Statistische Auswertungen mit *STACCADo* (1): Logfile-Profile

Logfile-Profile – Statistische Übersichten zu Teilkorpora *on demand*

Ein „Logfile-Profil" ist eine tabellarische Übersicht mit statistischen Daten zu allen Mitschnitten aus einem Teilkorpus des Dortmunder Chat-Korpus. Logfile-Profile zu beliebigen Teilkorpora lassen sich mit dem Korpusabfragewerkzeug *STACCADo* automatisch erzeugen.

Die Teilkorpora des Dortmunder Chat-Korpus enthalten einzelne Korpursdokumente. Die Korpusdokumente beinhalten einzelne Mitschnitte (engl. „Logfiles"), die um Annotationen zur Dokumentstruktur und zu typischen Stilelementen (z.B. Emoticons, Asterisk-Ausdrücken) sowie um statistische Metadaten angereichert sind. Die Annotationen und Metadaten bilden – neben dem Volltext der im Mitschnitt enthaltenen Chat-Beiträge – diejenigen Datentypen, über denen Suchanfragen definiert und mit *STACCADo* durchgeführt werden können.

Die Erzeugung von Logfile-Profilen greift v.a. auf die in den Dokumenten enthaltenen statistischen Metadaten zu. In einem Logfile-Profil werden zu jedem Logfile im gewählten Teilkorpus Daten zur Anzahl und Durchschnittslänge der Teilnehmerbeiträge, aber auch zur Anzahl und Verteilung von Emoticons („Smilies") und Asterisk-Ausdrücken aufgelistet.

Logfile-Profile eignen sich beispielsweise für den statistischen Vergleich von Teilkorpora, deren Mitschnitte aus Chat-Anwendungen in unterschiedlichen Zweckbereichen bzw. für unterschiedliche Diskursszenarien (z.B. ‚Freizeitkommunikation' versus ‚Moderierte Experten-/Prominentenbefragung') und mit unterschiedlich konfigurierten Chat-Umgebungen (z.B. ‚Standard-Chat-Umgebung' vs. ‚Chat-Umgebung mit technisch unterstützter Vorselektion der Teilnehmerbeiträge') stammen. Per Mausklick stellt *STACCADo* für die entsprechenden Teilkorpora die Anzahl und Länge der Beiträge, die Anzahl der enthaltenen Tokens und der im Mitschnitte dokumentierten Zeichen sowie die Gesamtzahl der enthaltenen Emoticons und Asterisk-Ausdrücken und deren Verhältnis zur Anzahl der Beiträge zusammen. Umständliches Auszählen von Hand oder mit den begrenzten Möglichkeiten der Statistik-Funktionen von Textverarbeitungsprogrammen erübrigt sich.

Im Folgenden beschreiben wir anhand zweier Beispiele – einem Logfile-Profil für ein Teilkorpus mit Plauder-Chats und einem Logfile-Profil für ein Teilkorpzus mit (technisch) moderierten Politiker-Befragungen –, welche Werte im Einzelnen in einem Logfile-Profil pro Mitschnitt spezifiziert werden und wie sie zu lesen sind.

Beispiele

A: Logfile-Profil zu einem Teilkorpus mit Plauder-Chats aus dem Chat-Angebot *unicum Space-Chat* (gekürzt)

Dateiname	TNOM	NOM Human	TNOT	NOT Human	TNOC	NOC Human	Ø-Länge Message	# Em.	Messages/ Em.	# Ast Ex.	Messages/ AstEx.
unicum_01-07-2003.xml	896	773	3692	2935	19571	15800	3,7968953	82	10,926829	161	5,5652175
unicum_19-02-2003.xml	2466	2060	10464	8588	61892	49366	4,168932	249	9,903614	458	5,3842793
unicum_03-03-2003.xml	3815	3051	17961	14359	110080	85751	4,706326	293	13,020478	642	5,9423676
unicum_21-02-2003_(1).xml	787	699	3566	2983	19703	16721	4,267525	105	7,4952383	144	5,4652777
unicum_12-02-2003.xml	1741	1437	8316	6545	47683	36854	4,554628	171	10,181287	267	6,5205994

B: Logfile-Profil zu einem Teilkorpus mit (technisch) moderierten Politiker-Chats aus dem WWW-Angebot von *politik-digital.de* (gekürzt)

Dateiname	TNOM	NOM Human	TNOT	NOT Human	TNOC	NOC Human	Ø-Länge Message	# Em.	Messages/ Em.	# Ast Ex.	Messages/ AstEx.
politik-digital_ Annette_Schavan_ Edelgard_Bulmahn_ 04-09-2002.xml	66	65	2712	2678	19919	19640	41,2	0	0	0	0
politik-digital_ Guenther_Beckstein_ 19-05-2004.xml	79	78	1923	1881	13548	13275	24,115385	1	79,0	0	0
politik-digital_ Walter_Riester_ 15-08-2002.xml	157	149	2861	2737	19936	19142	18,369127	0	0	0	0
politik-digital_ Gregor_Gysi_ Christoph_Schlingensief_ 06-09-2002.xml	105	104	2802	2766	18755	18464	26,596153	0	0	0	0
politik-digital_ Cornelia_Pieper_ 31-07-2002.xml	152	143	2192	2090	15263	14568	14,615385	0	0	0	0

Datentypen im Logfile-Profil

Die einzelnen *Spalten* der o.a. Logfile-Profile repräsentieren einzelne Datentypen, jede *Tabellenzeile* spezifiziert diese Datentypen für einen einzelnen Mitschnitt, der in der jeweils ersten Spalte über den Namen des entsprechenden Korpusdokuments identifizierbar ist.

Die Datentypen im Einzelnen:

a) Anzahl der Beiträge (*messages*), Wortformen (*tokens*) und Zeichen (*characters*)

- Die ***Total Number of Messages*** **(TNOM)** bezeichnet die Gesamtzahl aller Beiträge, die in diesem Logfile von den im Mitschnitt bezeigten Chattern produziert oder vom System generiert wurden.

  Beispiel: Im Plauder-Chat-Mitschnitt *unicum_01-07-2003.xml* sind insgesamt 896 Beiträge dokumentiert.

- Die ***Number of Human Messages*** **(NOM Human)** bezeichnet die Anzahl nur derjenigen Beiträge, die von den (menschlichen) Chattern produziert wurden. Sie berechnet sich aus der *Total Number of Messages* abzüglich der systemgenerierten Beiträge.

  Beispiel: Im Plauder-Chat-Mitschnitt *unicum_01-07-2003.xml* wurden 773 der insgesamt 896 Beiträge von menschlichen Chattern verfasst, d.h. die restlichen 123 Beiträge wurden vom Chatserver generiert.

- Die ***Total Number of Tokens*** **(TNOT)** bezeichnet die Gesamtzahl aller Wortformen, die im betreffenden Logfile enthalten sind. Der Ermittlung des Werts ist kein linguistischer Wortbegriff zugrunde gelegt; als *Wortformen* (*tokens*) werden solche Zeichenketten aufgefasst, die links und rechts durch Leer- oder Satzzeichen oder Zeilen-/Absatzwechsel begrenzt sind.

  Beispiel: Der Plauder-Chat-Mitschnitt *unicum_03-03-2003.xml* umfasst insgesamt 17.961 Tokens.

- Die ***Number of Human Tokens*** **(NOT Human)** ist die Anzahl all derjenigen Tokens, die von den Chattern (und nicht etwa vom System) wurden. Sie berechnet sich aus der *Total Number of Tokens* abzüglich der Tokens in systemgenerierten Beiträgen. Der Wert *NOT Human* kann beispielsweise dann von Interesse sein, wenn es darum gehen soll, Korpusausschnitte mit in etwa gleicher Anzahl an Sprachdaten zusammenzustellen (wenn man nicht gerade Systemmeldungen untersuchen möchte, wird man in einem solchen Fall auf eine vergleichbare Anzahl *menschlich* hervorgebrachter Tokens und nicht der insgesamt enthaltenen Tokens Wert legen).

  Beispiel: Im Plauder-Chat-Mitschnitt *unicum_12-02-2003.xml* wurden 6.545 der insgesamt 8.316 Tokens von menschlichen Chattern verfasst, d.h. die restlichen 1.771 Tokens wurden vom Chatserver generiert. Möchte man beispielsweise ausgehend von diesem Plauder-Chat-Mitschnitt einen Vergleich zwischen Plauder-Chats und (technisch) moderierten Politiker-Chats durchführen, findet man anhand der *NOT Hunman*-Werte in Logfile-Profil B (s.o.) mit den Dokumenten *politik-digital_Annette_Schavan_Edelgard_Bulmahn_04-09-2002.xml*, *politik-digital_Guenther_Beckstein_19-05-2004.xml* und *politik-digital_Cornelia_Pieper_31-07-2002.xml* schnell drei Mitschnitte, die mit ihren insgesamt 6.649 *Human Tokens* eine in etwa gleichgroße Zahl menschlich produzierter Sprachdaten beinhalten.[1]

- Die ***Total Number of Characters*** **(TNOC)** ist die Gesamtzahl aller Zeichen, aus denen die im betreffenden Mitschnitt dokumentierten Chat-Beiträge zusammengesetzt sind. Als „Zeichen" (*Characters*) werden hierbei Buchstaben, Zahlen, Leerzeichen und Sonderzeichen gewertet (also einzelne visuelle Einheiten, die durch Tastatureingabe am Bildschirm erzeugt werden können).

  Beispiel: Der Politik-Chat-Mitschnitt *politik-digital_Walter_Riester_15-08-2002.xml* umfasst insgesamt 19.936 Zeichen.

- Die ***Number of Human Characters*** **(NOC Human)** ist die Anzahl aller Zeichen, die von menschlichen Chattern eingegeben wurden. Sie berechnet sich aus der Total Number of Characters abzüglich der systemgenerierten Zeichen.

  Beispiel: Im Politik-Chat-Mitschnitt *politik-digital_Walter_Riester_15-08-2002.xml* wurden 19.142 der insgesamt 19.936 Zeichen von menschlichen Chattern verfasst, d.h. die restlichen 794 Zeichen wurden vom Chatserver generiert.

b) Durchschnittliche Beitragslänge

- Die durchschnittliche Beitragslänge in einem Logfile (**Ø-Länge Message**) ist der Quotient aus *NOT Human* und *NOM Human*. Diese Zahl kann als Referenzwert dienen, wenn Beiträge einzelner Chatter in diesem Logfile am „Durchschnittschatter" gemessen werden sollen. Vom System generierte Beiträge und Tokens fließen nicht in die Berechnung dieser Zahl ein.

  Beispiel: Die Beiträge in den (technisch) moderierten Politiker-Chats (Logfile-Profil B) sind zwischen 14,6 und 41,2 Tokens lang. Im Gegensatz dazu beträgt die durchschnittliche Beitragslänge in den Plauder-Chats (Logfile-Profil A) lediglich zwischen 3,8 und 4,7 Tokens – ein Unterschied, der vermuten lässt, dass sich das technisch-

1 Prinzipiell kann jeder Nutzer des Dortmunder Chat-Korpus, der sich die unter http://www.chatkorpus.uni-dortmund.de frei verfügbaren Korpusdokumente zusammen mit *STACCADo* auf seinen Rechner heruntergeladen hat, die Verzeichnisstrukturen, in denen die Teilkorpora und Korpusdokumente standardmäßig abgelegt sind (und die der Systematik des Dortmunder Chat-Korpus engtspfrechen) um weitere eigene Verzeichnisse erweitern, in denen er einzelne Mitschnitte aus den Teilkorpora oder Kopien davon für die Zwecke eigener empirischer Untersuchungen neu zusammenstellt. Die selbst hinzugefügten Verzeichnisse lassen sich ebenso wie die standardmäßig im Download enthaltenen Verzeichnisse in *STACCADo* als Teilkorpora auswählen, über welche Suchanfragen durchgeführt oder Statistiken (Logfile- oder Chatter-Profile) erzeugt werden sollen.

restriktive Setting der Beitragsselektion in den Politiker-Chats bei *politik-digital.de* erheblich auf den Umfang von Chatterbeiträgen auswirkt.[2]

c) Chattypische Elemente

- Die ***Anzahl der Emoticons*** in einem Korpusdokument (**# Em.**) bezeichnet die absolute Häufigkeit, mit der Emoticons in dem betreffenden Mitschnitt auftreten. Als Emoticons werden hierbei sowohl Zeichenketten (z. B. „ :-)" oder „(o:") als auch kleine Grafiken mit vergleichbarer Funktion gewertet.

  Beispiel: Im Politiker-Chat *politik-digital_Guenther_Beckstein_19-05-2004.xml* taucht insgesamt nur 1 Emoticon auf. Der Plauder-Chat-Mitschnitt *unicum_01-07-2003.xml* hingegen beinhaltet 82 Emoticons.

- Der **Quotient aus *TNOM* und *#Em.*** (**Messages/Em.**) gibt das Verhältnis von Beiträgen zu Emoticons an (*Lies:* „Durchschnittlich alle X Beiträge wird 1 Emoticon verwendet") und bezeichnet somit die Frequenz der Emoticon-Verwendung bezogen auf die Gesamtzahl der in einem Mitschnitt bezeugten Beiträge.

  Beispiel: Im Plauder-Chat-Mitschnitt *unicum_19-02-2003.xml* wurden in den insgesamt 2.466 Beiträgen 249 Emoticons verwendet, d.h. ein Emoticon tritt durchschnittlich alle 9,903614 Beiträge auf. Im Politik-Chat-Mitschnitt *politik-digital_Guenther_Beckstein_19-05-2004.xml* hingegen wurde in 79 Beiträgen nur 1 Emoticon verwendet.

- Die ***Anzahl der Asterisk-Expressions*** in einem Logfile (**# AstEx.**) ist die absolute Häufigkeit, mit der Ausdrücke in Asterisken (wie z.B. **grins**, **keuch**, **knuddel**, **tassekaffeerüberschieb**) in einem Mitschnitt bezeugt sind.

  Beispiel: Im Politiker-Chat *politik-digital_Guenther_Beckstein_19-05-2004.xml* wurden keine Asterisk-Expressions verwendet, im Plauder-Chat-Mitschnitt *unicum_03-03-2003.xml* dagegen 642 Stück.

- Der **Quotient aus *TNOM* und *#AstEx.*** (**Messages/AstEx.**) gibt das Verhältnis von Beiträgen zu Asterisk-Ausdrücken an (*Lies:* „Durchschnittlich alle X Beiträge wird 1 Asterisk-Ausdruck verwendet") und bezeichnet somit die Frequenz der Verwendung von Asterisk-Ausdrücken bezogen auf die Gesamtzahl der in einem Mitschnitt bezeugten Beiträge.

  Beispiel: Im Plauder-Chat-Mitschnitt *unicum_19-02-2003.xml* wurden in den insgesamt 2.466 Beiträgen 458 Asterisk-Ausdrücke verwendet, d.h. ein Asterisk-Ausdruck tritt durchschnittlich alle 5,3842793 Beiträge auf. Der Politiker-Chat *politik-digital_Gregor_Gysi_Christoph_Schlingensief_06-09-2002.xml* dagegen beinhaltet mit seinen insgesamt 105 Beiträgen keinen einzigen Asterisk-Ausdruck – ein Indiz dafür, dass dieses häufig als „chat-typisch" bezeichnete Stilelement zumindest in moderierten Chats mit technisch-restriktivem Setting eher die Ausnahme sind (eine Annahme, die sich anhand weiterer Analysen mit *STACCADo* über größeren Korpusausschnitten durchaus erhärten lässt).

Die tabellarisch ausgegebenen Logfile-Profile können problemlos in Tabellenkalkulationsprogramme (z.B. Microsoft Excel) übernommen und dort beliebig umsortiert, weiterverarbeitet und als Grundlage für weitergehende automatische Auswertungen oder die Erzeugung von Visualisierungen und Diagrammen genutzt werden.

2 Zum Setting von moderierten Politiker-Chats mit technisch-administrativer Vorselektion vgl. z.B. Beißwenger (2003: 220-224).