

Statistische Auswertungen mit *STACCADo* (2): Chatter-Profile

Chatter-Profile – Wer nimmt wie aktiv am Chat-Geschehen teil?

Ein „Chatter-Profil“ ist eine tabellarische Übersicht mit statistischen Daten zu allen in einem Teilkorpus oder in einem einzelnen Korpusdokument des Dortmunder Chat-Korpus bezugten Chat-Beteiligten. Chatter-Profile zu beliebigen Teilkorpora und Einzeldokumenten lassen sich mit dem Korpusabfragewerkzeug *STACCADo* automatisch erzeugen.

Die Teilkorpora des Dortmunder Chat-Korpus enthalten einzelne Korpusdokumente. 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 Chatter-Profilen greift speziell auf die in den Dokumenten enthaltenen Metadaten zum angenommenen Geschlecht und zu den statistischen Eigenschaften der Beiträge jedes einzelnen Chat-Beteiligten zu. In einem Chatter-Profil werden zu jedem Chatter im gewählten Teilkorpus oder Korpusdokument ein „estimated gender“, Daten zur Anzahl und Durchschnittslänge seiner Teilnehmerbeiträge sowie zu seinem Anteil am insgesamt von den beteiligten Chattern produzierten Beitrags- und Textaufkommen aufgelistet. Der individuelle Anteil am „Kommunikationsaufkommen“ wird hierbei sowohl als absoluter Wert als auch in Relation zum Beteiligungsanteil eines errechneten „Durchschnittschatters“ angegeben.

Logfile-Profile eignen sich beispielsweise für die Ermittlung derjenigen Chatter, die in einem Dokument oder Teilkorpus besonders aktiv oder besonders passiv in Erscheinung treten. Sie zeigen des weiteren, welche Chatter über- und welche Chatter unterdurchschnittlich lange Beiträge beisteuern und eignen sich für Untersuchungen zur Korrelation dieser Werte mit dem vermuteten Geschlecht der Beteiligten. *STACCADo* erzeugt Chatter-Profile für beliebige Teilkorpora und Dokumente per Mausklick. Umständliches Auszählen von Hand oder mit den begrenzten Möglichkeiten der Statistik-Funktionen von Textverarbeitungsprogrammen erübrigt sich.

Beispiel

Chatter-Profil für ein einzelnes Korpusdokument mit einem Plauder-Chat-Mitschnitt aus dem Chat-Angebot *unicum SpaceChat* (gekürzt)

Nickname	est. Gender	NOM	NOT	NOC	Ø-Länge Message	% Messages	% Tokens	% Characters	Länge/Ø-Chatter
marc30	male	83	341	1943	4,1084337	2,7204194	2,3748171	2,2658627	87,29598
stoeps	unknown	337	1221	6887	3,6231453	11,045559	8,503378	8,031394	76,98458
TomcatMJ	male	259	2843	19912	10,976834	8,48902	19,799429	23,22072	233,23575
Knuffine	female	80	225	1122	2,8125	2,622091	1,5669614	1,3084396	59,75999
nudelsuppenstern	unknown	46	328	1925	7,130435	1,5077024	2,2842817	2,2448719	151,50746
madcow2	unknown	45	294	1653	6,5333333	1,4749262	2,0474963	1,9276744	138,82024
Dussel	unknown	43	211	1245	4,9069767	1,409374	1,4694617	1,4518782	104,26343
Kosa	unknown	74	432	2463	5,8378377	2,4254344	3,0085661	2,8722699	124,04236
Lantonie	unknown	80	362	2082	4,525	2,622091	2,521067	2,42796	96,14719
Kool_Savas	unknown	52	172	1184	3,3076923	1,7043593	1,197855	1,3807418	70,28183
dieNachbarin	female	182	631	3537	3,467033	5,9652576	4,3944564	4,124733	73,6675
serIan	unknown	109	373	2058	3,4220183	3,5725992	2,597674	2,399972	72,71103

Nickname	est. Gender	NOM	NOT	NOC	Ø-Länge Message	% Messages	% Tokens	% Characters	Länge/Ø-Chatter
Tigerelse	female	187	614	3707	3,2834225	6,129138	4,2760634	4,322982	69,76615
quaki	unknown	78	188	1127	2,4102564	2,5565388	1,3092834	1,3142704	51,21312
Durchschnitt:		23,835938	112,17969	669,9297	4,706326				

Datentypen im Chatter-Profil

a) Vermutetes Geschlecht

- Das *estimatedGender* gibt an, ob anhand des Nicknames (eher) auf einen männlichen oder auf einen weiblichen Chatter zu schließen ist. Dass insbesondere in Plauder-Chats bisweilen sog. „gender switching“ (also spielerisches Experimentieren mit einer andersgeschlechtlichen „virtuellen Figur“) vorkommt, ist uns bewusst (wenngleich umstritten ist, wie häufig solches „gender switching“ – von anekdotischen Berichten abgesehen – tatsächlich auftritt und wie häufig ein einmal gewählter, andersgeschlechtlicher Nickname auch über einen längeren Zeitraum in einer Chat-Community beibehalten wird). Aus diesem Grunde kennzeichnen wir das als „male“ oder „female“ spezifizierte *gender* auch als lediglich „angenommen“ (*estimated*). Nicknames, die keine Rückschlüsse auf das Geschlecht des Chatters zulassen, werden mit dem Wert „unknown“ versehen. Einen Sonderfall bilden gender-Angaben in Teilkorpora, bei denen uns das Geschlecht der realen Personen, die hinter einzelnen Nicknames stehen, bekannt war oder offengelegt wurde.¹ Hier war eine verlässliche Genderzuordnung möglich.

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

- Die *Number of Messages (NOM)* bezeichnet die Gesamtzahl aller Beiträge, die ein bestimmter Chatter produziert und abgeschickt hat. Systemgenerierte Meldungen, die z. B. die An- und Abmeldung dieses Chatters anzeigen, fließen nicht in diese Zahl mit ein.

Beispiel: Die Chatterin *Tigerelse* hat im ausgewerteten Dokument 187 Beiträge produziert.

- Die *Number of Tokens (NOT)* bezeichnet die Gesamtzahl aller Wortformen, die ein bestimmter Chatter produziert und abgeschickt hat. 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: Die Beiträge der Chatterin *Tigerelse* hat im ausgewerteten Dokument umfassen insgesamt 614 Tokens.

- Die *Number of Characters (NOC)* ist die Gesamtzahl aller Zeichen, aus denen die Beiträge *eines* Chatters 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: Die Beiträge der Chatterin *Tigerelse* im ausgewerteten Dokument umfassen 3.707 Zeichen.²

1 Dies gilt beispielsweise für das Teilkorpus zum virtuellen Hoshuloseminar „Lexikographie und Wörterbuchbenutzung“.

2 Das bedeutet nicht (zwangsläufig), dass die Chatterin *Tigerelse* während ihrer Chat-Teilnahme auch tatsächlich nur 3.707 Tastaturanschläge getätigt hat. Neuere Untersuchungen zu den Teilnehmeraktivitäten beim Chatten (am Rechner bzw. vor dem Bildschirm) zeigen, dass nicht alles, was in die Textein-

c) Durchschnittliche Beitragslänge

- Die **durchschnittliche Länge der Beiträge eines Chatters (\emptyset -Länge Message)** ist der Quotient aus den für ihn ermittelten Werten *NOT* und *NOM*. Je größer dieser Wert, desto länger sind seine durchschnittlichen Beiträge.

Beispiel: Die Beiträge der Chatterin *Knuffine* sind durchschnittlich nur 2,8125 Wörter lang, während *Kosa* mit einer durchschnittlichen Beitragslänge von 5,8378377 mehr als doppelt so lange Beiträge schreibt.

d) Anteile am Beitragsaufkommen

- Der **prozentuale Anteil eines Chatters an der Gesamtmenge aller Beiträge (% Messages)** gibt an, wieviel Prozent aller abgeschickten Beiträge von diesem Chatter beigesteuert wurden. Je größer dieser Wert ist, desto mehr Beiträge hat der Chatter im Vergleich zu den übrigen Chattern verfasst. Die zur Berechnung verwendete Gesamtzahl aller Beiträge ist die Summe aller von den Chattern verfassten Beiträge, systemgenerierte Beiträge bleiben bei dieser Berechnung nicht unberücksichtigt.

Beispiel: Der Chatter *stoeps* hat im ausgewerteten Korpusdokument 11,045559% aller Beiträge verfasst – er hat damit den höchsten Redeanteil aller im Mitschnitt bezeugten Chatter.

- Der **prozentuale Anteil eines Chatters an der Gesamtmenge aller Tokens (% Tokens)** gibt an, wieviel Prozent aller von den Beteiligten realisierten Tokens vom betreffenden Chatter stammen. Je höher dieser Wert, desto mehr Tokens hat der Chatter im Vergleich zu den übrigen Chattern beigesteuert. Auch hier werden systemgenerierte Tokens nicht in die zugrundegelegte Gesamtzahl aller Tokens eingerechnet.

Beispiel: Der Chatter *TomcatMJ* hat im ausgewerteten Korpusdokument 19,799429% aller Tokens beigesteuert, obwohl er nicht die meisten Beiträge geschrieben hat (vgl. voriges Beispiel). Daraus lässt sich schließen, dass seine Beiträge durchschnittlich länger als die von *stoeps*, welcher die meisten Beiträge stellt.

- Der **prozentuale Anteil eines Chatters an der Gesamtmenge aller Zeichen (% Characters)** gibt an, wieviel Prozent aller von den Beteiligten verschickten Zeichen vom betreffenden Chatter stammen. Je höher dieser Wert ist, desto mehr Zeichen hat dieser Chatter im Vergleich zu den übrigen Chattern beigesteuert. Auch hier werden systemgenerierte Tokens nicht in die zugrundegelegte Gesamtzahl aller Tokens eingerechnet.

Beispiel: Der Chatter *TomcatMJ* hat im ausgewerteten Korpusdokument 23,22072% aller verschickten Zeichen beigesteuert – der Chatter *stoeps* hingegen nur 8,031394%. *TomcatMJ* hat also in etwa das Dreifache an Zeichen verschickt wie *stoeps*, obwohl *stoeps* insgesamt mehr Beiträge als *TomcatMJ* verfasst hat.

e) Verhältnis zum Durchschnittschatter

- Der Quotient aus Durchschnittslänge der Beiträge eines Chatters und der Durchschnittslänge aller Beiträge im Korpusteil (**Länge/ \emptyset -Chatter**) gibt an, **wie lang die Beiträge eines einzelnen Chatters im Vergleich zur durchschnittlichen Länge aller Beiträge im ausgewerteten Korpusdokument sind**. Weicht der hier angegebene Wert deutlich von 100% ab, so sind die Beiträge dieses Chatters deutlich länger bzw. kürzer als die Beiträge der anderen im Dokument bezeugten Chatter.

gabeformulare eingegeben wird, auch tatsächlich verschickt wird. Bisweilen werden Beitragsentwürfe und -fragmente auch wieder gelöscht, bedingt z.B. durch die Tatsache, dass sich der mit dem betreffenden Beitrag verfolgte Handlungsplan durch das zwischenzeitliche Eintreffen eines neuen Partnerbeitrags erledigt hat.

Beispiel: Das Verhältnis der Länge der Beiträge des Chatters *TomcatMJ* zur Durchschnittslänge beträgt 233,23575% - seine Beiträge sind somit ca. 2,3mal so lang wie die der anderen Chatter. Die Beiträge von *stoeps* hingegen sind um ca. 23% kürzer als die des errechneten „Durchschnittschatters“, obwohl *stoeps* prozentual die meisten Beiträge stellt.

Die tabellarisch ausgegebenen Chatter-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.