Marktbasiertes Crowdsourcing in der linguistischen Feldforschung

Fabian Kneißl

Institut für Informatik
Ludwig-Maximilians-Universität
München

Crowdsourcing Workshop

22. Januar 2013

Gliederung

Linguistische Feldforschung

Die Plattform „metropolitalia“

Marktbasiertes Crowdsourcing im Detail

Ausblick

Linguistische Feldforschung

Was wird erforscht?

- Dialekte und Varietäten
- Geographische Verbreitung
- Sprechereigenschaften (Alter, Geschlecht, Bildung...)

Warum Italienisch?

- Aktuelle Ausdifferenzierung der Sprache ausgehend von den großen Städten
- Ausprägung neuer Varietäten

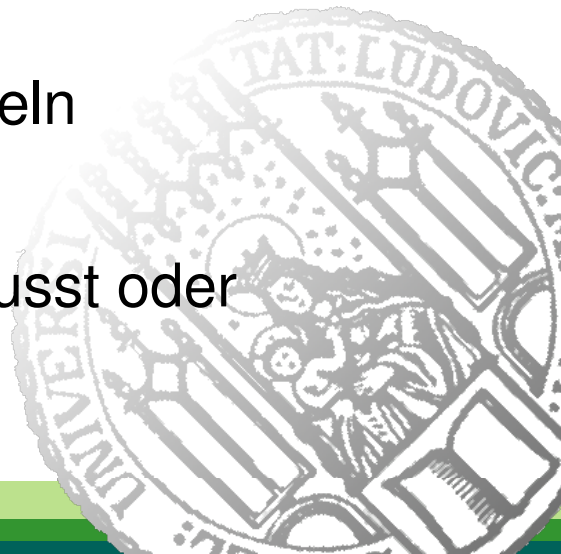
Probleme traditioneller linguistischer Feldforschung

Traditionelle linguistische Feldforschung ist

- teuer
- zeitintensiv
- möglicherweise voreingenommen

Warum?

- Forscher müssen **vor Ort** Informationen sammeln
- eine **große Menge** an Daten wird benötigt
- bei der Datenaufnahme sind Forscher oft (bewusst oder unbewusst) **voreingenommen**



Lösung dieser Probleme mit einer speziell dafür entwickelten Crowdsourcing-Plattform

Grundfunktion

Sammlung von sprachlichen Äußerungen und deren Charakteristika

Im Detail

- Nutzer-generierte Inhalte
- Punktesystem
- Suchfunktion
- Ergebnisdarstellung

Ablauf eines Spiels auf metropolitalia

http://www.metropolitalia.org

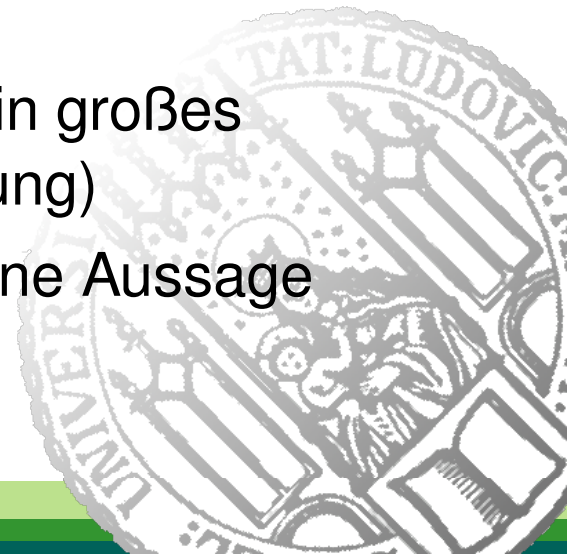
Bisher gesammelte Daten

Zeitraum: ab 1.3.2012 private Beta-Version, ab 1.8.2012 öffentlich

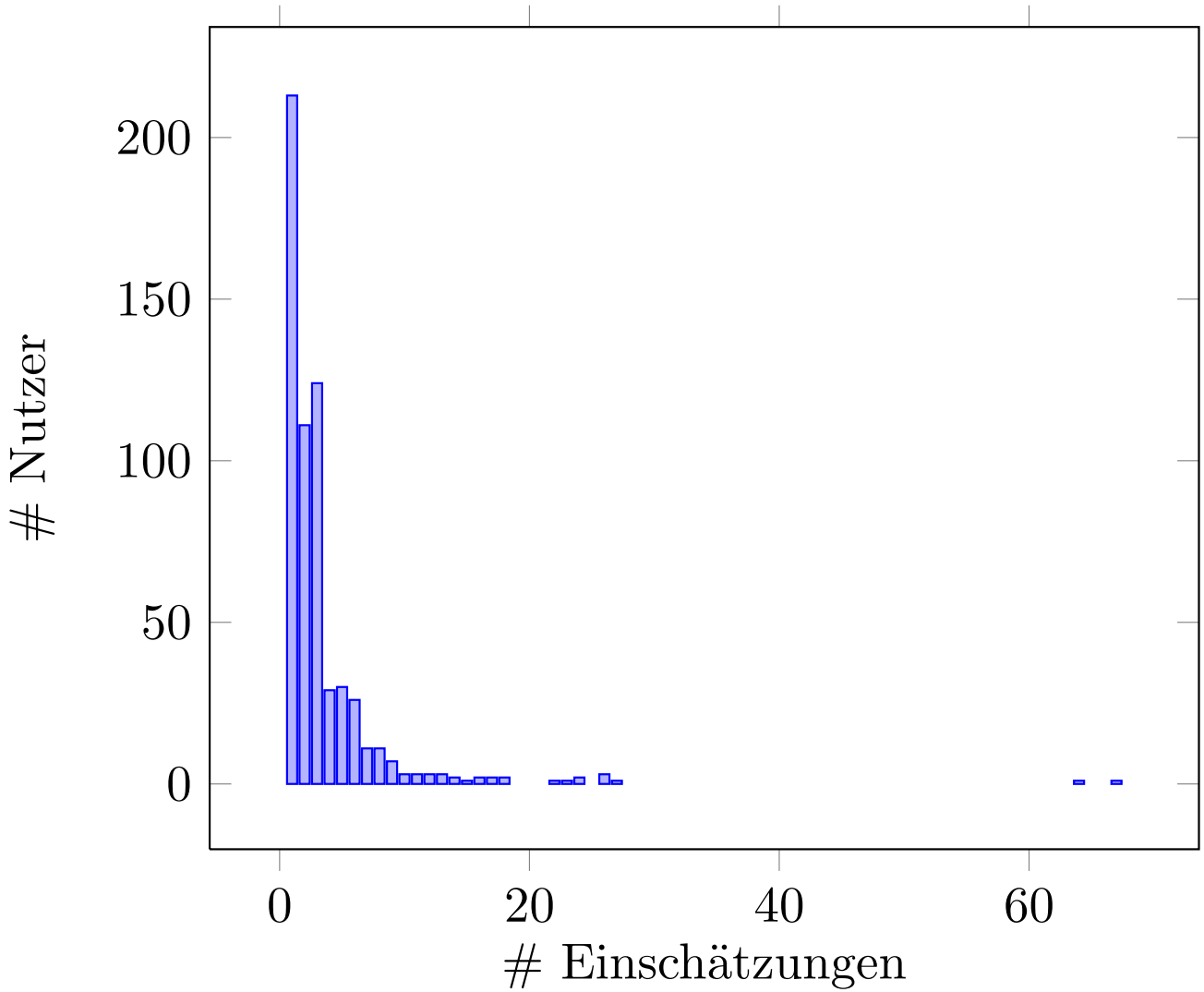
- 600 Nutzer
- 3500 Spielrunden
- 2100 geographische Einschätzungen
- 100 erstellte Aussagen

Folgerungen

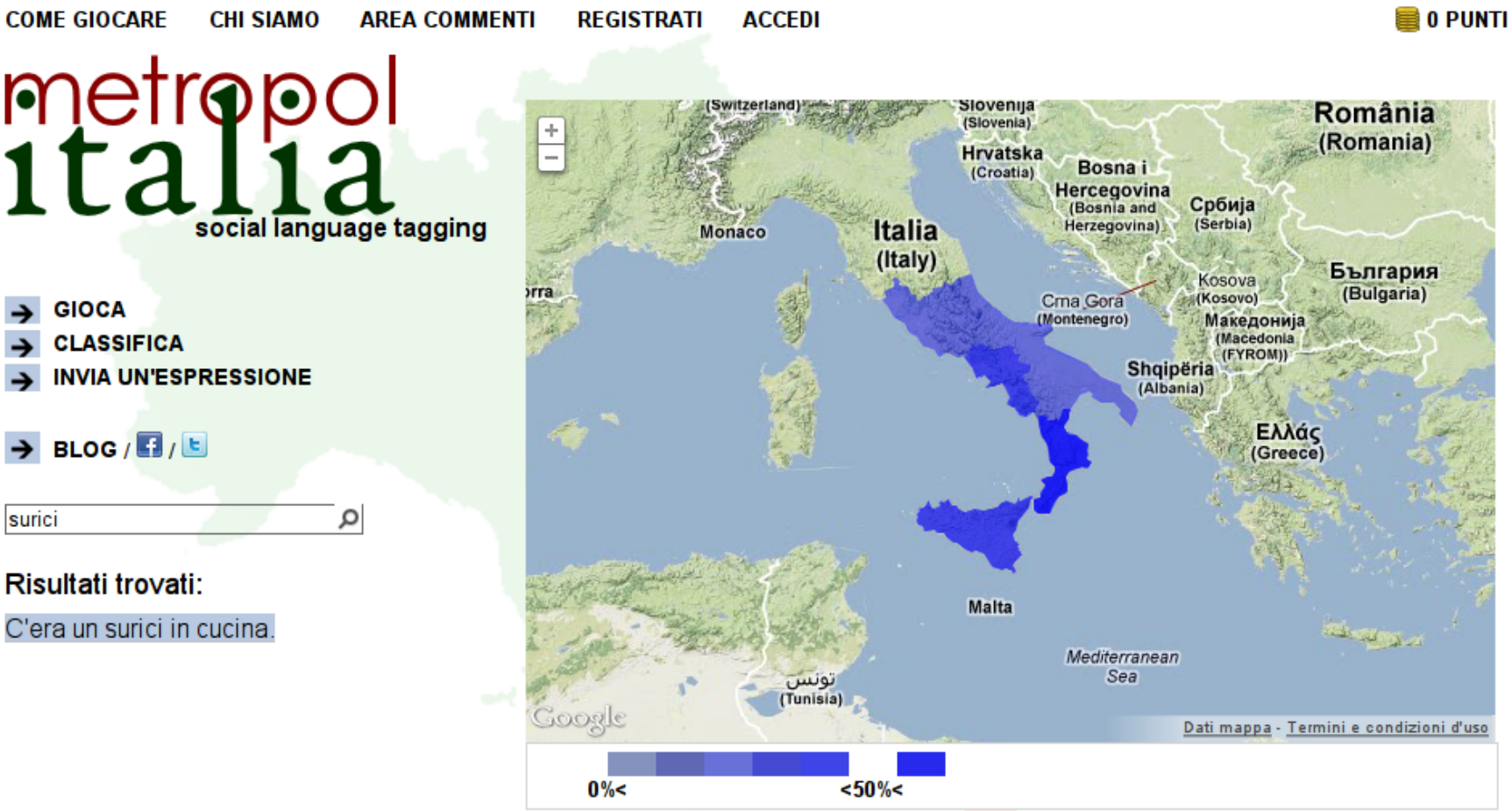
- 40% der Runden werden übersprungen
- Schätzung der Übereinstimmung für Nutzer kein großes Hindernis ($>$ 90% der Verortungen mit Schätzung)
- durchschnittlich jeder sechste Nutzer erstellt eine Aussage



Wie viele Einschätzungen geben Nutzer ab?

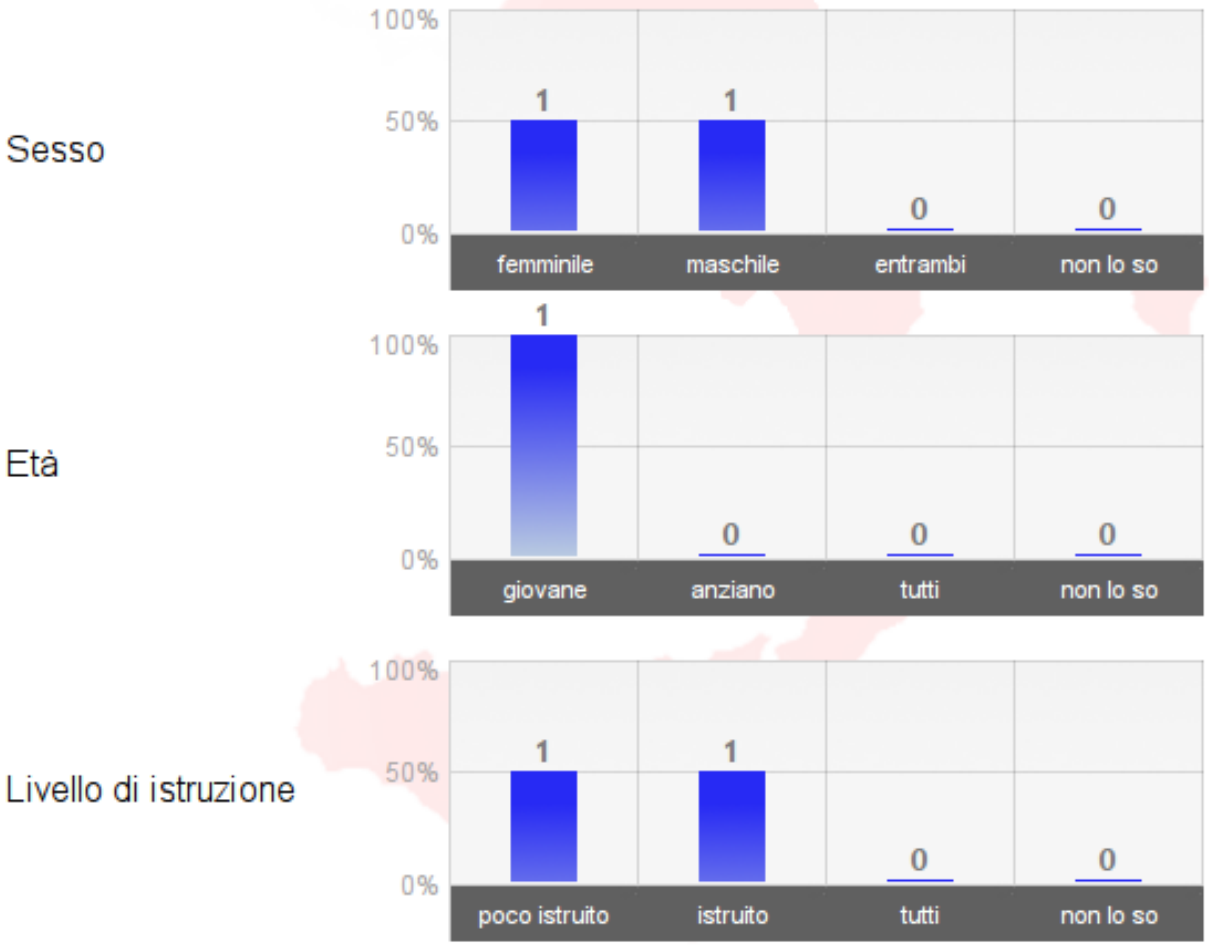


Detaillierte Daten von Aussagen



Detaillierte Daten von Aussagen

L'opinione dei giocatori:



Detaillierte Daten von Aussagen

Parole rilevanti:

C	'	era	un	surici	in	cucina	.	
0	0	0	0	5	0	0	0	di 5

3 su 7 (42%) puntano su: Sud>Calabria
2 su 7 (28%) puntano su: Sud
1 su 7 (14%) punta su: Sud>Sicilia
1 su 7 (14%) punta su: Sud>Campania

Abstrakter: Ein Markt mit symbolischen Gütern

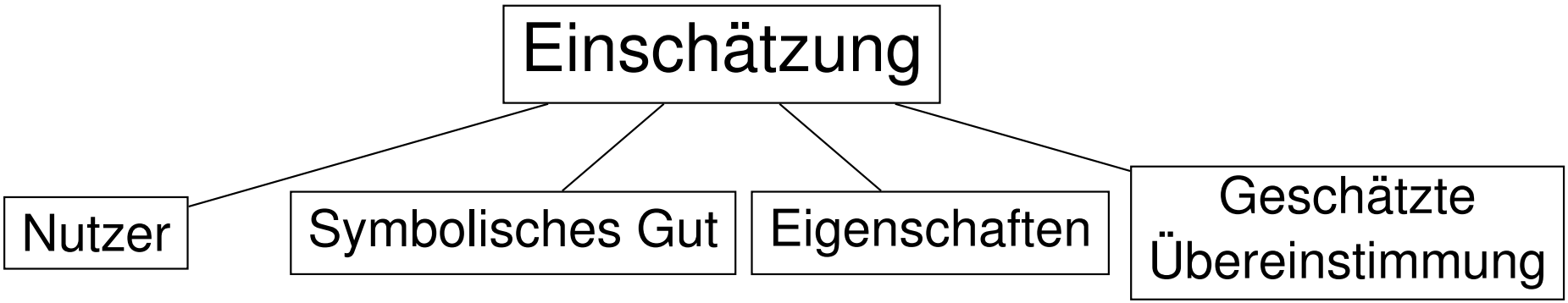
Markt-basierte soziale Software erlaubt Nutzern

- symbolische Güter zur Plattform beizutragen
- Einschätzungen zu bestimmten Eigenschaften dieser Güter abzugeben
- ihre eigenen Einschätzungen basierend auf Reaktionen im Markt anzupassen

In MetropolItalia

- symbolische Güter $\widehat{=}$ Aussagen in Varietät / Dialekt
- Eigenschaften $\widehat{=}$ geographische Einordnung, Sprechereigenschaften, charakteristische Wörter

Nutzer-Einschätzungen liefern detaillierte Informationen zu einem symbolischen Gut



Möglichkeiten:

- Einschätzungen nach Eigenschaft aggregieren
- Geschätzte Übereinstimmungen aggregieren
  $\Rightarrow$ Aussagen kann nach Grad der Nutzer-geschätzten Erkennungsrate sortiert werden
- Nutzer nach ihrem „Experten“-Grad für eine bestimmte Eigenschaft sortieren

Zusätzlich realisierbare Spielvarianten

Komplementäre Spiele

- Borsa Parole: Nutzer erhalten Punkte für mehrheitliche Übereinstimmung
  $\Rightarrow$ Sammlung weit verbreiteter Aussagen
- Poker Parole: Nutzer führen andere Nutzer bewusst in die Irre (also „bluffen“)
  $\Rightarrow$ Sammlung wenig weit verbreiteter Aussagen

Markt ausbauen

- Handel mit Aussagen

Zusammenfassung

- metropolitalia adressiert den Bedarf nach einer neuen Form von linguistischer Feldforschung
- Datenauswertung liefert optimistische Ergebnisse
- marktbasierte Spiele sammeln vielschichtige Daten

Weitere Ziele

- höhere Nutzeranzahl
- gesprochene Aussagen
- tiefgehendere Datenauswertung
- weitere marktbasierte Spiele
- …

Haben Sie Fragen?

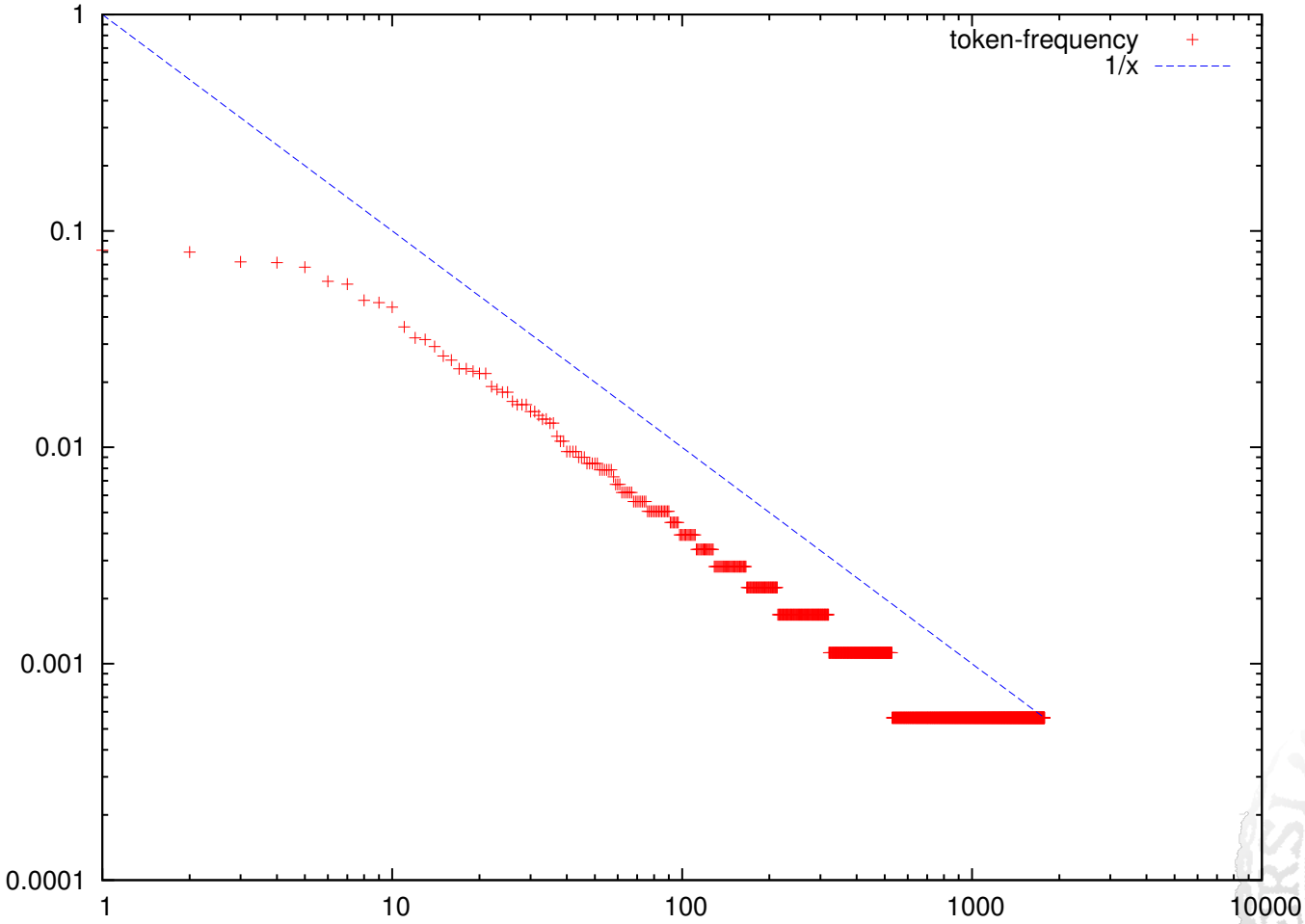
Vielen Dank für Ihre Aufmerksamkeit!

Gesammelte Daten

Zeitraum: ab 1.3.2012 private Beta-Version, ab 1.8.2012 öffentlich

- 593 (größtenteils anonyme) Nutzer
- 3446 Runden Borsa Parole
- 1896 Einschätzungen
- zusätzlich 173 Verortungen ohne Schätzung der Übereinstimmung
- 797 Sprechercharakterisierungen
- 1654 Hervorhebungen von Wörtern
- 110 erstellte Aussagen

Zipf-Verteilung der Token (log-log-Plot)



Wie viele Aussagen erstellen Nutzer?