Seminar Computational Sociolinguistics (CSL) — Part 1

Introduction to Computational Sociolinguistics

Henning Wachsmuth

henningw@upb.de

PADERBORN UNIVERSITY

Outline

- **Motivation**

- **Computational sociolinguistics (CSL)**

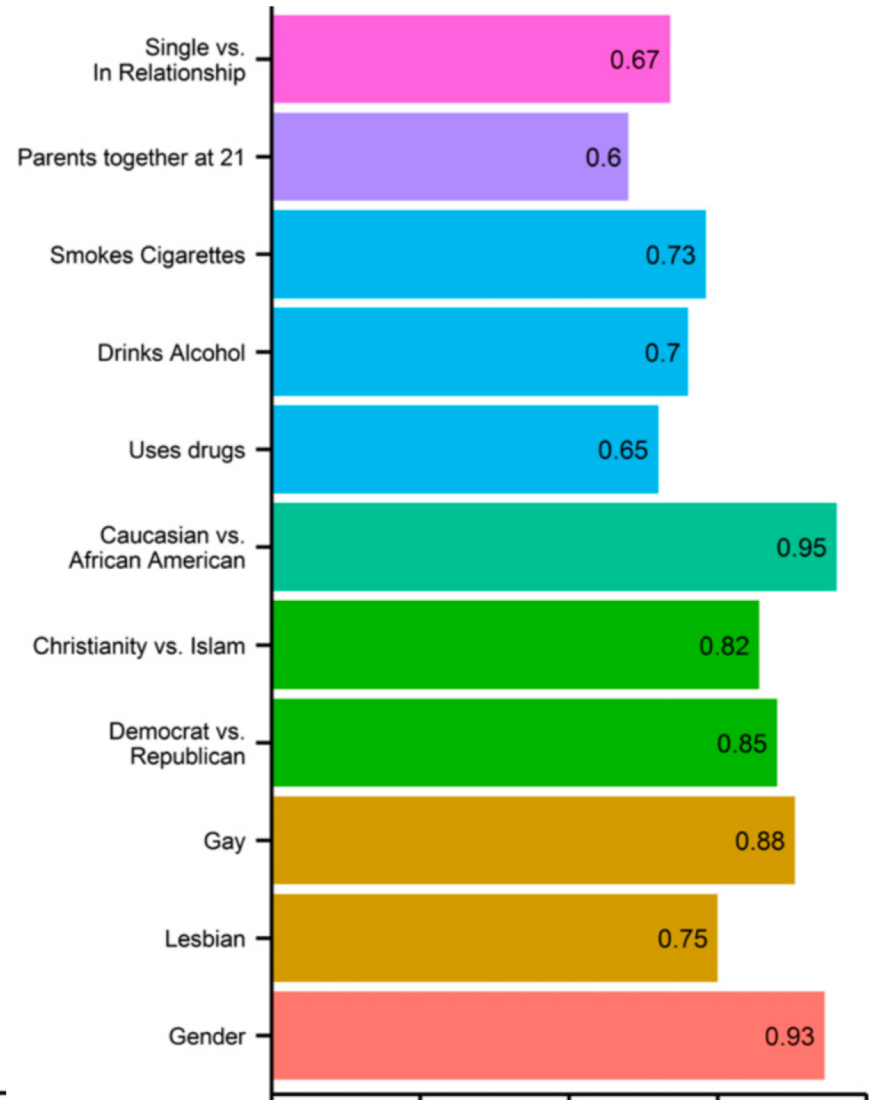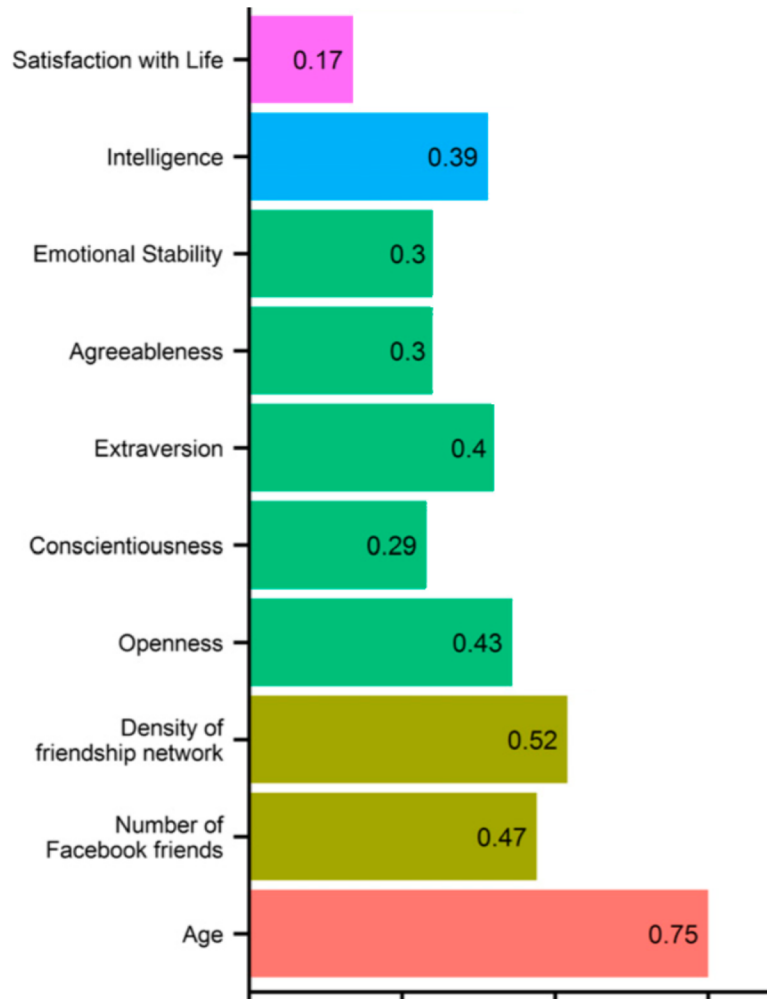- **CSL research of the CSS group**

- **CSL in this seminar**

Motivation

# Example: Predictiveness of Facebook likes (Kosinski et al., 2013)

- **What your Facebook likes reveal**

Attribute	Value
Satisfaction with Life	0.17
Intelligence	0.39
Emotional Stability	0.3
Agreeableness	0.3
Extraversion	0.4
Conscientiousness	0.29
Openness	0.43
Density of friendship network	0.52
Number of Facebook friends	0.47
Age	0.75

Attribute	Value
Single vs. In Relationship	0.67
Parents together at 21	0.6
Smokes Cigarettes	0.73
Drinks Alcohol	0.7
Uses drugs	0.65
Caucasian vs. African American	0.95
Christianity vs. Islam	0.82
Democrat vs. Republican	0.85
Gay	0.88
Lesbian	0.75
Gender	0.93

Example: Politicians' Twitter practices (Lietz et al., 2014)

- **Mentions of German politicians by other politicians**
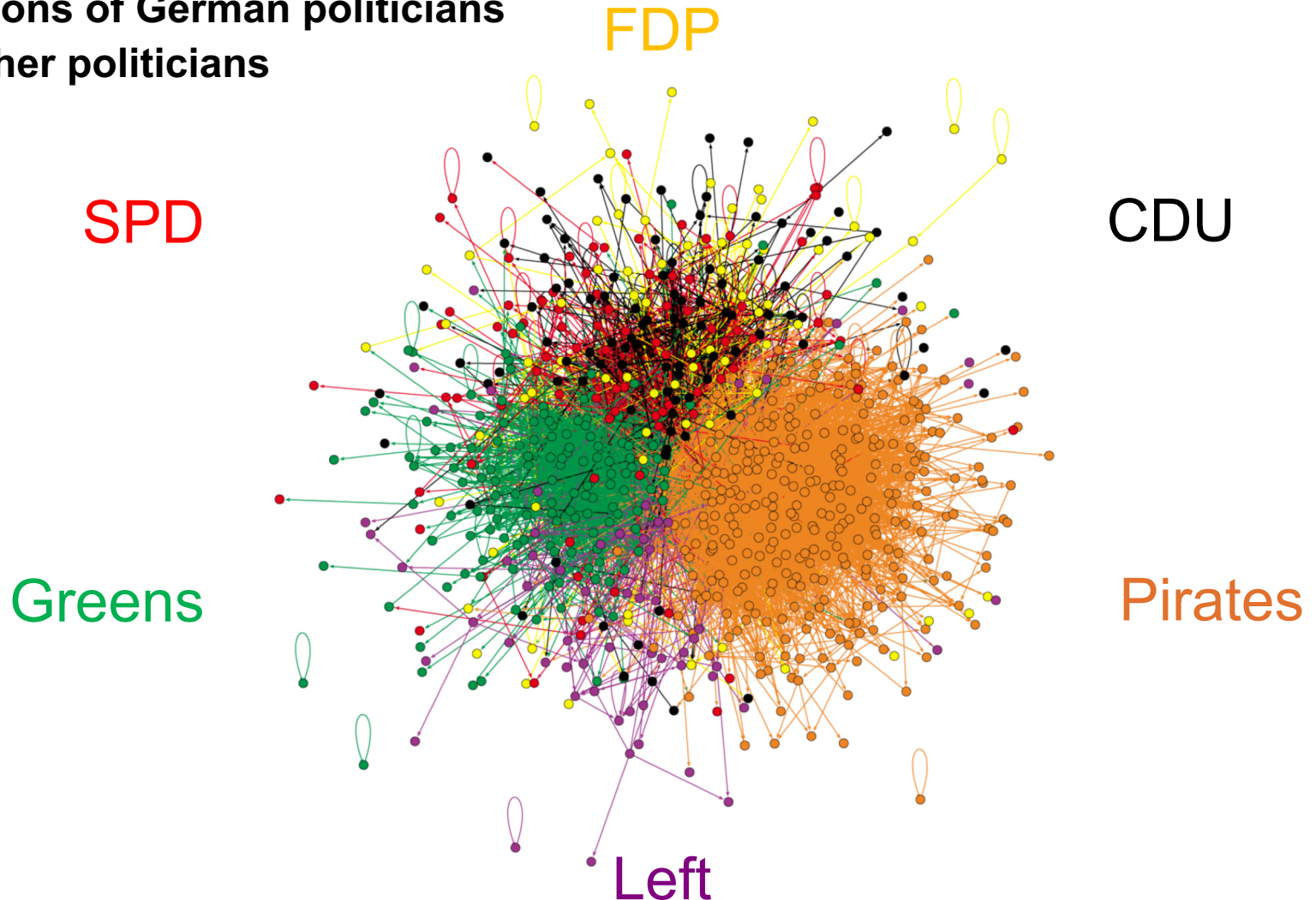
Example: Ethnicity-related police behavior (Voigt et al., 2017)
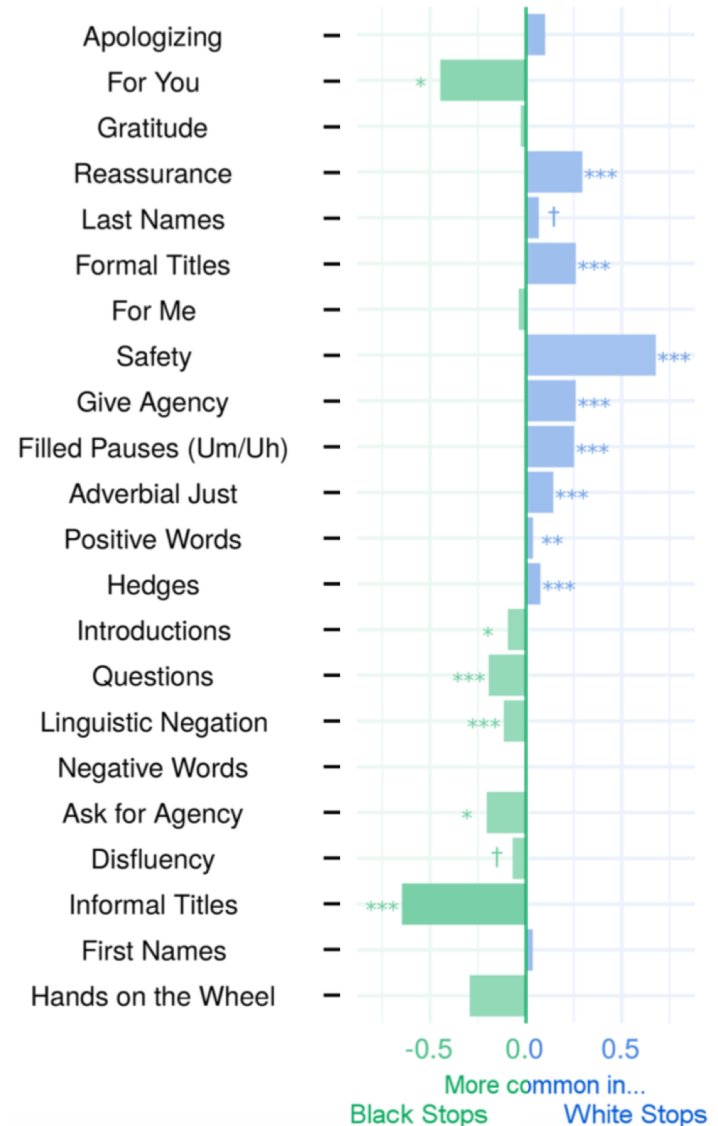
- **Language of US police officers towards black and white car drivers**

FIRST NAME ASK FOR AGENCY   QUESTIONS

[name], can I see that driver's license again?

It- it's showing suspended. Is that- that's you?

DISFLUENCY   NEGATIVE WORD   DISFLUENCY

APOLOGY   INTRODUCTION   LAST NAME

Sorry to stop you. My name's Officer [name] with the Police Department.

INFORMAL TITLE   ASK FOR AGENCY   ADVERBIAL "JUST"

All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick.

"HANDS ON THE WHEEL"

FORMAL TITLE   SAFETY   PLEASE

There you go, ma'am. Drive safe, please.

Apologizing
For You
Gratitude
Reassurance
Last Names
Formal Titles
For Me
Safety
Give Agency
Filled Pauses (Um/Uh)
Adverbial Just
Positive Words
Hedges
Introductions
Questions
Linguistic Negation
Negative Words
Ask for Agency
Disfluency
Informal Titles
First Names
Hands on the Wheel

-0.5   0.0   0.5

More common in...
Black Stops   White Stops

Computational Sociolinguistics (CSL)

An interdisciplinary research area

- **Two views of computational sociolinguistics (CSL)**
  - The intersection of computational linguistics and sociolinguistics
 - Computational social science on language data

computer science

computational social science

social science

CSL

computational linguistics
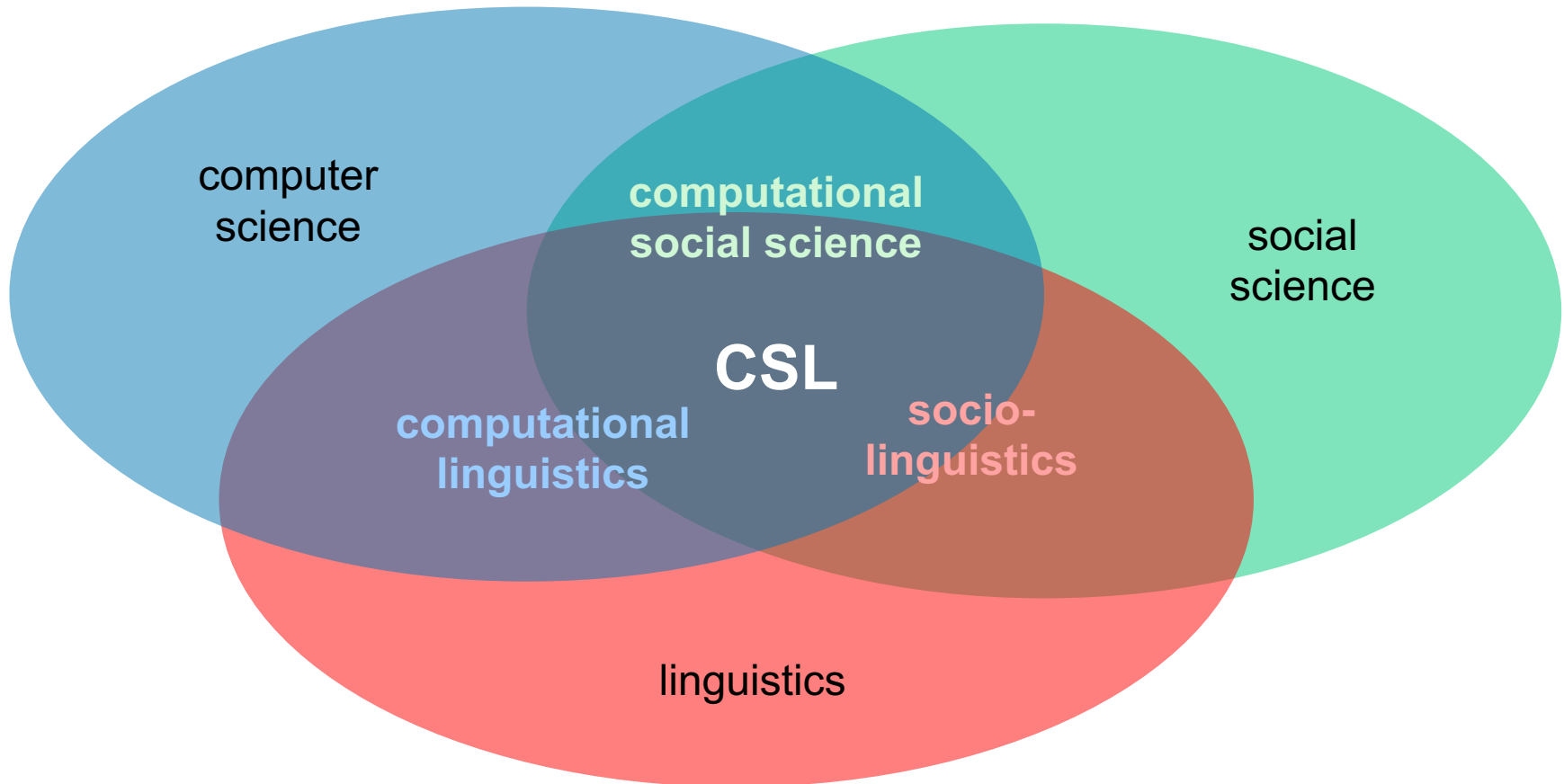
socio-linguistics

linguistics

Computational linguistics

- **Computational linguistics (CL)**

  - Studies the intersection of computer science and linguistics

  - Models for linguistic phenomena, based on knowledge and statistics (machine learning)

  - Technologies for natural language processing tasks
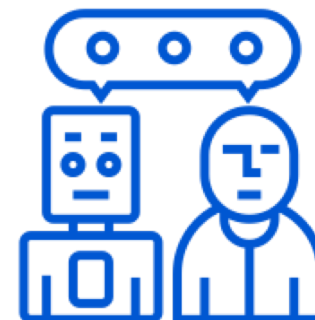
- **Natural language processing (NLP)**

  - Methods for understanding and generating speech and human-readable text

  - Targets various syntactic, semantic, and pragmatic tasks

  - From natural language to structured information, and vice versa

Analysis
Synthesis

- **Goals of research**

  - Creativity. Novelty of developed models and methods

  - Accuracy. Effectiveness in tackling tasks

  - Empirical research is often seen as stronger than theory

Sociolinguistics based on Nguyen et al. (2016)

- **Sociolinguistics (SL)**
  - Studies the mutual interaction of society and language
  - Relations between social variables and language use
  - Language variation across social groups, social contexts, and communicative situations

- **Language as a social phenomenon**
  - Social identity of speakers and listeners (gender, age, …) are inherently connected to language use
  - People can choose how to use language to achieve their goals
  - Analyzing language often requires to consider the people

- **Goals of research**
  - Validity. Extent to which research design isolates an issue to be studied
  - Reliability. Reproducibility of a result
  - Empirical research is seen as a means to support theory
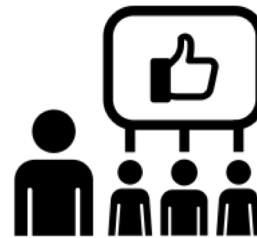
Computational social science

- **Computational social science (CSS)**
  - Studies questions from the social science through empirical data analysis
  - Insights into social phenomena and dynamics (primary)
  - Technologies to support social context (secondary)

- **Data** (among others)
  - Sociocultural key indicators
  - Social network structures
  - Online activities
  - News and social media texts

- **Analyses** (among others)
  - Statistical correlation analyses
  - Data mining
  - Natural language processing

Computational sociolinguistics based on Nguyen et al. (2016)

- **Computational sociolinguistics (CSL)**

  - Studies relations between language and society computationally based on data

  - Questions emerging from theory in sociolinguistics

  - Methods from computational linguistics

- **NLP in the context of CSL**

  - Data. Natural language texts, along with sociocultural meta-information
    Texts often come from news or discussions and posts on social media.

  - Methods. Primarily analysis (classification, regression, clustering, …), but also text generation may be involved

  - Applications. Tools with social dimensions (chatbots, writing support, …)

- **Mutual impact of involved fields**

  - SL → CL. Build more robust and well-grounded computational methods

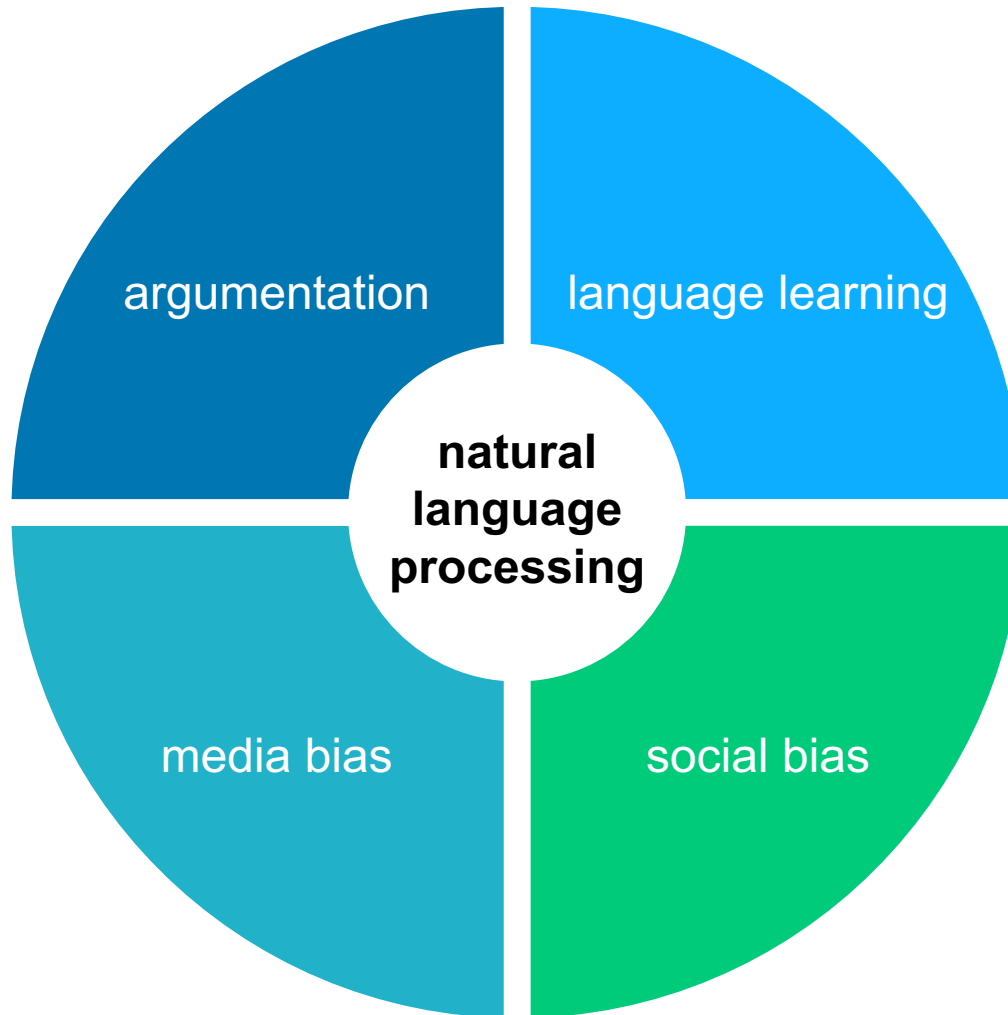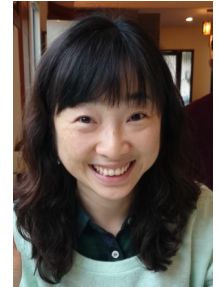 - CL → SL. Refine theoretical models, better understand social dynamics

CSL research of the CSS group

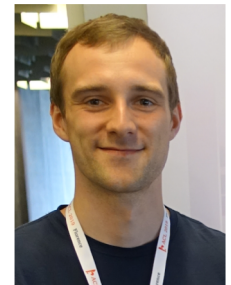# Computational sociolinguistics in the CSS group

Milad Alshomary

Mei-Hua Chen

Wei-Fan Chen

Maximilian Spliethöver

argumentation

language learning

natural language processing

media bias

social bias

Argumentation research

- **Argumentation**
  - The usage of arguments, along with rhetoric and dialectic, to persuade or agree with others
  - Arguments give reasons for claims controversial issues
  - Involved people impact the language and impact of arguments

- **Computational research**
  - Analysis and synthesis of argumentative natural language texts
  - Important for web search, debating systems, writing support, etc.
  - Sociolinguistic aspects include moral foundations and people's beliefs

- **Main research tasks**
  - Mining of argumentative units and relations
  - Assessment of stance, reasoning, and quality
  - Generation of units and arguments

If you wanna hear my view, I think that the EU should allow sea patrols in the Mediterranean Sea. Many innocent refugees will die if there are no rescue boats.

4 / 5

Media bias research

- **Media bias**
  - Polarized, tangential, and speculative information in news articles and other media reporting
  - Emerges from unjustified claims, omission, flawed logic, etc.
  - Media shapes the public opinion on contemporary issues

- **Computational research**
  - Analysis and synthesis of news articles with potential political bias
  - Important for news writing, search, and recommendation systems
  - Sociolinguistic aspects include fairness and neutrality towards people

- **Main research tasks**
  - Understand how bias is reflected in language
  - Detection of lexical and informational bias
  - Neutralization or flipping of media bias

Trump is making a huge mistake on Jerusalem

Trump is right in recognizing Jerusalem as Israel's capital

Social bias research

- **Social bias**
  - Prejudices, unequal treatment, and/or discrimination against certain social groups in society
  - Social groups include genders, ethnicities, and similar
 - Social bias may be implicit or explicit, conscious or unsconcious

- **Computational research**
  - Analysis and adjustment of datasets and language models in social contexts
  - Important for human-machine interaction in any sociotechnical system
  - Social bias in language is a core sociolinguistic concept

- **Main research tasks**
  - Understanding what social bias in language is
  - Detection of bias in texts and language models
  - Mitigation of bias in respective resources
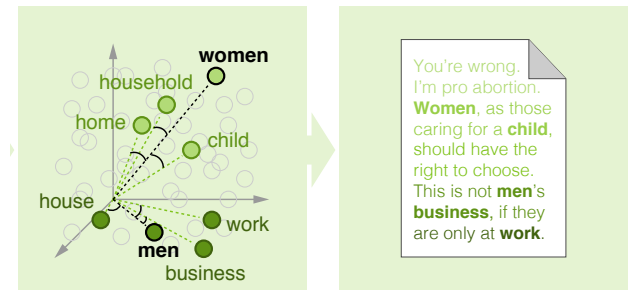
Language learning research

- **Language learning**
  - Learning a language requires dealing with complex linguistic expressions, writing strategies, etc.
  - Feedback by people (or systems) guides the learning process
 - Personal and social factors impact success of learners

- **Computational research**
  - Assessment and evaluation of learner texts as well as feedback generation
  - Important for writing evaluation/support systems
  - Sociolinguistic aspects include cultural and linguistic differences of learners

- **Main research tasks**
  - Analysis of grammar, style, and structure in text
  - Understanding of sociocultural factors within learning
  - Synthesis of suggestions for text improvements
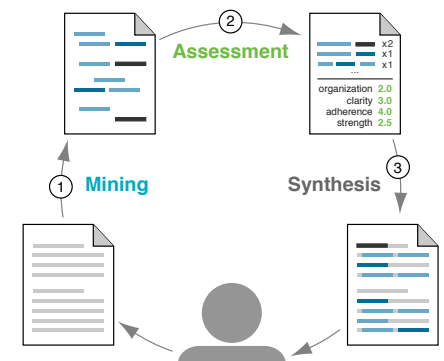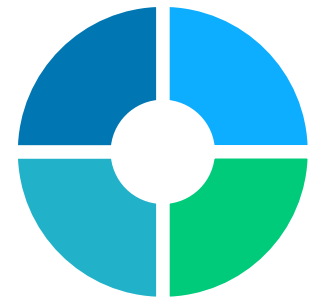
CSL in this seminar

This seminar

- **Frame of this seminar**
  - Basic ideas of computational sociolinguistics
  - State-of-the-art NLP research in this area
  - Connections to research at Paderborn University

- **Covered topics**
  - Argument mining, assessment, and generation
  - Analysis and mitigation of media bias and social bias
  - Computer-assisted language learning

- **Notice**
  - We take a broad view on computational sociolinguistics
  - Topics are selected according to our research interests
  - Basics of NLP are expected rather than taught

Concrete seminar topics: Next week

Conclusion

- **Computational sociolinguistics**
  - Studies relations between language and society computationally
  - Intersection of computational linguistics and sociolinguistics
  - Analysis and synthesis of texts from news and social media

- **This seminar**
  - State-of-the-art NLP research on computational sociolinguistics
  - Talks on argumentation, media/social bias, and language learning
  - Close connection to research in the CSS group

- **Next up**
  - Overview of concrete seminar topics with literature pointers
  - Topic preference choice and topic assignment
  - Basics of scientific presentation

References

- **Kosinski et al. (2013).** Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15):5802–5805, 2013.

- **Lietz et al. (2014).** Haiko Lietz, Claudia Wagner, Arnim Bleier, and Markus Strohmaier. When Politicians Talk: Assessing Online Conversational Practices of Political Parties on Twitter. In Proceedings of the Eighth International Conference on Weblogs and Social Media, 2014.

- **Nguyen et al. (2016).** Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, Franciska de Jong. Computational Sociolinguistics: A Survey. Computational Linguistics 42(3): 537–593, 2016.

- **Tsujii (2011).** Jun'ichi Tsujii. Computational Linguistics and Natural Language Processing. In Proceedings of the 12th International Conference on Computational linguistics and Intelligent Text Processing - Volume Part I, pages 52–67, 2011.

- **Voigt et al. (2017).** Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. Language from police body camera footage shows racial disparities in officer respect. Proceedings of the National Academy of Sciences, 114(25):6521–6526, 2017.