Seminar Natural Language Processing (NLP) — Part 1

Introduction to Computational Sociolinguistics

Henning Wachsmuth

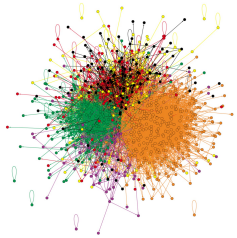
https://ai.uni-hannover.de

Natural Language Processing

Leibniz Universität Hannover

Outline

- **Motivation**



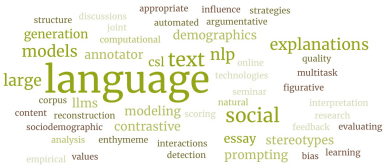
(Lietz et al., 2014)

- **Computational sociolinguistics (CSL)**

https://thenounproject.com

- **CSL research of the NLP Group**

- **CSL in this seminar**

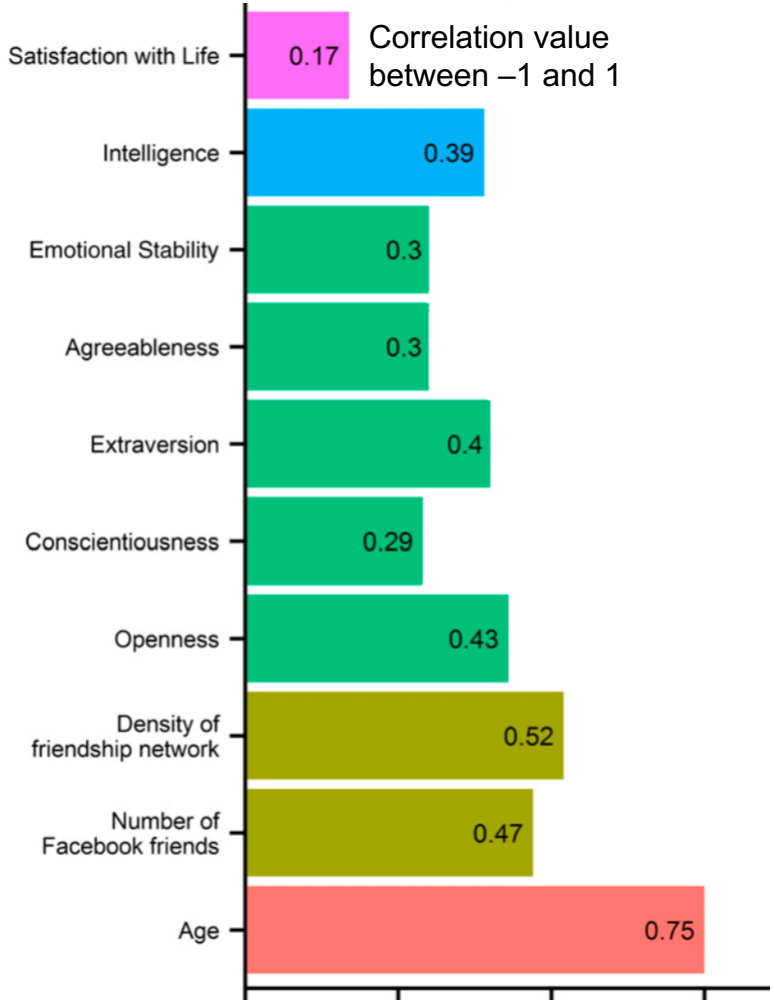


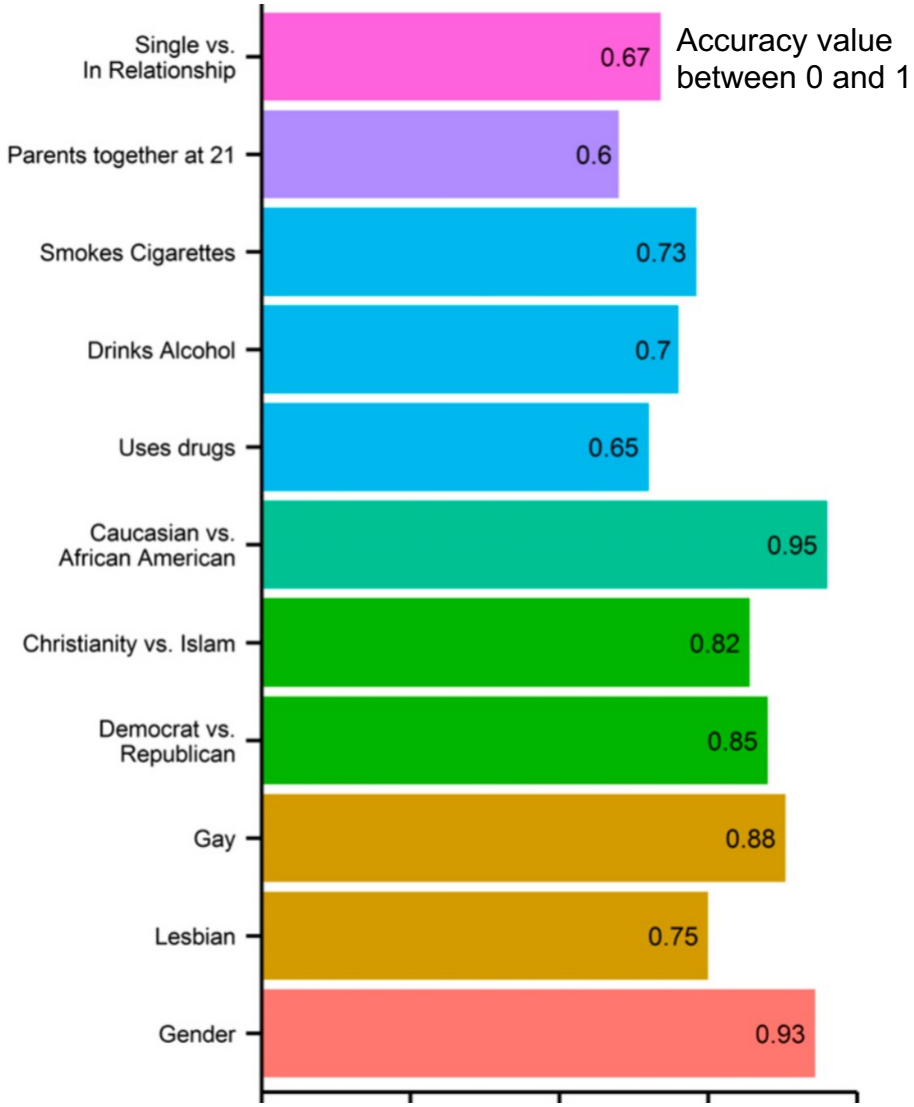
https://www.freewordcloudgenerator.com

Motivation

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

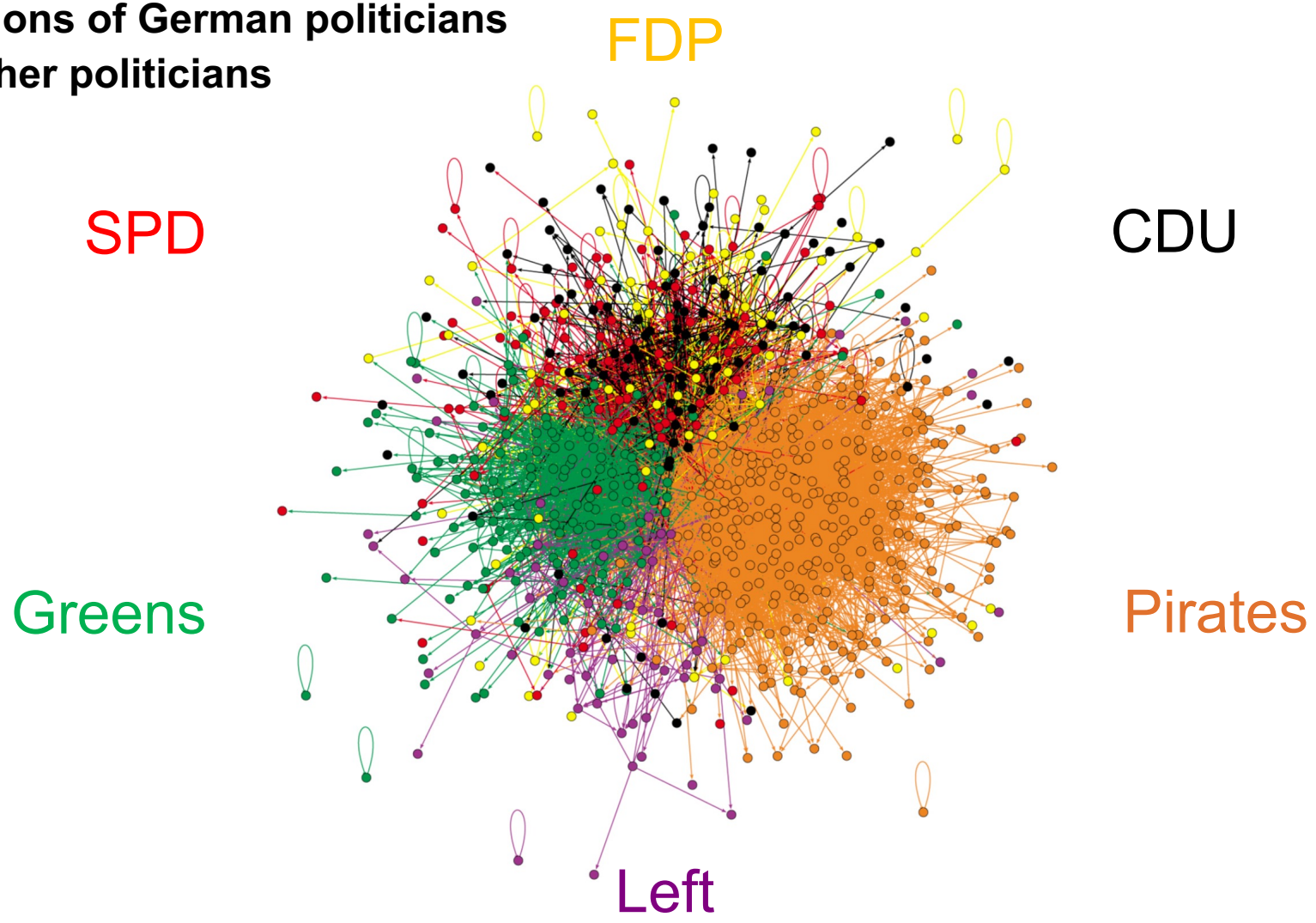
- **What social media likes reveal**





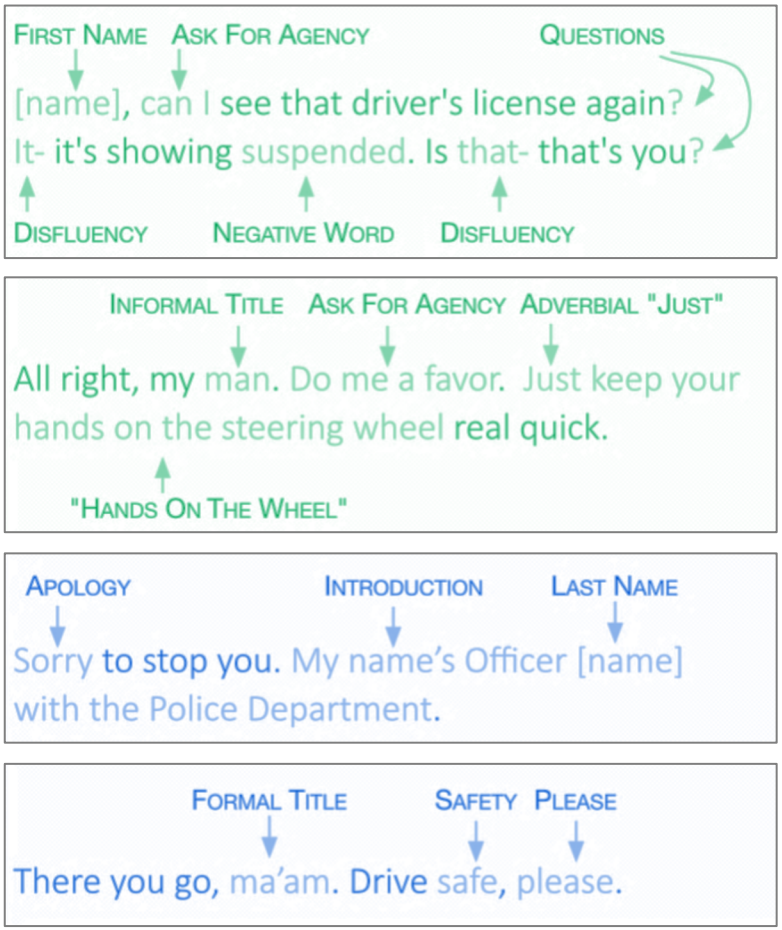
Example: Politicians‘ Twitter practices (Lietz et al., 2014)

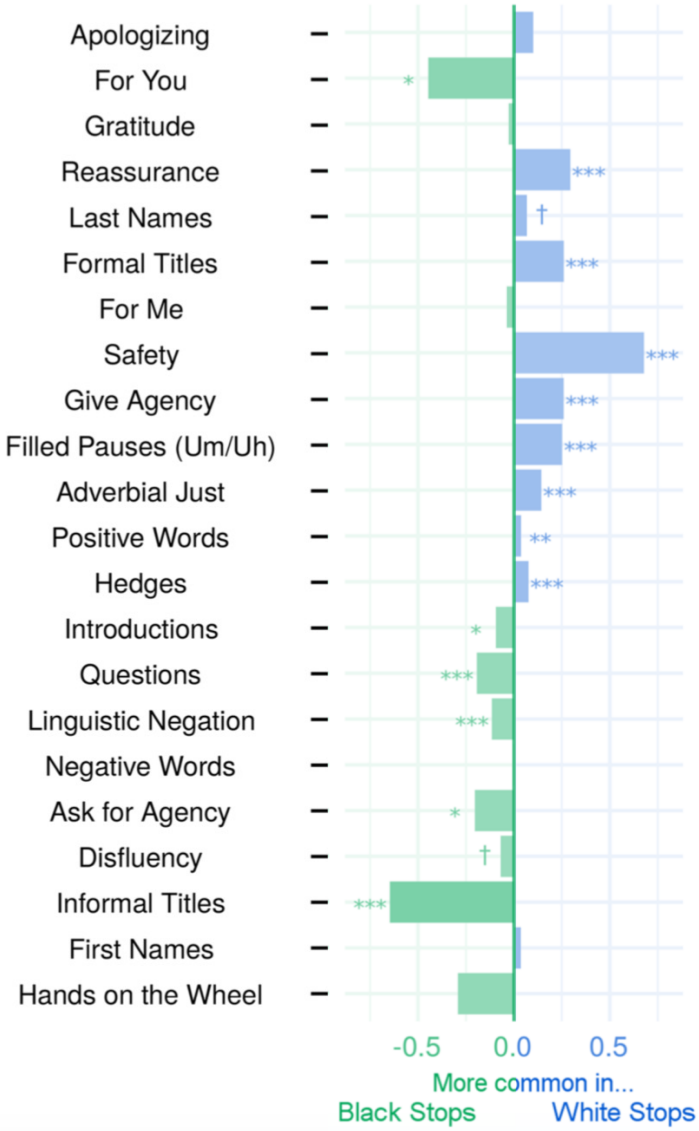
- **Mentions of German politicians by other politicians**



Example: Ethnicity-related police behavior (Voigt et al., 2017)

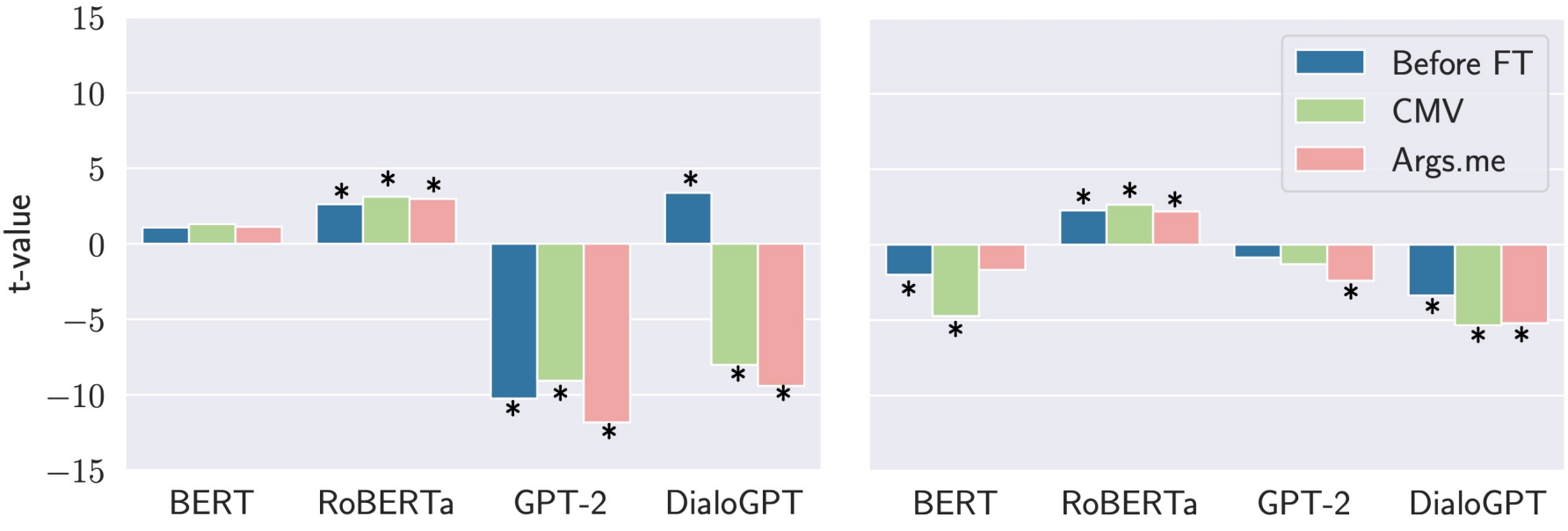
- **Language of US police officers towards black and white car drivers**





Example: Social bias in language models (Holtermann et al., 2022)

- **Social bias in language models before and after fine-tuning (FT) on subjective language**



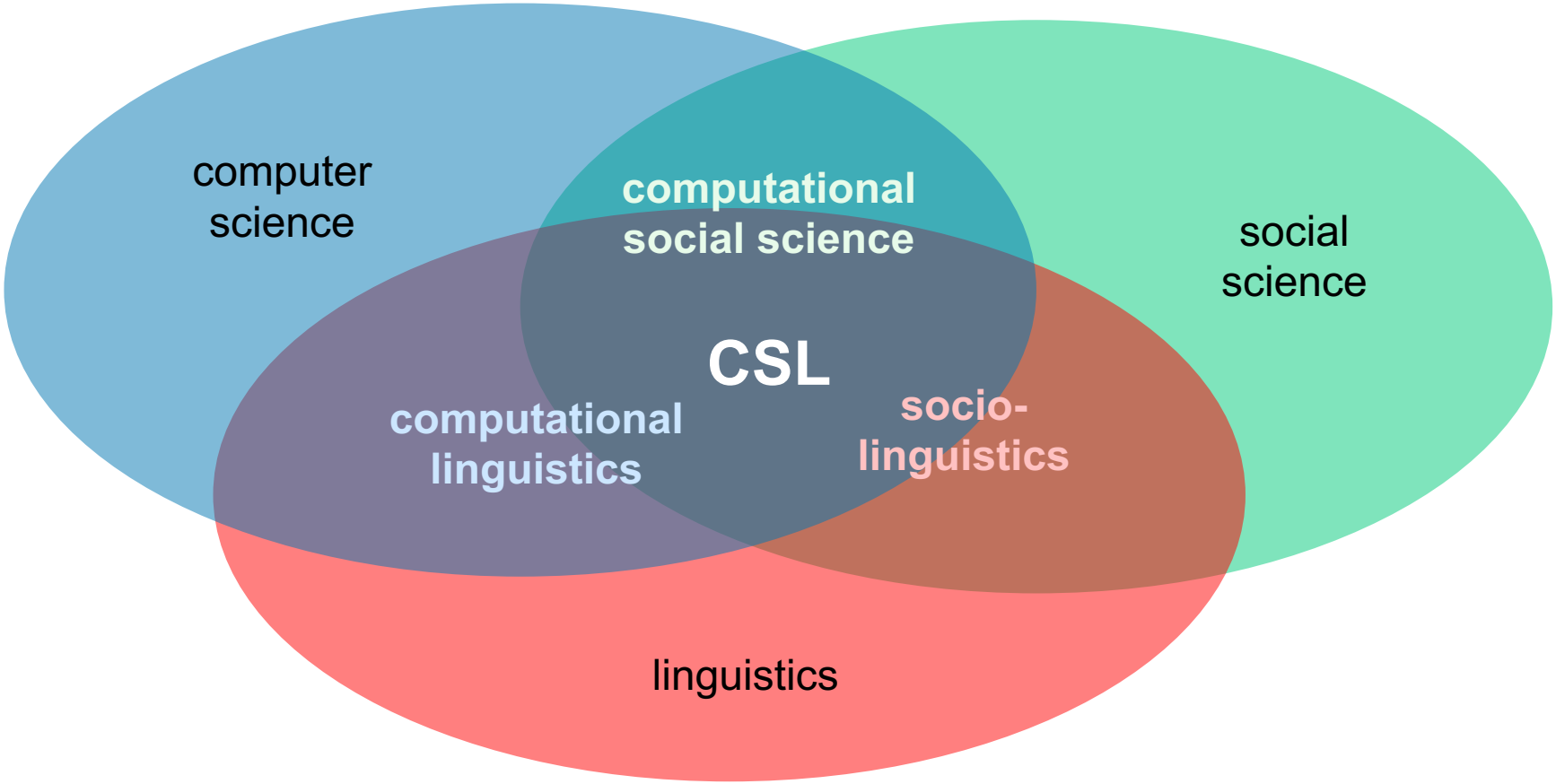
(a) LMB for Queerphobia

(b) LMB for Islamophobia

Computational Sociolinguistics (CSL)

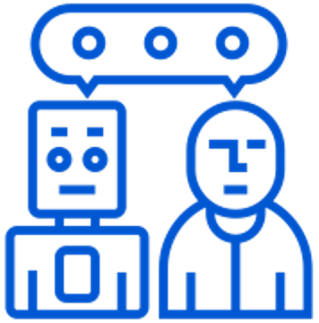
An interdisciplinary research area

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

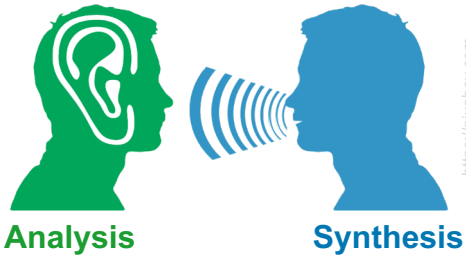


Computational linguistics based on Tsujii, (2011)

- **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
  - Various syntactic, semantic, and pragmatic tasks
  - From language to structured information, or vice versa



- **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

https://thenounproject.com

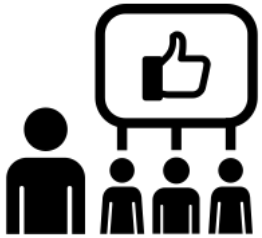
- **Language as a social phenomenon**
  - Social identity of speakers and listeners (gender, age, …) are inherently connected to language use
  - People can choose how to use language to achieve their goals
  - Analyzing language often requires to consider the people

- **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

All figures on this slide from https://thenounproject.com

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

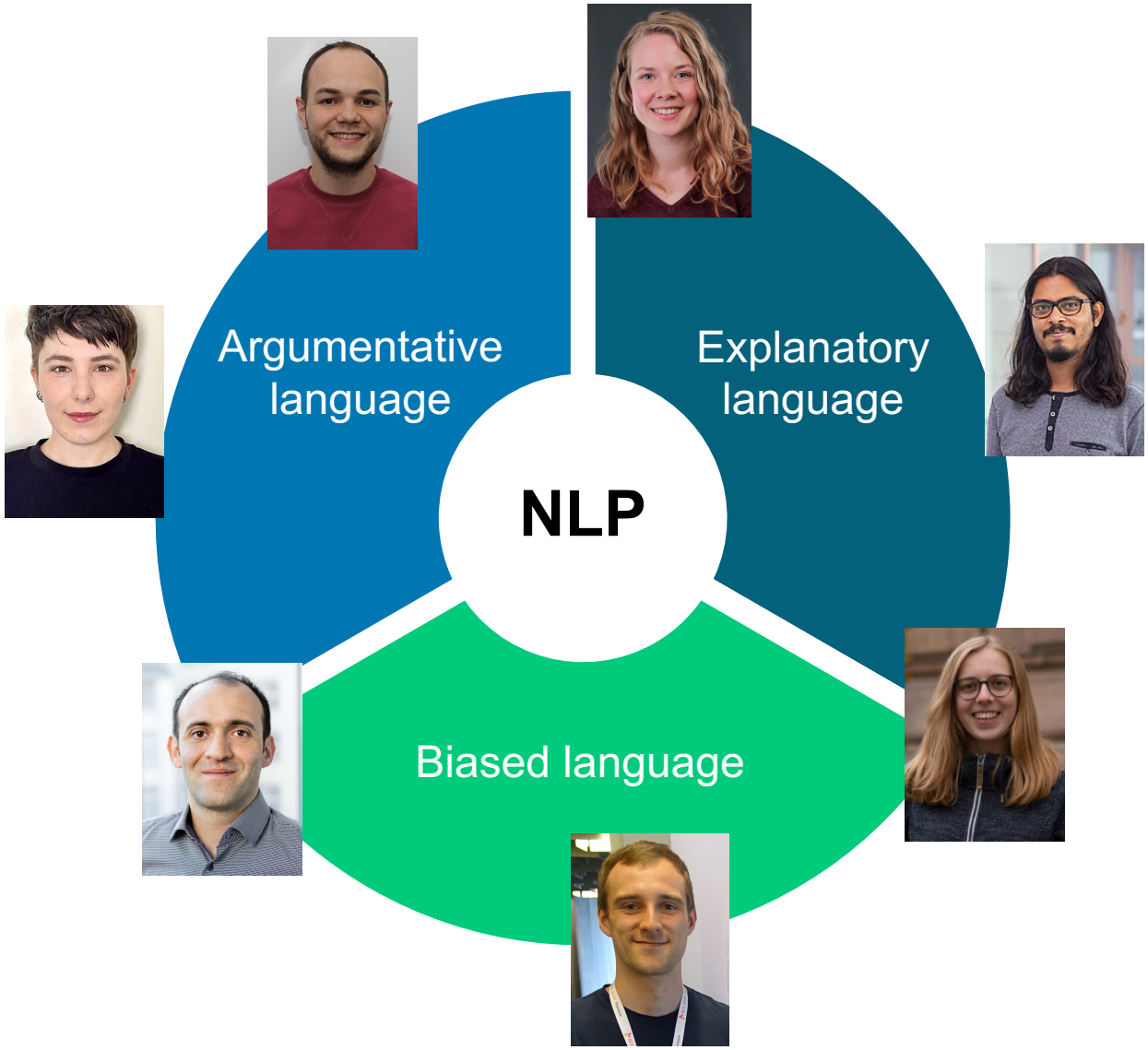
https://thenounproject.com

- **NLP in the context of CSL**
  - Data. Natural language texts, along with sociocultural meta-information
    Texts often come from news or from 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 NLP Group

CSL in the NLP Group



CSL research on argumentative language

- **Argumentative language**
  - Online discussions on controversial issues
  - Claims and arguments that express viewpoints
  - Persuasive essays written by learners

- **Analysis**
  - Modeling of the structure and style of the argumentative texts
  - Computational assessment of the quality of the texts
  - Automatic identification of toxicity in the texts

- **Generation**
  - Computational reconstruction of implicit argument parts
  - Computational generation of counterarguments
  - Computational generation of feedback to learner texts

CSL research on biased language

- **Biased language**
  - Language reflecting positive/negative views of social groups
  - Language reflecting stereotypes about the groups
  - Values and political views expressed through language

- **Analysis**
  - Empirical evaluation of social biases in language models
  - Computational analysis of priming effects on language models
  - Assessment of the quality of language model outputs

- **Generation**
  - Automatic generation of counterfactuals for bias mitigation
  - Prompting strategies for bias detection and mitigation
  - Influence of sociodemographic factors on generated text

CSL research on explanatory language

- **Explanatory language**
  - Explanations given to describe or justify phenomena
  - Linguistic devices used in explanations
  - Explanatory dialogues between two or more people

- **Analysis**
  - Modeling of structural patterns in explanatory texts and dialogues
  - Computational understanding of the functioning of explanations
  - Human and machine evaluation of explanation quality

- **Generation**
  - Computational generation of explanations
  - Computational adjustment of explanations to user language
  - Explaining machines that lead dialogues with users

CSL in this seminar

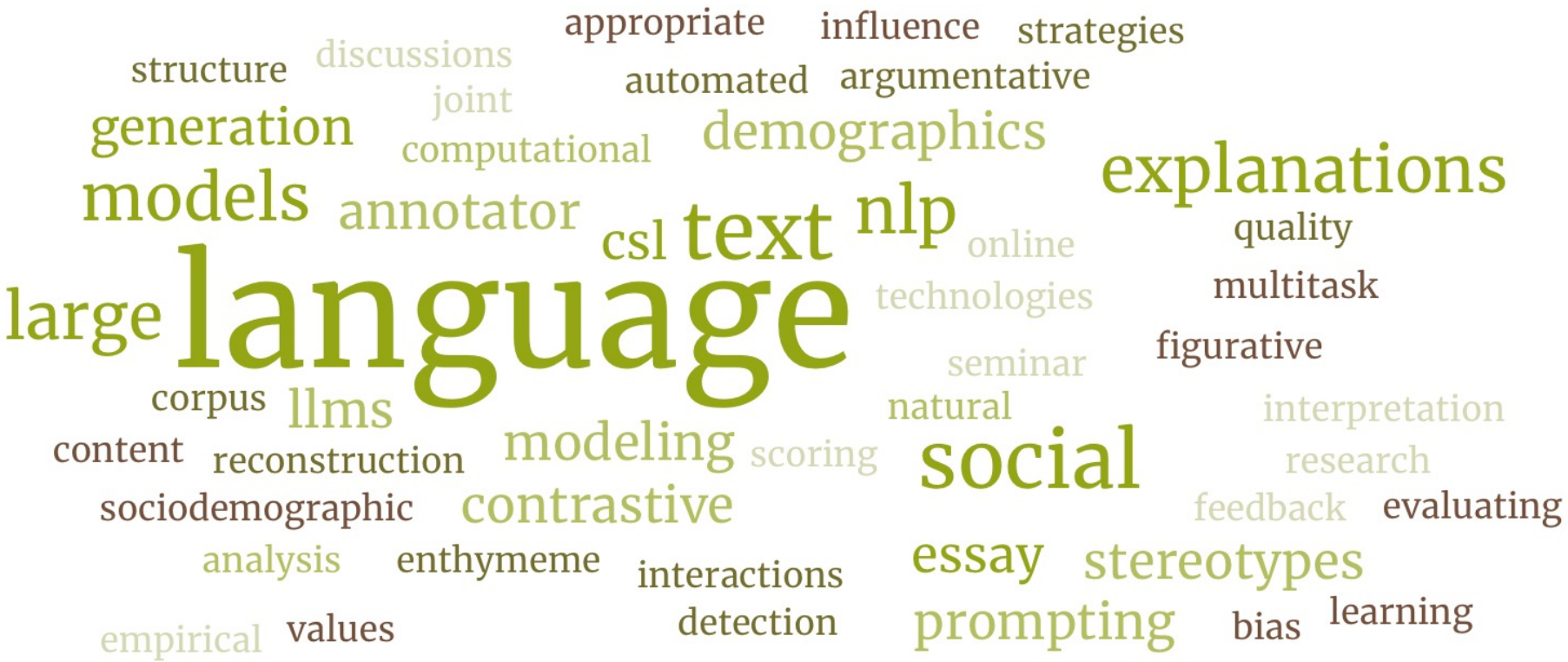
This seminar

- **Frame of this seminar**
  - Basics of NLP for computational sociolinguistics
  - State-of-the-art NLP research in this area
  - Connections to research at Institute of Artificial Intelligence

- **Covered topics**
  - Assessment and generation of argumentative language
  - Analysis and evaluation of biased language
  - Identification and understanding of explanatory language

- **Notice**
  - We take a broad view on CSL in this seminar
  - Topics are selected according to our research interests
  - Required basics of NLP to be acquired rather than taught

Concrete seminar topics: Next week



Conclusion

- **Computational sociolinguistics (CSL)**
  - Studies relations between language and society computationally
  - Intersection of computational linguistics and sociolinguistics
  - Analysis and synthesis of texts from social contexts

- **This seminar**
  - State-of-the-art NLP research on computational sociolinguistics
  - Learner, explanatory, subjective, and language model language
  - Close connection to research in the NLP Group

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

References

- **Holtermann et al. (2022).** Carolin Holtermann, Anne Lauscher, and Simone Ponzetto. Fair and Argumentative Language Modeling for Computational Argumentation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7841–7861, 2022.
- **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.