiertes Verfahren, das in vielen praktischen Fällen gute Näherungslösungen für kombinatorische Aufgabenstellungen liefert).

Die Kombination von multilingualem Chat und integrierter (auch multilingualer) Gesprächsunterstützung durch Begriffsassoziationen ist möglich und hat den in Abbildung 8 skizzierten Datenfluss:

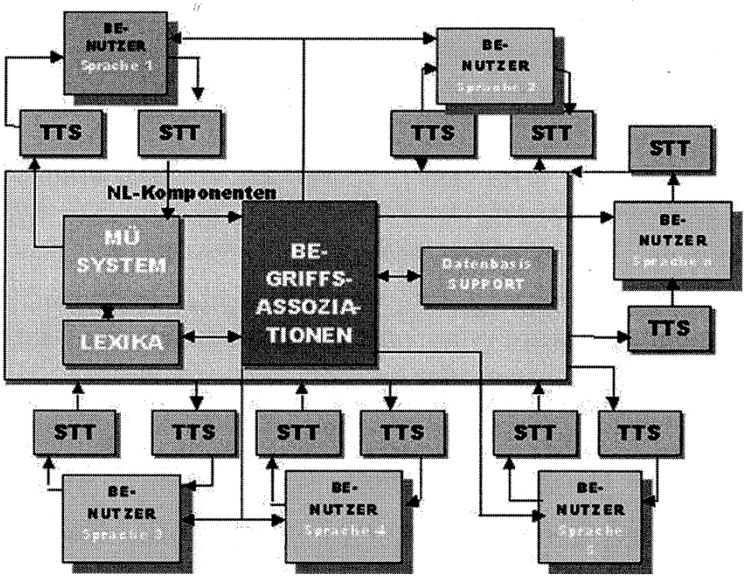


*Abb. 8: Multilingualer Chat und Begriffsassoziationen kombiniert*

Die Aufzeichnung der Diskussion in Sitzungen der Fachexperten erfolgt mittels des Systems VoicePro (Kernsoftware zur Erkennung von 1 Mio. unterschiedlichen Wortformen), das die Umwandlung von Sprach in Text (in den Abbildungen: STT) vornimmt. Anschließend erfolgt eine Extrahierung von auffälligen Themenbezügen, um damit automatisch ein erstes Themennetzwerk (in der Abbildung: Begriffsassoziationen) zu generieren. Aus dem Austausch verschiedener fachspezifischer Themen der Berater ermittelt das System VoicePro in Verbindung mit dem Assoziationsgenerator („Topic Builder") relevante Topics und erstellt daraus Vorschläge für die Strukturierung von Sub-Communities.

## 3.2.4 Bewertung und Ausblick

Der Ausbau des kollaborativen Anwendungsszenarios im Bezug auf die Mehrsprachigkeit sowie die kooperierende und intuitive Gesprächsunterstützung durch Begriffsassoziationen sind Bereiche von hoher Innovation und stellen Alleinstellungsmerkmale dar. Das Forschungsprojekt INVITE hat einen effektiven Rahmen bereitgestellt, um Spitzentechnologie im Hinblick auf die Realisierung einsetzbarer Produkte zu entwickeln. Dabei gab es auch eine Reihe von Detailverbesserungen in allen Bereichen der zum Einsatz kommenden Sprach- bzw. Übersetzungstools, die die Basis der Entwicklung bilden.

Der erreichte Entwicklungsstand des Fremdsprachen-Assistenten und des Systems zur Gesprächsunterstützung durch Begriffsassoziationen kann nach bisherigen Vergleichstests eine Spitzenstellung hinsichtlich Übersetzungsqualität, Skalierbarkeit und Adaptierbarkeit beanspruchen. Da die Prototypen des multilingualen Chats mit integrierter Sprachsynthese sowie der Generator von Begriffsassoziationen in anderen Kommunikationssystemen nicht ihresgleichen haben, können die erreichte Innovation sowie die Erfolgsaussichten für eine darauf aufbauende Entwicklung als sehr günstig eingeschätzt werden.

Die Verwertungsaktivitäten zeigten erste, erfreuliche Ergebnisse. So hat die Siemens AG nach erfolgreichem Abschluss umfangreicher Anwendertests entschieden, den Fremdsprachen-Assistenten im unternehmensweiten Intranet einzusetzen.

Anschlussaktivitäten mit beträchtlicher wirtschaftlicher Relevanz und technisch ehrgeiziger Zielsetzung bestehen in folgenden Bereichen:

- Erweiterung der Multimodalität: Integration von serverbasierter Spracherkennung mit weitgehender Sprecherunabhängigkeit unter besonderer Berücksichtigung der Anforderungen von mobilen Endgeräten.
- Ausbau der Wissensrepräsentation: Integration des Fremdsprachen-Assistenten mit Textanalyse- und Wissensextraktionsverfahren auf Basis großer Korpora.

## 3.2.5 Literatur

Butscher, R.: Report Spracherkennung – Chips mit Ohren. Bild der Wissenschaft 6/2003, Seite 96-102.

Heyer, G.; Quasthoff, U.; Wolff, C.: Möglichkeiten und Verfahren zur automatischen Gewinnung von Fachbegriffen aus Texten. In: Bullinger, H.-J. (Hrsg.): Proceedings Innovationsforum „Content-Management – Digitale Inhalte als Bausteine einer vernetzten Welt. Stuttgart: 2003-08-11.

Höge, H.: Project Proposal TC-STAR – Make Speech to Speech Translation Real. München: Siemens 2002.

Paper/LREC Conference, http://www.tc-star.org. Besucht: 14.03.2006.

Ihm, H.R.: Das große Spracherkennungsbuch. München: linguatec, 1999.

McCord, M.: English Slot Grammar – COMLEX Syntax Reference Manual. Proteus Project. New York: University 1990.

Quasthoff, U.: Deutscher Wortschatz im Internet. LDV-Forum 2, 1998.

Uchida, H.: Knowledge Description Language. Universität Tokio, Semantic Computing Workshop (Presentation), April 2003.

Wahlster, W. (Hrsg.): Verbmobil: Foundations of Speech-to-Speech Translation. Berlin-Heidelberg: Springer 2000.

Wengenmayr, R.: Rauschfreie Zone. In: Max-Planck-Forschung 2/2003, Seite 32–35.