NLP for Social Media

Lecture 7: Sociolinguistics & Language-usage based Studies

Monojit Choudhury

Microsoft Research Lab, monojitc@microsoft.com

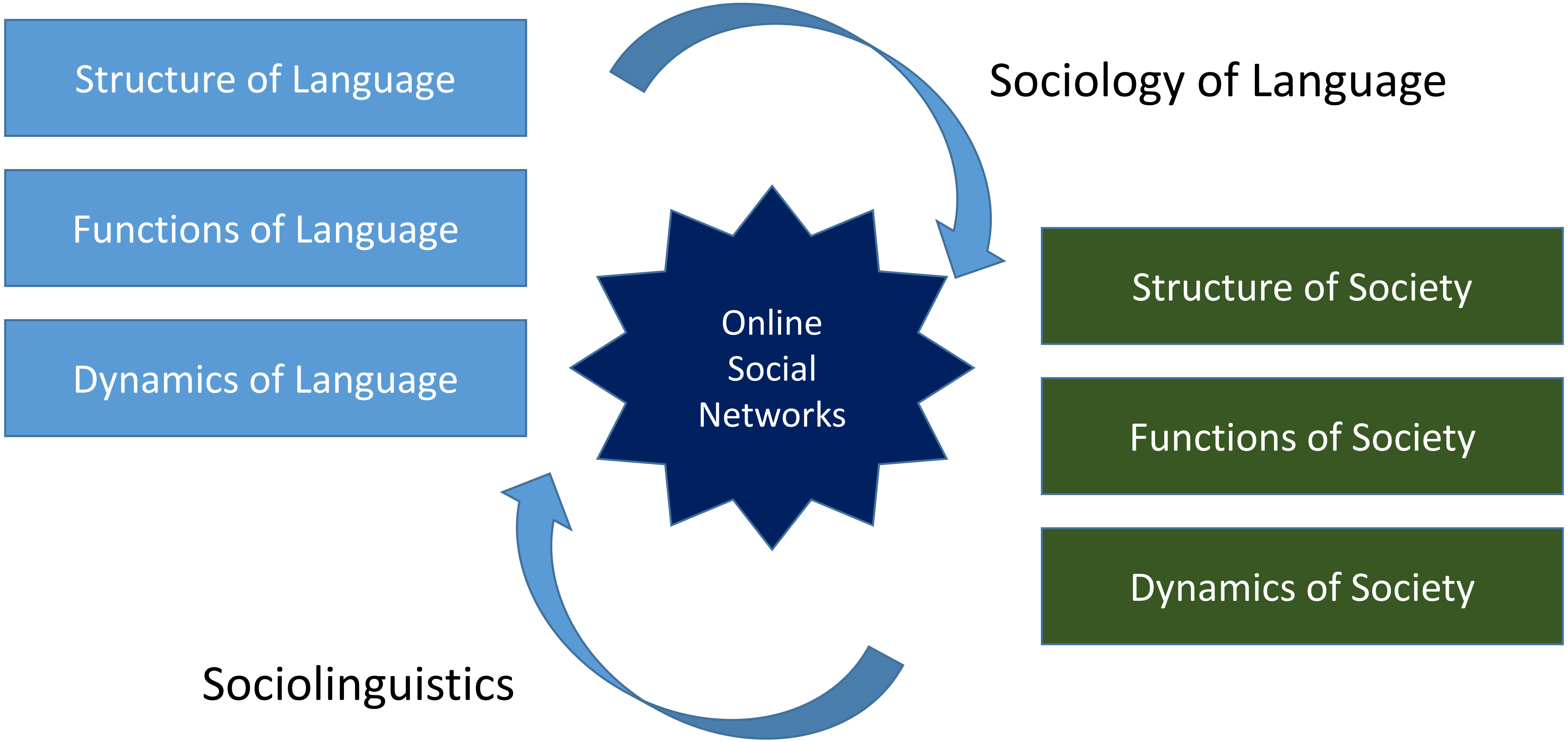
Language, individual and the society

Structure of Language

Functions of Language

Dynamics of Language

Interaction between Language & Society



Structure of Language

Functions of Language

Dynamics of Language

Sociology of Language

Online Social Networks

Structure of Society

Functions of Society

Dynamics of Society

Sociolinguistics

From an Individual’s perspective (node)

- Can we use NLP to predict individual’s
    - Moods and Mental state
    - Habits and Behavior
    - Demographic attributes – gender, ethnicity, region and language, education
    - Health: Mental, physical
    - Language acquisition

From a Relationship’s perspective (edge)

- Can we use NLP to predict
  - Dominance
  - Formality
  - Politeness
  - Threats, humiliation, stalking
  - Accommodation

From a group’s perspective (community)

- Dominance hierarchy
- Dialectal features (slangs, lingos)
- Homogeneity vs. language use
- Inclusivity
- New member dynamics
- Social ostracizing and outcasting

From Society’s perspective (whole network)

- Language evolution
  - Diffusion of linguistic innovation
  - Effect of Social influence on language change
- Prevalence of certain traits: *smoking, depression* or *swearing*
- Correlation between traits and demographic factors

Benefits & Caveats

- Large scale studies
- Effortless data collection
- Speech transcriptions not needed
- Automatic methods applicable (and necessary!)

- Caveats:
  - Potentially biased sample
  - No ways to generalize to non-OSN users
  - Representative of real linguistic data & communication?

Privacy Issues and Ethics

Example 1: Individual

Predicting Depression via Social Media. M De Choudhury, M Gamon, S Counts, E Horvitz. ICWSM 2013

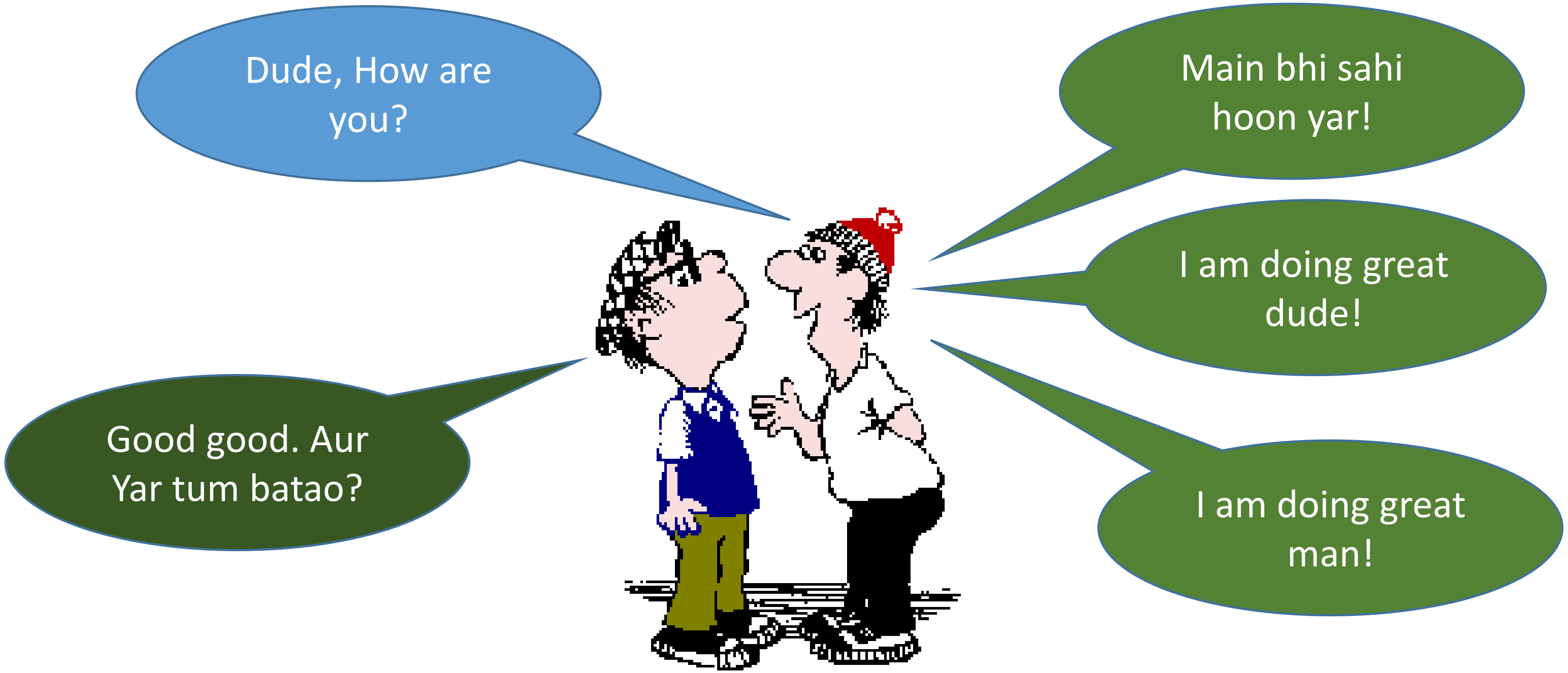
- Crowdsourcing to compile a set of Twitter users who report being diagnosed with clinical depression, based on a standard psychometric instrument.
- Through their social media postings over a year preceding the onset of depression, measure behavioral attributes relating to social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications.
- Leverage these behavioral cues to build a statistical classifier

“decrease in social activity, raised negative affect, highly clustered ego networks, heightened relational and medicinal concerns, and greater expression of religious involvement.

Example 2: Relation (and Group)

Mark my words! Linguistic style accommodation in social media

C Danescu-Niculescu-Mizil, M Gamon, S Dumais Proceedings of WWW 2011, 745-754

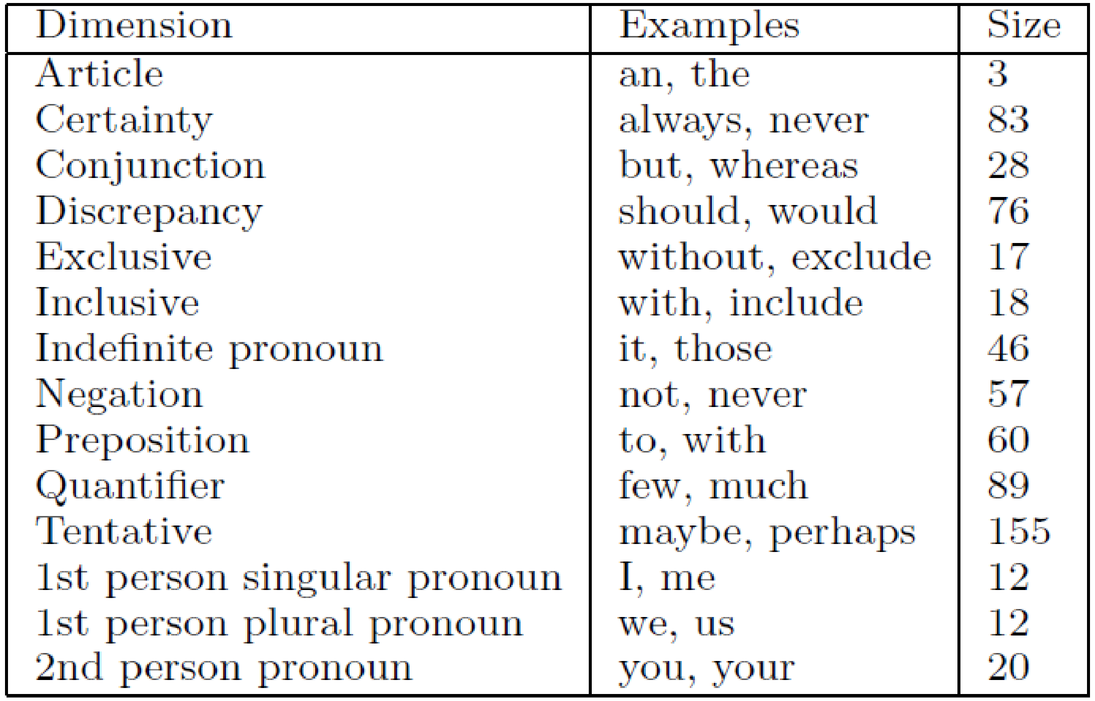


Symmetric Accommodation

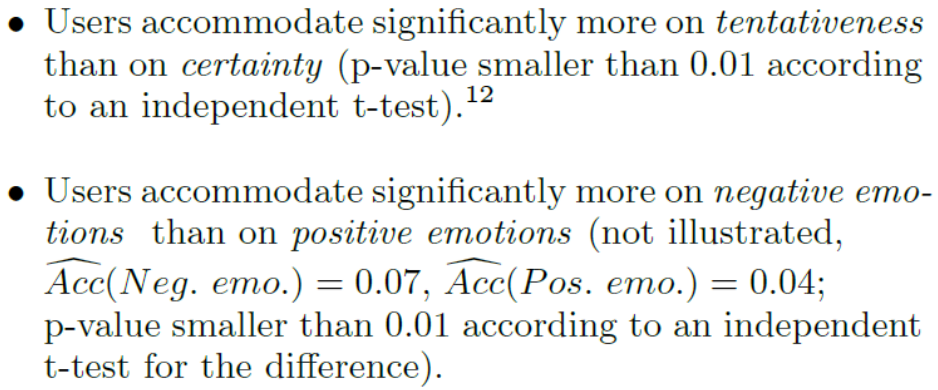
Default Asymmetry

Divergent Asymmetry

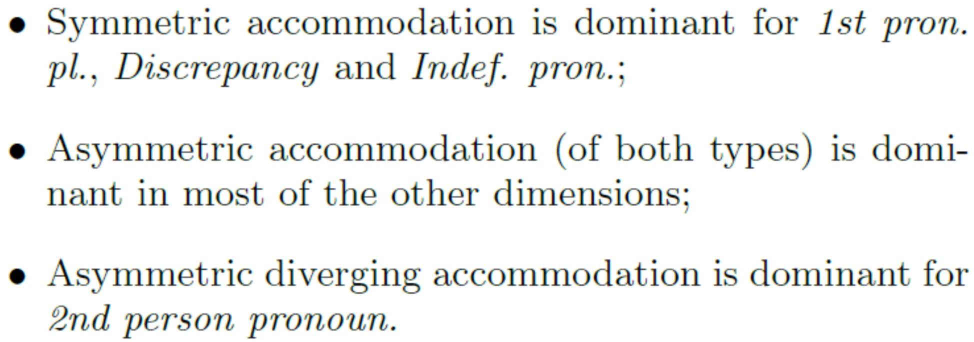
Example 2:



Dimension	Examples	Size
Article	an, the	3
Certainty	always, never	83
Conjunction	but, whereas	28
Discrepancy	should, would	76
Exclusive	without, exclude	17
Inclusive	with, include	18
Indefinite pronoun	it, those	46
Negation	not, never	57
Preposition	to, with	60
Quantifier	few, much	89
Tentative	maybe, perhaps	155
1st person singular pronoun	I, me	12
1st person plural pronoun	we, us	12
2nd person pronoun	you, your	20



- Users accommodate significantly more on *tentativeness* than on *certainty* (p-value smaller than 0.01 according to an independent t-test).[12]
- Users accommodate significantly more on *negative emotions* than on *positive emotions* (not illustrated, $\widehat{Acc}(Neg.\ emo.) = 0.07$, $\widehat{Acc}(Pos.\ emo.) = 0.04$; p-value smaller than 0.01 according to an independent t-test for the difference).



- Symmetric accommodation is dominant for *1st pron. pl.*, *Discrepancy* and *Indef. pron.*;
- Asymmetric accommodation (of both types) is dominant in most of the other dimensions;
- Asymmetric diverging accommodation is dominant for *2nd person pronoun.*

Example 3: Society

Cursing in English

Wang et al. CSCW 2014

#Bieber + #Blast = #BieberBlast: Early Prediction of Popular Hashtag Compounds,

S. K. Maiti et al. CSCW 2016

Now its your turn ☺

- Form groups of 8 to 10 students (based on physical proximity)
- Task: Come up with a research study idea (more details in next slide)
- Time: 20 min
- Each team present your idea (2 min per team) and receive feedback:
    - What is the objective of the study
    - Why is it interesting or useful
    - Why it’s challenging w/o social media
    - What data to be used
- We select the top 1 or 2 ideas (depending on votes and time) and develop the research strategy for those.

What kind of idea?

- Broad Objective: Use of language data from social media for a socio-linguistics study or language-based prediction of certain useful trait
- Desirable:
  - The study would require huge data collection and therefore, hard to run in real world w/o social media
  - Good use of NLP, but not so hard that current techniques fail to solve.
  - Use of social network properties
  - Of some practical use or interest ☺

References

- Predicting Depression via Social Media.M De Choudhury, M Gamon, S Counts, E Horvitz. ICWSM
- Predicting postpartum changes in emotion and behavior via social mediaM De Choudhury, S Counts, E Horvitz. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems
- What makes conversations interesting?: themes, participants and consequences of conversations in online social mediaM De Choudhury, H Sundaram, A John, DD Seligmann Proceedings of the 18th international conference on World wide web, 331-340
- Not all moods re created equal! a exploring human emotional states in social mediaMDCS Counts, M Gamon
- Major life changes and behavioral markers in social media: case of childbirthM De Choudhury, S Counts, E Horvitz
- Proceedings of the 2013 conference on Computer supported cooperative work

References

- Social media as a measurement tool of depression in populationsM De Choudhury, S Counts, E Horvitz. Proceedings of the 5th Annual ACM Web Science Conference, 47-5
- Social Media for Mental Illness Risk Assessment, Prevention and SupportM De Choudhury. Proceedings of the 1st ACM Workshop on Social Media World Sensors, 1-1
- Facebook Use and Disordered Eating in College-Aged WomenM Walker, L Thornton, M De Choudhury, J Teevan, CM Bulik, Journal of Adolescent Health 57 (2), 157-163
- Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity SuicidesM Kumar, M Dredze, G Coppersmith, M De Choudhury
- Proceedings of the 26th ACM Conference on Hypertext & Social Media, 85-94
- Characterizing Smoking and Drinking Abstinence from Social MediaA Tamersoy, M De Choudhury, DH Chau Proceedings of the 26th ACM Conference on Hypertext & Social Media, 139-148

References

- Echoes of power: Language effects and power differences in social interactionC Danescu-Niculescu-Mizil, L Lee, B Pang, J Kleinberg Proceedings of WWW 2012
- Mark my words! Linguistic style accommodation in social mediaC Danescu-Niculescu-Mizil, M Gamon, S Dumais Proceedings of WWW 2011, 745-754
- No country for old members: User lifecycle and linguistic change in online communitiesC Danescu-Niculescu-Mizil, R West, D Jurafsky, J Leskovec, C Potts Proceedings of WWW
- A computational approach to politeness with application to social factorsC Danescu-Niculescu-Mizil, M Sudhof, D Jurafsky, J Leskovec, C Potts Proceedings of ACL 2013
- How to Ask for a Favor: A Case Study on the Success of Altruistic RequestsT Althoff, C Danescu-Niculescu-Mizil, D Jurafsky Proceedings of ICWSM

References

- #Bieber + #Blast = #BieberBlast: Early Prediction of Popular Hashtag Compounds, S. K. Maiti et al. CSCW 2016
- Out of Vocabulary Words Decrease, Running Texts Prevail and Hashtags Coalesce: Twitter as an Evolving Sociolinguistic System. S K Maiti et al., HICSS 2016
- Cursing in English, Wang et al. CSCW 2014

Ideas

- Gender Diversity and Inclusion in different professions
- Internet.org
- Lingos for community Detection
- Product interesting or boring?
- Media’s linguistic style and topic and its effect on user’s opinion
- Phrases and reactions in media