Darja Fišer

2 RAZISKOVANJE RAČUNALNIŠKO POSREDOVANE KOMUNIKACIJE

2.1 Raziskovanje računalniško posredovane komunikacije – izročki

Opredelitev področja

- Kaj je RPK:
  - email, chat, pogovorne strani, družbena omrežja (Herring 2001)
- Koga je RPK najprej zanimal:
  - pragmatika
  - analiza diskurza
  - sociolingvistika
- Koga RPK zanima danes:
  - Podatkovna analitika
    - oglaševanje
    - podatkovno novinarstvo
  - Procesiranje šumnih besedil
    - strojno prevajanje
    - analiza sentimenta
  - Analiza vedenja uporabnikov / skupnosti
  - Variantnost
    - dialektologija
  - Analiza leksikalnih sprememb
    - ortografske spremembe
    - neologizmi
    - pomenski premiki
  - Večjezičnost
    - mešanje jezikov
    - preklapljanje med jeziki / pisavami
  - Pismenost
    - usvajanje 1. jezika
    - usvajanje tujega jezika
- Glavni problemi RPK
  - metodologija zbiranja in označevanja gradiva
  - metodologija analize gradiva

Zbiranje gradiva

- Zbiranje gradiva z interneta vse prej kot trivialno
  - velikost & reprezentativnost korpusa
  - procesiranje korpusa
  - ločevanje med žanri
  - vrsta & količina potrebnih kontekstualnih informacij
  - etična vprašanja (anonimnost & varovanje zasebnosti)
- Slaba praksa
  - majhni, ad-hoc vzorci besedil
- Problemi
  - standardizirane smernice za načrtovanje korpusov RPK
  - pomanjkanje javno dostopnih korpusov RPK (Beißwenger & Storrer 2008)

Analiza gradiva

- Popularne številne metode in pristopi tradicionalne lingvistike:
  - pragmatika
  - konverzacijska analiza
  - sociolingvistika
  - žanrska analiza
  - etnografska analiza
- Predmet raziskav:
  - raba jezikovnih sredstev za vzpostavljanje stika, vzdrževanje interakcije in gradnjo identitete v računalniških omrežjih
- Slaba praksa:
  - kritična refleksija problemov in izzivov ob uporabi tradicionalnih pristopov na nova okolja in okoliščine RPK
- Ključni elementi:
  - interakcijska koherenca, participatorni okvir, intertekstualnost, jezikovna identiteta, skupnost (Herring 2004)

Osnove RKP

SORODNI PROJEKTI

JANES

DeRiK

IJS

Projekt: Deutsches Referenzkorpus zur internetbasierten Kommunikation (DeRiK)

Kurzbeschreibung

Das Vorhaben, das in Kooperation mit dem Berliner Akademieprojekt "Digitales Wörterbuch der deutschen Sprache" durchgeführt wird, zielt auf den Aufbau eines ausgewogenen Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation (IBK). Ausgewogenheit wird in zweierlei Hinsicht angestrebt: Einerseits sollen für jedes Kalenderjahr gleich große Teilkorpora erhoben werden, andererseits sollen die einzelnen Teilkorpora einen ausgewogenen Querschnitt der jeweils populärsten IBK-Genres repräsentieren.

Zu Projektbeginn (Ende 2010) wurde an der TU Dortmund ein **Testdatenset** (112.000 Tokens) erhoben, das Sprachdaten aus diversen korpusrelevanten Genres (E-Mails, Mailinglisten, Forendiskussionen, Chats, ICQ, Kommunikation in sozialen Netzwerken, Videoplattformen, Online-Tauschbörsen, Online-Games, Wikipedia-Diskussionen etc.) und Nutzungskontexten umfasst. Dieses Datenset bildet das Testbett für die Überprüfung von Methoden zur automatischen Aufbereitung der Korpusdaten.

Im ersten Projektabschnitt (Dezember 2010 bis September 2011) wurde für die Repräsentation der korpusrelevanten IBK-Genres ein **Repräsentationsschema auf Basis der Formate der** Text Encoding Initiative (TEI) (Version TEI-P5) erarbeitet. Das Schema wurde im Oktober 2011 im Rahmen der internationalen Konferenz "Philology in the Digital Age" (*2011 Annual Conference and Members' Meeting of the TEI Consortium*) vorgestellt und ist auf der Projektseite **XML Schema for the Representation of CMC Genres in TEI** dokumentiert. Eine ausführliche Bergründung und Darstellung des Schemas wurde 2012 in einem Forschungsartikel für das *Journal of the Text Encoding Initiative (jTEI)* publiziert.

Gegenwärtig befindet sich das Projekt in der Phase der Datenerhebung. Parallel dazu werden Verfahren für die linguistische Aufbereitung der Daten und für ihre Integration in das Korpus entwickelt.

JANES

DiDi

IJS

didi

In the project *DiDi* we have analysed the linguistic strategies employed by users of social network sites (SNS). The data analysis focused on South Tyrolean users and we investigated how they communicate with each other. In regions of the German speaking area where dialect is frequently used in different communicative contexts, regional and social codes are often also used in written computer mediated communication. Another interesting but more general aspect of the new media is connected to the emerging linguistic and social practices (*new literacy*). One of the main research questions in DiDi was whether people of different age use language on SNS in a similar way or in an age-specific manner.

The purpose of the study was:

1. to record the contemporary language use of South Tyrolean German in the new media (cf. the DiDi Corpus)
2. to describe the everyday usage of language of South Tyrolean SNS users with L1 German with respect to their choice of languages and varieties as well as with respect to their usage of specific cmc phenomena.

JANES

CoMeRe

IJS

A propos de CoMeRe

Site du projet CoMeRe, corpus de communication médiée par les réseaux.

CoMeRe a pour objectif, à l'horizon 2014 ,de créer un noyau de corpus de communication médiée par les réseaux (*Computer Mediated Communication* – CMC) en français. Chaque corpus rassemblera un ensemble de conversations intervenant sur la Toile et les réseaux. Nous nous intéressons à une variété de systèmes de communication synchrone ou asynchrone, mono ou multimodaux (éventuellement) : blogues, tweets, SMS / textos, courriels , clavardage, forums, etc.

Les corpus et leurs métadonnées seront structurés suivant des formats standard : TEI (*Text Encoding Initiative*), CLARIN, OLAC. La banque de corpus sera diffusée en accès libre en 2014 sur le site Ortolang. L'assemblage des corpus se fera sur les serveurs de la MSH (Maison des Sciences de l'Homme) de Clermont-Ferrand et du Laboratoire de Recherche sur le Langage (LRL). Le travail s'effectue avec partenariat européen sur la TEI (groupe d'annotation TEI-CMC) avec relation avec l'infrastructure DARIAH. Ce noyau de corpus sera intégré au futur « Corpus de référence du français »

Les membres du projet CoMeRe appartiennent au groupe de travail « Nouvelles formes de communication » du consortium Corpus-écrits. Le projet a reçu l'appui de Corpus-écrits et de Ortolang.

JANES

What's up, Switzerland?

IJS

Language, Individuals and Ideologies in mobile messaging

What's up, Switzerland?

- Project duration: 36 months (1/1/2016 – 31/12/2018)
- Overall Research Questions:
  1. What do Swiss WhatsApp messages look like? What has changed overall between Swiss SMS and Swiss WhatsApp messages, and why (as regards linguistic structures, use of images in a broad sense, spelling, register-specific style, individualization vs. accomodation)
  2. What is said / done by the individual users and the media in/on WhatsApp messages and chats, in relation to the findings for question 1?
- Subprojects: The project consists of four subprojects:
  - Language(s) of WhatsApp: Verbal Periphrases and Argument Drop
  - Language Design in WhatsApp: Icono/Graphy
  - Individuals in WhatsApp
  - The Cultural Discourses and Social Meanings of Mobile Communication

Osnove RPK

RAZISKAVE NA KORPUSU JANES

Raba vejice v spletni slovenščini

- V čem raba vejice v USV odstopa od norme, ali se ta odstopanja bistveno razlikujejo od tradicionalnih besedil, ali na odstopanja vplivata stopnja formalnosti komuniciranja in interaktivnost medija? (Popič et al., v tisku)
  - označevanje:
    - 500 naključnih tvitov (250 T1L3 + 250 T3L3), 2 anotatorja + kuriranje
    - razvoj & testiranje tipologije
  - rezultati:
    - nestandardna raba vejice na Twitterju je vezana predvsem na skladenjsko rabo
    - kot najbolj problematični sta kategoriji manjkajoče vejice v pristavčnih strukturah (novo) in odvisnikih (univerzalno)
    - uporabniki Twitterja z rabo odvečne vejice nimajo težav (presenetljivo)

Primerjava z govorjeno slovenščino

- Ali je spletna slovenščina res zapisana govorjena slovenščine? (Zwitter & Fišer 2015)
  - analiza:
    - triangulacija ključnih prvin v pisni, govorjeni in spletni slovenščini
  - rezultati:
    - Janes sicer je med govorom in standardnim pisnim jezikom, a veliko bliže pisni slovenščini
    - podobnosti z govorjeno slovenščino:
      - os. zaim. v im. (*jaz*), gl. 2. os. v sed. (*veš*) v vlogi diskurznih označevalcev, kaz. zaim. v vlogi besed. aktualizatorjev (*un ta rdeč*)
      - izrazi interakcije (*lej*), deiktični izrazi (*tale*), izrazi nestandardne izreke/zapisa (*vidla*)
    - razlike z govorjeno slovenščino:
      - potek načrtovanja in tvorjenja besedil (ekspresivni izrazi *eee* v govoru, *lol* v spletnih žanrih)
      - prostorska oddaljenost udeležencev (specifična raba svojilnih in kazalnih zaimkov v spletnih žanrih *tale, tvoj*)

Strategije krajšanja

- Pogostost & načini krajšanja slovenskih tvitov (Goli at al., v tisku)
- analiza:
  - 800 naključnih tvitov (400 T1L1 + 400 T3L3), 2 anotatorja + kuriranje
  - razvoj & testiranje tipologije
- rezultati:
  - trend krajšanja je zelo pogost (90 % vseh tvitov)
  - v nestandardnih tvitih bistveno več krajšanja kot v standardnih
  - močno prevladujejo strategije krajšanja na ortografskem nivoju (87 %), na leksikalnem nivoju smo zabeležili približno 11,5 %, na skladenjskem pa nekaj manj kot 1,5 % odstotka vseh krajšanj
  - najpogosteje rabljena strategija krajšanja je opuščanje presledkov pri ločilih, sledi izpuščanje samoglasnikov na koncu besede, najredkeje pa sta rabljena izpusta predloga in zaimka

Kolokacije v spletni slovenščini

- Ali se raba splošnega besedišča na družbenih omrežjih v čem razlikuje od standardne slovenščine?
  - analiza:
    - nove kolokacije splošnega besedišča
    - kolokacije spletnega besedišča
  - rezultati:
    - neformalni kolokatorji (*nategovati ljudi, frej dan*)
    - aktualne tematike (*feminizacija moških, transspolna oseba, privatizacija vode*)
    - pomenski premiki (*brisanje zgodovine na računalniku*)
    - tujejezične prvine (*rimejk filma, startup podjetje*)
    - terminologija (*evklidski prostor, prva/glavna/spletna/desna stran bloga*)
    - frazeologija (*muca jezik papala*)

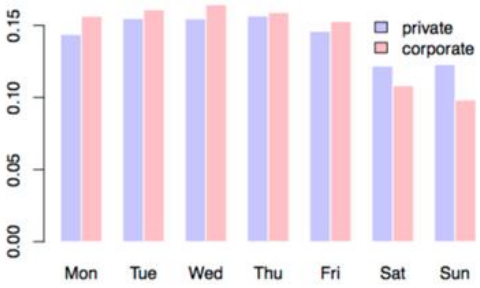
Terminologija v spletni slovenščini

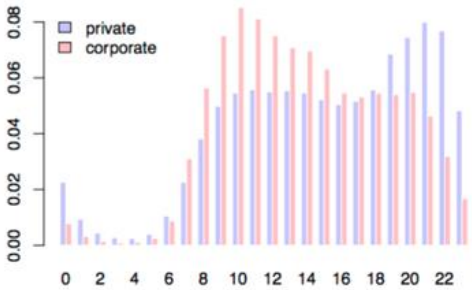
- Kakšna je raven specializiranosti, oblika terminov in stopnja nestandardnosti terminov na forumih? (Vintar 2015)
  - analiza:
    - primerjava moderiranih & nemoderiranih forumov
    - objave strokovnjakov & laikov
  - rezultati:
    - terminološka bogatost foruma ni povezana z stopnjo (ne)standardnosti izrazja (avtomobilizem izstopa tako po št. terminov kot po deležu pogovornih, žargonskih in nestandardno zapisanih izrazov, medovernet pa prav tako uporablja veliko terminologije, a z majhnimi odstopanji od standarda)
    - razpravljalci na forumu se izražajo bolj standardno, če je forum moderiran
    - strokovna področja se močno razlikujejo med seboj po terminoloških posebnostih spletne komunikacije (kratice, ki so se ustalile znotraj tega spletnega žanra: *ZM – zadnja menstruacija, G – ginekolog, KT – kontracepcijske tabletke*)

Statistična analiza tvitov glede na metapodatke

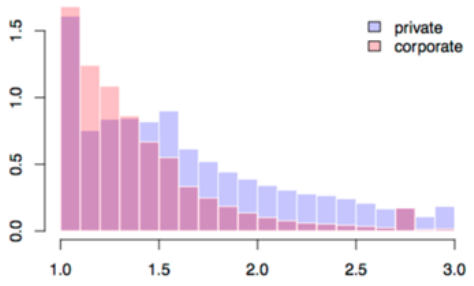
- Ali se vedenje različnih tipov uporabnikov družbenega omrežja Twitter razlikuje glede na dinamiko objavljanja, stopnjo standardnosti & sentiment sporočil? (Ljubešić & Fišer, v tisku)

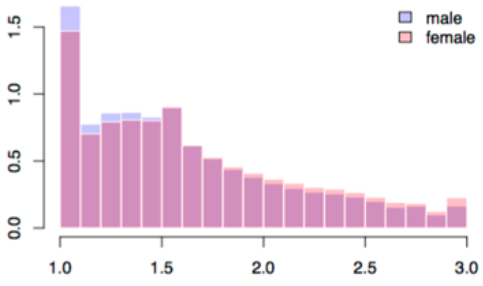
	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%

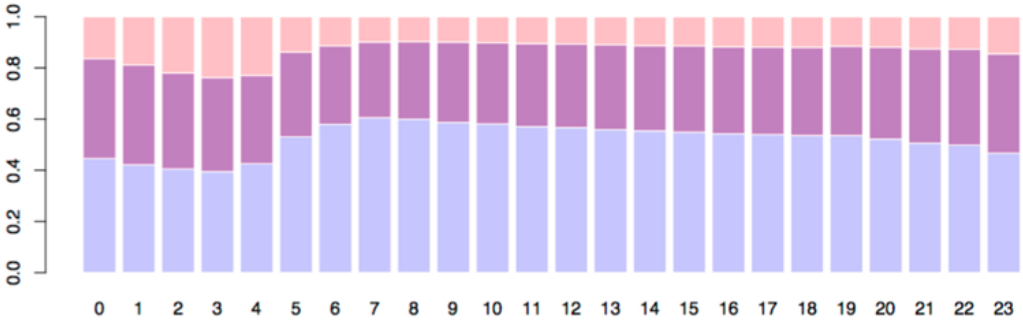




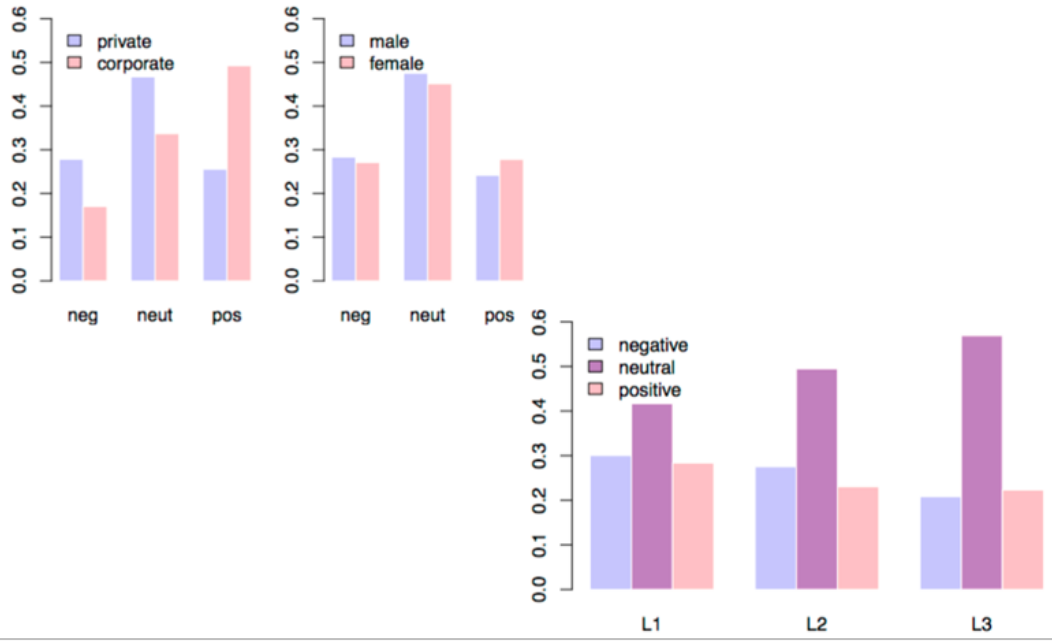
Statistična analiza tvitov glede na metapodatke







Statistična analiza tvitov glede na metapodatke



Osnove RPK

SORODNE RAZISKAVE

No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities

Cristian Danescu-Niculescu-Mizil
Stanford University
Max Planck Institute SWS
cristiand@cs.stanford.edu

Robert West
Stanford University
west@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

- spletne skupnosti so dinamične, interakcijske norme se spreminjajo
- jezikovne spremembe – inovacije, ki postanejo norma – so bistveni del tega dinamičnega procesa
- jezikovne spremembe omogočajo tako možnost individualnega izražanja kot tudi vzpostavljanje kolektivne identitete
- dovzetnost za jezikovne spremembe v spletnih skupnosti ima dve fazi:
  - v fazi učenja uporabniki na jezikovno inovativen način prevzemajo jezik skupnosti
  - v konzervativni fazi uporabniki svoj jezik nehajo spreminjati in razvijajoče se norme v skupnosti jih prehitijo

No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities

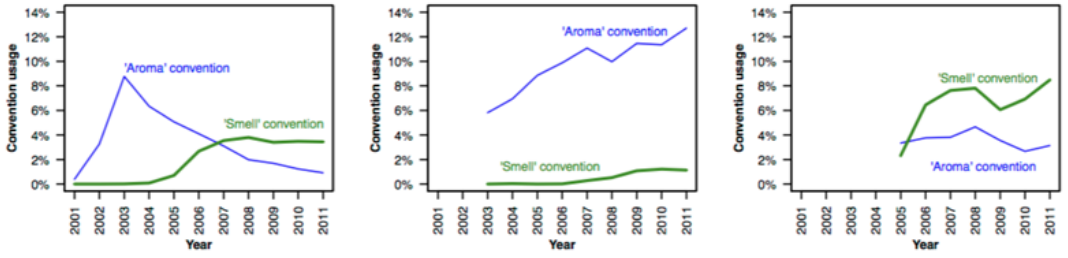
Cristian Danescu-Niculescu-Mizil
Stanford University
Max Planck Institute SWS
cristiand@cs.stanford.edu

Robert West
Stanford University
west@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu



(a) 'Aroma' was the dominant convention by 2003, but it was supplanted by 'S' (for 'Smell') around 2007.

(b) Users who joined in 2003 hung on to the 'Aroma' convention of their youth.

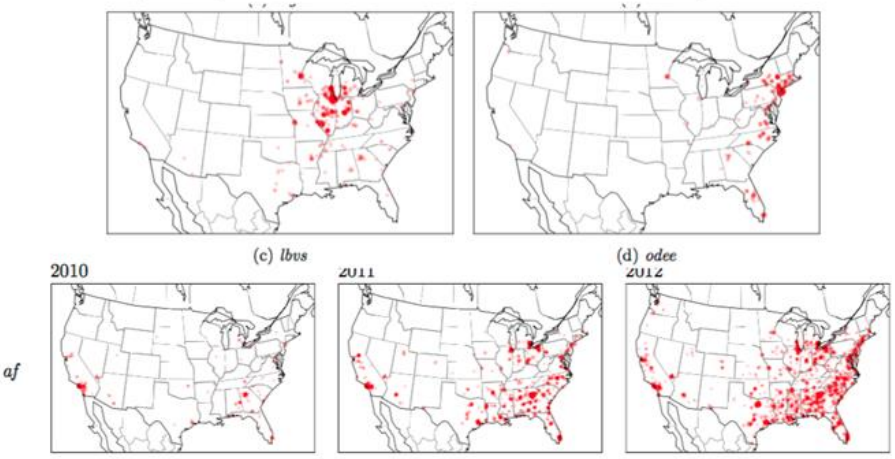
(c) Users who joined in 2005 were more receptive to the emerging 'S' norm.

Identifying regional dialects in online social media

Jacob Eisenstein

August 6, 2014

- avtomatsko odkrivanje značilnih leksikalnih variabel pri uporabnikih družbenega omrežja Twitter iz istih geografskih območij
- razširitev pristopa na diahrono analizo leksikalnih sprememb



Hateful Symbols or Hateful People?
Predictive Features for Hate Speech Detection on Twitter

Zeerak Waseem
University of Copenhagen
Copenhagen, Denmark
csp265@alumni.ku.dk

Dirk Hovy
University of Copenhagen
Copenhagen, Denmark
dirk.hovy@hum.ku.dk

- ročna anotacija korpusa 17.000 rasističnih, seksističnih ali nevtralnih tvitov
- analiza korpusa in iskanje značilk za avtomatsko prepoznavanje sovražnega govora
- najbolje delujejo značilke na nivoju znakov
- sociodemografski podatki (regija uporabnika) in kompleksnejše mere (dolžina besede), ne izboljšajo prepoznavanja, edina izjema je spol

	All	Racism	Sexism	Neither
Men	50.08%	33.33%	50.24%	50.92%
Women	02.26%	0.00 %	02.28%	01.74%
Unidentified	47.64%	66.66%	47.47%	47.32%

JANES

These 6 charts show how much sexism Hillary Clinton faces on Twitter

IJS

By Rebekah Tromble and Dirk Hovy February 24

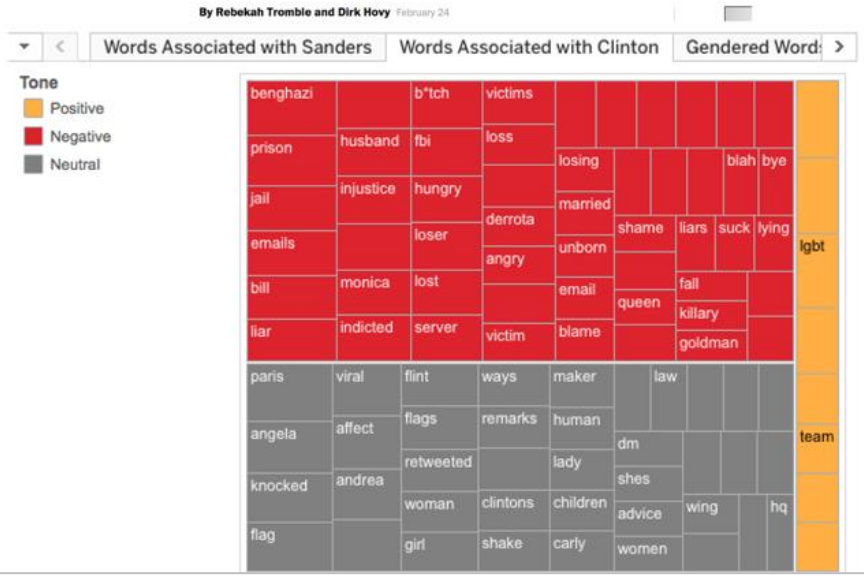
- analiza 100,000 tvitov z omembo @HillaryClinton in/ali @BernieSanders
- statistična analiza vseh besed, ki se pojavijo z njunimi imeni vsaj 50x
- podrobna analiza 100 najmočneje povezanih besed z vsakim kandidatom v naključno izbranih tvitih

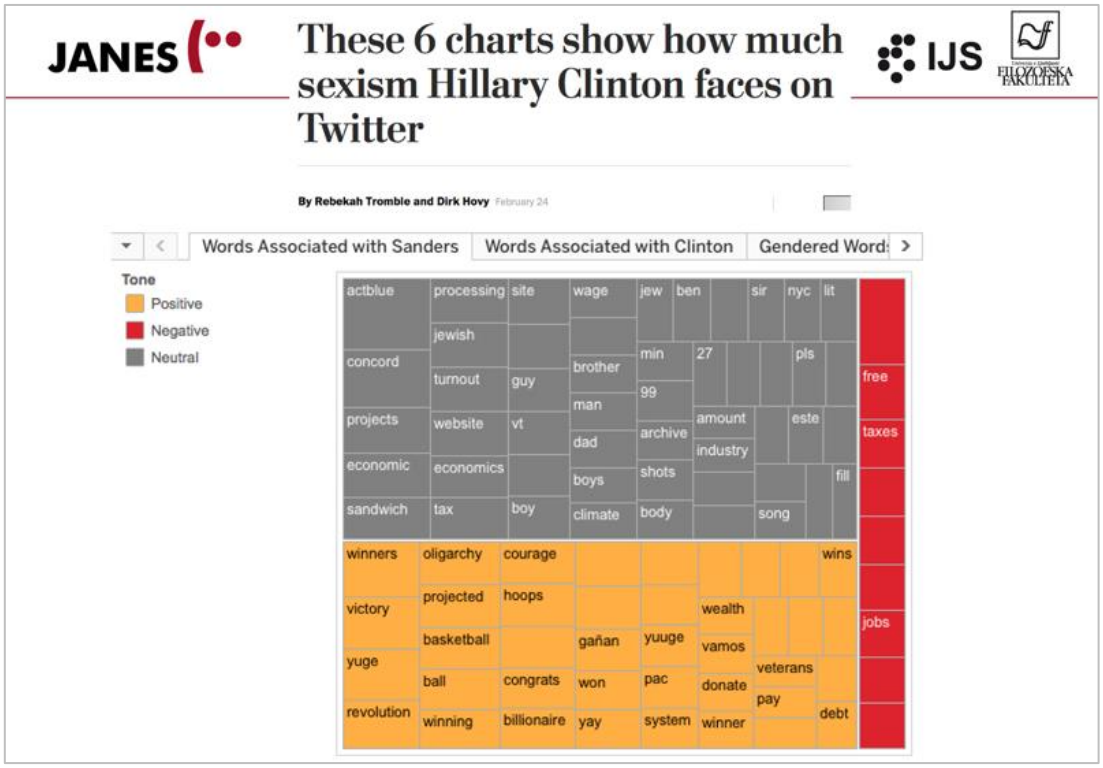
JANES

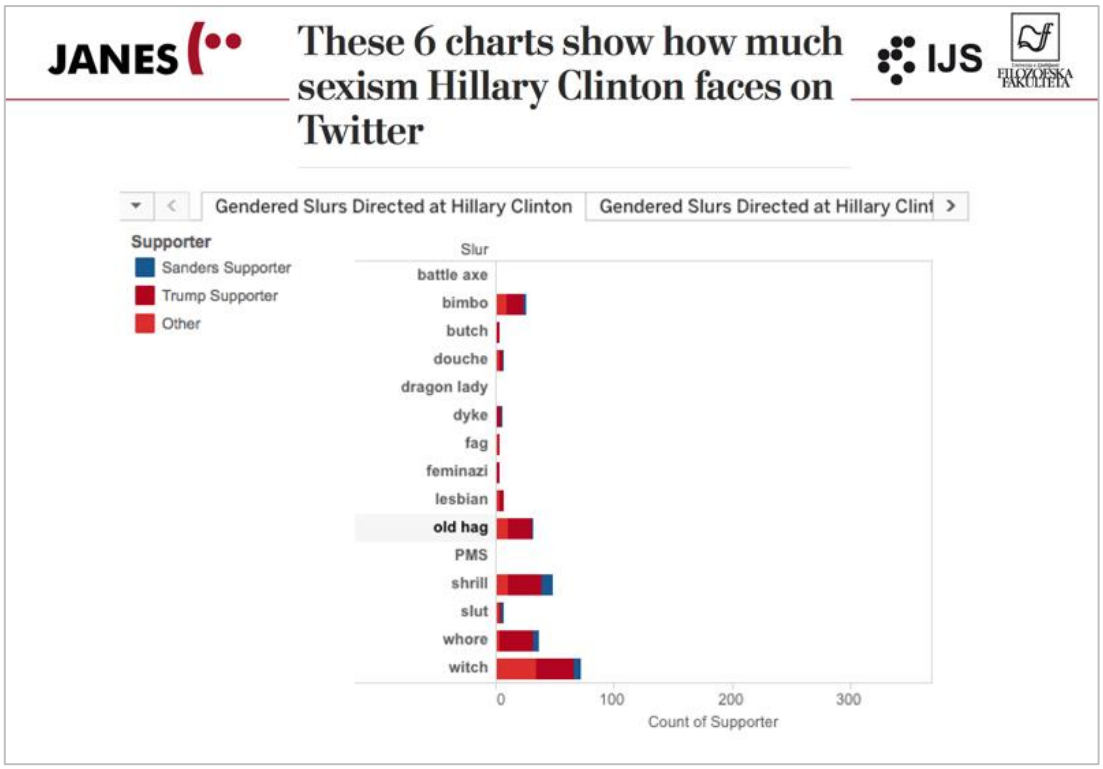
These 6 charts show how much sexism Hillary Clinton faces on Twitter

IJS

By Rebekah Tromble and Dirk Hovy February 24




JANES
These 6 charts show how much sexism Hillary Clinton faces on Twitter
IJS
By Rebekah Tromble and Dirk Hovy February 24
Words Associated with Sanders
Words Associated with Clinton
Gendered Words
Tone
Positive
Negative
Neutral
actblue
processing
site
wage
jew
ben
sir
nyc
lit
jewish
concord
turnout
guy
brother
min
27
pls
free
99
man
projects
website
vt
amount
este
taxes
archive
dad
industry
economic
economics
shots
fill
boys
sandwich
tax
boy
climate
body
song
winners
oligarchy
courage
wins
victory
projected
hoops
wealth
basketball
gañan
yuuge
vamos
jobs
yuge
ball
congrats
won
pac
donate
veterans
pay
revolution
winning
billionaire
yay
system
winner
debt

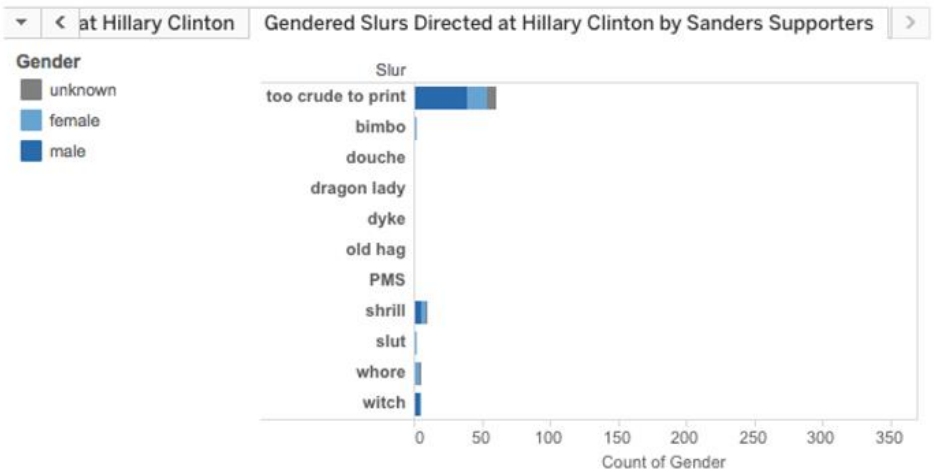

JANES
These 6 charts show how much sexism Hillary Clinton faces on Twitter
IJS
Gendered Slurs Directed at Hillary Clinton
Gendered Slurs Directed at Hillary Clint
Supporter
Sanders Supporter
Trump Supporter
Other
Slur
battle axe
bimbo
butch
douche
dragon lady
dyke
fag
feminazi
lesbian
old hag
PMS
shrill
slut
whore
witch
0
100
200
300
Count of Supporter

JANES

These 6 charts show how much sexism Hillary Clinton faces on Twitter

IJS

By Rebekah Tromble and Dirk Hovy February 24



JANES

Raba emojijev po svetu (Ljubešič & Fišer, 2016)

IJS

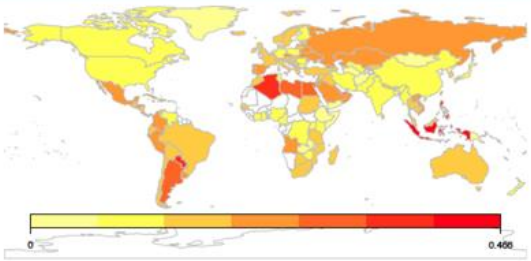


Figure 1: Percentage of tweets containing emojis

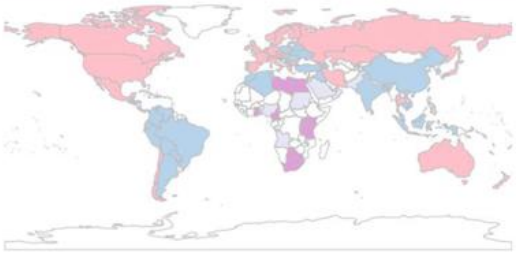


Figure 2: Clustering countries as emoji probability distributions

cluster	most distinctive emojis
"first world"	
"second world"	
"third world"	
"fourth world"	

correlation	Life expectancy at birth	Total tax rate	Trade in services	GDP per capita
positive				
negative				