In: Eibl, Maximilian; Wolff, Christian; Womser-Hacker, Christa (Hg.): Designing Information Systems. Festschrift für Jürgen Krause. Konstanz: UVK Verlagsgesellschaft mbH, 2005. S. 127 – 143.

Sprachtechnologie in einem Informationssystem

Gregor Thurmair

Einleitung

Das im Folgenden beschriebene System ist das Ergebnis eines Förderprojektes der EU namens SENSUS[1] und ähnlicher Arbeiten. Ziel der Projekte war die Erstellung der Architektur eines Informationssystems, das die Arbeit öffentlicher Behörden bei der Drogenbekämpfung unterstützt. Schwerpunkt des Projekts waren Aspekte der Informationsgewinnung aus Texten, der Multilingualität und der Visualisierung. Der folgende Beitrag gibt eine Beschreibung des erstellten Systems und der Erfahrungen, die damit gemacht worden sind.

1 Systembeschreibung

Folgend einer Bedarfsanalyse, sollte das System zwei Szenarien unterstützen:

- **Input-Szenario**: eingehende Daten, speziell fremdsprachliche, sollten mit intelligenten Verfahren analysiert werden können.
- **Retrieval-Szenario**: Benutzer sollten multilingual und multimodal in den gespeicherten Daten suchen können.

1.1 Input-Szenario

Hier sollen eingehende Informationen möglichst vollautomatisch inhaltlich erschlossen und aufbereitet werden. Ziel dabei ist die Erhöhung des Durchsatzes; d.h. der Weg einer Information von ihrem Eingang bis zum Bearbeiter soll signifikant verkürzt werden.

1.1.1 Input-Quellen

Als typische Medien von eingehenden Informationsquellen werden betrachtet:

- **Papier**; immer noch das Hauptmedium im Behördenbereich.
- **elektronische Dokumente**, je nach Herkunft in verschiedenen Formaten und Zeichensätzen

[1] SENSUS ist ein von der EU gefördertes Projekt, cf. www.sensus-int.de. Partner waren Unternehmen, Forschungsinstitute (CBS, ILSP) und behördliche Partner aus dem Polizei- und Nachrichtenbereich aus 8 europäischen Ländern. Details vgl. Bodenkamp 2000.

- **Multimedia-Dokumente** (Audio- und Videomaterial); dafür wurden im Projekt keine eigenen Aktivitäten etwa des Audio Mining entfaltet, für die Weiterbearbeitung wurden nur die Meta-Informationen verwendet.

Alle Input-Quellen werden in ein **einheitliches Zielformat** gebracht (**USDF** genannt: Unified Sensus Document Format, s. u.). **Elektronischer Input** wird mit Filtern bearbeitet, die die nativen Formate (HTML, ASCII, RTF) in USDF konvertieren. **Papier-Input** wird mit einem OCR-Erkenner[2] zunächst in elektronisches Format und dann ebenfalls nach USDF gebracht. Während versucht wurde, den Verlust an Retrievalqualität durch Erkennungsfehler[3] mittels Fuzzy-Techniken auszugleichen, stellt eine Übersetzung oder Informationsextraktion aufgrund der begrenzten Robustheit gegenwärtiger Verfahren höhere Anforderungen an die Erkennungsqualität, sodass man eine Komponente zur Bereinigung des OCR-Outputs ins Auge fassen muss, wenn man linguistisch basierte Weiterverarbeitungen plant.

Die **Dokumente**, die dem Projekt zur Verfügung standen, stammten einerseits aus offenen Quellen (Agenturmeldungen), andererseits wurden spezielle Dokumente angefertigt, die in der Struktur der internen Kommunikation nachgebildet waren. Insgesamt standen mehrere tausend Dokument und Berichte zur Verfügung, in mehreren Sprachen (Englisch, Deutsch, Französisch, Spanisch, Italienisch, Schwedisch u. a.) und Zeichensätzen (u. a. Kyrillisch, Griechisch). Manche der Berichte wurden übersetzt, um parallele Korpora zu haben.

Aufgrund der begrenzten Länge der einzelnen Dokumente (etwa eine halbe Seite) lag das Datenvolumen im Projekt an der unteren Grenze dessen, was statistisch basierte Verfahren erfordern; andererseits ist im praktischen Einsatz sofort mit sehr großen Dokumentenvolumina zu rechnen.

1.1.2 Input-Analyse-Komponenten

Die Eingabe-Texte, wurden dann von verschiedenen Analysetools bearbeitet.

1.1.2.1 Architektur

Für die Kommunikation dieser Tools wurde eine **Whiteboard**-Architektur gewählt[4]; dabei wird das Dokument von den einzelnen Komponenten annotiert und so mit zusätzlicher Intelligenz versehen. Das **USDF**-Format wurde entwik-

[2] Das OCR-Paket und das Standard-Textretrieval mit Fuzzy-Suche wurde von der Firma Zylab beigesteuert (ZyLAB Technologies (2004). ZyLAB Homepage. http://www.zylab.com [Zugriff September 2004]).

[3] Vgl. dazu Weber & Hechtbauer 1998.

[4] Vgl. Boitet & Seligman 1994, Neumann 2001; zu XML-basierten Analyseergebnissen auch Buitelaar et al. 2003.

kelt, um den Austausch der Information zu organisieren; es ist XML-basiert und besteht aus drei Teilen:

- einem *Header*, der generelle Informationen zum Dokument enthält (Location, Sprache, Topic, Encoding);
- einem *Body*, der den eigentlichen Text enthält, wie er von den Formatfiltern geliefert wird; dieser Abschnitt folgt dem OText-Standard (Thurmair 1997) und markiert wesentliche Layout- und Inhaltselemente (Heading, ListElement, Fontwechsel u. dgl.);
- einem *Footer*, der weitere Analyseergebnisse enthält, wie Indexterme, Ergebnisse der Informations-Extraktion, Summariser-Information u. dgl.

Auf diese Weise ist eine flexible Kombination verschiedener relativ autonomer Analysekomponenten implementierbar.

1.1.2.2 Sprachenerkennung

Die erste Komponente der Analyse ist ein Sprachen-Erkenner. Er identifiziert Sprache und Zeichencode. Er wurde implementiert gemäß dem bei Cowie et al. 1998 beschriebenen Algorithmus. Die erkannte Sprache wird in den USDF-Header geschrieben und steuert alle Folgekomponenten.

1.1.2.3 Themenerkennung

Nach der Sprachenerkennung wird ein Topic-Erkenner eingesetzt, um herauszufinden, ob es sich beim fraglichen Text um ein Dokument aus dem Drogenbereich handelt oder nicht.

Als Klassifikator wurde zunächst, angesichts der geringen Menge von Trainingsdaten und der Einfachheit der Taxonomie, ein stichwortbasierter Ansatz gewählt. In einem zweiten Schritt wurde ein Classifier entwickelt, der auf Support-Vektor-Maschinen-Technologie basiert (Goller et al. 2000).

Der Topic-Erkenner ermittelt das Topic des Dokuments und trägt es in den USDF-Header ein. Nicht drogenrelevante Texte werden auf diese Weise ausgefiltert, sodass die aufwendigeren Verfahren nur den relevanten Texten zugute kommen; so wird eine Überbelastung des Gesamtsystems verhindert.

1.1.2.4 Übersetzung

Die Benutzer hatten jederzeit die Möglichkeit, sich den Dokument-Text online maschinell übersetzen zu lassen mit dem Ziel, die Relevanz eines Textes zu beurteilen. Das Ergebnis der Übersetzung wurde nicht gespeichert, die Texte wurden nur in der jeweiligen Dokumentsprache indexiert.

1.1.2.5 Indexierung
Danach wurden zwei Verfahren eingesetzt, um die Texte weiter zu erschließen:

- Standard-Volltext-Indexing für eine Standard-Textsuche
- Grundformen-Indexing, die für eine Term-Übersetzung erforderlich ist; dabei wurde für jede Sprache ein eigener Index generiert, um sprachliche Überlappungen und damit Ballast bei der Suche zu vermeiden.

Die Ergebnisse wurden in einem kommerziellen IR-System abgelegt.

1.1.2.6 Informations-Extraktion
Dann wurden Verfahren der **Informations-Extraktion** eingesetzt, um relevante Informationsobjekte in Texten zu identifizieren. Es wurden folgende Objekttypen analysiert:

- Personen
- Orte
- Maßeinheiten
- Transportmittel
- Datumsangaben
- Firmen
- Institutionen
- Emails und URLs
- Telefon- und Faxnummern
- Waffen
- Fluglinien und Flughäfen

Die Technologie ist regelbasiert und verwendet Finite-State-Technologie. Sie wurde in den genannten acht Sprachen implementiert.

Die Ergebnisse der Extraktion werden im USDF-Footer gesammelt und dann in einer relationalen Datenbank abgelegt, damit sie mit anderen Objekten verknüpft und so von Data-Mining-Verfahren über automatisch aus Texten extrahierten Informationsobjekten benutzt werden können.

Über die Erkennung von Informationselementen hinaus wurde auch an der Analyse von Relationen zwischen den Elementen gearbeitet, speziell an einer Repräsentationsform für diese Relationen[5]. Die entsprechenden Komponenten konnten jedoch nicht mehr integriert werden, was sich als Nachteil speziell bei der graphischen Suche erwiesen hat.

1.1.2.7 Ergebnis der Analyse
Ergebnis der Analyse ist ein voll instantiiertes USDF-Dokument, das zum eigentlichen Text die extrahierten Informationen (Sprache, Topic, Indexterme, Informationsobjekte) enthält. Sie werden in zwei Repositories abgelegt:

- In einem **IR-System**, in dem über die Texte gesucht werden kann.
- In einer **strukturierten Datenbank**, in der einzelne Elemente verknüpft, visualisiert und gesucht werden können.

[5] Es handelt sich um eine quasilogische Repräsentation namens SSRL, die aus syntaktischen Analysebäumen deriviert wird, vgl. Ritzke 2000, 2001.

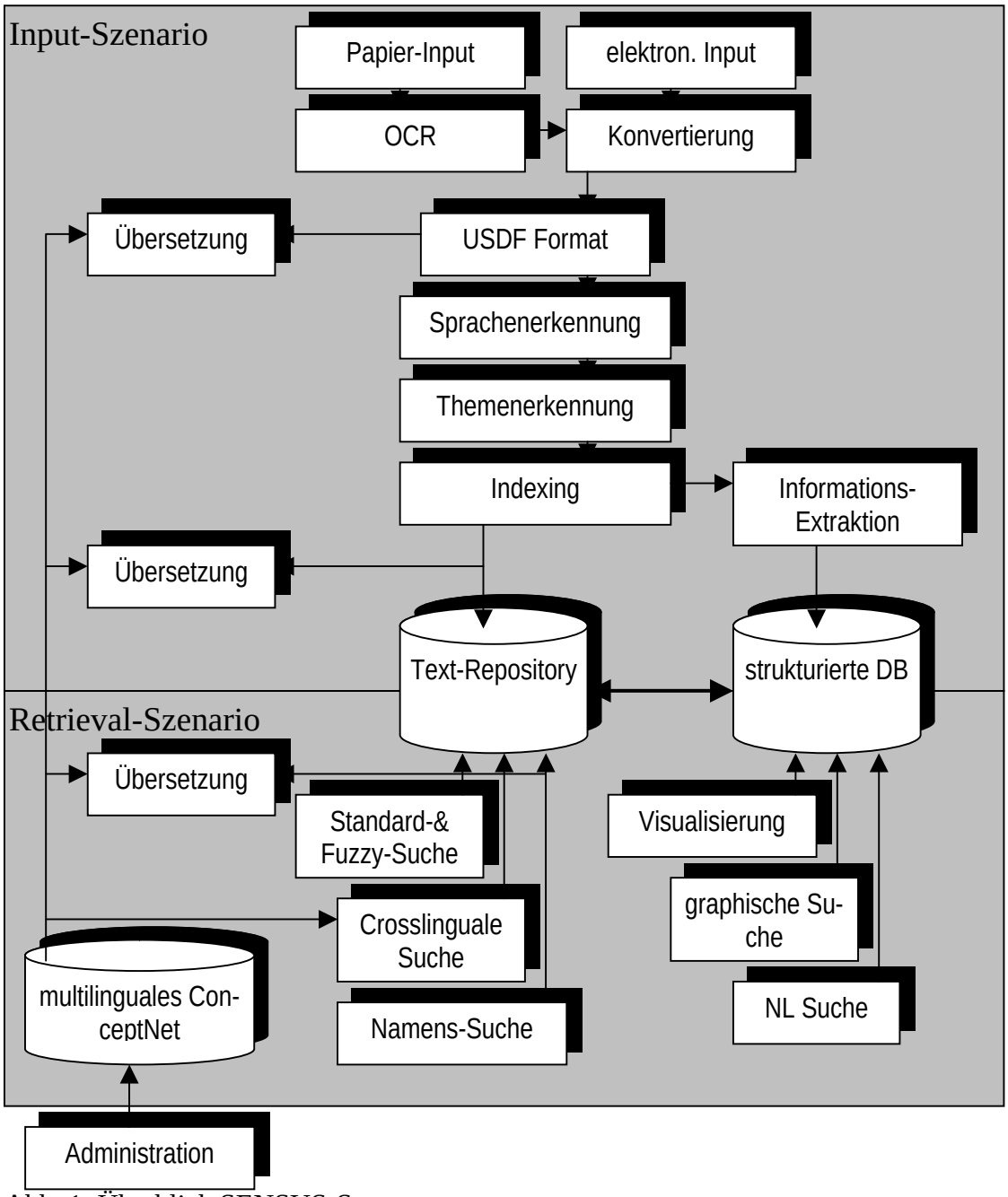

Abb. 1: Überblick SENSUS-System

Die beiden Repositories sind miteinander verbunden: Zu jedem Informationsobjekt ist angegeben, in welchem Dokument es gefunden wurde.

1.2 Retrieval-Szenario

Dieses Szenario soll es gestatten, die Ergebnisse der Textaufbereitung zu *visualisieren* und darin zu *suchen*.

Die folgenden Suchmöglichkeiten werden im System angeboten:

Suche in **strukturierten** Daten:
- graphische Anzeige der Informationsobjekte aus einem Dokument
- graphische Suche nach Objekten und ihren Verknüpfungen
- natürlichsprachliche Suche

Suche in **Text**-Daten:
- Standard-Suche (Boolesche Suche) mit Fuzzy-Search-Optionen
- erweiterte Textsuche mit Übersetzung und Expansion der Suchfrage
- Namens-Suche nach ähnlichen Namen

Die Benutzer wählen zunächst aus, in welchem Bestand sie suchen wollen, und können danach die spezielle Such-Art selektieren.

1.2.1 Visualisierung des Dokument-Inhalts

Diese Option gestattet die Anzeige aller Informationsobjekte und Relationen zu einem selektierten Dokument. Die Objekte werden über Ikons der verschiedenen Objekttypen repräsentiert, die mit Text annotiert sind. Sie werden über ebenfalls textlich annotierte Links verbunden, etwa *<person>_fährt_<auto>*, *<person>_trägt_<hut>*. Der Fortschritt bei SENSUS besteht darin, solche Graphiken automatisch aus Textinhalten zu gewinnen. [6]

1.2.2 Graphische Suche

Zur geschilderten Visualisierung der Dokumentinhalte empfiehlt sich, im Sinne der Konsistenz der Modalitäten, auch eine graphische Suche. Dabei werden Teilnetze manuell vorgegeben (etwa *<person>_besitzt_<auto>*), und im System wird eine solche Relation mit den beteiligten Objekten gesucht[7].

Die Suche erzeugt alle Objekte, die den gesuchten Graphen enthalten. Durch Klicken auf eines dieser Objekte kann dieses als Suchfokus genommen und erneut expandiert werden usw.: So ist eine graphische Navigation in den Daten möglich. Gleichzeitig gibt es immer eine Referenz auf die Texte, aus denen Objekte und Links extrahiert wurden.

[6] Existierende Tools visualisieren solche Informationen, sind aber nicht in der Lage, sie aus Dokumenten zu extrahieren. wie *Analyst's Notebook* von i2, vgl. i2 Ltd. (2004). I2 Homepage, http://www.i2.co.uk [Zugriff September 2004].

[7] Diese Option ist in der gewählten Anwendung von großer Bedeutung (etwa: Welche Beziehung besteht zwischen zwei Personen?).

1.2.3 Natürlichsprachliche Suche

Das System bietet auch die Möglichkeit der natürlichsprachlichen Suche in strukturierten Daten. Dazu wurde eine Text-to-SQL-Komponente erstellt. Sie besteht aus einer morphosyntaktischen Analyse, einer Abbildung in eine quasi-logische Repräsentation[8], und deren Konvertierung nach SQL. Sie wurde in zwei Sprachen (Spanisch und Deutsch) realisiert.

Die Benutzer konnten die Informationsobjekte abfragen, Links auf deren Dokumente anzeigen und diese ggf. in ihre Muttersprache übersetzen.

Natürlichsprachliche Interaktionsformen sind nützlich, wenn sie auf ein pragmatisches Maß beschränkt werden wie im vorliegenden Fall; dort sind sowohl die Domäne als auch die suchbaren Daten relativ restringiert.

Ein Seitenzweig der Entwicklung war die Generierung von natürlichsprachlichen Antworten[9], um auf diese Weise, ebenfalls zur Wahrung der Konsistenz der Modalität, eine rein natürlichsprachliche Interaktion zu ermöglichen, die ggf. mit Speech-Frontends versehen werden kann (etwa zwecks Anfragen via Telefon aus dem Polizeifahrzeug). Diese Komponente ist jedoch nicht ins System integriert worden.

1.2.4 Standard-Textsuche

Neben der Suche in strukturierten Daten bietet das System auch Suche in Text-Repositories. Die Suche konnte auf Sprachen und nach Topic (z.B. Suche nur in den *Drogen*-Texte auf *Spanisch*) eingegrenzt werden, um die Filter zu nutzen, die bei der Analyse des Input-Szenario instantiiert worden waren. Die Suche selbst ist eine Standard-Volltextsuche, basierend auf einem Booleschen Retrievalmodell.

Ein spezielles Merkmal dieses Systems ist die Unterstützung von OCR-Dokumenten:

- Einerseits können mit mehrstufiger Fuzzy-Suche Texte zu fehlerhaft erkannten Suchbegriffen dennoch gefunden werden
- Außerdem werden die Suchbegriffe direkt in den tif-Dokumenten hervorgehoben, sodass das (zuweilen wunderliche) Ergebnis der OCR-Analyse für die Benutzer transparent bleibt.

1.2.5 Erweiterte cross-linguale Suche

Neben der Standard-Volltextsuche wurde auch eine Suche implementiert, die Suchfragen-Expansion und Übersetzung anbietet.

---

[8] Dafür wurde ebenfalls, wie bei der Extraktion, die SSRL-Notation verwendet
[9] Vgl. Ritzke 2000.

Zur Suchfragen-Expansion wurde auf eine linguistische Ressource zurückgegriffen, die als **multilingual ConceptNet** die Terminologie der Anwendung modelliert. Mithilfe von Term-Extraktionsverfahren (Thurmair 2003) wurden die wesentlichen Text-Begriffe identifiziert, dann in eine begriffliche Hierarchie gebracht (neuerdings *Ontologie* genannt, cf. Fensel 2001), und mithilfe thesaurus-orientierter Links (v. a. Synonym, Ober- / Unterbegriff) verknüpft. Für die einzelnen Knoten der Hierarchie wurden Übersetzungen gesucht. Dieses multilinguale ConceptNet wurde zur Suchzeit abgefragt.

Zum Zeitpunkt der Suche können die Benutzer Zielsprache(n) und Fachgebiete bestimmen (letzteres dient zur Disambiguierung im Fall von mehrfachen Übersetzungen), und dann ihre Suche in natürlicher Sprache (der Anfrage-Sprache) eingeben.

Das System ermittelt die relevanten Begriffe und **übersetzt** und **expandiert** sie auf der Seite der Dokumentsprachen gemäß dem Inhalt des ConceptNet[10]. Die gefundenen Begriffe werden dem Benutzer zur Verifikation angeboten. Dabei wurden folgende Beobachtungen gemacht:

- Die Suchqualität ist wesentlich bestimmt von der Behandlung der Mehrwortbegriffe. Einerseits geben deren Einzelterme oft völlig unbrauchbare Übersetzungen (*Nuclear* (-> *Kern*-?) *Power* (->*Macht*?) *Plant* (->*Pflanze*?)) und ebenso unbrauchbare zielsprachige Dokumente, wenn die korrekte Übersetzung *Kernkraftwerk* nicht erkannt wird. Andererseits sind Mehrwortbegriffe in der Dokumentensprache mehr als nur „und"-verknüpft; sie folgen einer linguistischen Struktur.
- Für die Qualität der Suche ist auch der Bezug auf das Textkorpus wesentlich: Wenn *drug* nach *Drogen* übersetzt wird, in den Texten jedoch v. a. von *Betäubungsmitteln* die Rede ist, wird die Suche beeinträchtigt.
- Ein Problem stellt die Behandlung nicht übersetzbarer Begriffe dar: Im Fall von Eigennamen (*Clinton*), sollen sie in die Suche eingeschlossen werden, im Fall von sonstigen Suchbegriffen (*Verbrennungsmotor*) würden sie die fremdsprachliche Suche massiv stören. Vor allem deshalb wurde im System ein Benutzer-Feedback vorgesehen (ansonsten wenig hilfreich, wenn der Benutzer die Fremdsprache nicht beherrscht).

Bei der **Suche** werden deutsche Fragen auf deutsche, englische Fragen auf englische Dokumente usw. abgebildet. Das System behandelt somit die ein-

[10] Die Expansion mittels fester Hierarchien wird als etwas weniger effektiv angesehen als die Expansion mithilfe der Trefferdokumente; das hängt jedoch stark davon ab, wie "textnah" diese Hierarchien erstellt worden sind.

zelsprachlichen Anfragen – und auch das Ranking der Ergebnisse! – als voneinander unabhängig.

Das Problem bei Ranking-Verfahren ist, dass sie zumeist auf Ähnlichkeiten von begrifflichen Kontexten rekurrieren, und damit sprachabhängig sind. Ansätze zum Ranking multilingualer Treffer bestanden darin, die Suche mehrfach abzusetzen und dabei die Suchfragenlogik stufenweise abzuschwächen. Dadurch entstehen sprachunabhängige Trefferklassen, die ein gewisses Ranking gestatten[11].

Die gefundenen Dokumente werden angezeigt und können mithilfe der verfügbaren Übersetzungstechnologie interaktiv übersetzt werden.
Auf diese Weise können z.B. deutsche Anfragen mit deutschen Ergebnissen beantwortet werden, auch wenn der Dokumentenbestand multilingual ist.

1.2.6 Namenssuche

Eine der Anforderungen war es, die Suche nach Personennamen zu unterstützen, indem verschiedene Schreibungen für einen Namen gefunden werden sollten (*Meier Mair Mayer Maier*). Existierende Verfahren (Soundex, Kölner Phonetik) sind nicht verwendbar, da sprachabhängig.

Im vorliegenden System wurde deswegen, basierend auf dem Vorschlag von Jörg 1999, ein Ansatz entwickelt, der regelbasiert eine kombinierte graphemische und phonetische Normalisierung von Namen vorsieht, und aus dieser Normalform die verschiedenen möglichen Varianten generiert. (*Henderson -> [hEnd6zn] -> Hendersson Hennderson Hendersohn* usw.) (Barrio-Alvers 2000). Dem Benutzer wurden die so expandierten Namen zur Auswahl angeboten; die gewählten Namen wurden mit "oder"-Verknüpfung zur Suche gegeben. Dieses Verfahren wurde für elf Sprachen implementiert. [12]

1.2.7 Ergebnis

Das Ergebnis der Implementierung lässt sich wie folgt zusammenfassen:

1. Im Bereich der **strukturierten Daten** ist die Qualität der Suche sehr abhängig von der Qualität der Extraktionskomponenten; da diese nicht

[11] Dieser Ansatz wurde aber nicht mehr implementiert. Eine andere Option, die Dokumente zurückzuübersetzen und dann nur in der Zielsprache das Ranking durchzuführen, hängt sehr von der Übersetzungsqualität und der Terminologie des Übersetzungssystems ab.

[12] Ein solcher Ansatz produziert relativ viele Varianten, speziell bei längeren Namen. Recherchen mit den jeweils 20 besten Hypothesen ergaben allerdings, dass die meisten der Varianten (d.h. Personen mit Namen in dieser Schreibweise) tatsächlich existieren; sodass man über einen Tradeoff zwischen vernünftiger Größe der Suchanfrage und Vollständigkeit der Treffermenge nachdenken muss. In späteren Projekten wurde die Normalform nicht expandiert, sondern gespeichert und auf Basis der Normalformen verglichen.

voll entwickelt worden ist, zeigten die Tests Defizite in den Suchmöglichkeiten.
Multimodale Aspekte (graphische Suche, natürlichsprachliche Suche) werden vom System unterstützt, mit guter Resonanz für die Graphik (Konsistenz der Modalität); auf Seiten der natürlichen Sprache würde eine Antwort-Generierungskomponente zusätzliche Möglichkeiten bieten.
Der Link der strukturierten Objekte auf die Texte gestattet den Benutzern, die Ergebnisse der Extraktion immer zu überprüfen und ggf. zu korrigieren[13].

2. Im Bereich der **Textdaten** kann die linguistische Verarbeitung optimiert werden durch eine Komponente, die die Ergebnisse des OCR-Imports zumindest partiell korrigiert.
   Die Text-Suchmöglichkeiten sind ausreichend und ermöglichen gute Ergebnisse. Bei der crosslingualen Suche hängt die Qualität entscheidend von der Qualität des ConceptNet ab; wenn diese Ressource inadäquat und ohne Bezug auf die realen Texte aufgebaut ist, lassen die Ergebnisse sofort deutlich nach.
3. Im Hinblick auf die Integrationsaspekte sind viele Technologien erfolgreich eingebunden worden, aber es bleiben offene Punkte:
   Im Bereich der Multimodalität sind v. a. Aspekte der Transparenz gegenüber der Datenhaltung (strukturierte vs. Text-Daten) offen geblieben. Das macht sich an zwei Stellen bemerkbar:
   - Die Benutzer konnten **natürlichsprachliche Suchanfragen** eingeben, müssen sie aber entweder an die strukturierten Daten oder an die Textdaten richten; beide Komponenten agieren unabhängig voneinander, verwenden verschiedene Technologien, suchen in verschiedenen Repositories usw.; dieser Umstand ist eigentlich nicht intuitiv.
   - Als Folge davon ergeben sich Friktionen etwa bei **gemischten Suchanfragen** wie: *Fischer*$_{[Person]}$ *zu Weinexporten aus Verona*$_{[Ort]}$. Während die strukturierte Suche die Person und den Ort als Objekte finden, aber nichts über *Weinexport* wissen wird, wird die Textsuche auch die *Fischer* als Beruf und *Verona* als Person als Ergebnis bringen.

Eine engere Verknüpfung von (strukturierten) Informationsobjekten und Texten würde in solchen Fällen die Suchergebnisse noch verbessern.

[13] Dieser Umstand ist in der gewählten Applikation erheblich, weil man personenbezogene Daten nicht irrtümlich falsch weiterverarbeiten will.

Im Bereich der **Multilingualität** ist die Integrationsanstrengung ziemlich erfolgreich gewesen: Benutzer konnten eingehende fremdsprachliche Dokumente sofort übersetzen und ihre Relevanz beurteilen; die Analysen sind sprachspezifisch aufgebaut, um Quereinflüsse zu vermeiden; die crosslinguale Suche ist mit einem korpusbezogenen multilingualen ConceptNet zur Suchfragenübersetzung optimiert worden, und die Ergebnisse der Suche können sofort in die Anfragesprache rückübersetzt werden.

Die Eigenschaft, zu jeder Zeit ein fremdsprachliches Dokument in die eigene Sprache übersetzen zu können, ist einer der Haupt-Fortschritte des Systems.

2 Übersetzungstechnologien

2.1 Ansatz

Das gesamte System ist so konzipiert, dass es im Kern multilingual ist; das bedeutet, dass im Prinzip auf jeder Stufe der Verarbeitung eine Übersetzung möglich sein soll. Die Übersetzung soll dabei wie ein Displaymodus benutzt werden können, d.h. so wie die Benutzer im Editor den Font, sollen sie auch die Anzeigesprache wechseln können. Das bedeutet auf der Systemseite, dass die Übersetzungstechnologie ständig präsent sein muss.

Hauptzweck der Übersetzung ist dabei das Verstehen des Dokument-Inhalts, nicht so sehr ein perfektes zielsprachliches Dokument. Für diesen restringierten Zweck eignen sich maschinelle Verfahren gut.

2.2 Technologien

Im System wurden drei Technologien implementiert:

1. Als **Translation Memory** Technologie wurde TrAid (Piperidis et al., 1998) eingebunden; dabei werden einzelne Segmente (meist Sätze) mit ihren Übersetzungen gespeichert und zur Laufzeit nachgeschlagen, ggf. mit einem Faktor von Fuzziness. Diese Technik kann erfolgreich verwendet werden, wenn es sich um repetitive Texte handelt; das war bei der gegebenen Applikation (News, Polizeiberichte) aber eigentlich weniger der Fall.
2. In den Sprachrichtungen, in denen sie verfügbar war, wurde **maschinelle Übersetzung** angeboten; ausgewählt wurde das T1-System mit Deutsch, Englisch und einer speziellen Adaption auf Spanisch-Englisch. Zu diesem Zweck wurde die Adaptierbarkeit des MT-Systems an diese Domäne untersucht, und die lexikalischen Ressourcen entsprechend aktualisiert (Car-

denas 2000). Es zeigte sich, dass bereits über die Adaption des Lexikons erhebliche Verbesserungen der Übersetzungsqualität erzielbar sind[14].

3. In den Sprachrichtungen, in denen keine maschinelle Übersetzung verfügbar war, wurde eine Technologie verwendet, die als **Term Substitution** im Wesentlichen einzelne Termini übersetzt und die Übersetzung in den Text einschießt. Auf diese Weise erhält der Benutzer immerhin einen Eindruck, ob der Text für sein Interessenprofil relevant ist. Die notwendigen terminologischen Bestände wurden erstellt und zur Laufzeit nachgeschlagen.

Die Übersetzungstechnik der Term Substitution wurde von den Benutzern gegenüber den anderen bevorzugt; wohl deshalb, weil die zentralen Begriffe eines Textes präsentiert werden, ohne dass er ganz gelesen werden muss.[15]

2.3 Aspekte der Multilingualität

1. Der Aufbau multilingualer **Ressourcen** spielt eine Schlüsselrolle in einem multilingualen Informationssystem. Verfügbare Fachglossare repräsentieren die in den realen Texten auftretende Terminologie nur sehr partiell[16]. Dieser Umstand würde eine cross-linguale Suche erheblich behindern (weil viele Suchbegriffe keine Übersetzung hätten); er zeigt, dass in multilingualen Informationssystemen nur dann gute Ergebnisse erzielt werden können, wenn die verwendeten Termini aus den realen Textkorpora stammen.
2. Ein zweiter Aspekt liegt in der **Konsistenz** der Ressourcen. Das System darf nicht in den verschiedenen Komponenten verschiedene Übersetzungen anbieten und etwa für *Drogen* bei der Suche *drugs* und bei der Übersetzung *narcotics* wählen: In einem inkonsistenten Datenbestand verlieren die Benutzer das Zutrauen zu den implementierten Lösungen. Dieser Umstand erfordert es, die Ressourcen zwischen den verschiedenen Komponenten auszutauschen. Als **Austauschformat** wurde auf OLIF zurückgegriffen[17].
3. Ein dritter Aspekt liegt in der **Pflege** der Ressourcen. Für ein multilinguales und multifunktionales System ist es zentral, die Ressourcen an nur einer Stelle zu pflegen und über Compiler den Einzelkomponenten (z.B.

[14] Dies bestätigt auch Weber 2003.

[15] Es stellte sich allerdings heraus, dass die Technologie dazu verleitet, Texte als relevant zu betrachten, die zwar nur wenige relevante Termini enthalten, diese aber die einzig verständlichen, da übersetzten, sind; insofern sind die relevanten und die übersetzten Termini getrennt auszuweisen.

[16] Zentrale Textbegriffe fehlten, und nur 20 % der Glossarbegriffe fanden sich im Text.

[17] Open Lexicon Interchange Format, vgl. OLIF Consortium (2003). Open Lexicon Interchange Format Homepage. http://www.olif.net [Zugriff September 2004].

MT-Lexika) zuzuteilen. Es gab aber kein Tool, das multilinguale Ontologien mit Begriffshierarchien, mehrsprachigen Termini an deren Knoten, und linguistischen Annotationen zu den einzelnen Termini zu pflegen gestattet[18]; deshalb wurde eine eigene Komponente entwickelt (Jackson et al., 2002). Im Endausbau wurde damit eine Begriffshierarchie von mehr als 16000 Begriffen codiert, in elf Sprachen und mit insgesamt über 200000 Termini.

3 Evaluierung

3.1 Evaluierung durch die Benutzer

Das System wurde an vier Standorten in vier Ländern von den End-Benutzern evaluiert, v. a. bezüglich seiner Benutzbarkeit (Lewandowski 2000). Die Haupteindrücke waren:

- Erstmals eine breite Palette von Sprachtechnologie zur Unterstützung der Benutzer in diesem Feld. Jedoch:
- störende Limitierungen in den einzelnen Komponenten; z.B.: zuwenig Textformate werden unterstützt; die Qualität der Informationsextraktion ist in den verschiedenen Sprachen unterschiedlich.
- unklare oder fehlende Integration, z.B.: Unterschiedliche Ergebnisse bei der Standard- und der erweiterten Textsuche; OCR-Texte nicht übersetzbar wegen der hohen Fehlerrate.
- Probleme im Handling und im Interface.

Die meisten dieser Punkte verdanken sich dem Charakter des Systems als Prototyp und ließen sich in ausgereifteren Folgeversionen beheben. Drei Punkte sind allerdings aufgefallen:

1. Die Benutzer bevorzugen eine Term Substitution vor einer maschinellen Vollübersetzung, wegen ihres „abstract-bildenden“ Effekts.
2. Die Benutzer hatten Schwierigkeiten mit dem Konzept einer natürlichsprachlichen Suche in strukturierten Datenbeständen; mangels Erfahrung war unklar, wie man dort erfolgreiche Fragen stellen muss[19]; für diese Daten wurde eine graphische Suche wesentlich eher akzeptiert.
3. Suche in der Muttersprache in multilingualen Beständen und Anzeige der Ergebnisse in der Muttersprache ist ein überzeugendes Konzept.

[18] Vgl. Zwickl 2000; so auch Gómez-Perez 2003.
[19] Es wurde dann eine Serie von "typischen" Beispielfragen als Hilfe mit angeboten

Insgesamt war die Reaktion der Benutzer gemischt; einige Systemaspekte wurden sehr positiv aufgenommen (Topic Identifier, Namenssuche, Fuzzy Suche, Übersetzung).

3.2 Vergleich mit anderen Ansätzen (CLEF)

Im Vergleich zu Ansätzen crosslingualen Retrievals wie z.B. CLEF[20] sind als wesentliche Punkte festzuhalten:

- Eine systematische Auswertung der Qualität der Suche wie in CLEF ist im Projekt nicht unternommen worden. Speziell das Problem der Suchfragenübersetzung und Validierung wurde nicht behandelt, es stellt sich auch in dieser konkreten Applikation so nicht.
- Im Unterschied zu CLEF ist das System nicht nur in der Domäne eingeschränkt (folgt also eher dem GIRT-Paradigma), sondern betont die Wichtigkeit der korpus- und applikationsbezogenen Ressourcen. Es folgt damit Erfahrungen aus der maschinellen Übersetzungen, wonach der Fokus auf eine spezifische Anwendung die Systemqualität verbessert.

Andererseits enthält das System Aspekte, die bei CLEF nicht thematisiert werden, etwa die Suche in strukturierten Beständen, oder die Suche mit Informationselementen.

3.3 Evaluierung der Konzeption

Das Ziel des Projektes war es, die Architektur eines Informationssystems zu finden, die alle sprachtechnologisch verfügbaren Mittel aufbietet, um schnelles Verstehen und Verarbeiten von Informationen zu ermöglichen und sowohl Multilingualität als auch Multimodalität zu unterstützen.

Dieses Ziel ist erreicht worden; es gibt nur wenige Beispiele von Systemen, die eine solche Vielzahl von linguistischen Komponenten in einem Informationssystem integrieren konnten. Die gewählte Whiteboard-Architektur der Text-Analyse hat sich als sehr flexibles Mittel der Integration bewährt; auf Seiten der Suche sind mächtige Möglichkeiten geschaffen worden; und auch die ständige Verfügbarkeit von Übersetzungshilfen trägt zum Gesamterfolg bei.

Das Ergebnis dieser Architektur ist, abgesehen von den immer möglichen und notwendigen Verbesserungen in Einzelkomponenten und Oberflächen, eine

[20] Cf. Kluck & Womser-Hacker 2000, Braschler et al., 2000, Peters & Braschler 2002.

vertiefte Sicht auf die Anforderungen, die sich aus der geschilderten Systemintegration wiederum ergeben; davon sei noch kurz gesprochen.

1. Im Bereich der **Eingabe**verarbeitung müsste eine Verbesserung vor allem bei der Qualität der Front-End-Verfahren ansetzen und geeignete Korrektursysteme entwickeln. Andere Formen der Eingabe (gesprochener Input, Audio-Mining, Multimedia-Input) können die Mächtigkeit des Systems erhöhen; sie erfordern jedoch wieder spezielle integrative Lösungen.
2. Im Bereich der Dokument-**Analyse** muss die Kapazität der Extraktion und flexiblen Weiterverarbeitung gesteigert werden. Dazu zählen etwa:
   - Möglichkeit des Dokument-Filterns nach spezifischen Benutzerprofilen, die auf Topic und Information Extraction beruhen
   - Fragen der Integration der Ergebnisse der Extraktion in generische Strukturen wie z.B. Topic Maps (ISO 13250) oder Semantic Web, sodass die Analyseergebnisse mit bereits existentem Wissen verknüpft werden können (etwa in Data Mining Applikationen)
   - Fragen der Informationskomprimierung, einerseits durch automatisches Abstracting (basierend auf der Generierung eines Texts aus den extrahierten Informationen (Ritzke 2000), andererseits durch Verfahren der Visualisierung von Dokumentinhalten (Widdows et al. 2002)
3. Im Bereich der **Suche** stellt sich die Anforderung einer besseren Integration der strukturierten und der Text-Suche, v. a. die Möglichkeit der semantischen Attribuierung von Suchbegriffen[21] ("*Fischer*" als <*Person*> usw.); im Gegenzug wird allerdings der Suchdialog deutlich komplizierter, was zusätzliche Anforderungen an die Benutzerschnittstelle stellt.
   Konsistenz und Transparenz der Modalität ist ebenfalls zu verbessern; Konsistenz insofern, als eine natürlichsprachliche Anfrage ein ebensolches Ergebnis liefern soll, (z.B. für Speech-Frontends), was eine Komponente zur Textgenerierung erfordert; Transparenz insofern, als eine Anfrage sowohl im Textteil als auch im strukturierten Teil suchen soll, was ein Problem der Fusion der jeweiligen Treffer schafft.
4. In der Frage der **Ressourcen** sind entsprechende Extraktions- und Pflegekomponenten zu entwickeln, weil eine gute Qualität bei crosslingualer Suche erfordert, dass die Ressourcen korpusbasiert und anwendungsspezifisch aufgebaut werden können.
5. Schließlich gebiert die Integration des Systems als ganzes in die jeweiligen **Applikationen** neue Anforderungen, etwa in Hinblick auf die verwendeten Backend-Systeme (existierende Datenbanken mit einer erheblichen Zahl von Einträgen), oder im Hinblick auf Mobilität (Abfrage des Systems, per Sprache, aus dem Polizeifahrzeug). Solche Aspekte haben übli-

[21] Der kritische Faktor hier ist die Qualität der Dokumentanalyse, die den Anforderungen der Suche entsprechen muss.

cherweise erheblichen Einfluss auf die Architektur des Informationssystems selbst.

6. Ein Desiderat geblieben ist eine sorgfältige **Evaluierung** des Systemverhaltens. Der im Projekt gewonnene Eindruck, dass sich die Qualität der Interaktion und der Information durch den Einsatz der geschilderten Verfahren verbessern lässt, ist nicht systematisch validiert worden; dafür fehlten im Projekt die Ressourcen. Dies bleibt neuen Kontexten vorbehalten.

4 Literaturverzeichnis

Barrio-Alvers, L. (2000). Similarities Detector. Sensus Report.

Bodenkamp, St. (2000). SENSUS Final Report, Sensus Report.

Boitet, Chr., Seligman, M. (1994). The "Whiteboard" Architecture: A Way to Integrate Heterogeneous Components of NLP Systems. Proc. COLING 1994.

Braschler, M., Peters, C., Schäuble, P. (2000). Cross-Language Information Retrieval (CLIR) Track, Overview. Proc. TREC-8, 2000.

Buitelaar, P., Declerck, Th., Sacaleanu, B., et al. (2003). A Multi-Layered, XML-Based Approach to the Integration of Linguistic and Semantic Annotations; Proc. EACL 2003.

Cardenas, D. (2000). Untersuchung zur lexikalischen Vertikalisierung eines automatischen Übersetzungssystems am Beispiel Polizeiberichte. Dipl.A. FH München.

Cowie, J., Ludovik, E., Zacharski, R. (1998). "An Autonomous, Web-based, Multilingual Corpus Collection Tool." In: Proceedings of the Natural Language Processing and Industrial Applications, Moncton, Canada, 1998.

Fensel, D. (2001). Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Berlin: Springer.

Goller, Chr., Löning, J., Will, Th., et al. (2000). Automatic Document Classification: A thorough Evaluation of various Methods. In: Proc. ISI 2000.

Gómez Pérez, A. (ed.) (2002). A Survey on Ontology Tools. Ontoweb Report (Deliverable 1.3, IST-2000-29243).

Gómez-Perez, A., Manzano-Macho, D. (2003). A survey of ontology learning methods and techniques. Ontoweb Report 2003.

Jackson, A., Lewandowski, M., Thurmair, Gr., et al. (2002). ConceptManager, Pflege multilingualer Ontologien im crosslingualen Retrieval. In: Proc. ISI, Regensburg.

Jackson, P., Moulinier, I. (2002). Natural Language Processing for Online Applications, Text Retrieval, Extraction and Categorization. Amsterdam: J. Benjamins.

Jörg, M. (1999). Doppelgänger gesucht, Ein Programm für kontextsensitive phonetische Textumwandlung, c't, Heft 25.

Kluck, M., Womser-Hacker, Chr. (2002). Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. In: Proc. LREC 2002, Gran Canaria.

Krause, J. (1993). "A Multilayered Empirical Approach to Multimodality: Towards Mixed Solutions of Natural Language and Graphical Interfaces." In: Maybury, M., (ed.) (1993). Intelligent Multimedia Interfaces. MIT Press.

Lewandowski, M. (2000). Sensus User Test. Sensus Report.

Neumann, G. (2001). Whiteboard, Project Sheet. (Slides) DFKI 2001.

Peters, C., Braschler, M. (2002). “The Importance of Evaluation for Cross-Language System Development: the CLEF Experience.” In: Proc. LREC 2002, Gran Canaria.

Piperidis, S., Malavazos, C., & Triantafyllou, Y. (1998). “TrAID : a memory-based translation-aid framework.” In: Proc. Natural Language Processing and Industrial Applications Conference, 1998, Moncton, Canada.

Rapp, R. (1997). "Text-Detektor, Fehlertolerantes Retrieval ganz einfach" In: c’t, Heft 4.

Ritzke, J. (2000). SEN-DNL-Gen: Dynamic Natural Language generation within the SENSUS System Environment. Sensus Report.

Ritzke, J. (2001). Information Extraction: tree2ssrl, Syntactic Semantic tree to SSRL. Sensus Report.

Roppel, St. (1998). Visualisierung und Adaption. Konstanz: UVK.

Thurmair, Gr. (1997). “Language Technology for Cross-Language Text Retrieval.” In: Proc. HCI München.

Thurmair, Gr. (1997). “Exchange Interfaces for Translation Tools.” In: Proc. MT Summit 6, San Diego.

Thurmair, Gr. (2003). “Making Term Extraction Tools Usable.” In: Proc. EAMT Dublin.

Weber, M., Hechtbauer, A. (1996). Konkrete Anwendung des Vector Space Modells zum Textretrieval in WING. WING-IIR Arbeitsbericht 72.

Weber, N. (2003). "MÜ-Lexikografie." In: Proc. GLDV, Anhalt.

Widdows, D., Cederberg, S., Dorow, B. (2002). “Visualisation Techniques for Analysing Meaning.” In: Proc. 5th Int. Conf. on Text, Speech and Dialogue, Brno 2002.

Zwickl, J. (2000). Terminologiearbeit zum Thema internationale Kriminalität, Erstellung einer Ontologie für das europäische SENSUS-Projekt. Dipl.A. Univ. Saarbrücken, 2000.