technische universität dortmund

Dortmunder Chat-Korpus

www.chatkorpus.tu-dortmund.de

MANUAL zum Suchwerkzeug *STACCADo*
(Search Tool for Annotated Chat Corpus Analyses Dortmund)

Konzeption und Programmierung:

Bianca Selzam <selzam@hytex.info>

Projektleitung:

Michael Beißwenger und **Angelika Storrer**

Dokumentation:

Bianca Selzam und **Michael Beißwenger**

Version 1.0
(Release März 2006)

1 Einleitung 3
1.1 Was ist STACCADo? 3
1.2 Wie ist das Dortmunder Chat-Korpus aufbereitet? 3
2 Nutzungsbedingungen und Installation 4
2.1 Nutzungsbedingungen 4
2.2 STACCADo und Korpus installieren 4
2.3 Installation von Java 5
2.4 Problembehebung 5
3 Benutzung von *STACCADo 1.0* 7
3.1 Grundeinstellungen 8
3.1.1 Auswahl des zu durchsuchenden Korpusteils 8
3.1.2 Festlegung der Ausgabedatei und deren Ordner 9
3.1.3 Einstellen der Kontextgröße f. die Belegausgabe („Message-Kontext“) 10
3.2 Suchanfrage 10
3.2.1 Formulierung eines Suchstrings 10
3.2.2 Suche nach „typischen“ Elementen in Chat-Beiträgen 13
3.2.3 Filter nach Beitrags-Typen und Beitrags-Produzenten 14
3.3 Suchoptionen 15
3.3.1 Suche mit Belegstellen 15
3.3.2 Suche ohne Belegstellen 16
3.3.3 Chatter-Profil erstellen 17
3.3.4 Logfile-Profil erstellen 19
3.3.5 Nur Textausgabe 20
3.4 Ergebnisse 21
4 Weiterverarbeitung der statistischen Daten 23
4.1 Kopieren einer Tabelle 23
4.2 Einfügen einer Tabelle in Microsoft Excel 24
4.3 Sortieren nach einer Spalte 24
4.4 Ausblenden unerwünschter Spalten 25
4.5 Erzeugen von Diagrammen 25
4.5.1 Auswahl des Diagrammtyps 26
4.5.2 Auswahl des Datenbereichs 26
4.5.3 Beschriftung des Diagramms 27
4.5.4 Platzierung des Diagramms 27
5 Aufbereitung und Annotation der Korpusdaten 29
5.1 Teilautomatische Aufbereitung des Dokumentenbestandes 29
5.2 Die XML-Struktur 30
5.2.1 Element *head* 30
5.2.2 Element *body* 32

1 Einleitung

1.1 Was ist STACCADo?

STACCADo ist eine GUI-basierte Java-Anwendung, die speziell für die Formulierung und Durchführung von Suchanfragen über dem Datenbestand des Dortmunder Chat-Korpus programmiert wurde. Hierbei erlaubt ***STACCADo*** nicht nur die rein stringbasierte Suche, sondern auch die Verwendung Boolescher Operatoren sowie die Suche nach in der XML-Struktur der Korpusdokumente gesondert ausgezeichneten „chat-typischen" Elementen wie Emoticons, Adressierungen, Nicknames oder deklarativen Handlungs- und Zustandszuschreibungen in Asterisken (**lach**, **knuddel**, **dichmalganzliebtröst**). Neben einer Standard-Suche mit Ausgabe der Belegstellen sind auch statistische Auswertungen zu einzelnen Teilkorpora oder Logfiles („Logfile-Profile") und zu den darin dokumentierten Chattern („Chatter-Profile") sowie eine Textausgabe einzelner Korpusdokumente möglich.

1.2 Wie ist das Dortmunder Chat-Korpus aufbereitet?

Die Basis für das Dortmunder Chat-Korpus bilden die Mitschnitte diverser Chats, so wie sie durch Archivierung client- oder serverseitiger Logfiles aus den jeweiligen Chat-Anwendungen bezogen werden konnten. Diese „Rohdaten" wurden zunächst in ein einheitliches HTML-Format überführt und daraus schrittweise in eine XML-Struktur überführt.
Die XML-Struktur modelliert die in den Logfiles dokumentierten Chat-Ereignisse als Abfolgen von Chat-Beiträgen (*messages*), die jeweils

- einem bestimmten Chat-Teilnehmer (oder dem Chat-System) als Produzent zugeordnet sind, der im Regelfall durch die automatische Voranstellung seines Teilnehmernamens (*nickname*) kenntlich gemacht ist;
- sich einem bestimmten *message type* zuordnen lassen (unterschieden werden *messages* vom Typ „utterance", die Kommunikationsbeiträge in direkter „Rede" darstellen, von *messages* vom Typ „action", mit denen Zuschreibungen aus einer fiktiven Außensicht realisiert werden, und systemgenerierten *messages*) (vgl. 3.5.3 und 5.2.2);
- verschiedene „chat-typische" Stilelemente beinhalten können, die in der XML-Struktur gesondert ausgezeichnet wurden (Emoticons, Erwähnungen von Nicknames, Adressierungen, Asterisk-Ausdrücke). (vgl. 3.5.2 und 5.2.2)

Die XML-Struktur enthält des weiteren statistische Daten zu den in den Logfiles dokumentierten Chattern sowie Anzahl und Umfang der von ihnen produzierten Beiträge.

Eine kompakte Beschreibung der in die Korpusdokumente eingebrachten XML-Annotationen sowie der zugrunde gelegten XML-Struktur bietet Kapitel 5 dieses Handbuchs.

Eine Übersicht über den Dokumentenbestand des Dortmunder Chat-Korpus bietet die Bestandsübersicht unter http://www.chatkorpus.uni-dortmund.de.

2 Nutzungsbedingungen und Installation

2.1 Nutzungsbedingungen

STACCADo wird zusammen mit einem frei nutzbaren Ausschnitt aus dem *Dortmunder Chat-Korpus* als ZIP-Datei oder auf CD-ROM zur Verfügung gestellt. Mit dem Download des ZIP-Archivs <Chat-Korpus.zip> von unserer Website und mit der Extraktion der enthaltenen Dateien oder mit der Nutzung der auf CD-ROM zur Verfügung gestellten Daten auf Ihrem Rechner erklären Sie sich mit den folgenden Nutzungsbedingungen einverstanden:

1. Die zur freien Nutzung zur Verfügung gestellten Korpusdokumente bleiben Eigentum des Projekts „Dortmunder Chat-Korpus“ am Institut für deutsche Sprache und Literatur der Universität Dortmund. Eine Verwertung zu kommerziellen Zwecken bedarf einer vorherigen Genehmigung.
2. Bei Zitationen aus dem Datenbestand im Rahmen von Publikationen ist die Herkunft der Daten mit einem Hinweis auf das „Dortmunder Chat-Korpus“ zu belegen.

2.2 STACCADo und Korpus installieren

Entpacken Sie das ZIP-Archiv <Chat-Korpus.zip> mit einem ZIP-Programm (z. B. dem Freeware-Tool *QuickZip*) oder überspielen Sie den kompletten Inhalt der gelieferten CD-ROM in ein Verzeichnis auf Ihrem Rechner. Folgende Unterverzeichnisse und Dateien werden automatisch im Zeilverzeichnis angelegt:

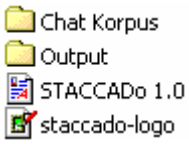


Abb. 1: Verzeichnisstruktur nach dem Entpacken

Das Verzeichnis <Chat Korpus> beinhaltet die Korpusdokumente im XML-Format. Die Dateien sind nach Zweckbereichen des Chat-Einsatzes sowie ggf. Subtypen und Chat-Angeboten auf verschiedene Unterordner verteilt. Beim Starten von ***STACCADo*** wird das Verzeichnis <Chat-Korpus> automatisch als dasjenige Verzeichnis erkannt, über welchem – sofern nicht näher spezifiziert – eine Suche ausgeführt werden soll („Eingabeordner“).

Das Verzeichnis <Output> ist das Verzeichnis, in welches ***STACCADo*** standardmäßig die Dateien mit den Sucheregebnissen ablegt („Ausgabeordner“). Wird dieses Verzeichnis verschoben oder umbenannt, erkennt ***STACCADo*** nicht mehr automatisch, wohin Suchergebnisse gespeichert werden sollen und wählt stattdessen das Verzeichnis <Eigene Dateien> auf Ihrem Rechner.

Die Datei <staccado-logo.gif> beinhaltet das Logo des Suchwerkzeugs, das nach dem Start in der linken oberen Ecke der Benutzeroberfläche angezeigt wird. Diese Grafik sollte immer im selben Verzeichnis wie die Datei <STACCADo 1.0.jar> liegen, da sie sonst auf der ***STACCADo***-Eingabemaske nicht mehr angezeigt werden kann.

Die ***STACCADo***-Programmdatei selbst trägt den Namen <STACCADo 1.0.jar>. Das Programm kann durch Doppelklick auf den Dateinamen gestartet werden (siehe hierzu die ausführlichen Erläuterungen in Abschnitt 3).

2.3 Installation von Java

Um ***STACCADo*** 1.0 ausführen zu können, muss auf Ihrem PC eine aktuelle Java-Version (1.4.X oder höher) installiert sein. Meistens findet sich diese im Ordner <C:/Programme/Java/>.

Falls Sie feststellen, dass Sie eine veraltete oder auch gar keine Version besitzen, besuchen Sie folgende Webseite:

http://java.sun.com/j2se/corejava/index.jsp

Klicken Sie hier im blauen Kasten auf der rechten Seiten auf eine der Versionen, die mit „J2SE“ beginnen und eine ausreichend hohe Versions-Nummer besitzen:

Abb. 2: Download-Link zu den Java-Versionen

In Abb. 2 sind zwei Java-Versionen zu sehen, die beide aktuell genug sind, um ***STACCADo*** zum Laufen zu bringen. Sowohl „J2SE 5.0“ als auch „J2SE 1.4.2“ wären hier auswählbar.

Entscheiden Sie sich für eine der Versionen und klicken Sie auf den Link. Sie gelangen nun auf eine Downloadseite, wo mehrere Java-Technologien zum Herunterladen angeboten werden. Suchen Sie den Link zum Download der „Java Runtime Environment (JRE)“, da dieses Programm völlig ausreicht, um Java-Applikationen laufen zu lassen.

Abb. 3: Download der JRE-Installationsdatei

Klicken Sie auf den Link zum Download und wählen Sie die Installationsdatei, die zu Ihrem Betriebssystem passt. Vergessen Sie nicht, vorher den Radio-Button „Accept Licence Agreement“ anzuklicken.

Nach dem Download der Datei muss diese nur noch ausgeführt werden. Nach der erfolgreichen Installation von Java ist eventuell ein Neustart Ihres Systems erforderlich.

2.4 Problembehebung

Eventuell kann der Fall auftreten, dass auf Ihrem PC der Dateityp „JAR“ nicht mit dem richtigen Programm verknüpft ist. Beim Doppelklick auf die Datei <STACCADo 1.0.jar> sollte sich eigentlich die grafische Benutzeroberfläche von ***STACCADo*** öffnen.

Manche PCs sind so konfiguriert, dass sie JAR-Dateien wie ZIP-Archive behandeln und sie deshalb auch mit einem ZIP-Programm öffnen wollen. In diesem Fall müssen Sie JAR-

Dateien mit dem Programm <javaw.exe> verknüpfen. Falls Ihr Betriebssystem Microsoft Windows ist, gehen Sie hierzu bitte wie folgt vor:

- Öffnen Sie den Windows Explorer und klicken Sie in der Menüleiste auf „Extras“ → „Ordneroptionen“ → „Dateitypen“.
- Suchen Sie in der Liste unter „Erweiterungen“ den Eintrag „JAR“. In der unteren Hälfte des Fensters steht das Programm, mit dem Dateien dieses Typs standardmäßig geöffnet werden.
- Um ein anderes Programm auszuwählen, klicken Sie auf „Ändern“.

Falls diese Option deaktiviert ist, verfügen Sie nicht über die nötigen Administratorrechte. Bitte melden Sie sich in diesem Fall an Ihrem PC als Administrator ein, um die Änderung durchzuführen.

Suchen Sie nun die Datei <javaw.exe>. Diese findet sich im Ordner <[Java-Verzeichnis]/bin>. Das Java-Verzeichnis ist auf den meisten Computern standardmäßig <C:/Programme/Java/>. Nachdem Sie die Datei ausgewählt haben, bestätigen Sie Ihre Auswahl und schließen Sie das Fenster.

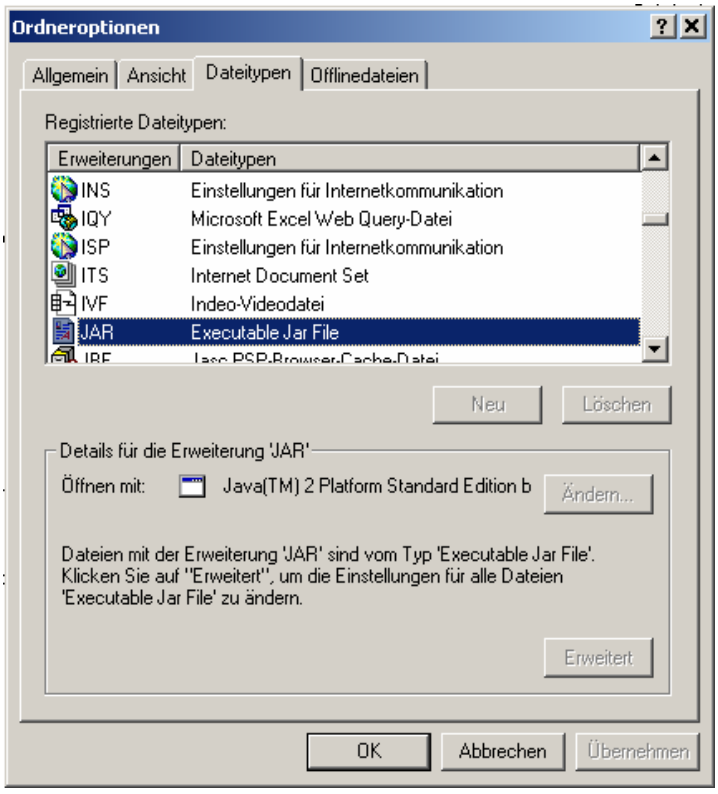


Abb. 4: Zuweisen eines Programms zu einem Dateitypen

3 Benutzung von *STACCADo 1.0*

Um die grafische Benutzeroberfläche von ***STACCADo*** aufzurufen, klicken Sie bitte doppelt auf die Datei <STACCADo 1.0.jar>. Anschließend öffnet sich folgender Startbildschirm:

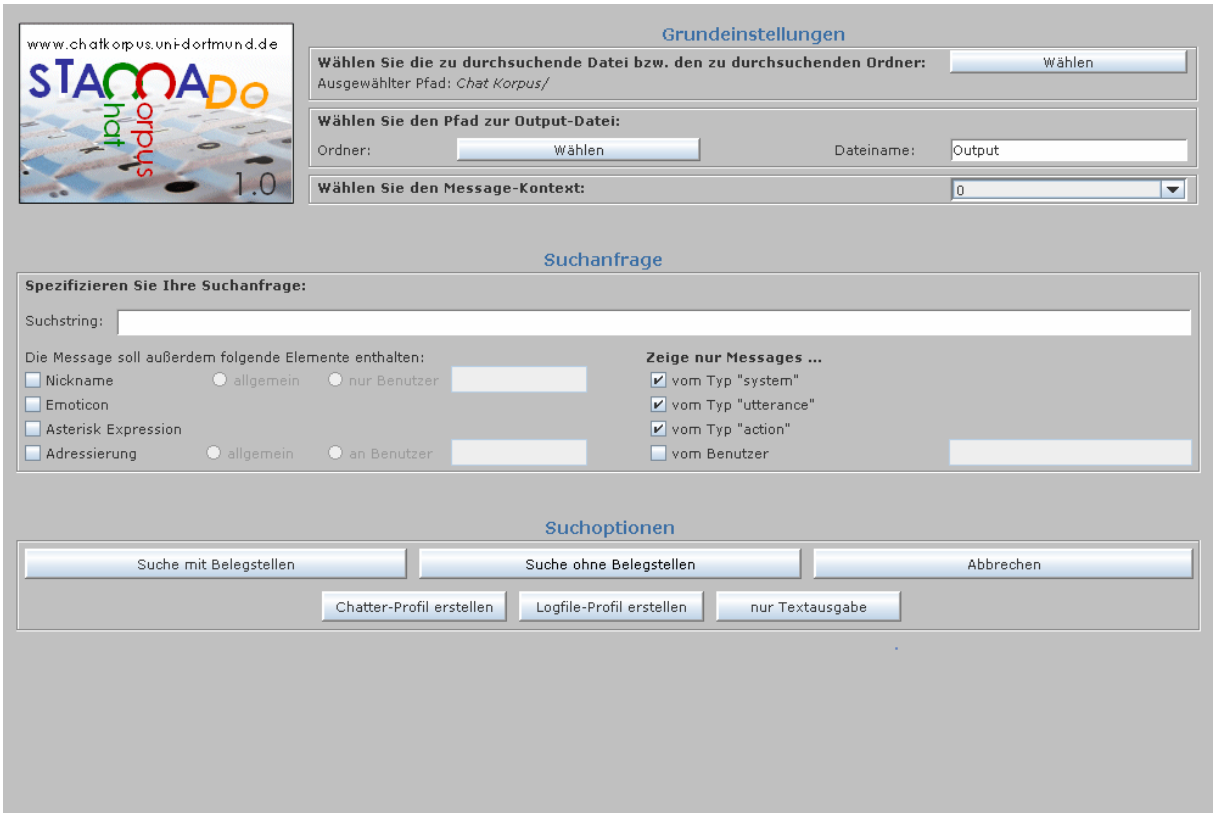


Abb. 5: Startbildschirm von „*STACCADo* 1.0“

Die Oberfläche von ***STACCADo*** ist in vier Bereiche unterteilt:

- **Grundeinstellungen:** In diesem Bereich werden der Eingabe- und Ausgabeordner gewählt, der Name der Ausgabedatei vergeben und die Kontextgröße für die Ausgabe von Belegstellen („Message-Kontext“) eingestellt. (Näheres siehe 3.1)
- **Suchanfrage:** Hier werden die Details der Suche festgelegt. Neben der bloßen Eingabe eines Suchausdrucks („Suchstrings“) ist auch die Beschränkung auf bestimmte Typen von Chat-Beiträgen möglich sowie eine Begrenzung der Suche auf Messages, die bestimmte Elemente (z.B. die Erwähnung eines Nicknames, das Vorkommen mindestens eines Emoticons, eines Asterisk-Ausrucks oder einer Adressierung) enthalten. Diese Details sind nur für die Suche mit und ohne Belegstellen notwendig. Falls ein „Chatter“- oder „Logfile-Profil“ erstellt oder die reine Textausgabe generiert werden soll, haben die hier zu treffenden Festlegungen keine Bedeutung.
- **Suchoptionen:** Dieser Bereich enthält sechs Buttons zur Auswahl der verschiedenen Such- bzw. Statistikfunktionen, die ***STACCADo*** über den Dokumentenbestand ausführt: „Suche mit Belegstellen“, „Suche ohne Belegstellen“, „Chatter-Profil erstellen“, „Logfile-Profil erstellen“ und „nur Textausgabe“. Außerdem ermöglicht der „Abbrechen“-Button zu jeder Zeit den Abbruch eines laufenden Suchvorgangs.
- **Ergebnisse:** Dieser Abschnitt wird erst dann angezeigt, nachdem eine Suche gestartet wurde. Neben der Anzahl aller gefundenen Logfiles werden auch die aktuell durchsuchte Datei, die Anzahl der bisher gefundenen Treffer sowie der aktuelle Fortschritt bei der Abarbeitung des Suchvorgangs angezeigt.

3.1 Grundeinstellungen

3.1.1 Auswahl des zu durchsuchenden Korpusteils

Abb. 6: Bereich zur Auswahl des Eingabeordners bzw. der Eingabedatei

Um den Teil des Korpus zu spezifizieren, den Sie durchsuchen möchten, klicken Sie auf den Button mit der Aufschrift „Wählen“. Daraufhin öffnet sich ein Dialogfenster, das Sie nach dem Eingabeordner bzw. der Eingabedatei fragt:

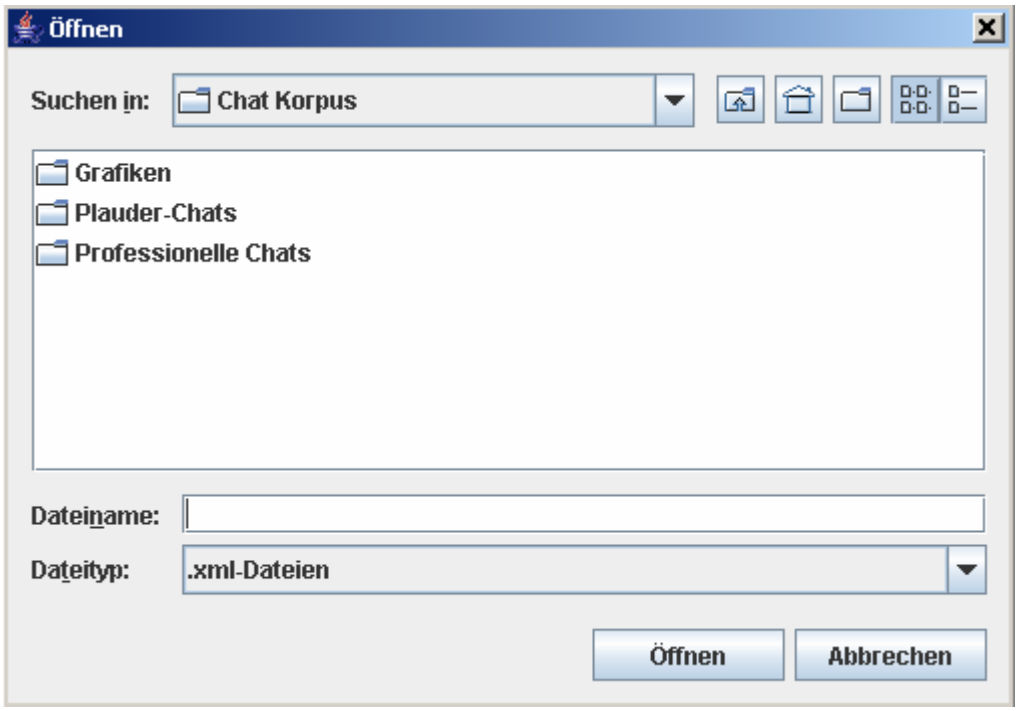


Abb. 7: Auswahl des Eingabeordners bzw. der Eingabedatei

a) Um den kompletten Inhalt eines Ordners mit allen Unterordnern auszuwählen, wechseln Sie per Doppelklick in diesen Ordner und klicken Sie anschließend auf „Abbrechen“.
b) Falls Sie nur eine einzelne Datei durchsuchen möchten, markieren Sie diese und klicken Sie anschließend auf „Öffnen“. Eine Auswahl der Datei durch einen Doppelklick ist ebenfalls möglich.

Nachdem ein Eingabeordner bzw. eine Eingabedatei ausgewählt wurde, erscheint der Pfad des ausgewählten Korpusteils unterhalb des „Wählen“-Buttons:

Abb. 8: Zum Durchsuchen ausgewählter Korpusteil

3.1.2 Festlegung der Ausgabedatei und deren Ordner

Abb. 9: Bereich zur Festlegung von Ausgabedatei und Ausgabeordner

In diesem Bereich können Sie einen Namen für die Datei festlegen, in welcher ***STACCADo*** die Suchergebnisse speichert („Ausgabedatei“) und angeben, in welchem Ordner sie gespeichert werden soll. Die von ***STACCADo*** erzeugten Ausgabedateien sind grundsätzlich im HTML-Format und lassen sich mit einem gängigen WWW-Browser (z.B. *Mozilla Firefox*, *Microsoft Internet Explorer*) betrachten.
Standardmäßig speichert ***STACCADo*** alle HTML-Dateien, die Sie erstellen, im Ordner <Output>. <Output> ist auch der Standard-Name für die Ausgabedatei. Es ist empfehlenswert, den Namen der Ausgabedatei bei jeder neuen Suche zu ändern, um das Überschreiben bereits vorhandener Dateien mit demselben Namen zu vermeiden.

Um einen anderen Ausgabeordner als den voreingestellten zu wählen, klicken Sie auf den Button „Wählen“.

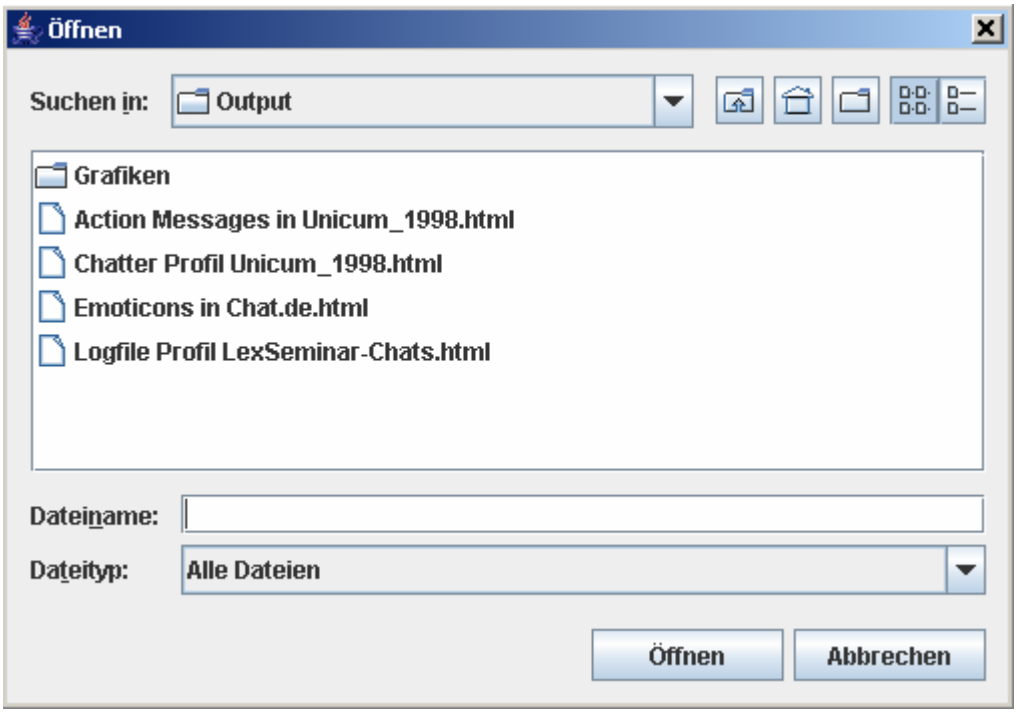


Abb. 10: Auswahl des Ausgabeordners

Die Auswahl eines anderen Ordners erfolgt wie bei der Wahl des Ausgabeordners, indem Sie in den gewünschten Ordner wechseln und anschließend auf „Abbrechen“ klicken.

3.1.3 Einstellen der Kontextgröße für die Belegausgabe („Message-Kontext“)

Abb. 11: Bereich zum Einstellen der Kontextgröße

Standardmäßig gibt ***STACCADo*** bei einer Suche *mit* Belegstellen immer den jeweiligen Chat-Beitrag („message“) als Kontext eines Treffers aus. Bei einer Suche nach dem Vorkommen von „Hallo“ werden also in der Ergebnisdatei jeweils die kompletten Chat-Beiträge wiedergegeben, in denen der Suchausdruck „Hallo“ enthalten ist. Jeder Beleg wird zudem mit einer Belegangabe versehen, aus der hervorgeht, welchem Korpusdokument der Beleg entstammt.

Wird mehr Kontext als lediglich die einen Beleg einbettende „message“ gewünscht, so kann die Kontextgröße für die Belegausgabe vor Ausführung der Suchanfrage manuell festgelegt werden. Geben Sie in das in Abb. 11 abgebildete Eingabefeld die Zahl derjenigen Messages an, die Sie vor und nach einem Treffer in der Ausgabedatei angezeigt bekommen möchten. Die maximale Kontextgröße ist auf 20 Messages vor und nach dem Treffer begrenzt.

Das Einstellen der Kontextgröße ist nur erforderlich, wenn Sie eine Suche mit Belegstellen durchführen. In allen anderen Suchoptionen wird die Zahl, die Sie hier festlegen, nicht berücksichtigt.

3.1 Suchanfrage

3.2.1 Formulierung eines Suchstrings

Abb. 12: Textfeld zur Eingabe des Suchstrings

Der „Suchstring“ ist derjenige Ausdruck, nach dessen Vorkommen im Korpus gesucht werden soll. Der Suchstring kann einfach sein und lediglich aus einer Zeichenfolge bestehen, für die Treffer gefunden werden sollen; er kann aber auch komplex sein und Muster definieren, deren Vorkommen im Datenbestand überprüft werden soll. Ein Beispiel für einen einfachen Suchstring wäre: „Suche alle Vorkommnisse der Zeichenfolge <hallo>“. Ein Beispiel für ein Suchmuster wäre: „Suche alle Beiträge, in denen eine der Zeichnfolgen <sag>, <rede> <sprech> oder <sproch> enthalten ist“.
Um das Chat-Korpus auf das Vorkommen eines bestimmten Suchausdrucks bzw. Suchmusters zu durchsuchen, tragen Sie den Suchstring in das Textfeld ein. Hierbei stehen Ihnen folgende Sonderzeichen, die untereinander alle kombinierbar sind, zur Verfügung:

1. **Das Anführungszeichen <"> markiert eine Wortgrenze**[1]. Hiermit können Sie festlegen, ob der von Ihnen gesuchte String von Leerzeichen oder Interpunktionszeichen begrenzt sein soll oder auch inmitten eines anderen Worts gefunden werden soll.

[1] Gemeint ist hier nicht eine „Wortgrenze“ im streng linguistischen Sinne, sondern eine Grenze, die dadurch markiert ist, dass einer Buchstabensequenz ein Leer- oder Interpunktionszeichen vorausgeht oder nachfolgt.

Beispiel für eine Suche mit Wortgrenzenmarkierung:	
Suchstring:	"oben"
Was wird gefunden?	**oben**
Was wird nicht gefunden?	angehoben, erhoben, obengenannt

Falls der gesuchte String dagegen auch inmitten anderer Zeichenfolgen gefunden werden soll, lassen Sie die Anführungszeichen einfach weg.

STACCADo interpretiert einen Suchstring ohne Wortgrenzenmarkierung automatisch so, dass die Treffer auch echte Teilmengen anderer Zeichenfolgen (Wortformen) sein dürfen.

Beispiel für eine Suche ohne Wortgrenzenmarkierung:	
Suchstring:	oben
Was wird gefunden?	**oben**, angeh**oben**, **oben**genannt, geh**oben**en
Was wird nicht gefunden?	Oben, Obengenannt, Obendrein

Mit Hilfe des Anführungszeichens lassen sich auch Suchen nach Vorkommen eines Suchtrings unmittelbar vor oder nach einer Wortgrenze formulieren. Um alle Vorkommen eines Suchstrings unmittelbar *nach* einer Wortgrenze zu finden, setzen Sie zu Beginn des Suchstrings eine Wortgrenze und lassen diese am Ende weg.

Beispiel für eine Suche nach Vorkommen eines Suchstrings *nach* einer Wortgrenze (= nach Wortformen, die mit dem Suchstring beginnen):	
Suchstring:	"oben
Was wird gefunden?	**oben**, **oben**genannt, **oben**auf, **oben**drein
Was wird nicht gefunden?	angehoben, gehobenen

Analog hierzu können Sie eine Suche nach dem Vorkommen eines Suchstrings unmittelbar *vor* einer Wortgrenze formulieren, indem Sie das Ende des Suchstrings mit einem Anführungszeichen markieren.

Beispiel für eine Suche nach Vorkommen eines Suchstrings *vor* einer Wortgrenze (= nach Wortformen, die auf den Suchstring enden):	
Suchstring:	oben"
Was wird gefunden?	**oben**, angeh**oben**, Pr**oben**, aust**oben**, l**oben**
Was wird nicht gefunden?	obengenannt, angehoben, gehobenen

Bitte beachten Sie, dass die hier gefundenen Wortteile nicht notwendigerweise echte Teilmengen vom gesamten Wort sein müssen. Eine Suche nach dem Wortanfang *oben* wird zum Beispiel immer auch das Wort *oben* selbst finden – analog gilt dies auch beim Wortende!

2. **Das Dollarzeichen <$> ist ein Platzhalter für exakt ein Zeichen**. Sie können beliebig viele dieser Platzhalter in einem Suchstring verwenden.

Beispiel für eine Suche mit dem $-Platzhalter:	
Suchstring:	au$en
Was wird gefunden?	t**au$en**d, k**au$en**, st**au$en**, gl**au$en**, l**au$en**
Was wird nicht gefunden?	aussen, brauchen, behaupten

3. **Der *ODER*-Operator ermöglicht die Suche nach mehreren Suchstrings, wobei *mindestens einer* von ihnen pro durchsuchte Message gefunden werden muss.** Die einzelnen Suchstrings werden hierbei durch das Zeichen **<|>** abgetrennt. Sie können beliebig viele Suchstrings zu einer ODER-Verknüpfung zusammenfassen.

Beispiel für eine Suche mit dem *ODER*-Operator <|>:	
Suchstring:	ich|du|er|sie|es
Was wird gefunden?	n**ich**t, b**er**eits, H**er**r, di**eses**, **sie**hst
Was wird nicht gefunden?	Ich, Du, Er, Sie, Es

4. **Der *UND*-Operator ermöglicht die Suche nach gemeinsamen Vorkommen zweier oder mehrerer Suchstrings in ein- und derselben Message.** Die einzelnen suchstrings werden hierbei durch die Zeichenfolge **<AND>** (in Großbuchstaben) mit einander verknüpft. Wenn Sie den AND-Operator verwenden, liefert ***STACCADo*** nur dann einen Treffer, wenn jeder einzelne der Suchstrings mindestens einmal pro Message gefunden werden konnte.

Beispiel für eine Suche mit dem *UND*-Operator <AND>:	
Suchstring:	ichANDduANDerANDsieANDes
Was wird gefunden?	Meine Beiträge kamen n**ich**t **du**rch. Für AiP´s int**eressie**rt s**ich** kein**er** mehr.
Was wird nicht gefunden?	m**ich** int**er**es**sie**rt nur, wie **du** zum op wurdest

5. **Der Code <\d> (für *engl.* „digit“) ermöglicht die Suche nach Ziffern**, ohne diese explizit aufzählen müssen (z. B. 0|1|2 usw.).

Beispiel für eine Suche nach Ziffern mit dem Code <\d>:	
Suchstring:	\d Uhr
Was wird gefunden?	6.0**0 Uhr,** 16.3**0 Uhr,** 15:0**0 Uhr**
Was wird nicht gefunden?	6:00 uhr, 16.30. Uhr

6. **Der Code <(?i)> legt fest, dass bei der Suche nach Treffern zu einem Suchausdruck die Groß- oder Kleinschreibung ignoriert werden soll.** Der Code bezieht sich immer nur auf den unmittelbar nachfolgenden Suchausdruck. Soll ein komplexer Suchstring ohne Beachtung der Groß- und Kleinschreibung gefunden werden, so setzen Sie den String in runde Klammern () und schreiben Sie den Code <(?i)< direkt vor die Klammer.

Beispiel für eine einfache Suche ohne Groß-/Kleinschreibung:	
Suchstring:	(?i)hallo
Was wird gefunden?	**hallo**, **Hallo**, **hallo**ooo, **Hallo**ween

Beispiel für eine komplexe Suche ohne Groß-/Kleinschreibung:	
Suchstring:	(?i)(hallo|hi|servus)
Was wird gefunden?	**hallo**, **Hallo**oo, **hi**, **Hi**, **hi**er, ver**hi**ndern, **servus**, **Servus**

Wichtiger Hinweis zur Suche nach Sonderzeichen:

STACCADo betrachtet einige Sonderzeichen als dafür „reserviert“, als Operatoren bei der Formulierung komplexer Suchanfragen zu fungieren. **Falls jedoch nach genau diesen Zeichen gesucht werden soll, müssen die Zeichen mit einem vorangehenden Backslash <\> eingeleitet werden.** Daran „erkennt“ ***STACCADo***, dass das Zeichen hier nicht als reserviertes Zeichen (z.B. „$“ als Platzhaltersymbol in Suchstrings) verwendet ist, sondern selbst als Suchausdruck (oder ein Teil davon) zu interpretieren ist.
Falls nach mehreren dieser reservierten Zeichen gesucht werden soll, muss vor jedes einzelne Zeichen der Backslash <\> eingegeben werden.

Die folgenden Zeichen sind hiervon betroffen:

& | * ? + $ „ ( ) [ ] { } ^ \

Alle anderen Sonderzeichen können ohne den Backslash \ im Textfeld eingegeben werden.

3.2.2 Suche nach „typischen“ Elementen in Chat-Beiträgen

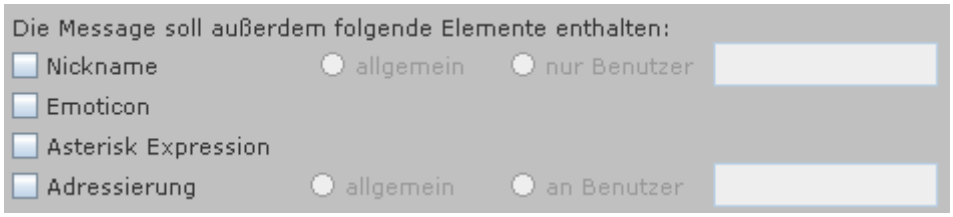


Abb. 13: Filter für die Suche nach chat-typischen Elementen

In den Korpusdokumenten sind verschiedene „chat-typische“ Elemente von Chat-Beiträgen durch XML-Annotationen als solche gekennzeichnet. Sie können daher mit ***STACCADo*** gezielt nach Vorkommen der betreffenden Elemente suchen. Im einzelnen ist eine Suche nach den folgenden Elementen möglich:

- **Nickname:** Als *Nicknames* sind Erwähnungen der Namen anderer Chatter in den Messages ausgezeichnet. Da auch Koseformen und Abkürzungen von Nicknames bei der Korpusaufbereitung manuell der jeweiligen Nickname-Grundform zugeordnet wurden, werden auch Variationen eines Nicknames als Vorkommen seiner Erwähnung gefuinden.
- **Emoticon:** Als *Emoticons* sind solche Elemente gekennzeichnet, die – entweder mit Hilfe von Sonderzeichen oder anhand kleiner Grafik-Ikons – typisierte Gesichtsausdrücke nachbilden, um auf diese Weise mimische (im Falle der Grafik-Icons bisweilen auch gestische) Ausdrucksformen zu emulieren.
- **Asterisk Expression:** Zuschreibungen von Aktion und Emotion, die zwischen Asterisken („Sternchen“) stehen (z.B. **lach**, **knuddel**, **kaffetasserüberschieb**).
- **Adressierungen:** Sprachliche Ausdrücke, mit denen durch Nennung des Nicknames eines anderen Chatters deutlich gemacht wird, dass sich der Beitrag an ihn gerichtet ist oder sich auf einen vom Adressaten produzierten Vorbeitrag bezieht (*@stoeps*, *schrödi:*, *an anne26:*). Bei der Korpusaufbereitung wurde der Adressat einer Adressierung jeweils noch einmal gesondert als Datum vermerkt (dies u.a. deshalb, weil – z.B. aus Ökonomiegründen – bisweilen unter Verwendung von Nickname-Varianten anstelle der Nickname-Grundformen adressiert wird). Hierdurch ist gewährleistet, dass sämtliche Adressierungen an denselben Adressaten als solche ausgezeichnet sind – ganz unabhängig davon, mit welcher Namensvariante er im Rahmen der Adressierung tatsächlich benannt wird.

Falls Sie nur Treffer wünschen, die mindestens eines dieser Elemente beinhalten, aktivieren Sie das Häkchen vor dem entsprechenden Element. Die Optionen können beliebig miteinander kombiniert werden. Werden mehrere Optionen gewählt (z.B. *Adressierung* und *Emoticon*), muss von jedem der betreffenden Elemente mindestens eines in einer Message enthalten sein, damit diese als Treffer gewertet wird.
Die Suche nach Nicknames und Adressierung kann optional auf einen bestimmten Benutzer eingeschränkt werden. Aktivieren Sie hierzu die Option „nur Benutzer“ und tragen Sie den entsprechenden Namen in das jeweilige Textfeld ein. Ist die Option „nur Benutzer“ aktiviert und es wurde kein Nickname eingetragen, gibt ***STACCADo*** eine Fehlermeldung aus. Wird keine Beschränkung vorgenommen, ist automatisch die Option „allgemein“ aktiv. Hierbei werden alle Nicknames und Adressierungen, die in den Logfiles gefunden werden, als Treffer gezählt, und zwar unabhängig davon, wessen Nickname erwähnt wird bzw. wer der Addressat einer Addressierung ist.

3.2.3 Filter nach Beitrags-Typen und Beitrags-Produzenten

Abb. 14: Filter nach Message-Typen und Benutzern

Bei der Aufbereitung der Chat-Logfiles wurde jeder Message einer der folgenden drei Typen zugewiesen:

- Bei ***system*** **messages** handelt es sich um vom Chat-System automatisch generierte Beiträge.
- Messages vom typ ***utterance*** sind solche Chat-Beiträge, die von ihren Produzenten in direkter "Rede" verfasst wurden. Die überwiegende Mehrzahl aller von menschlichen Nutzern produzierten Chat-Beiträge fällt in diese Kategorie.
- Bei ***action*** **messages** handelt es sich um Chat-Beiträge, die – in der Regel unter Verwendung bestimmter Prozessierungsanweisungen für das Chat-System – Zuschreibungscharakter haben, indem der Chatter Aussagen über sich selbst bzw. über die durch seinen Nickname symbolisierte "virtuelle Person" aus einer fiktiven Außensicht tätigt (z.B. *sebi03217 setzt sich aufs Sofa*, *ineli blickt sich fragend um*).

Um die Suche auf bestimmte Message-Typen zu begrenzen, aktivieren bzw. deaktivieren Sie die Häkchen vor den gewünschten Typen. Standardmäßig sind alle drei Typen aktiviert, d. h. alle drei Typen werden bei der Abarbeitung einer Suchanfrage berücksichtigt.

Die vierte Option bietet die Möglichkeit, sich nur Beiträge anzeigen zu lassen, die von einem bestimmten Benutzer produziert wurden. Wenn Sie hier ein Häkchen setzen, müssen Sie den Namen des Benutzers in das Textfeld eintragen, da andernfalls eine Fehlermeldung generiert wird. Groß- und Kleinschreibung werden bei der Eingabe des Benutzernamens nicht berücksichtigt. Diese Option ist mit der Einschränkung der Suche auf bestimmte Message-Typen kombinierbar.

3.3 Suchoptionen

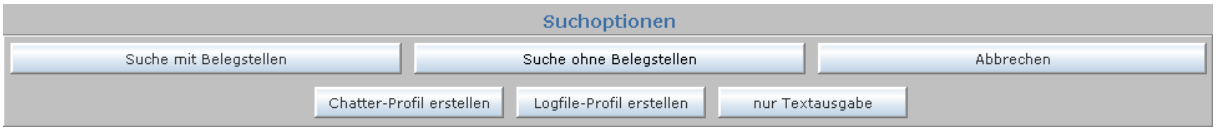


Abb. 15: Verfügbare Suchoptionen

STACCADo bietet fünf verschiedene Suchoptionen:

- **Suche mit Belegstellen:** Durchsuchen des gewählten Korpusteils gemäß der zuvor spezifizierten Suchanfrage mit Ausgabe der Belegstellen in den Logfiles.
- **Suche ohne Belegstellen:** Durchsuchen des gewählten Korpusteils gemäß der zuvor spezifizierten Suchanfrage mit statistischer Ausgabe zur Häufigkeit der Treffer und deren Produzenten.
- **Chatter-Profil erstellen:** Statistische Angaben zu allen Chattern, die im durchsuchten Korpusteil als Beitragsproduzenten dokumentiert sind.
- **Logfile-Profil erstellen:** Statistische Angaben zu allen Logfiles, die im durchsuchten Korpusteil enthalten sind.
- **nur Textausgabe:** Textausgabe des ursprünglichen Logfiles.

Der Button „Abbrechen“ ermöglicht zu jeder Zeit den Abbruch einer aktuell in Abarbeitung befindlichen Suchanfrage. ***STACCADo*** führt im Falle eines Abbruchs die Suche im aktuell durchsuchten Dokument noch zu Ende und erzeugt eine Ausgabedatei, die alle Ergebnisse enthält, die bis zum Abbruch gefunden wurden. Anschließend ist ***STACCADo*** wieder bereit für eine erneute Suchanfrage.

3.3.1 Suche mit Belegstellen

Alle Logfiles des gewählten Korpusteils werden gemäß der zuvor definierten Suchanfrage durchsucht. Die Ausgabedatei listet zuerst die Details der Suchanfrage auf, damit diese anhand der Ausgabedatei leicht rekonstruierbar sind. Anschließend wird jede Fundstelle mit Angabe der laufenden Nummer des Trefferbeitrags im Korpusdokument, des Dateinamens und des Korpusteils, in welchem sich das Dokument befindet, ausgegeben. Wurde in den Grundeinstellungen ein Message-Kontext größer als null eingestellt, werden vor und nach der Belegstelle die entsprechenden Beiträge mit ausgegeben.

Die Ausgabedatei berücksichtigt keine Farben, die evtl. in den Originallogfiles von den Chattern verwendet wurden. Der Beitrag, der den oder die Treffer enthält, wird blau markiert. Die gefundenen Treffer *innerhalb* des Beitrags werden fett gesetzt und rot markiert, um ein schnelles Auffinden zu ermöglichen. Am Ende der Ausgabedatei wird außerdem die Gesamtzahl der Treffer angegeben.

Wichtig:
Manche Logfiles enthalten kleine Grafiken, die von den Chattern als Emoticons eingesetzt wurden. Die entsprechenden Bilddateien liegen im Unterordner <Grafiken> des Standardausgabeordners <Output>. Wenn Sie die Ausgabedateien in einen anderen Ordner kopieren oder verschieben, kopieren Sie immer auch den kompletten Ordner <Grafiken> mit! Andernfalls können diese Grafiken bei der Anzeige der HTML-Ausgabedateien in einem WWW-Browser nicht mit angezeigt werden!

Beispiel:

Suchkriterien:

Zu durchsuchender Korpusteil:	Chat Korpus/
Suchstring:	(?i)(hallo wie)
Spezielle Elemente:	
Zeige nur Messages vom Typ:	System Action Utterance
Message Kontext:	3

Beleg: **Message** Nr. 116 aus **Dokument** bluewin_DJ_Tatana_11-04-2002.xml im **Teilkorpus** Professionelle Chats / Medienkontext / Bluewin /

113	**DJ Tatana**	nein leider nicht...aber wer weiss vielleicht werde ich bald mal dort auflegen..
114	**DJ5**	hey finde deine sound echt cool, ebenso den von dj antoine, wann mach ihr zusammen etwas?
115	**DJ Tatana**	keine ahnung es ist nichts geplant...
116	**Dj_kombination**	frage an dj_tatana: **hallo wie** haben sie eigentlich ire kariere begonnen?
117	**DJ Tatana**	habe vor fast acht jahren angefangen aufzulegen...und das hat sich in den jahren entwickelt..
118	**simsalabim**	wiso ist auf deinem album der song words nicht in der gleichen version wie im Clip?
119	**moderator**	tatana muss schnell ihren hunger stillen....

Beleg: **Message** Nr. 28 aus **Dokument** bluewin_Steve_Lee_12-12-2000.xml im **Teilkorpus** Professionelle Chats / Medienkontext / Bluewin /

25	**Steve Lee**	Nächstes Jahr feiern wir unser zehnjähriges Jubiläum! Ein kleiner Rekord!
26	**Punk**	wie alt bist du?
27	**Steve Lee**	Muss mal zählen... Werde nächstes Jahr stolze 38. Habe aber noch nicht im Sinn, in Pension zu gehen...
28	**Thomas1**	**Hallo wie**so soll ich Gotthard hören?
29	**Steve Lee**	Wenn Du gute Musik magst, kommst Du nicht daran vorbei...!
30	**Caramelle1**	Was hältst du von unserer Schweizer Britney Spears TAMY?
31	**Steve Lee**	Ehrlich gesagt, kenne ich sie noch nicht. Schick mir doch mal ein Foto!

Gesamtzahl der Treffer: 2

3.3.2 Suche ohne Belegstellen

Alle Logfiles des gewählten Korpusteils werden gemäß der zuvor definierten Suchanfrage durchsucht. Die Ausgabedatei listet jedoch nicht explizit die Treffer mit den Belegstellen auf, sondern zeigt nach Angabe der gewählten Suchkriterien eine absteigend geordnete Liste mit allen durchsuchten Korpusdokumenten an. Zu jedem Korpusdokument wird jeweils die darin gefundene Trefferanzahl angegeben. Weiterhin werden diese Treffer nach den Chattern aufgeschlüsselt, die die jeweiligen Trefferbeiträge produziert haben. Auch diese Liste ist absteigend nach der Trefferanzahl sortiert.

Beispiel:

Suchkriterien:

Zu durchsuchender Korpusteil:	Y:\CHAT KORPUS***STACCADo*** 1.0 + Release Korpus\Chat Korpus\Plauder-Chats\Ausserhalb Medienkontext\Unicum
Suchstring:	(?i)hallöle
Spezielle Elemente:	
Zeige nur Messages vom Typ:	System Action Utterance

Die Treffer verteilen sich wie folgt auf die durchsuchten Logfiles:

Dateiname:	**Treffer:**
unicum_03-03-2003.xml	3
unicum_30-06-2003.xml	2
unicum_01-07-2003.xml	1
unicum_19-02-2003.xml	1
unicum_20-02-2003.xml	0
unicum_15-06-2004.xml	0
unicum_11-02-2003.xml	0
unicum_1998.xml	0
unicum_21-02-2003_(1).xml	0
unicum_21-02-2003_(2).xml	0
unicum_23-06-2004.xml	0
unicum_12-02-2003.xml	0
Summe	**7**

Die gefundenen Messages wurden von folgenden Chattern produziert:

Name:	**Anzahl:**
Isadora	2
pool	1
lovebaby	1
Your_Misery	1
line	1
CUB	1
Summe	**7**

3.3.3 Chatter-Profil erstellen

Ein Chatter-Profil wird immer für alle Chatter erstellt, die im ausgewählten Korpusteil als Beitragsproduzenten in Erscheinung treten. Zu jedem Chatter werden die folgenden statistischen Daten in einer Tabelle aufgelistet:

- **est. Gender:** das Geschlecht ("male" oder "female"), das sich auf Grundlage des Nicknames für den Chatter vermuten lässt. Im Falle, dass sich *kein* Geschlecht schätzen lässt, ist das estimated Gender mit dem Wert „unknown“ belegt.
- **NOM:** die *Number of Messages*, d. h. die Anzahl der Beiträge, die der betreffende Chatter produziert hat.
- **NOT:** die *Number of Tokens*, d. h. die Anzahl der laufenden Wortformen, die der betreffende Chatter produziert hat.
- **NOC:** die *Number of Characters*, d. h. die Anzahl der Tastaturanschläge, die der betreffende Chatter produziert und abgeschickt hat.
- **Ø-Länge Message:** Die Anzahl der Tokens des betreffenden Chatters geteilt durch die Anzahl seiner Beiträge.
- **% Messages:** Gibt an, wie viel Prozent aller Beiträge im gewählten Korpusteil vom betreffenden Chatter stammen.
- **% Tokens:** Gibt an, wie viel Prozent aller Tokens im gewählten Korpusteil vom betreffenden Chatter stammen.
- **% Characters:** Gibt an, wie viel Prozent aller im gewählten Korpusteil dokumentierten Tastaturanschläge vom betreffenden Chatter stammen.
- **Länge/Ø-Chatter:** Diese Zahl zeigt die Durchschnittslänge der Beiträge eines Chatters im Verhältnis zur allgemeinen Beitrags-Durchschnittslänge in Prozent. Je näher dieser Wert an 100 liegt, desto „durchschnittlicher“ lang waren die Beiträge des betreffenden Chatters.

Bei der Berechnung der letzten fünf Werte wurde die Anzahl der vom System automatisch erzeugten Beiträge nicht miteinbezogen. Die Werte basieren also lediglich auf Auswertungen der von *menschlichen Nutzern* produzierten Beiträge.

Am Ende dier Tabelle wird der Durchschnittswert für die Spalten „NOM“, „NOT“ und „NOC“ angezeigt sowie die Durchschnittslänge der Beiträge. Am Ende jeder Ausgabedatei ist eine Legende zu den in den Spaltenüberschriften verwendeten Abkürzungen wiedergegeben.

Beispiel:

Durchsuchter Korpusteil: Y:\CHAT KORPUS***STACCADo*** 1.0 + Release Korpus\Chat Korpus\Plauder-Chats\Ausserhalb Medienkontext\Unicum\unicum_1998.xml

Nickname	est. Gender	NOM	NOT	NOC	Ø-Länge Message	% Messages	% Tokens	% Characters	Länge/Ø-Chatter
Raebchen	unknown	88	474	3327	5,3863635	12,643679	12,613092	13,304808	99,75808
McMike	male	112	580	4162	5,178571	16,091955	15,433741	16,644005	95,909676
ineli26	female	183	964	5981	5,26776	26,293104	25,651943	23,918259	97,561485
Matrose	unknown	59	333	2184	5,644068	8,477012	8,861096	8,733904	104,53089
adelheid	female	90	600	4284	6,6666665	12,931034	15,96594	17,13189	123,469925
janos	male	4	20	123	5,0	0,57471263	0,532198	0,49188194	92,60245
Interseb	male	1	3	7	3,0	0,14367816	0,07982969	0,02799328	55,561466
Monk	unknown	95	514	3293	5,4105263	13,649425	13,677488	13,16884	100,20559
Mi1kaQ	female	13	59	294	4,5384617	1,8678161	1,5699841	1,1757178	84,05453
Gangster	unknown	39	165	1065	4,230769	5,6034484	4,3906336	4,258978	78,35591
katja*	female	12	46	286	3,8333333	1,7241379	1,2240553	1,1437255	70,99521
Durch-schnitt:		**63,272728**	**341,63635**	**2273,2727**	**5,3994255**				

Legende:

est. Gender: das Geschlecht, das sich (falls möglich) auf Grundlage des Nicknames für den Chatter vermuten

lässt. Mögliche Geschlechter sind "male", "female" und "unknown".
NOM: die *Number of Messages*, d. h. die Anzahl der Messages, die dieser Chatter produziert hat.
NOT: die *Number of Tokens*, d. h. die Anzahl der laufenden Wortformen, die dieser Chatter produziert hat.
NOC: die *Number of Characters*, d. h. die Anzahl der Anschläge, die dieser Chatter produziert hat.
Ø-Länge Message: Die Anzahl der Tokens dieses Chatters geteilt durch seine Messages.
% Messages: Gibt an, wie viel Prozent aller Messages im gewählten Korpusteil von diesem Chatter stammen.
% Tokens: Gibt an, wie viel Prozent aller Tokens im gewählten Korpusteil von diesem Chatter stammen.
% Characters: Gibt an, wie viel Prozent aller Zeichen im gewählten Korpusteil von diesem Chatter stammen.
Länge/Ø-Chatter: Diese Zahl zeigt die Durchschnittslänge der Messages eines Chatters im Verhältnis zur allgemeinen Durchschnittslänge der Messages in Prozent. Je näher dieser Wert an 100 liegt, desto "durchschnittlicher" lang waren die Messages eines Chatters.

3.3.4 Logfile-Profil erstellen

Analog zum Chatter-Profil lassen sich mit dieser Suchoption statistische Daten zu allen Logfiles im gewählten Korpusteil erzeugen. Hierbei werden zu jedem Logfile in einer tabellarischen Übersicht die folgenden Daten ausgegeben:

- **TNOM:** Die *Total Number of Messages*, d. h. die Gesamtzahl aller Beiträge im betreffenden Logfile.
- **NOM Human („Number of human messages"):** Die Anzahl der von menschlichen Nutzern produzierten Beiträge, die sich aus der Differenz zwischen *TNOM* und der Anzahl der systemgenerierten Beiträge ergibt.
- **TNOT:** Die *Total Number of Tokens*, d. h. die Gesamtzahl aller Wortformen im betreffenden Logfile.
- **NOT Human („Number of human tokens"):** Die Anzahl der von menschlichen Nutzern produzierten Tokens, die sich aus der Differenz zwischen *TNOT* und der Anzahl der systemgenerierten Tokens ergibt.
- **TNOC:** Die *Total Number of Characters*, d. h. die Gesamtzahl aller Zeichen (inkl. Leer- und Sonderzeichen) im betreffenden Logfile.
- **NOC Human:** Die Anzahl der von menschlichen Nutzern produzierten Zeichen, die sich aus der Differenz zwischen *TNOC* und der Anzahl der systemgenerierten Zeichen ergibt.
- **Ø-Länge Message:** Der Wert gibt die Durchschnittslänge aller von menschlichen Nutzern produzierten Beiträge an. Er errechnet sich aus der Anzahl aller Wortformen geteilt durch die Anzahl aller Beiträge. Servergenerierte Beiträge und Wortformen werden hierbei zuvor herausgerechnet.
- **# Em.:** Die Gesamtzahl aller im betreffenden Logfile vorkommenden Emoticons.
- **Messages/Em.:** Gibt an, alle wie viele Beiträge im Durchschnitt ein Emoticon vorkommt.
- **# AstEx.:** Die Gesamtzahl aller im betreffenden Logfile vorkommenden Asterisk-Ausdrücke („AsteriskExpressions").
- **Messages/AstEx.:** Gibt an, alle wie viele Beiträge im Durchschnitt ein Asterisk-Ausdruck vorkommt.

Am Ende jeder Ausgabedatei ist eine Legende zu den in den Spaltenüberschriften verwendeten Abkürzungen wiedergegeben.

Beispiel:

Durchsuchter Korpusteil: Y:\CHAT KORPUS***STACCADo*** 1.0 + Release Korpus\Chat Korpus\Plauder-Chats\Ausserhalb Medienkontext\Unicum

Dateiname	TNOM	NOM	TNOT	NOT	TNOC	NOC	Ø-Länge	#	Messages/	#	Messages

		Human		**Human**		**Human**	**Message**	**Em.**	**Em.**	**AstEx.**	**/AstEx.**
unicum_20-02-2003.xml	865	740	3923	3313	22851	18886	4,477027	51	16,960785	118	7,3305087
unicum_01-07-2003.xml	896	778	3677	2931	19436	15869	3,767352	82	10,926829	161	5,5652175
unicum_15-06-2004.xml	1093	896	4127	3205	24206	18094	3,577009	52	21,01923	126	8,674603
unicum_19-02-2003.xml	2466	2070	10434	8586	61662	49549	4,147826	249	9,903614	458	5,3842793
unicum_11-02-2003.xml	562	506	3132	2794	18073	16143	5,521739	37	15,189189	72	7,8055553
unicum_03-03-2003.xml	3815	3061	17931	14352	109840	85882	4,688664	293	13,020478	642	5,9423676
unicum_1998.xml	758	696	4101	3758	27244	25006	5,3994255	8	94,75	128	5,921875
unicum_21-02-2003_(1).xml	787	701	3560	2981	19700	16809	4,2524962	105	7,4952383	144	5,4652777
unicum_21-02-2003_(2).xml	531	455	2242	1827	12783	10322	4,0153847	31	17,129032	69	7,695652
unicum_30-06-2003.xml	995	853	4981	4005	28844	24118	4,6951933	133	7,481203	120	8,291667
unicum_23-06-2004.xml	972	678	4284	2878	26246	16590	4,2448378	47	20,680851	78	12,461538
unicum_12-02-2003.xml	1741	1444	8292	6532	47416	36920	4,5235457	171	10,181287	267	6,5205994

Legende:
TNOM: Die Total Number of Messages, d. h. die Gesamtzahl aller Messages in diesem Logfile.
NOM Human: Anzahl der von Menschen erzeugten Messages (d. h. ohne Systemmeldungen).
TNOT: Die Total Number of Tokens, d. h. die Gesamtzahl aller Tokens in diesem Logfile.
NOT Human: Anzahl der von Menschen erzeugten Tokens (d. h. ohne Systemmeldungen).
TNOC: Die Total Number of Characters, d. h. die Gesamtzahl aller Zeichen in diesem Logfile.
NOC Human: Anzahl der von Menschen erzeugten Zeichen (d. h. ohne Systemmeldungen).
Ø-Länge Message: Die Anzahl aller Tokens geteilt durch die Anzahl aller Messages. Hierbei werden die servergenerierten Messages und Tokens herausgerechnet.
 # Em.: Die Anzahl aller in diesem Logfile vorkommenden Emoticons.
Messages/Em.: Gibt an, alle wie viel Messages durchschnittlich ein Emoticon vorkommt.
 # AstEx.: Die Anzahl aller in diesem Logfile vorkommenden AsteriskExpressions.
Messages/AstEx.: Gibt an, alle wie viel Messages durchschnittlich eine AsteriskExpression vorkommt.

3.3.5 Nur Textausgabe

Die Funktion „nur Textausgabe“ führt keine Suche über den Logfiles aus, sondern schreibt alle Dateien des gewählten Korpusteils unter Angabe des zugehörigen Dateinamens komplett in eine HTML-Datei. Auf diese Weise lassen sich die ursprünglichen Logfiles in Gänze sichten.

Wichtig:
Manche Logfiles enthalten kleine Grafiken, die von den Chattern als Emoticons eingesetzt wurden. Die entsprechenden Bilddateien liegen im Unterordner <Grafiken> des Standardausgabeordners <Output>. Wenn Sie die Ausgabedateien in einen anderen Ordner kopieren oder verschieben, kopieren Sie immer auch den kompletten Ordner <Grafiken> mit! Andernfalls können diese Grafiken bei der Anzeige der HTML-Ausgabedateien in einem WWW-Browser nicht mit angezeigt werden!

Beispiel (gekürzt):

Zu durchsuchender Korpusteil: Y:\CHAT KORPUS***STACCADo*** 1.0 + Release Korpus\Chat Korpus\Plauder-Chats\Ausserhalb Medienkontext\Unicum

Dateiname: unicum_20-02-2003.xml

1 **system** JustChat 4.0r0.204 (55.204) developed by Medium.net.
2 **system** Du betrittst den Raum.
3 **Emon** wow... ich hab hier sooo viel bücher rumliegen... und sooo wenig lust die zu bearbeiten...
4 **Emon** :
5 **Emon** : .
6 **system** jawara verdrückt sich in einen anderen Raum: suki_dakara_suki
7 **Denza** tja Emon
8 **Emon** denza, kann ich dich einstellen ;)
9 **Emon** ?
10 **olli13** wie komme ich denn hier in einen anderen raum???
11 **Denza** emon nö lass mal
12 **Emon** lohn: ein paar peitschenhiebe pro stunde
13 **Emon** /j raumname
14 **Emon** @lolli
15 **olli13** yo-ich probier das mal
16 **system** george verlässt den Raum.
17 **system** dddd betritt den Raum.
18 **system** olli13 geht in einen anderen Raum: eckzimmer
19 **Emon** na also :)
20 **Emon** hat ja geklappt
21 **system** dddd verlässt den Raum.
22 **system** stoeps kommt aus dem Raum mal_wieder_mit_stoeps_allein herein.
23 **stoeps** ree
24 **Emon** re

3.4 Ergebnisse

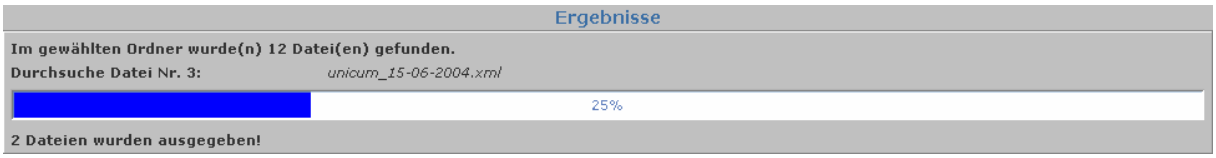


Abb. 16: Anzeige der bisher gefundenen Ergebnisse

Der Bereich „Ergebnisse“ wird auf der Benutzeroberfläche von ***STACCADo*** erst dann angezeigt, wenn eine Suche gestartet wurde. Neben der Anzahl aller im gewählten Korpusteil gefundenen Logfiles wird auch der Name des aktuell durchsuchten Logfiles ausgegeben. Ein Fortschrittsbalken zeigt anschaulich den Verlauf der Suche.

Die Trefferanzeige in der letzten Zeile hängt von der Art der Suche ab.

- Bei der Suche mit oder ohne Belegstellen wird hier die Anzahl der bisher gefundenen Treffer gezeigt.
- Bei der Erzeugung eines Chatter- und Logfile-Profils informiert die Anzeige, wie viele Logfiles bereits durchsucht wurden.
- Bei der Option „nur Textausgabe“ wird die Anzahl aller bisher in die Ausgabedatei geschriebenen Logfiles genannt.

Wurde die Suche erfolgreich abgeschlossen (d. h. mindestens ein Treffer wurde gefunden oder über alle im gewählten Korpusteil vorhandenen Logfile-Dokumente wurde ein Chatter-Profil, Logfile-Profil oder eine Textausgabe erstellt) oder vorzeitig abgebrochen, so öffnet sich ein Pop-up-Fenster mit der Frage, ob die Ausgabedatei direkt im WWW-Browser geöffnet werden soll.

Abb. 17: Abschluss-Dialog bei erfolgreicher Suche

Bei Klick auf „Ja“ wird die Ausgabedatei direkt im Browser geöffnet. Der Klick auf „Nein“ oder „Abbrechen“ dagegen schließt das Fenster und ermöglicht die Formulierung einer neuen Suchanfrage.

Wurde *kein* Treffer im gewählten Korpusteil gefunden, werden Sie durch folgenden Hinweis darüber informiert:

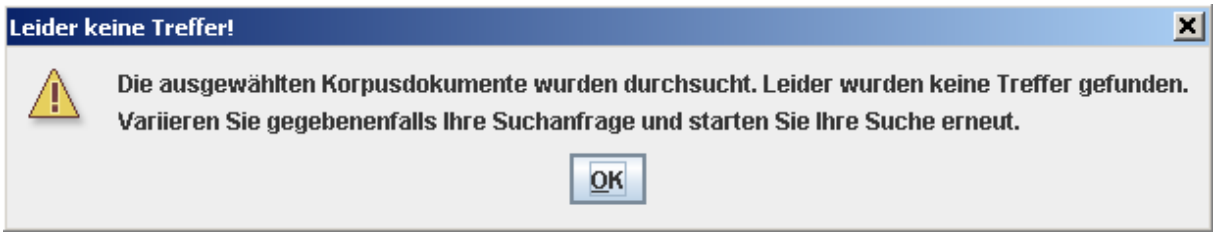


Abb. 17: Abschluss-Dialog bei erfolgloser Suche

4 Weiterverarbeitung der statistischen Daten

STACCADo bietet drei Suchoptionen zur Erzeugung von statistischen Daten an:

- Suche ohne Belegstellen (vgl. 3.3.2)
- Chatter-Profil (vgl. 3.3.3)
- Logfile-Profil (vgl. 3.3.4)

Die Ausgabedateien zu diesen Suchoptionen bestehen hauptsächlich aus Tabellen, die jedem Logfile bzw. Chatter bestimmte Werte zuordnen. Um solche Daten effizient weiterverarbeiten zu können, bieten sich Tabellenkalkulationsprogramme wie z. B. *Microsoft Excel* an. Anhand dieses Programms soll hier exemplarisch gezeigt werden, welche Möglichkeiten der Weiterverarbeitung sich dort erschließen.

4.1 Kopieren einer Tabelle

Um eine Tabelle aus der Ausgabedatei im HTML-Format in Microsoft Excel zu importieren, markieren Sie die komplette Ausgabedatei, indem Sie die Tastenkombination „Strg + A“ drücken. Alternativ dazu können Sie auch mit der Maus den kompletten Text samt Tabelle markieren, indem Sie die linke Maustaste gedrückt halten und komplett über die Seite ziehen.

Durchsuchter Korpusteil: Y:\CHAT KORPUS\STACCADo 1.0 + Release Korpus\Chat Korpus\Plauder-Chats\Ausserhalb Medienkontext\Unicum

Dateiname	TNOM	NOM Human	TNOT	NOT Human	TNOC	NOC Human	Ø-Länge Message	# Em.	Messages/Em.	# AstEx.	Messages/AstEx.
unicum_20-02-2003.xml	865	740	3923	3313	22851	18886	4,477027	51	16,960785	118	7,3305087
unicum_01-07-2003.xml	896	778	3677	2931	19436	15869	3,767352	82	10,926829	161	5,5652175
unicum_15-06-2004.xml	1093	896	4127	3205	24206	18094	3,577009	52	21,01923	126	8,674603
unicum_19-02-2003.xml	2466	2070	10434	8586	61662	49549	4,147826	249	9,903614	458	5,3842793
unicum_11-02-2003.xml	562	506	3132	2794	18073	16143	5,521739	37	15,189189	72	7,8055553
unicum_03-03-2003.xml	3815	3061	17931	14352	109840	85882	4,688664	293	13,020478	642	5,9423676
unicum_1998.xml	758	696	4101	3758	27244	25006	5,3994255	8	94,75	128	5,921875
unicum_21-02-2003_(1).xml	787	701	3560	2981	19700	16809	4,2524962	105	7,4952383	144	5,4652777
unicum_21-02-2003_(2).xml	531	455	2242	1827	12783	10322	4,0153847	31	17,129032	69	7,695652
unicum_30-06-2003.xml	995	853	4981	4005	28844	24118	4,6951933	133	7,481203	120	8,291667
unicum_23-06-2004.xml	972	678	4284	2878	26246	16590	4,2448378	47	20,680851	78	12,461538
unicum_12-02-2003.xml	1741	1444	8292	6532	47416	36920	4,5235457	171	10,181287	267	6,5205994

Legende:
TNOM: Die Total Number of Messages, d. h. die Gesamtzahl aller Messages in diesem Logfile.
NOM Human: Anzahl der von Menschen erzeugten Messages (d. h. ohne Systemmeldungen).
TNOT: Die Total Number of Tokens, d. h. die Gesamtzahl aller Tokens in diesem Logfile.
NOT Human: Anzahl der von Menschen erzeugten Tokens (d. h. ohne Systemmeldungen).
TNOC: Die Total Number of Characters, d. h. die Gesamtzahl aller Zeichen in diesem Logfile.
NOC Human: Anzahl der von Menschen erzeugten Zeichen (d. h. ohne Systemmeldungen).
Ø-Länge Message: Die Anzahl aller Tokens geteilt durch die Anzahl aller Messages. Hierbei werden die servergenerierten Messages und Tokens herausgerechnet.
 # Em.: Die Anzahl aller in diesem Logfile vorkommenden Emoticons.
Messages/Em.: Gibt an, alle wie viel Messages durchschnittlich ein Emoticon vorkommt.
 # AstEx.: Die Anzahl aller in diesem Logfile vorkommenden AsteriskExpressions.

Abb. 18: Markierte Seite eines Logfile-Profils

Kopieren Sie den Inhalt der Seite entweder durch Drücken der Tastenkombination „Strg + C“ oder durch Drücken der rechten Maustaste. In diesem Fall öffnet sich ein Pop-up-Fenster, in welchem Sie „Kopieren“ bzw. „Copy“ auswählen.

4.2 Einfügen einer Tabelle in Microsoft Excel

Öffnen Sie das Programm Microsoft Excel. Klicken Sie in der neuen leeren Datei <Mappe1> die Zelle an, in welche Sie die eben kopierten Daten einfügen möchten (standardmäßig ist die Zelle A1 markiert). Fügen Sie dort den Inhalt der Ausgabedatei ein, indem Sie entweder die Tastenkombination „Strg + V“ drücken oder mit der rechten Maustaste das Pop-up-Menü aktivieren, wo u. a. auch die Option „Einfügen“ zur Verfügung steht.

Microsoft Excel - Mappe1

Datei Bearbeiten Ansicht Einfügen Format Extras Daten Fenster ? Adobe PDF — Frage hier eingeben

Arial 12 — 100%

A1 — Durchsuchter Korpusteil: Y:\CHAT KORPUS\STACCADo 1.0 + Release Korpus\Chat Korpus\Plauder-Chats\Ausserhalb Medienkontext\Unicu

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Durchsuchter Korpusteil:** Y:\CHAT KORPUS\STACCADo 1.0 + Release Korpus\Chat Korpus\Plauder-Chats\Ausserhalb Medienkontext\Unicum											
2												
3	**Dateiname**	**TNOM**	**NOM Human**	**TNOT**	**NOT Human**	**TNOC**	**NOC Human**	**Ø-Länge Message**	**# Em.**	**Messages/Em.**	**# AstEx.**	**Messages/AstEx.**
4	unicum_20-02-2003.xml	865	740	3923	3313	22851	18886	4,477027	51	16,960785	118	7,3305087
5	unicum_01-07-2003.xml	896	778	3677	2931	19436	15869	3,767352	82	10,926829	161	5,5652175
6	unicum_15-06-2004.xml	1093	896	4127	3205	24206	18094	3,577009	52	21,01923	126	8,674603
7	unicum_19-02-2003.xml	2466	2070	10434	8586	61662	49549	4,147826	249	9,903614	458	5,3842793
8	unicum_11-02-2003.xml	562	506	3132	2794	18073	16143	5,521739	37	15,189189	72	7,8055553
9	unicum_03-03-2003.xml	3815	3061	17931	14352	109840	85882	4,688664	293	13,020478	642	5,9423676
10	unicum_1998.xml	758	696	4101	3758	27244	25006	5,3994255	8	94,75	128	5,921875
11	unicum_21-02-2003_(1).xml	787	701	3560	2981	19700	16809	4,2524962	105	7,4952383	144	5,4652777
12	unicum_21-02-2003_(2).xml	531	455	2242	1827	12783	10322	4,0153847	31	17,129032	69	7,695652
13	unicum_30-06-2003.xml	995	853	4981	4005	28844	24118	4,6951933	133	7,481203	120	8,291667
14	unicum_23-06-2004.xml	972	678	4284	2878	26246	16590	4,2448378	47	20,680851	78	12,461538
15	unicum_12-02-2003.xml	1741	1444	8292	6532	47416	36920	4,5235457	171	10,181287	267	6,5205994
16												

Abb. 19: Logfile-Profile nach dem Einfügen in Excel

4.3 Sortieren nach einer Spalte

Die Ausgabetabellen der Chatter- und Logfile-Profile sind nicht geordnet, da sie von ***STACCADo*** in derjenigen Reihenfolge in die Ergebnisdatei geschrieben werden, in welcher die einzelnen Verzeichnisse des Korpusteils durchsucht wurden. Je nach Untersuchungsfrage kann es jedoch sehr hilfreich sein, die Ausgabedateien nach einer bestimmten Spalte sortieren zu lassen.

Um die Tabelle nach einer bestimmten Spalte zu sortieren, wählen Sie in der Excel-Menüleiste im Punkt „Daten“ die Option „Sortieren“ aus. Im nun erscheinenden Fenster wählen Sie die Spalte aus, nach welcher die Tabelle sortiert werden soll. Standardmäßig ist die Sortierrichtung „Absteigend“ ausgewählt, was Sie selbstverständlich zu „Aufsteigend“ ändern können.

Falls in zwei oder mehr Zellen der gewählten Spalte derselbe Wert gefunden werden sollte, können Sie zwei weitere Spalten angeben, nach denen in solchen Fällen sortiert werden soll.

Abb. 20: Auswahl der Sortierrichtung und -spalte

Wenn Sie alle Einstellungen vorgenommen haben, klicken Sie anschließend auf „OK“. Ihre Tabelle wird nun Ihren Wünschen entsprechend sortiert sein und sich somit besser zur Weiterverarbeitung eignen.

4.4 Ausblenden unerwünschter Spalten

Für die Auswertung der Chatter- und Logfile-Profile sind vermutlich nicht alle Spalten von gleichem Interesse. Um die Übersichtlichkeit zu erhöhen, können aktuell irrelevante Spalten bei Bedarf ausgeblendet werden.

Markieren Sie die komplette Spalte bzw. die kompletten Spalten, die ausgeblendet werden sollen, durch Anklicken des Buchstabens bzw. der Buchstaben (z. B. „A“, „B“ oder „C“) der Spaltenüberschrift. Nebeneinander liegende Spalten lassen sich durch Ziehen der Maus bei gedrückter linker Maustaste markieren – andernfalls halten Sie die Taste „Strg“ gedrückt, wenn Sie nacheinander die auszublendenden Spalten markieren.

Wurden alle unerwünschten Spalten markiert, klicken Sie mit der rechten Maustaste auf einen der markierten Buchstaben in den Spaltenüberschriften. Wählen Sie nun im sich öffnenden Pop-up-Fenster die Option „Ausblenden“. Die markierten Spalten werden ab sofort nicht mehr angezeigt. Die Buchstaben der Spaltenüberschriften werden hierbei allerdings nicht verändert.

Um sich die ausgeblendeten Spalten zu einem späteren Zeitpunkt wieder anzeigen zu lassen, markieren Sie die komplette Tabelle durch Eingeben der Tastenkombination „Strg + A“, klicken auf eine beliebige Zelle in der Tabelle mit der rechten Maustaste und wählen im erscheinenden Pop-up-Menü die Option „Einblenden“. Nun werden wieder alle Spalten zusammen mit den zuvor ausgeblendeten Spalten angezeigt.

4.5 Erzeugen von Diagrammen

Microsoft Excel bietet die Möglichkeit, in wenigen Schritten aus mindestens zwei Tabellenspalten ein Diagramm zu erzeugen, das in vielen Fällen anschaulicher sein kann als die reinen Zahlen in den Tabellenzellen.

Wenn Sie ein Diagramm aus zwei oder mehr Spalten Ihrer Excel-Tabelle erzeugen möchten, wählen Sie in der Excel-Menüleiste im Punkt „Einfügen“ die Option „Diagramm“ aus. Sofort öffnet sich automatisch der Diagramm-Assistent, der Sie beim Erstellen des Diagramms unterstützt.

4.5.1 Auswahl des Diagrammtyps

Wählen Sie in der linken Spalte einen der zur Verfügung stehenden Diagrammtypen (z. B. „Säule“) sowie in der rechten Spalte einen entsprechenden Diagrammuntertypen aus, der Ihren Vorstellungen entspricht. Klicken Sie anschließend auf „Weiter >“.

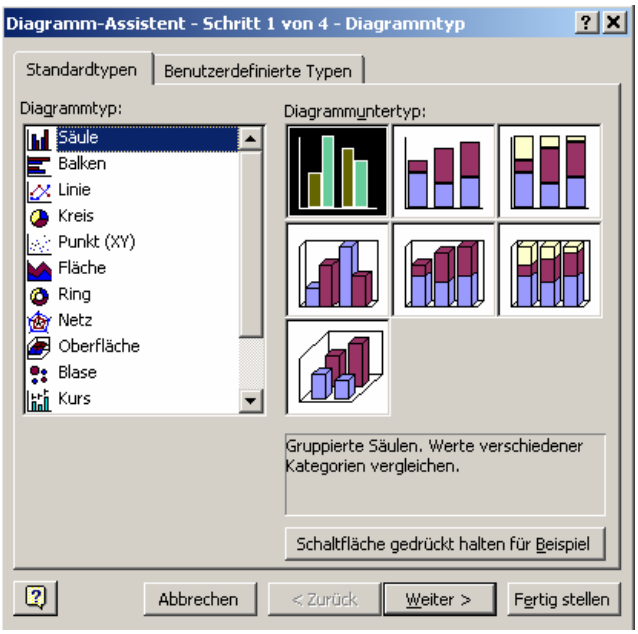


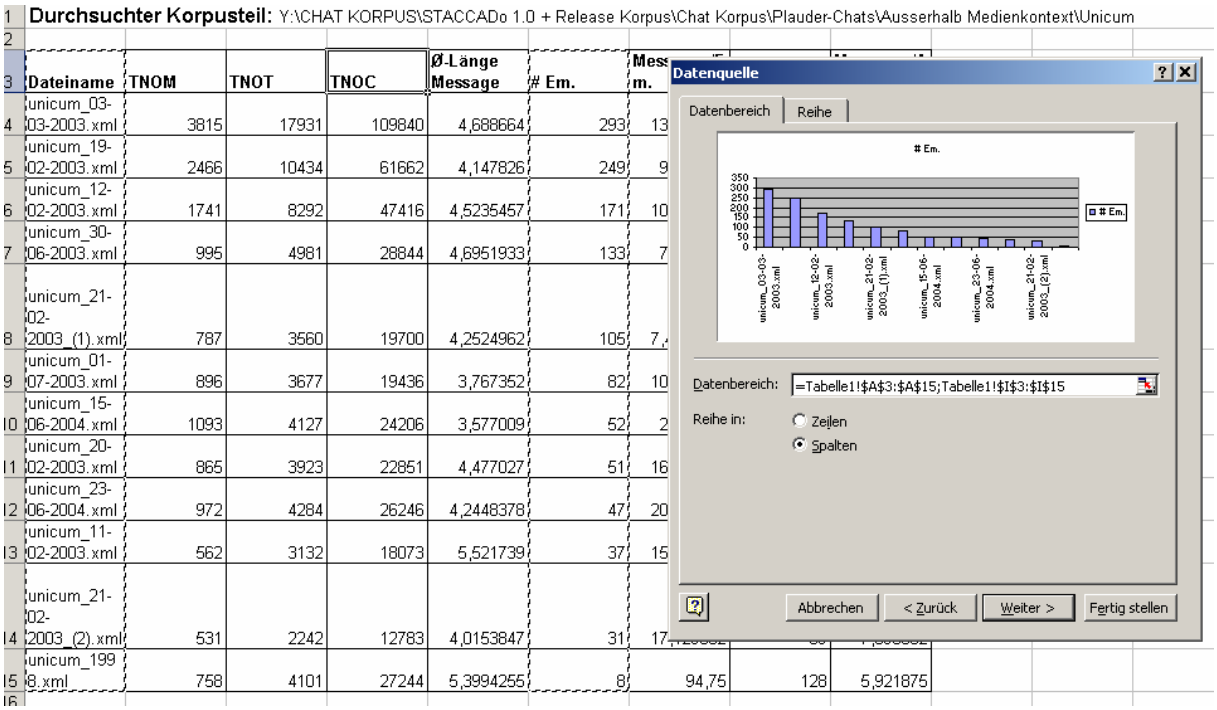
Abb. 21: Diagrammassistent von MS Excel: Auswählen des Diagrammtyps

4.5.2 Auswahl des Datenbereichs

Definieren Sie nun die Daten, über welche das Diagramm erstellt werden soll. Standardmäßig ist die komplette Tabelle markiert. Markieren Sie nur die Zellen, deren Werte auch in das Diagramm eingetragen werden sollen. Um mehrere Zellen zu markieren, die nicht nebeneinander liegen, halten Sie die Taste „Strg“ gedrückt und markieren Sie die Zellen nacheinander.

Hinweis:
Falls Sie komplette Spalten durch Anklicken der Buchstaben in den Spaltenüberschriften markieren möchten, achten Sie darauf, dass in keinen Zeilen dieser Spalten unerwünschte Daten liegen (z. B. die Legende des Chatter- oder Logfile-Profils). Andernfalls werden diese Daten ebenfalls in das Diagramm eingebunden, was zu unerwünschten Effekten in der Darstellung führen kann.

Wenn Sie alle Daten, aus denen das Diagramm erzeugt werden soll, ausgewählt haben, klicken Sie auf den Button „Weiter >“.



Dateiname	TNOM	TNOT	TNOC	Ø-Länge Message	# Em.	Mess m.		
unicum_03-03-2003.xml	3815	17931	109840	4,688664	293	13		
unicum_19-02-2003.xml	2466	10434	61662	4,147826	249	9		
unicum_12-02-2003.xml	1741	8292	47416	4,5235457	171	10		
unicum_30-06-2003.xml	995	4981	28844	4,6951933	133	7		
unicum_21-02-2003_(1).xml	787	3560	19700	4,2524962	105	7		
unicum_01-07-2003.xml	896	3677	19436	3,767352	82	10		
unicum_15-06-2004.xml	1093	4127	24206	3,577009	52	2		
unicum_20-02-2003.xml	865	3923	22851	4,477027	51	16		
unicum_23-06-2004.xml	972	4284	26246	4,2448378	47	20		
unicum_11-02-2003.xml	562	3132	18073	5,521739	37	15		
unicum_21-02-2003_(2).xml	531	2242	12783	4,0153847	31	17		
unicum_1998.xml	758	4101	27244	5,3994255	8	94,75	128	5,921875

Abb. 22: Auswahl des Datenbereichs

4.5.3 Beschriftung des Diagramms

Nun können Sie die X- und Y-Achse des Diagramms beschriften sowie dem Diagramm einen Titel geben. Tragen Sie die gewünschten Bezeichnungen in die vorgesehenen Felder ein und klicken Sie anschließend auf den Button „Weiter >“.

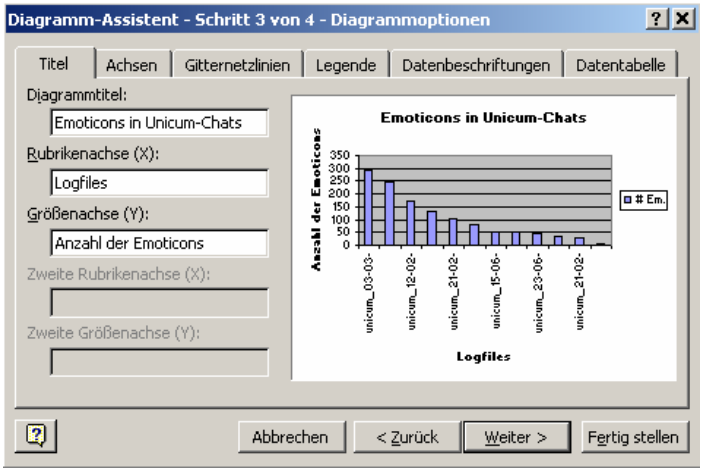


Abb. 23: Auswahl des Datenbereichs

4.5.4 Platzierung des Diagramms

Im letzten Schritt können Sie entscheiden, ob das Diagramm in Ihre vorhandene Excel-Tabelle oder in ein neues Tabellenblatt eingefügt werden soll. Der Übersichtlichkeit halber ist die Option „Als neues Blatt“ zu empfehlen.

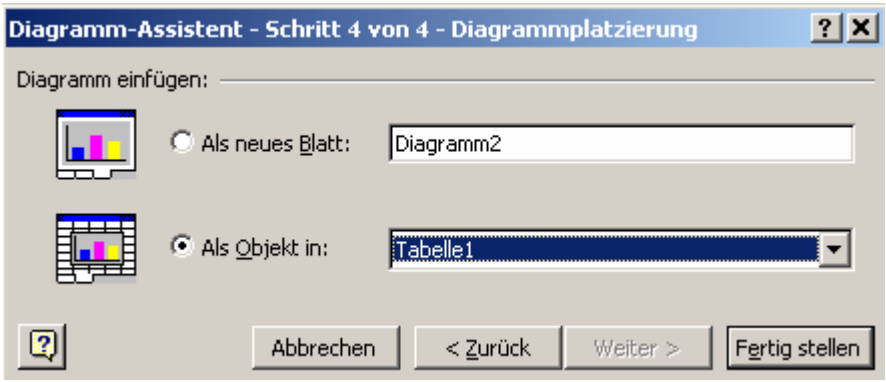


Abb. 24: Auswahl der Diagrammplatzierung

Klicken Sie anschließend auf „Fertig stellen“. Ihr Diagramm wird nun erzeugt und in das gewählte neue Blatt bzw. die ausgewählte Tabelle eingefügt.

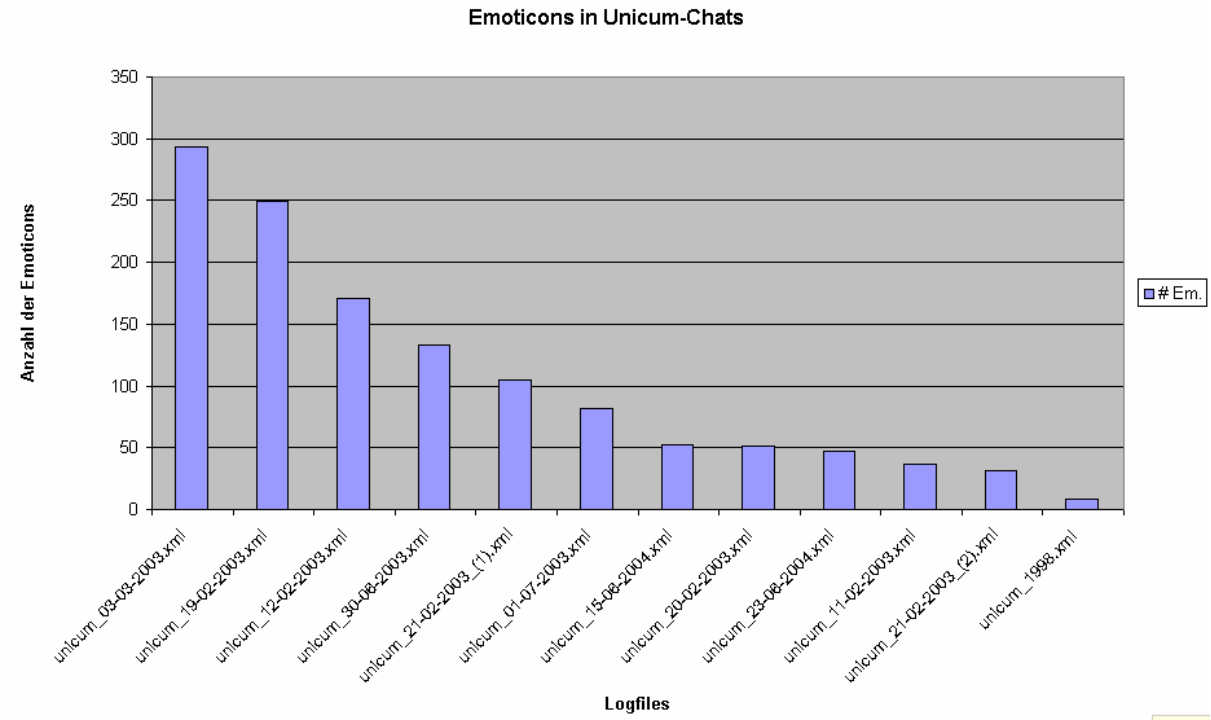


Abb. 25: Fertiges Diagramm zur Anzahl der Emoticons in Unicum-Chats

Um das Diagramm in anderen Programmen wie *Microsoft Word* oder *PowerPoint* weiterverwenden zu können, markieren Sie das Diagramm mit der Tastenkombination „Strg + A“, kopieren es mit den Tasten „Strg + C“ in die Zwischenablage und fügen es im gewünschten Word- oder PowerPoint-Dokument mit den Tasten „Strg + V“ oder einem Rechtsklick und der Auswahl der Option „Einfügen“ bzw. „Paste“ wieder ein.

5 Aufbereitung und Annotation der Korpusdaten

Dieses Kapitel soll dazu dienen, die Details der XML-Annotation der Logfiles zu erläutern. Für eine erfolgreiche Benutzung von ***STACCADo*** ist das Kapitel 4 nicht notwendig – es bietet nur die Hintergrundinformationen, auf welche Weise und nach welcher DTD die Logfiles bearbeitet wurden, um sie mit ***STACCADo*** durchsuchbar zu machen.

5.1 Teilautomatische Aufbereitung des Dokumentenbestandes

Die Basis für das Dortmunder Chat-Korpus bilden die Mitschnitte diverser Chats, so wie sie von den jeweiligen Chat-Anwendungen bezogen werden konnten. Die Ausgangsdaten sind im Falle serverseitig erzeugter Logfiles häufig im TXT-Format, im Falle clientseitig erzeugter Logfiles entweder HTML (bei Speicherung des Bildschirmsprotokolls über einen Browser) oder (bei der Copy & Paste-Sicherung des Bildschirmprotokolls) DOC bzw. RTF. Diese „Rohdaten“ wurden zunächst mit Hilfe des MS Office-Konverters in ein HTML-Format überführt, sofern das Ausgangsformat nicht bereits HTML war. Manuell wurden anschließend irrelevante HTML-Tags entfernt, um den teilweise recht unübersichtlichen Office-Code zu vereinfachen. Durch eine Suchen & Ersetzen-Routine wurden die Tags für die formale Grundeinheit unserer Modellierung (*message*) automatisch eingefügt und das Dokument als XML-Datei gespeichert.

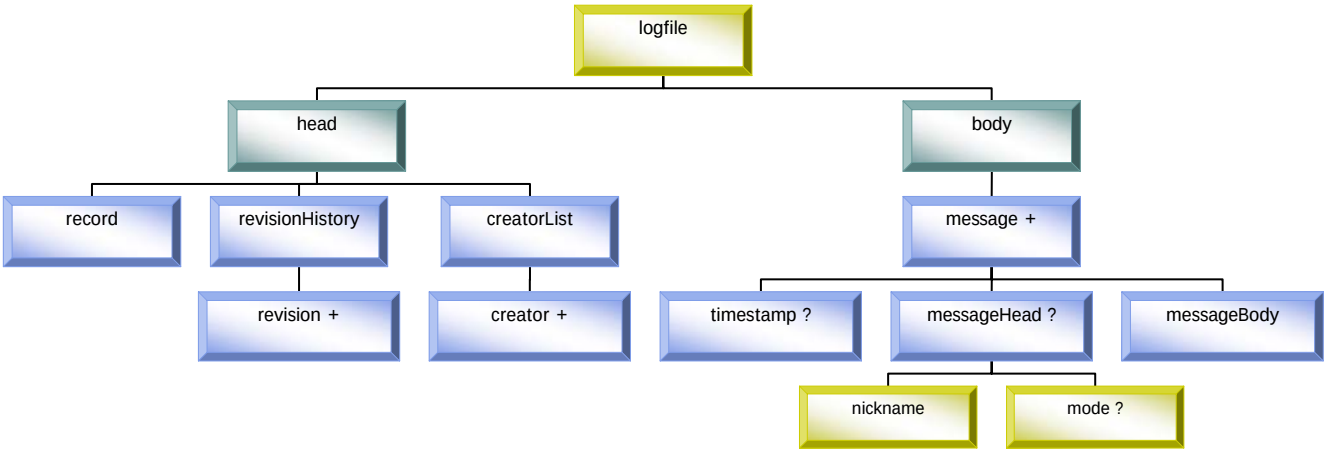
In Anbetracht der signifikanten Unterschiede zwischen der Organisation mündlicher Gespräche und der Handlungskoordination im Chat haben wir uns dafür entschieden, in den Korpusdokumenten nicht **Gesprächsbeiträge (*Turns*)**, sondern **Chat-Beiträge (*messages*)** als Grundeinheiten unserer Modellierung anzunehmen. Unter einem ***Chat-Beitrag*** verstehen wir solche Teilnehmeräußerungen, die im Display aufgrund jeweils eines vorangehenden und eines nachfolgenden Absatzreturns als Einheiten isolierbar sind, die vom betreffenden Produzenten durch Ausführung eines Sendeakts als Einheit an den Chat-Server übermittelt und von diesem in das Display der Adressatenrechner übermittelt wurden. Der *Chat-Beitrag* stellt somit eine lediglich formale Einheit dar; über seine Funktion oder den ihm von Seiten des Produzenten beigemessenen Handlungswert ist damit noch nichts ausgesagt.

Mit dem selbst entwickelten Java-Werkzeug ***Logfile2XML*** wurden die Dokumente nach Einfügung der *message*-Tags automatisch vorannotiert: Bestimmte Attribute und Attributwerte zum Element *message* wurden eingefügt sowie Emoticons und Asterisk-Ausdrücke unterhalb der *message*-Ebene ausgezeichnet. Die somit geschaffene rudimentäre XML-Struktur wurde in einem weiteren Schritt von Hand weiterbearbeitet, um automatisch nicht zweifelsfrei identifizierbare Elemente wie Adressierungen (*address*) oder im Text einer *message* erwähnte (nicht selten abgekürzte) Nicknames (*nickname*) auszuzeichnen und kleinere Unzulänglichkeiten der maschinellen Vorannotation nachzukorrigieren.

Die fertig annotierte Datei wurde zuletzt in das ebenfalls selbst entwickelte Java-Werkzeug ***ExtendedHead*** eingelesen. ***ExtendedHead*** generiert automatisch statistische Daten zum Inhalt eines Dokuments und schreibt diese – ebenfalls in Form einer XML-Struktur – in das Element *head* der XML-Struktur. Weitere Metadaten, die sich nicht maschinell auslesen lassen, für die Dokumentation der Korpusdaten jedoch wünschenswert sind, wurden anschließend von Hand ergänzt (z. B. das *estimatedGender* oder die *creatorList*; s.u.).

5.2 Die XML-Struktur

Den annotierten XML-Dateien unseres Korpus liegt folgende Struktur zugrunde:



Das Wurzelelement *logfile* hat als Kinder *head* und *body*. Der *head* eines enthält Metadaten zum Mitschnitt, eine "Revision History" des Korpusdokuments sowie statistische Daten zum im Logfile dokumentierten Kommunikationsaufkommen. Im *body* hingegen steht mit den Chat-Daten der eigentliche Content des Mitschnitts.

5.2.1 Element *head*

Die direkten Kindelemente von *head* sind *record*, *revisionHistory* und *creatorList.* Ihre Funktionen und Attribute werden im Folgenden erläutert.

a) Element *record*

record enthält Informationen zum Chat-Angebot und dessen Aufzeichnung sowie statistische Daten.

Attribut	*Wertebereich*	*obligatorisch* od. *fakulatativ* ?
plattformName	Name des Chat-Angebots	obligatorisch
plattformURL	URL des Chat-Angebots, z.B. "http://www.unicum.de/chat"; falls nicht ermittelbar, dann "unknown" eintragen	obligatorisch
recDate	Aufzeichnungsdatum in der Form YYYY-MM-DD, z.B.: "2003-09-21" für "21. September 2003". Ist das Datum nicht bekannt, wird das Attribut mit dem Wert "unknown" belegt.	obligatorisch
recStart	Starzeitpunkt der Aufzeichnung in der Form HH-MM, z.B. "19-25" für "19 Uhr 25 Minuten". Ist der Zeitpunkt nicht bekannt, wird das Attribut mit dem Wert "unknown" belegt.	obligatorisch
recEnd	Starzeitpunkt der Aufzeichnung in der Form HH-MM, z.B. "21-17" für "21 Uhr 17 Minuten". Ist der Zeitpunkt nicht bekannt, wird das Attribut mit dem Wert "unknown" belegt.	obligatorisch
recBy	Name des/der Aufzeichnenden (sofern bekannt), ansonsten "unknown"	obligatorisch

TNOM ("total number of messages")	Anzahl der messages im Dokument-Body	obligatorisch
TNOT ("total number of tokens")	Anzahl der laufenden Wortformen im Logfile (zu ermitteln durch "Wörter zählen" in der Word-Ausgangsversion)	obligatorisch
view	Nickname des Chatters, dessen Sicht auf das Kommunikationsgeschehen im Logfile dokumentiert ist (nur, sofern relevant)	fakultativ

b) Element *revisionHistory*

Die *revisionHistory* dokumentiert die verschiedenen Aufbereitungs- und Bearbeitungsschritte eines Dokuments. Sie wird bei jedem Bearbeitungsvorgang bzw. bei jeder Änderung des Dokuments aktualisiert.

Das Element *revisionHistory* enthält beliebig viele Kindelemente des Typs *revision*.

Element *revision*

revision dokumentiert einen Bearbeitungsvorgang am Dokument. Bei jedem Bearbeitungsvorgang eines Dokuments bzw. bei jeder Änderung in einem Dokument wird der *revisionHistory* ein neues Kindelement *revision* hinzugefügt. Die Elemente *revision* werden laufend durchnummeriert. Inhalt jedes *revision*-Elements ist eine Kurzbeschreibung der vorgenommenen Änderungen bzw. Überarbeitungsschritte.

Attribut	*Wertebereich*	*obligatorisch* od. *fakulatativ* ?
no	Laufende Nummer des Vorkommens des Elements **revision**.	obligatorisch
by	Name des Bearbeiters/der Bearbeiterin.	obligatorisch

Annotationsbeispiel:

```
<revision no="6" by="Bianca Selzam">
      Metadaten hinzugefügt, Durchnumerierung der messages vorgenommen
</revision>
```

c) Element *creatorList*

Die *creatorList* enthält eine Liste sämtlicher im Mitschnitt aktiver Chatter, die in beliebig vielen Kindelementen vom Typ *creator* kodiert sind.

Element *creator*

Ein Element *creator* repräsentiert einen Chatter in einem Logfile, der mindestens einen Beitrag aktiv verfasst hat.

Attribut	*Wertebereich*	*obligatorisch* od. *fakulatativ* ?
name	Nickname des Chatters bzw. "system" für das System (im Falle, dass das Logfile systemgenerierte Beiträge enthält)	obligatorisch
estimatedGender	geschätztes Geschlecht: "male" / "female" / "unknown" / "system"	obligatorisch
NOM ("number of messages")	Anzahl der vom betreffenden Chatter produzierten Messages (also derjenigen	obligatorisch

	messages im Body, für welche er als creator fungiert)	
NOT ("number of tokens")	Anzahl der laufenden Wortformen sämtlicher messages des betreffenden Chatters	obligatorisch
role	Angabe einer Kommunikantenrolle (falls für den betreffenden Chat relevant), z.B. "moderator", "celebrity"	fakultativ

Annotationsbeispiel:

```
<creatorList>
      <creator name="system" estimatedGender="system" NOM="8" NOT="48"/>
      <creator name="Martin" estimatedGender="male" NOM="3" NOT="31"/>
      <creator name="Nina" estimatedGender="female"  NOM="7" NOT="64"/>
      <creator name="Ludwig" estimatedGender="male" NOM="15" NOT="122"
            role="moderator"/>
</creatorList>
```

5.2.2 Element *body*

Der *body* der XML-Datei enthält den eigentlichen Mitschnitt. Er enthält beliebig viele Elemente vom Typ *message*, deren Unterelemente und Aufbau nun beschrieben werden.

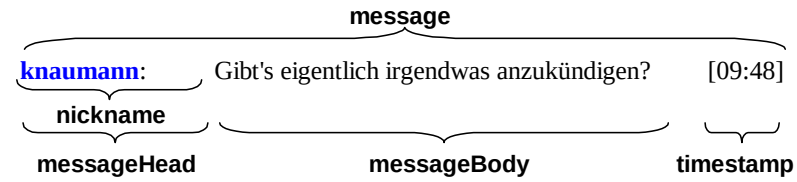
a) Element *message*

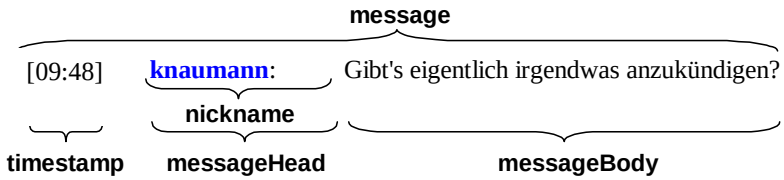
Die Kategorie *message* beschreibt solche Einheiten, die von einzelnen Chattern durch Ausführung einer Verschickungshandlung (z.B. durch Betätigen der Eingabetaste oder Mausklick auf einen Sendebutton) an den Chat-Server aufgegeben wurden und die in den Logfiles jeweils einzeln als Produkte eines bestimmten Urhebers ausgewiesen werden (in aller Regel durch automatische Voranstellung des Teilnehmer-Nicknames sowie durch vorangehenden und nachfolgenden Absatzreturn).

Attribut	*Wertebereich*	*obligatorisch* od. *fakulatativ* ?
id	Laufende Nummer der *message*.	obligatorisch
type	Zuweisung eines der Subtypen *utterance* ("Äußerungsbeiträge"), *action* ("Beiträge mit Zuschreibungscharakter") oder *system* ("Systemmeldungen"). In den IRC-Chats tritt zusätzlich der Typ *bot* auf ("automatisch generierte Beiträge eines Chat-Robots").	obligatorisch
creator	Produzent der Message. Bei Systemmeldungen wird "system" angegeben.	obligatorisch
color	Farbe des Beitrags	fakultativ

a) Element *timestamp*

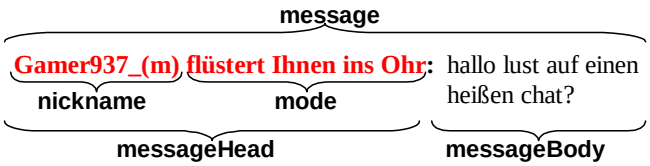
Dieses (fakultative) Element gibt den Zeitpunkt der Entgegennahme eines Beitrags durch den Chat-Server an. Der timestamp kann entweder message-initial oder message-final stehen:





b) Element *messageHead*

Der *messageHead* umfasst diejenigen Teile einer *message*, der (a) vom System automatisch generiert wurden und (b) die Funktion haben, den Produzenten des Beitrags anzuzeigen sowie ggf. den Äußerungsmodus zu benennen, der vom Produzenten für den Beitrag gewählt wurde (z.B. "Flüster"-Modus).



c) Element *messageBody*

Eine *message* beinhaltet immer einen *messageBody*. Dieser umfasst denjenigen Teil der *message*, der die vom betreffenden Teilnehmer eingegebene Zeichenfolge (und somit den "Beitrag" im engeren Sinne) wiedergibt.

Adressierungen innerhalb des Elements *messageBody* werden durch das XML-Element *address* gekennzeichnet. Das zugehörige Attribut *addressee* zeigt, welcher andere Chat-Teilnehmer als Adressat gewählt wurde. Da Nicknames in Adressierungen häufig abgekürzt werden (z.B. „anton“ anstatt „anton23“) und bisweilen aus Flüchtigkeit auch Tippfehler enthalten („atnon“ statt „anton“), ist die Belegung des Attributs *addressee* obligatorisch. Während das Element *address* diejenige Zeichenfolge markiert, die im Beitrag als Adressierung fungiert, wird als Wert zu *addressee* die originäre Form des Nicknames angegeben.

Annotationsbeispiel:

```
<message id="20" type="utterance" creator="tourteam" color="#CC0000">
      <timestamp>
            17:02:06
      </timestamp>
      <messageHead>
            <nickname>tourteam</nickname>
      </messageHead>
      <messageBody>
            <address addressee="anton23">atnon:</address> wir müssen und
            noch ein paar tage gedulden
      </messageBody>
</message>
```

Wenn andere Chatter nicht direkt adressiert, sondern im Rahmen eines Teilnehmerbeitrags erwähnt werden, so werden diese Erwähnungen mit dem Element *nickname* ausgezeichnet, und zwear zunächst unabhängig davon, ob der betreffende Teilnehmer tatsächlich mit seinem Nickname, mit seinem realweltlichen Namen oder einem anderen sprachlichen Ausdruck genannt wird (z.B. „Torsten“ anstelle von „Rocky19“ oder „Grenzwall“ anstelle von „Limes“). Falls der verwendete Ausdruck vom Nickname abweicht, wird zum Attribut *baseform* der im Chat verwendete Nickname des betreffenden Teilnehmers als Wert angegeben.

Annotationsbeispiel:

```
<message id="339" type="utterance" creator="quaki" color="#D62994">
      <messageHead>
            <nickname>quaki</nickname>
      </messageHead>
      <messageBody>
            <nickname baseform="limes">der grenzwall</nickname> is schon
            wieda da heheh
      </messageBody>
</message>
```

„Netspeak“-Elemente wie Emoticons und Handlungsbeschreibungen in Asterisken werden durch die Tags *emoticon* und *asteriskExpression* markiert. Diese können auch ineinander verschachtelt sein.

Annotationsbeispiel:

```
<message id="599" type="utterance" creator="TomcatMJ" color="#003388">
      <messageHead>
            <nickname>TomcatMJ</nickname>
      </messageHead>
      <messageBody>
            <asteriskExpression>*hängematte in baum aufspann und mich
            reinleg*</asteriskExpression><emoticon>:-)</emoticon>
      </messageBody>
</message>
```