1 Statistics

1.1 Dialog und Followerzahl

1.1.1 Verteilung der Followerzahl

Während des untersuchten Zeitraums zwischen dem 1. April und dem 30. April 2013 waren 1.577.083 unterschiedliche Accounts auf Twitter aktiv. Dabei zeigte sich zum einen eine extreme Ungleichverteilung hinsichtlich der Follower: Die obersten 5 % der User vereinen mit 4.549.895 99,9 % der Follower auf sich, während bereits die nächste Stufe der 90 - 95% follower-stärksten User im Schnitt nur noch 1.302 Follower aufweisen. Der Median betrug 126 Follower. Die Zahlen zeigen, dass das unterste Drittel der User seine Follower aus dem persönlichen Umfeld rekrutiert. Dieser Vermutung sollte sich natürlich auch in der Art und Weise einer dialogischen Kommunikation niederschlagen. So war es für uns naheliegend, die letzten 5 Prozent der User nicht zu betrachten, da die extrem hohe Followerzahl von über 4 Mio Dialoge praktisch komplett ausschließt. Stattdessen scheint die vorletzte Kohorte mit durchschnittlich 1.302 Followern noch tatsächliche Dialoge (*real dialogs* durchzuführen.

%	Follower	%	Follower	%	Follower	%	Follower
0-5	1	25-30	37	50-55	154	75-80	406
5-10	3	30-35	55	55-60	185	80-85	528
10-15	7	35-40	77	60-65	223	85-90	750
15-20	12	40-45	100	65-70	267	90-95	1.302
20-25	22	45-50	126	70-75	325	95-100	4.549.895

Tabelle 1: Followerzahl

1.2 Dialog und Followerzahl

1.2.1 Antwortzeit in Abhängigkeit von der Followerzahl

Ein Beispiel für unsere *real dialogs* - These ist die durchschnittliche Antwortzeit, die eine getweetete Frage erzielte. Wir setzten diese Zeit mit der Followerzahl des Users in Verbindung und erhielten eine Kurve, die darauf schließen lässt, das *real dialogs* nur bis zu einer Größenordung von ca. 1.000 Followern stattfinden. Man kann erkennen, dass erwartungsgemäß die Zeit bis zur Beantwortung einer Frage bei sehr wenigen Followern hoch ist. Ab ca. 100 Followern jedoch pegelt sich die Beantwortungszeit einer Frage auf einen Wert um 80 Minuten ein. Ab einer Followerzahl von 1.000 schnellt die Zeit nach oben bzw. lässt keinen linearen Zusammenhang mit der Followerzahl erkennen. Dies verwundert, sollte doch bei einer hohen Zahl von Followern eine schnellere Beantwortung erwartet werden. Wir deuten die Zahlen als einen Hinweiß darauf, dass in diesen Größenordnungen eben kein persönlicher Kontakt des Users zu seinen Followern vorhanden ist

und die Follower einer aufgeworfenen Frage gegenüber indifferent reagieren, da einerseits ohnehin eine große Menge an Antworten vorhanden sein dürfte und sich der Follower zudem der Tatsache bewusst ist, sich eben in keinem *real dialog* zu befinden.

1.2.2 Breite des Dialogbaums in Abhängigkeit von der Followerzahl

Einen weiteren Hinweis darauf, ob es sich um einen *real dialog* handelt versprachen wir uns vom Blick auf die Breite des Dialogbaums. Unsere Überlegung war, dass der dialogische Charakter einer Twitter-Konversation durch einen „schlanken" Baum eher abgebildet wird, als durch einen „breiten", da Breite nur darauf hindeutet, dass eine Frage von vielen Followern gelesen wurde und entsprechend singulär beantwortet wird, d.h., ohne nochmals auf andere Antworten einzugehen. Der Blick auf die Daten bestätigte unsere Vermutung: Zunächst kann man ablesen, dass eine übergroße Zahl von Dialogen genau aus 2 Tweets bestehen - einem initialem Tweet und einer Antwort. Ebenso jedoch kann abgelesen werden, dass mit zunehmender Followerzahl eines Users die Dialogbäume breiter werden. Betont werden muss, dass hier nicht betrachtet wird, wie viele unterschiedliche Teilnehmer sich an einem Dialog beteiligen, sondern wie viele Follower der Dialog-Eröffner hat. Es bleibt jedoch bei der Feststellung, dass eine größere Followerzahl mit einer größeren Breite eines Dialogbaums einhergeht und somit ein Dialog gewissermaßen mäandert.

Breite des Dialogbaums	Durchschnittliche Follower	Anzahl
1	1.506	307.745
2	2.524	33.297
3	3.979	4.323
4	5.270	890
5	8.022	263
6	10.583	83
7	20.603	40

Tabelle 2: My caption

1.3 Dialoglänge und Dialogdauer

Twitter ist ein extrem schnelllebiges Medium. Beiträge anderer User können sehr schnell aus dem Blickfeld der Follower geraten. Damit läuft eine erwünschte Konversationseröffung eines Users Gefahr, nicht wahrgenommen zu werden. Diese missglückten Versuche konnten von uns natürlich nicht nachgewiesen werden. Trotzdem konnten wir die Dauer von Konversation messen, indem wir die Differenz zwischen Dialogeröffnung und dem letzten Tweet des Dialoges zum Zeitpunkt der Erstellung des Korpus maßen. Die Dialoglänge selbst ermittelten wir, indem wir die Länge des längsten Pfades des Dialogbaums

ermittelten. Man kann erkennen, dass die übergroße Zahl von Dialogen eine Lebensdauer von weniger als 3 Stunden hat und zudem kaum Dialoge stattfinden, die länger als einen Tag dauern.

Dialoglänge	Zeitdifferenz in Min	Anzahl Dialoge
2	120	170.298
3	164	83.479
4	199	39.734
5	223	22.168
6	256	11.640
7	305	7.109
8	308	4.074
9	345	2.799
10	405	1.671
11	392	1.089
12	447	674
13	410	522
14	593	373
15	552	236

Tabelle 3: My caption

1.4 Dialog und Fragen

Besonders interessiert hat uns, ob Twitter genutzt wird, um Fragen zu stellen und zu hoffen, diese beantwortet zu bekommen.

Eine zentrale Aufgabe stellte dabei die Identifizierung von Fragen dar. Wir gingen davon aus, dass aufgrund der Beschränkung auf 140 Zeichen viele Twitter-Nutzer auf die Angabe von Fragezeichen verzichten würden:

Wieso flüstern die so maaaan (ID: 318495612649762816)

Daher untersuchten wir einen Tweet zusätzlich darauf, ob er sogenannte w-Wörter (*wer, wie, was, wodurch ...*) enthielt (satzinitial sowie satzintern) und ermittelten folgende Verteilung hinsichtlich der Attribute Fragezeichen und w-Wörter:

Fragezeichen	2.374.174
Tweet mit w-Wort	588.322
Tweet mit initialem w-Wort mit Fragezeichen	92.077
Tweet mit initalem w-Wort ohne Fragezeichen	51.465

Tabelle 4: My caption

Aus dieser Verteilung kann abgelesen werden, dass der potentielle Zugewinn durch das Betrachten von Tweets mit initalen w-Wörtern und ohne Fragezeichen gering wäre, zumal eine eindeutige Klassifizierung als Frage dann immer noch nicht gegeben wäre:

Wen du endlich ewachsen bist und mich in Ruhe lässt lass es mich wissen ansonsten viel spaß noch! Wirst schon sehen! (ID: 318501739030536192)

Da es bei einer so großen Datenmenge aus Geschwindigkeitsgründen nicht möglich ist, Fragen akkurat und linguistisch korrekt anhand syntaktischer Merkmale zu identifizieren, mussten wir uns einer Näherungslösung bedienen. Jeweils 100 zufällig gewählte Tweets, die entweder als wh-question, als Tweet mit Fragezeichen oder als Nicht-Frage getagged waren, überprüften wir händisch und ermittelten folgende Ergebnisse:

	wh_question=1	question_mark=1	is_question=0
richtig	73	100	100
falsch	27	0	0

Tabelle 5: My caption

Wir mussten erkennen, dass nicht nur der Anteil potentieller Fragen, bei denen auf ein Fragezeichen verzichtet wurde, gering war, sondern gleichzeitig auch die Überprüfung hinsichtlich der Existenz von w-Wörtern eine Fehlerquote von 27 % aufwies. Gleichzeitig war in unserer Stichprobe jeder der untersuchten Tweets mit Fragezeichen tatsächlich eine Frage. Das Ergebnis dieses Test widersprach unserer Erwartung, dass auf den Einsatz von Fragezeichen verzichtet würde. Das Attribut `question_mark` stellte sich als zuverlässigstes Indiz zur Fragenerkennung heraus. Wir beschränkten uns daher auf dieses Attribut, um Tweets als Fragen zu identifizieren.

1.4.1 Länge von normalen Tweets und Fragen