# Ein Email-Korpus zur Entwicklung und Evaluierung der Analysekomponente eines Terminvereinbarungssystems

Thierry Declerck, Judith Klein
{declerck;klein}@dfki.de

**Abstract:** This paper describes the development of a diagnostic tool, implemented as an annotated Email corpus database, for the evaluation of the analysis component of a German dialogue system, which has been realized as a NL front end to autonomous agent systems for appointment scheduling that use Email as their communication means. A shallow parsing approach has been adopted as the NL server's module. Methodology and measures applied for the evaluation must be understood in the context of the role played by the NL component within the overall system.

## 1 Hintergrund

Diese Arbeit beschreibt den Einsatz eines annotierten Korpus deutscher Emails zur Unterstützung einer empirisch fundierten linguistischen Verarbeitung der Analysekomponente eines Terminvereinbarungssystems, das als Testanwendung zum wiederverwendbaren Server für Dialogsysteme im Projekt Cosma (Cooperative Schedule Management Agent)[1] entwickelt wurde. Um die Subsprache der Terminplanungsdomäne zu spezifizieren und den Abdeckungsgrad der Analysekomponente feststellen zu können, wurden typische sprachliche Konstruktionen des Email-Korpus identifiziert und mit syntaktischen und semantischen Informationseinheiten annotiert. Email-Texte und Annotationen wurden in einem relationalen Datenbankschema klassifiziert, so daß eine linguistische Auswertung der Sprachdaten leicht unter unterschiedlichen Fragestellungen durchgeführt werden konnte.
Für die Entwicklung der anwendungsspezifischen Analysekomponente wurde ein robuster Shallowparsing Ansatz gewählt, da es bei der natürlichsprachlichen Eingabe via Email nicht ungewöhnlich ist, — ganz im Gegenteil — unvollständige oder fehlerhafte Texte vorzufinden. Zudem stand die effiziente Verarbeitung der für die Terminvereinbarung relevanten Email-Fragmente — bei gleichzeitiger Vernachlässigung irrelevanter Textteile — im Vordergrund, um eine erfolgreiche und reibungslose Kommunikation mittels des Dialogsystems zu gewährleisten. Daher wurde für die natürlichsprachliche Analyse das System smes (Saarbrücker Message Extraction System)[2], ein Informationsextraktionssystem für das Deutsche, gewählt, das ein Kernsystem zum robusten Shallowparsing und den Einsatz benutzerdefinierter endlicher Automaten zur anwendungsorientierten, selektiven Syntaxverarbeitung kombiniert.
Die Weiterentwicklung von smes basiert auf den Ergebnissen diagnostischer Evaluierungszyklen, die auf der Basis der Email-Texte des Referenzkorpus durchgeführt wurden. Die Evaluierung der Analysekomponente ist im Zusammenhang mit der speziellen Anwendung zu verstehen: Die Leistungen des Shallowparsers werden im Kontext des Dialogsystems zu Terminvereinbarungen evaluiert.

[1] Das Projekt Cosma [Busemann et al.1997] ist vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMB+F) gefördert worden (FKZ ITW 9402).
[2] Das System wurde von Günter Neumann entwickelt [Neumann1997].

## 2 Auswahl und Annotierung des Korpus

Als Basis für das Korpus wurden 160 Email-Nachrichten aus einer vorliegenden Sammlung von mehreren Hundert Emails mit Terminvereinbarungen manuell ausgewählt. Obwohl keine ausgearbeiteten Selektionskriterien vorlagen, wurde bei der Auswahl versucht, eine möglichst große Vielfalt typischer linguistischer Strukturen, die in der Anwendungsdomäne vorkommen, anhand der selektierten Testdaten abzudecken.
Aufbauend auf den TSNLP-Richtlinien zur Annotierung von Sprachdaten[3] wurde ein Annotationsschema entworfen, das nicht nur speziell der systemspezifischen Verarbeitung in smes nutzt, sondern eine allgemeine linguistische Charakterisierung der Texte widerspiegelt.
Das Annotationsschema umfaßt neben syntaktisch-semantischen Informationen einzelner Konstituenten, d.h. der Verben und ihrer zugehörigen Komplemente und Adjunkte, auch strukturelle und funktionale Informationen zu den syntaktischen Konstruktionstypen. Zur Annotierung dieser Information wurden die in TSNLP erarbeiteten Kategorie- und Funktionslisten eingesetzt.[4] Da der Schwerpunkt von TSNLP auf der syntaktischen Beschreibung lag, während für COSMA auch semantische Informationen von großer Bedeutung sind, wurden die ursprünglichen Listen um Informationen erweitert, die semantische Merkmale der Terminvereinbarungsdomäne abbilden. Hierzu gehört die detaillierte Klassifikation der Präpositional- und Nominalphrasen, um die für die Termindomäne wichtigen Temporalausdrücke adäquat zu beschreiben, wie Abbildung 1 zeigt:

| Annotierung | Beispiel |
|---|---|
| **Präpositionalphrasen** *Wie wäre es ...* | |
| PP_temp | *in dieser Woche?* |
| PP_temp-date | *am 4.11?* |
| PP_temp-day | *am Montag?* |
| PP_temp-time | *um 10?* |
| **Nominalphrasen** *Ich komme ...* | |
| NP_temp | *diese Woche?* |
| NP_temp-date | am Montag, *den 4.11.* |
| NP_temp-day | *Montag*, 14.00. |
| NP_temp-time | am Montag, *14.00.* |

Abbildung 1: Annotierung temporaler PPs und NPs

Da das semantische Modul von COSMA Informationen zur Klassifizierung von Kommunikationsprimitiven benötigt, sind die Email-Nachrichten außerdem mit entsprechenden Informationen zu Negation, Modus und Modalität annotiert.
Um die für die Anwendungsdomäne erforderliche linguistische Abdeckung zu bestimmen, sind die annotierten Email-Texte in einem relationalen Datenbanksystem aufbereitet, so daß die vorkommenden linguistischen Konstruktionstypen gezielt herausgefiltert werden können. Zudem erlaubt der Einsatz des relationalen Datenbankschemas die einfache Modifikation und Ergänzung des Korpus um neue Email-Texte und weitere Annotationen. Zwei (indirekt) miteinander verbundene Datenbanksysteme wurden eingesetzt, um das Material zu organisieren:

[3] Siehe z.B. [Lehmann et al.1996].
[4] Die Listen sind in [Estival et al.1995] angegeben.

- **Email-DiTo**, eine einfache Volltextdatenbank, aus der die vollständigen Email-Texte abgerufen werden können.[5]

- **Email-DB**, eine elaborierte Faktendatenbank zum Speichern der annotierten Email-Fragmente, die das Abrufen spezieller linguistischer Konstruktionen erlaubt.[6]

Mit Hilfe von Datenbankabfragen wurde das Email-Korpus ausgewertet, so daß zum Beispiel alle Textfragmente zusammengestellt werden konnten, die temporale PPs enthalten. Aus der Gesamtanalyse ergab sich eine Prioritätenliste, an der sich die Entwicklung der Grammatikkomponente des Systems orientierte.

## 3 Ein Shallowparsing-Ansatz

Als Analyse-Modul wurde das Informationsextraktionssystem smes eingesetzt, das ein Kernsystem zum Shallowparsing und den Einsatz spezialisierter endlicher Automaten kombiniert. Die Trennung zwischen Kernsystem und einem benutzerdefinierten System von Automaten erlaubt eine modulare Beschreibung der Grammatik für sprachliche Subdomänen. Der Kern des Systems besteht aus:

- einem Tokenizer zur Erkennung der Eingabe, wobei eine Menge regulärer Ausdrücke zur Identifikation der Fragmentmuster (z.B. Wörter, Zeitausdrücke usw.) eingesetzt wird,

- einer Komponente zur schnellen lexikalischen und morphologischen Verarbeitung (basierend auf dem Morphix-Tool [Finkler and Neumann1988]),

- einem Shallowparsing-Modul, das auf einer Menge endlicher Automaten basiert,

- einer flexiblen Ausgabe-Komponente, die die extrahierten und analysierten Ergebnisse kombiniert und in unterschiedlichen Ausgabeformaten repräsentieren kann.

Auf der Basis der morphologischen Wortanalyse der erkannten sprachlichen Ausdrücke identifizieren und extrahieren endliche Automaten relevante Email-Fragmente mit ihren morphosyntaktischen und kategorialen Informationen. Die extrahierten Textteile werden von einem lexikalisch gesteuerten bidirektionalen Shallowparser verarbeitet: Ausgehend von dem Verb (und seinem assozierten Verbautomaten) werden zuvor erkannte und extrahierte NPs und PPs aufgesammelt, wobei Zeitausdrücke zum Zweck der weiteren Verarbeitung mit spezifischen temporalen Merkmalen versehen werden. Die in Cosma definierten Verbautomaten sammeln zusätzlich zu der verbalen Subkategorisierung Informationen über Negation, Modus und Modalität, da diese für die nachfolgende semantische Interpretation der Email-Teile gebraucht werden. Das (partielle) Ausgabe-Format für die temporale und modalitätsbezogene Information des Satzes *Ich kann am Montag nicht kommen* wird in den Abbildungen 2 und 3 gezeigt.

Die für Cosma notwendige Subsprache der Email-basierten Terminvereinbarungsdomäne erfordert von dem Parser sowohl die korrekte Behandlung vielfältiger sprachlicher und numerischer Zeitausdrücke als auch die Verarbeitung z.T. eigenwilliger syntaktischer Konstruktionstypen. Nicht selten werden Textfragmente extrahiert, in denen das Verb fehlt (elliptische Konstruktionen oder auch Schreibfaulheit, die beim Mailschreiben oft festgestellt werden kann). Der

[5] Es handelt sich um eine angepaßte Version des DiTo-Datenbanksystems [Nerbonne et al.1993].

[6] Es handelt sich um eine angepaßte Version des TSDB-Datenbanksystems [Oepen et al.1995].

$$\begin{bmatrix} \texttt{TYPE} & < \texttt{date\_expr}, \texttt{pp} > \\ \texttt{SEM} & \begin{bmatrix} \texttt{DATE-PREP} & \texttt{am} \\ \texttt{WEEKDAY} & \texttt{1} \end{bmatrix} \end{bmatrix}$$

Abbildung 2: smes-Ausgabe für die temporale PP *"am Montag."*

$$\begin{bmatrix} \texttt{TYPE} & \begin{bmatrix} \texttt{FIRST} & \begin{bmatrix} \texttt{MODAL} & \texttt{koenn} \\ \texttt{NEG} & \texttt{t} \\ \texttt{QUES} & \texttt{no} \end{bmatrix} \\ \texttt{REST} & < \texttt{komm} > \end{bmatrix} \end{bmatrix}$$

Abbildung 3: smes-Ausgabe von Negation, Modus und Modalität für *"Ich kann am Montag nicht kommen."*

eingesetzte Prototyp des smes-Kernsystems, der zunächst nicht in der Lage war, bestimmte Arten unvollständiger Eingaben zu verarbeiten oder verteilte temporale Informationen zu extrahieren, konnte aufgrund seiner modularen Architektur leicht an die anwendungsspezifischen Anforderungen von COSMA angepaßt werden: Aus verblosen Email-Fragmente, denen der für die bidirektionale Suche erforderliche verbale lexikalische Anker fehlt, oder aus speziellen Funktionsverbgefügen (z.B. *Zeit haben*) können jetzt die für COSMA relevanten Informationseinheiten extrahiert werden. Allein autretende Modalverben (z.B. *Ich kann aber am Montag nicht!*) werden im Unterschied zu Modalverben in Verbalgruppen als Ankerverben behandelt. Und Email-Fragmente mit multiplen temporalen Ausdrücken werden korrekt verarbeitet.

## 4 Evaluierung von smes in COSMA

Ziel der Evaluierung von smes war es, ein Profil der erreichten linguistischen Abdeckung zu erhalten und die Schwachstellen der syntaktischen Analyse aufzudecken. Da sich die Evaluierung von smes auf die Bewertung des Shallowparsers als einer separaten Komponente des Gesamtsystems COSMA bezieht, wurde ein diagnostischer Ansatz im Rahmen der zyklischen Evaluierung unter Anwendung der Glass-Box-Methode gewählt. Als Testmaterial wurde ein Email-Korpus – kombiniert aus bereits annotierten und neuen Email-Texten – eingesetzt, das als Grundlage für die Bewertung der Analysekomponente diente.
Da die Email-Texte der Korpusdatenbank nicht direkt zur Systementwicklung eingesetzt worden waren, sondern dazu dienten, die Subsprache der Terminvereinbarungsdialoge und damit die Anforderungen an die syntaktische Verarbeitung allgemein zu definieren, konnten Auszüge der Sammlung als Testmaterial herangezogen werden. Mit Hilfe eines Zufallsgenerators wurden 3 x 15 Email-Nachrichten als erstes Testset ausgewählt. Zusätzlich wurde diesselbe Anzahl neuer, nicht im Korpus annotierter Emails zufällig zusammengestellt, um sicherzustellen, daß die Evaluierung des Shallowparsers wirklich entwicklungsunabhängig verläuft.
In der Testphase wurden die Email-Texte einzeln in smes eingespielt und die Analyse-Ergebnisse anschließend manuell überprüft. Da die Evaluierung im Zusammenhang mit der Einbindung von smes in COSMA gesehen werden muß, also anwendungsspezifisch ist, und sich eher an Performanzvorgaben als an der rein linguistischen Kompetenz orientiert, waren folgende Kriterien

zur Qualität der Ergebnisse relevant:

- **Verbtemplate**: korrekte Anbindung eines Automaten an einen lexikalischen Verbeintrag und korrekte Verarbeitung der abhängigen syntaktischen Struktur
- **Modalität**: korrekte Verarbeitung von Modalverben und Ausgabe der besonderen Satzmodalität
- **Temporale Ausdrücke**: korrekte Erkennung und Zuordnung temporaler Informationseinheiten zu den entsprechenden syntaktischen Domänen

Je nach Erfüllung der Kriterien wurden folgende Qualitätsprädikate unterschieden[7]:

- **Gut**: korrekte Analyse oder minimale Fehler, die aber das wichtigste Kriterium, die temporale Information, nicht betreffen und nicht zu Fehlinterpretationen der semantischen Analyse führen.
- **Mittel**: mittelschwere Fehler, d.h. die temporale Information wird so ausgegeben, daß einen korrekte Weiterverarbeitung im Rahmen der semantischen Analyse möglich ist. Obwohl die syntaktische Analyse nicht vollständig oder nicht ganz korrekt ist, führt das Ergebnis nicht zu semantischen Fehlinterpretationen.
- **Schlecht**: gravierende Fehler und Lücken in der syntaktischen Analyse, so daß eine korrekte Interpretation des Email-Fragments nicht möglich ist.

Die Ergebnisse des ersten Durchlaufs der diagnostischen Evaluierung wurden als Feedback für die Weiterentwicklung der Automaten genutzt. Als Hauptpunkte, die zu einer schnellen Leistungssteigerung führen sollten, wurden (i) Verarbeitung der Prädikation (Kopula-Konstruktionen), (ii) Ergänzung und Modifikation bestimmter Verbtemplates und (iii) Erkennung bestimmter lexikalischer Terminausdrücke identifiziert. Nach der Verbesserungsphase wurde ein zweiter Testzyklus durchgeführt, an dessen Ergebnissen eine deutliche Leistungssteigerung für die anvisierten Problemfelder abgelesen werden konnte. Das Ergebnis beider Testläufe ist in Tabelle 4 zusammengefaßt:

| **Evaluierung** | **gut** | **mittel** | **schlecht** |
|---|---|---|---|
| **Erster Testlauf** | | | |
| Datenbank-Set | 51,1 % | 37,8 % | 11,1 % |
| Neue Emails | 28,9 % | 20,0 % | 51,1 % |
| **Zweiter Testlauf** | | | |
| Datenbank-Set | 73,2 % | 13,3 % | 11,1 % |
| Neue Emails | 62,1 % | 11,1 % | 26,6 % |

Abbildung 4: Ergebnisse der diagnostischen Evaluierung von smes

[7] Die Zuordnung der Extraktionsergebnisse soll künftig auf den speziell für die Informationsextraktion definierten Maßen *recall* und *precision* basieren.

## 5 Zusammenfassung und Ausblick

Aufgrund des annotierten Referenzkorpus konnte die relevante Subsprache spezifiziert und damit die erforderliche syntaktische Abdeckung von smes bestimmt werden. Die auf den Email-Texten basierenden Testläufe der diagnostischen Evaluierung gaben Auskunft über Abdeckungsgrad bzw. Lücken in smes und trugen so entscheidend zur Verbesserung dieser anwendungsspezifischen Analysekomponente bei.
Die hier vorgestellte Arbeit wird im Rahmen des Projekts PARADIME[8] fortgesetzt. Das Kernsystem von smes wird erweitert, auch im Hinblick auf die Verarbeitung neuer Textdomänen, wie z.B. Wirtschaftsmeldungen. Die Evaluierung der Leistungen von Grammatiken, die im Rahmen von smes definiert sind, wird dabei verstärkt allgemeine linguistische Phänomene betreffen (wie z.B. grammatische Funktionen). Auch das Informationsextraktionssystem als solches wird einer grundlegenden Evaluierung unterzogen werden. Quantitative Bewertungen, wie "recall" und "precision" (im Sinne von MUC[9]) werden hier im Vordergrund stehen. Ein erster Teil dieser Arbeit ist bereits in ([Neumann et al.1997]) dokumentiert.

## 6 Danksagung

Die hier vorgestellte Arbeit wurde im Rahmen des Projekts PARADIME vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMB+F) gefördert (FKZ ITW 9704). Bei Stephan Busemann möchten wir uns für seine anregenden Diskussionsbeiträge besonders bedanken.

[8] Das Projekt PARADIME wird vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMB+F) gefördert (Kontrakt FKZ ITW 9704.)
[9] Siehe [MUC1995]

## Literatur


[Busemann et al.1997] Stephan Busemann, Thierry Declerck, Abdel Kader Diagne, Luca Dini, Judith Klein, and Sven Schmeier. 1997. Natural language dialogue service for appointment scheduling agents. In *Proceedings of the 5th Conference on Applied Natural Language Processing, ANLP-97*, pages 25–32, Washington DC.

[Estival et al.1995] Dominique Estival, Sabine Lehmann, Kirstin Falkedal, Hervé Compagnion, Lorna Balkan, Frederik Fouvry, Doug Arnold, Eva Dauphin, Sylvie Régnier-Prost, Véronika Lux, Judith Klein, Judith Baur, Klaus Netter, and Stephan Oepen. 1995. The construction of test material. Technical report, ISSCO, University of Essex. Report to LRE 62-089.

[Finkler and Neumann1988] Wolfgang Finkler and Günter Neumann. 1988. Morphix: A fast realization of a classification–based approach to morphology. In *Proceedings der 4. Österreichischen Artificial–Intelligence Tagung, Wiener Workshop Wissensbasierte Sprachverarbeitung*, pages 11–19, Berlin. Springer.

[Lehmann et al.1996] Sabine Lehmann, Stephan Oepen, Dominique Estival, Kirstin Falkedal, Hervé Compagnion, Lorna Balkan, Frederik Fouvry, Doug Arnold, Eva Dauphin, Sylvie Régnier-Prost, Véronika Lux, Judith Klein, Judith Baur, and Klaus Netter. 1996. Tslnp – test suites for natural language processing. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, pages 711–716, Copenhagen.

[MUC1995] 1995. *Proceedings of the Sixth Message Understanding Conference, MUC-6*. Morgan Kaufman.

[Nerbonne et al.1993] John Nerbonne, Klaus Netter, Abdel Kader Diagne, Judith Klein, and Ludwig Dickmann. 1993. A diagnostic tool for german syntax. *Machine Translation*, 8(1-2):85–109.

[Neumann et al.1997] Günter Neumann, Rolf Backofen, Judith Baur, Markus Becker, and Christian Braun. 1997. An information extraction core system for real world german text processing. In *Proceedings of the 5th Conference on Applied Natural Language Processing, ANLP-97*, pages 209–216, Washington DC.

[Neumann1997] Günter Neumann. 1997. Methoden zur intelligenten informationsextraktion im internet. In *Proceedings of the 20th European Congress Fair of Technical Communications*, Hamburg. forthcoming.

[Oepen et al.1995] Stephan Oepen, Klaus Netter, Judith Baur, Tom Fettig, Judith Klein, and Fred Oberhauser. 1995. The tsnlp database: From tsct(1) to tsdb(1). Technical report, DFKI, Saarbrücken. Report to LRE 62-089.