# semtracks - Finding Patterns in Discourse

Klaus Rothenhäusler
rothenhaeusler@semtracks.com

Heidelberg Center for American Studies
University of Heidelberg

23. 4. 2009 / Freiburg

# Introduction

- project title

  **Tracking Meaning on the Surface: A Data-Driven Approach to Semantic Imprints of Texts**

- funded by *Innnovationsfond Frontier* at University of Heidelberg
- interdisciplinary research team: linguists, historians, geographers and a computational linguist

# Introduction

- project title

  **Tracking Meaning on the Surface: A Data-Driven Approach to Semantic Imprints of Texts**

- funded by *Innnovationsfond Frontier* at University of Heidelberg
- interdisciplinary research team: linguists, historians, geographers and a computational linguist

# Introduction

- project title

  **Tracking Meaning on the Surface: A Data-Driven Approach to Semantic Imprints of Texts**

- funded by *Innnovationsfond Frontier* at University of Heidelberg
- interdisciplinary research team: linguists, historians, geographers and a computational linguist

# Introduction – Case Study

- original (and future):
  *Perceptions of America after 9/11*
- coverage of presidential campaign in 2008
- now following Bundestagswahl

# Introduction – Case Study

- original (and future):
  *Perceptions of America after 9/11*
- coverage of presidential campaign in 2008

- now following Bundestagswahl

# Introduction – Case Study

- original (and future):
  *Perceptions of America after 9/11*
- coverage of presidential campaign in 2008


- now following Bundestagswahl

# Introduction – Case Study

- original (and future):
  *Perceptions of America after 9/11*
- coverage of presidential campaign in 2008


- now following Bundestagswahl

# Background

- *semantic imprints* (*Semantische Prägung*)
  - premise: language is not only mirroring but constructing reality
  - construction via habitual linguistic patterns
  - observable *on the surface* and statistically identifiable
  - no deep semantic knowledge required
- specific configurations of pattern occurrence hint at pragmatic, social and cultural function
- goal: text analysis tool for social and cultural studies to position a given text within such categories

# Background

- *semantic imprints* (*Semantische Prägung*)
  - premise: language is not only mirroring but constructing reality
  - construction via habitual linguistic patterns
  - observable *on the surface* and statistically identifiable
  - no deep semantic knowledge required
- specific configurations of pattern occurrence hint at pragmatic, social and cultural function
- goal: text analysis tool for social and cultural studies to position a given text within such categories

# Background

- *semantic imprints* (*Semantische Prägung*)
  - premise: language is not only mirroring but constructing reality
  - construction via habitual linguistic patterns
  - observable *on the surface* and statistically identifiable
  - no deep semantic knowledge required
- specific configurations of pattern occurrence hint at pragmatic, social and cultural function
- goal: text analysis tool for social and cultural studies to position a given text within such categories

# Background

- *semantic imprints* (*Semantische Prägung*)
  - premise: language is not only mirroring but constructing reality
  - construction via habitual linguistic patterns
  - observable *on the surface* and statistically identifiable
  - no deep semantic knowledge required
- specific configurations of pattern occurrence hint at pragmatic, social and cultural function
- goal: text analysis tool for social and cultural studies to position a given text within such categories

# Background

- *semantic imprints* (*Semantische Prägung*)
  - premise: language is not only mirroring but constructing reality
  - construction via habitual linguistic patterns
  - observable *on the surface* and statistically identifiable
  - no deep semantic knowledge required
- specific configurations of pattern occurrence hint at pragmatic, social and cultural function
- goal: text analysis tool for social and cultural studies to position a given text within such categories

# Background

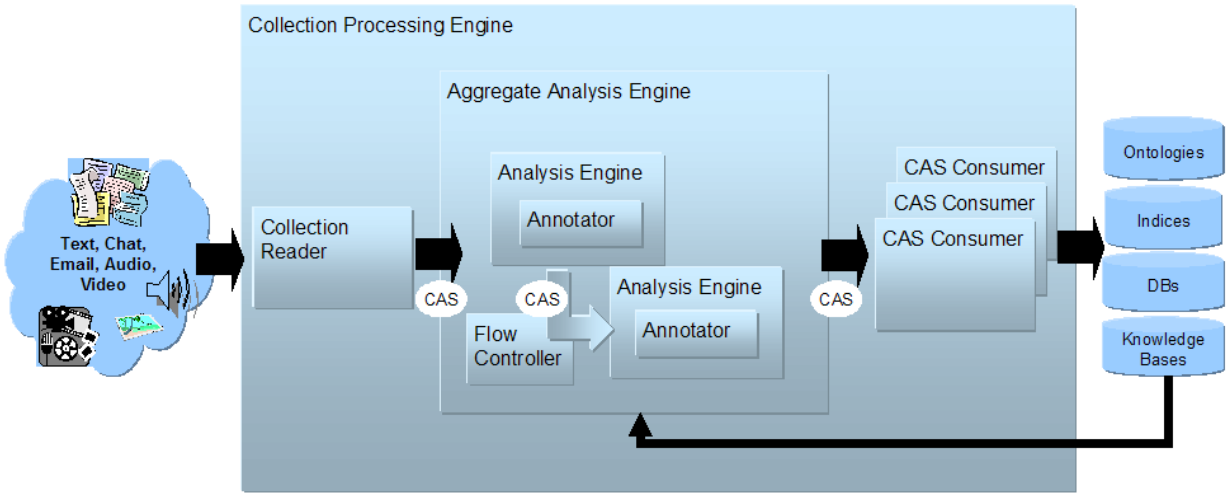
- *semantic imprints* (*Semantische Prägung*)
  - premise: language is not only mirroring but constructing reality
  - construction via habitual linguistic patterns
  - observable *on the surface* and statistically identifiable
  - no deep semantic knowledge required
- specific configurations of pattern occurrence hint at pragmatic, social and cultural function
- goal: text analysis tool for social and cultural studies to position a given text within such categories

# Background

- *semantic imprints* (*Semantische Prägung*)
  - premise: language is not only mirroring but constructing reality
  - construction via habitual linguistic patterns
  - observable *on the surface* and statistically identifiable
  - no deep semantic knowledge required
- specific configurations of pattern occurrence hint at pragmatic, social and cultural function
- goal: text analysis tool for social and cultural studies to position a given text within such categories

# Linguistic Analysis – Architecture

- pipeline architecture based on Unstructured Information Management Architecture
  An Apache Incubator Project




- allows for multimodal representations

# Linguistic Analysis

- tokenization (identifying words), includes
  - markup removal
  - boilerplate identification (e.g. navigational parts of a web page)
  - encoding issues
- part of speech (*Wortart*) tagging
- annotation of certain word classes, e.g.
  - *itensifiers*
    *I completely agree . . .*
  - *hedges*
    *I'm quite sure . . .*
- sentence and clause annotation
  - hybrid approach: rule based and statistic (memory based learner)
  - including *tense*, *mood* and *Aktionsart* identification

# Statistical Analysis

- data sink: 

- inverted index for storing texts along with linguistic annotations
- advanced querying
  - regular expressions over text and annotations
  - constraining search to annotated regions
- efficient implementation of
  - buzzword analysis
  - n-gram analysis
  - collocation analysis

# Statistical Analysis

- data sink: CWB The IMS Open Corpus Workbench
- inverted index for storing texts along with linguistic annotations
- advanced querying
  - regular expressions over text and annotations
  - constraining search to annotated regions
- efficient implementation of
  - buzzword analysis
  - n-gram analysis
  - collocation analysis

# Statistical Analysis

- data sink: 

- inverted index for storing texts along with linguistic annotations
- advanced querying
  - regular expressions over text and annotations
  - constraining search to annotated regions
- efficient implementation of
  - buzzword analysis
  - n-gram analysis
  - collocation analysis

# Statistical Analysis

- data sink: CWB The IMS Open Corpus Workbench
- inverted index for storing texts along with linguistic annotations
- advanced querying
  - regular expressions over text and annotations
  - constraining search to annotated regions
- efficient implementation of
  - buzzword analysis
  - n-gram analysis
  - collocation analysis

# Statistical Analysis

- data sink: CWB The IMS Open Corpus Workbench
- inverted index for storing texts along with linguistic annotations
- advanced querying
  - regular expressions over text and annotations
  - constraining search to annotated regions
- efficient implementation of
  - buzzword analysis
  - n-gram analysis
  - collocation analysis

# Statistical Analysis

- data sink: 

- inverted index for storing texts along with linguistic annotations
- advanced querying
  - regular expressions over text and annotations
  - constraining search to annotated regions
- efficient implementation of
  - buzzword analysis
  - n-gram analysis
  - collocation analysis

# Statistical Analysis

- data sink: 

- inverted index for storing texts along with linguistic annotations
- advanced querying
  - regular expressions over text and annotations
  - constraining search to annotated regions
- efficient implementation of
  - buzzword analysis
  - n-gram analysis
  - collocation analysis

# Statistical Analysis

- data sink: 

- inverted index for storing texts along with linguistic annotations
- advanced querying
  - regular expressions over text and annotations
  - constraining search to annotated regions
- efficient implementation of
  - buzzword analysis
  - n-gram analysis
  - collocation analysis

# Statistical Analysis

- data sink: 

- inverted index for storing texts along with linguistic annotations
- advanced querying
  - regular expressions over text and annotations
  - constraining search to annotated regions
- efficient implementation of
  - buzzword analysis
  - n-gram analysis
  - collocation analysis

# Example: Explorative Analysis

corpora (text collections)

- Presidential Campaign 2008
  - all speeches from both candidates
  - tv debates
- Merkel vs Steinmeier
  - all speeches in current legislative period

# Example: Explorative Analysis

corpora (text collections)

- Presidential Campaign 2008
  - all speeches from both candidates
  - tv debates
- Merkel vs Steinmeier
  - all speeches in current legislative period

# Example: Explorative Analysis

corpora (text collections)

- Presidential Campaign 2008
  - all speeches from both candidates
  - tv debates
- Merkel vs Steinmeier
  - all speeches in current legislative period

# Example: Explorative Analysis

corpora (text collections)

- Presidential Campaign 2008
  - all speeches from both candidates
  - tv debates
- Merkel vs Steinmeier
  - all speeches in current legislative period

# Example: Explorative Analysis

corpora (text collections)

- Presidential Campaign 2008
  - all speeches from both candidates
  - tv debates
- Merkel vs Steinmeier
  - all speeches in current legislative period

# Example: Buzzword Analysis

- words appearing significantly more often in one corpus than in another
- constrained by parts of speech

# Example: Buzzword Analysis

- words appearing significantly more often in one corpus than in another
- constrained by parts of speech
- Merkel vs Steinmeier: adjectives

# Example: Buzzword Analyis (Adjectives)



- Merkel: evaluative and intensifying adjectives
- gender specific language

# Example: Buzzword Analysis

- words appearing significantly more often in one corpus than in another
- constrained by parts of speech
- Obama vs McCain: personal pronouns

# Example: Buzzword Analyis (Pronouns)

McCain

| pronoun | significance level ($\chi^2$) | relative frequency factor |
|---|---|---|
| my | $< 0.0001$ | 1.66 |
| their | $< 0.0001$ | 1.19 |
| he | $< 0.001$ | 1.26 |
| I | $< 0.01$ | 1.08 |

Obama

| pronoun | significance level ($\chi^2$) | relative frequency factor |
|---|---|---|
| we | $< 0.0001$ | 1.4 |
| you | $< 0.0001$ | 1.5 |
| us | $< 0.0001$ | 1.33 |
| yourself | $< 0.001$ | 6.17 |
| they | $< 0.05$ | 1.1 |

- McCain: self references and references to opponents
- Obama: appeal to collectivity

# Example: N-Gram Analysis

statistically significant sequences of $n$ consecutive words or parts of speech

# Example: N-Gram Analysis

Merkel

- subjective interpretation and
  Ich glaube, dass
  Ich vermute, dass
  Ich hoffe, dass
- factual construction of reality
  Ich weiß, dass
  Sie wissen, dass
  Wir wissen, dass
- consecutive continuations
  Deshalb werden wir
  Deshalb haben wir
  Das heißt

Steinmeier

- genitive objects – nominalisations
  Autorität des Gouverneursrates der IAEO
  Erklärung des Botschafters des Irans
  Klärung des Status des Kosovo
- consecutive prepositional groups
  für die Bekämpfung der Intoleranz
  angesichts der Tätergruppen von New York
  für die Bedeutung von Innovation
- enumerations
  den amerikanischen Freunden und Partnern
  gesellschaftliche und kulturelle Austausch
  Gewicht, Stimme und Handlungsfähigkeit
  Verhalten, Augenmaß und Vernunft

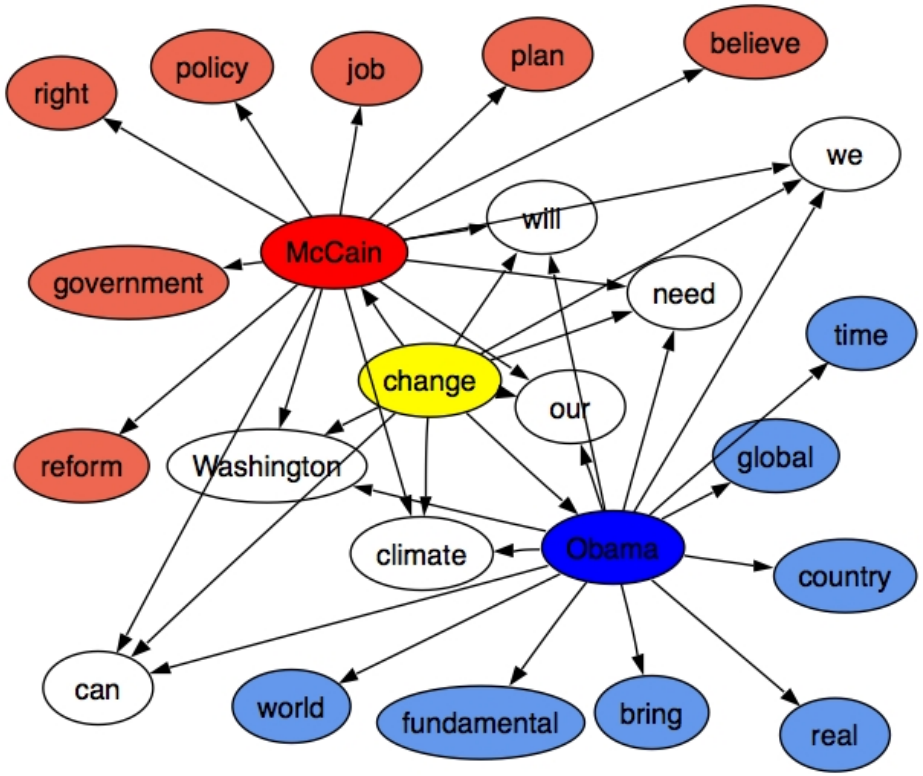
- Merkel: emotional emphasis and subjective style
- Steinmeier: competent expert using prefab formulations (repetitions)

# Example: Collocation Analysis

collocation two words occuring together more often than would be expected by chance, e.g.

- Examples
  - *strong* not *powerful tea*
  - *handsome man* but *beautiful woman*
  - *to get sick* but *to fall ill*

# Example: Collocation Analysis




- Obama: *fundamental* socio-cultural change
- McCain: *reform* in politics (Washington) and economics

# Summary

- results of explorative data analysis
  - semantically loaded entities can be identified
  - provides hints for further investigation

# Future Research, or: Why do we Need the BFG?

- bigger corpora
- advanced linguistic analysis, e.g. syntactic parsing ($O(N^3)$)
- general approach
  - analyse textual data
  - construct feature representation
  - train a classifier (supervised or unsupervised)
- open questions:
  - What kind of features are useful, i.e. what linguistic analyses are necessary?
  - Do we need different representations for different catgorisations?

# References

- semtracks: http://semtracks.com/
- semtracks political tracker blog: http://semtracks.com/index.php?id=Political%20Tracker
- Apache UIMA: http://incubator.apache.org/uima/
- Corpus Workbench: http://cwb.sourceforge.net/