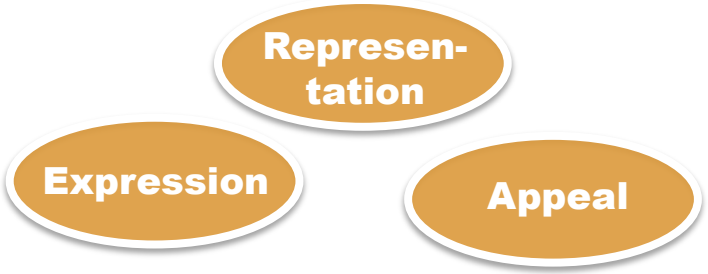
Back to the Roots of Genres: Text Classification by Language Function

Henning Wachsmuth and Kathrin Bujna

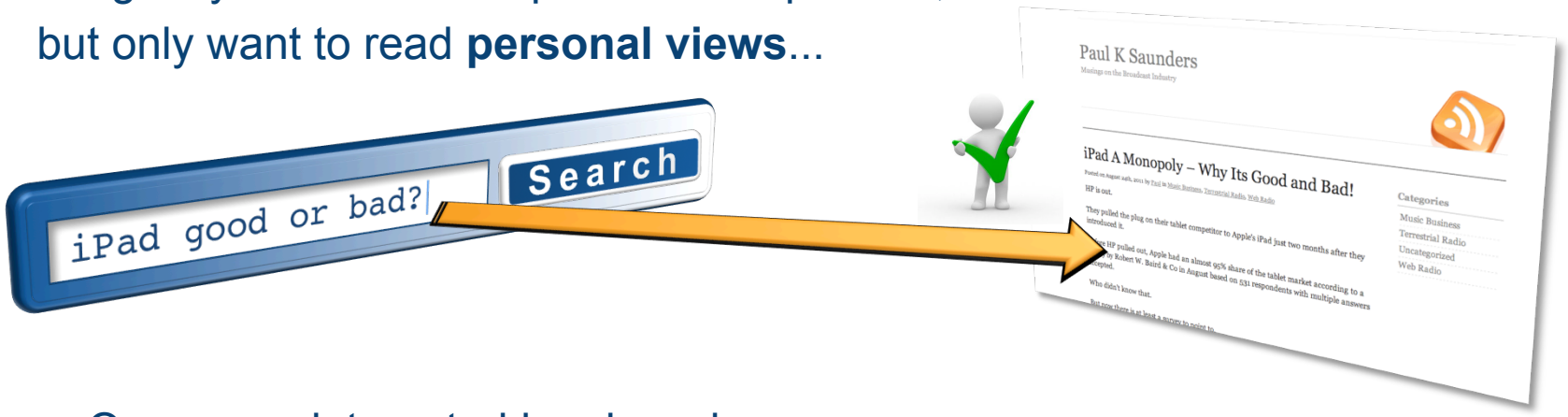
presented at IJCNLP on November 10, 2011



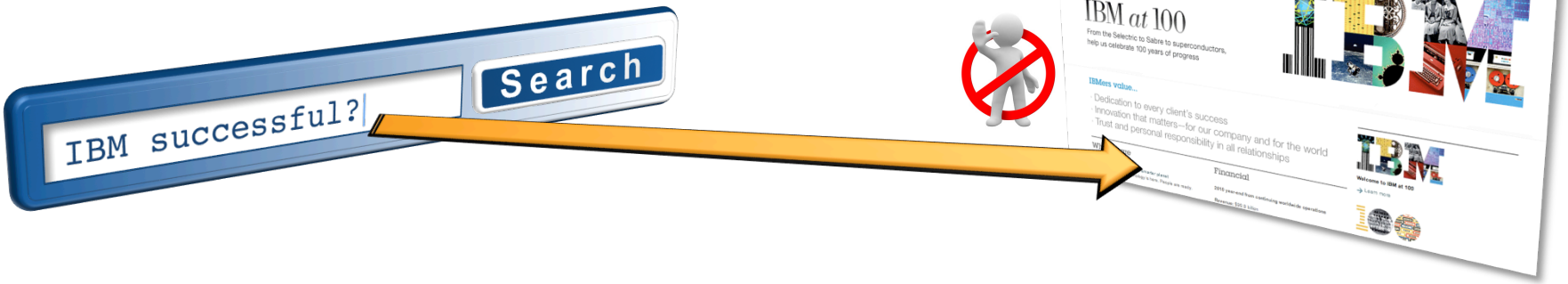
UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Motivation: Filter search results

- Imagine you search for opinions on a product, but only want to read **personal views**...



- ... Or you are interested in a brand, but do not want **commercial texts** on that brand...



- ... Such filtering could be approached with **genre identification**, but...

Motivation: Problems with genres

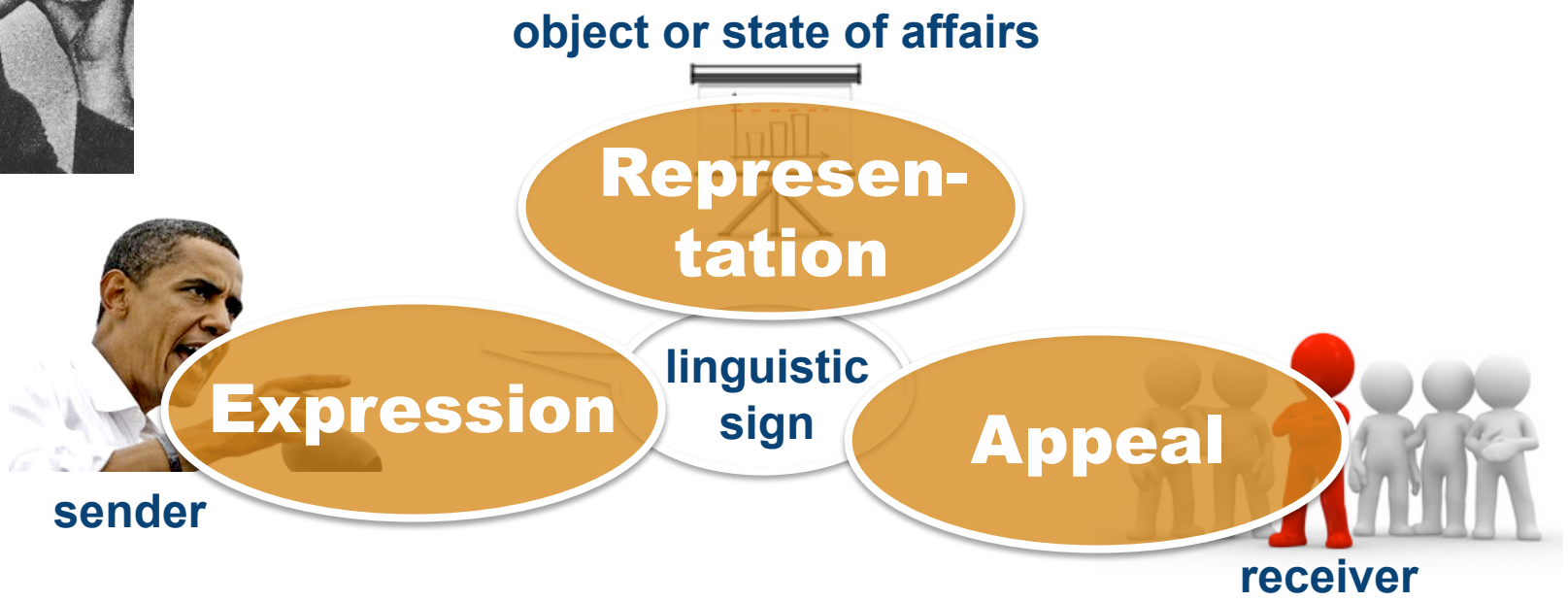
- Unlike many renowned classification tasks, **genre identification mixes different aspects** of both texts and documents

- There is a **missing common understanding of genres**
  - As a consequence, several genre classification schemes exist
  - Different approaches are badly comparable (see Sharoff et. al., 2010)
  - The task itself is unclear

- In contrast, we focus on **one single aspect of genres:** language functions

The functions of natural language

- In 1934, the psychologist **Karl Bühler** introduced one of the most influential attempts to categorize the functions of natural language

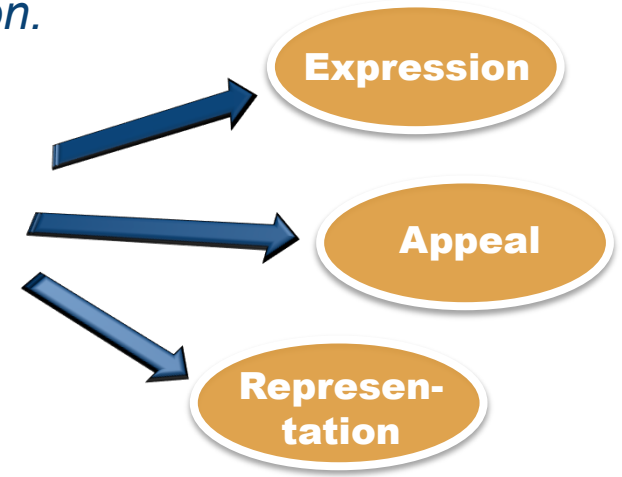


- Later on (1971), the linguist **Katharina Reiß** carried the three language functions over to text

Language Function Analysis (LFA)

- We introduce the **new task** “Language Function Analysis“ (LFA)

  Given a text, decide whether its predominant language function is 1) expression, 2) appeal, or 3) representation.



- **Properties** of LFA
  - Very general
  - Addresses one single aspect of genres
  - Can be used for document filtering purposes

- So, yet another classification scheme?
  - LFA is not meant to solve genre identification, but might help to better understand genres
  - **Question:** How can we identify the language function of a text?

A text corpus for LFA

Freely available at
http://infexba.upb.de

- For evaluation, we built a **German text corpus** in cooperation with industry
  - Contains separated text collections of **two product domains**:

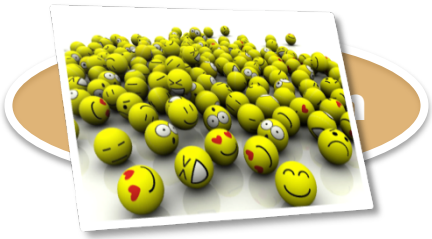
Music

2713 well-written promotional texts and reviews

Smartphones

2093 blog posts of varying quality and style

- Each text is **manually classified** by language function and sentiment polarity
  - Many details about the annotation process in the paper
  - We mapped the language functions to product-related classes:



personal texts

commercial texts

informational texts

A machine learning approach to LFA

- Our approach to LFA relies on **supervised machine learning** classification
  - Experiments with features from different research areas
  - Organization into 6 feature groups

(Simple) Genre

part-of-speech distribution and text statistics

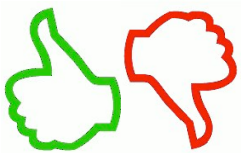
Text type

frequency of entities and some parts-of-speech

Writing style

most common words and trigrams

Sentiment

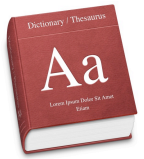


sentiment polarity and emoticons

Core trigrams

most discriminative trigrams

Core terms



most discriminative terms

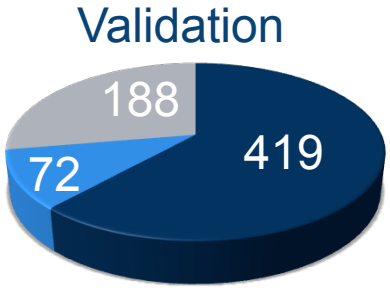
Evaluation

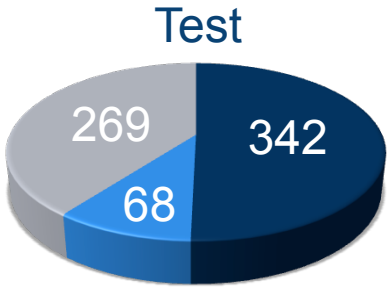
Source code and feature files at **http://infexba.upb.de**

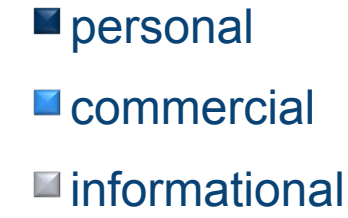
- We **evaluated LFA for both corpus domains** based on the 6 feature groups

  - We used linear multi-class support vector machines in all experiments
  - Text classification often suffers from domain dependency, so we also evaluated out-of-domain classification

- We **splitted the corpus** into training, validation, and test sets





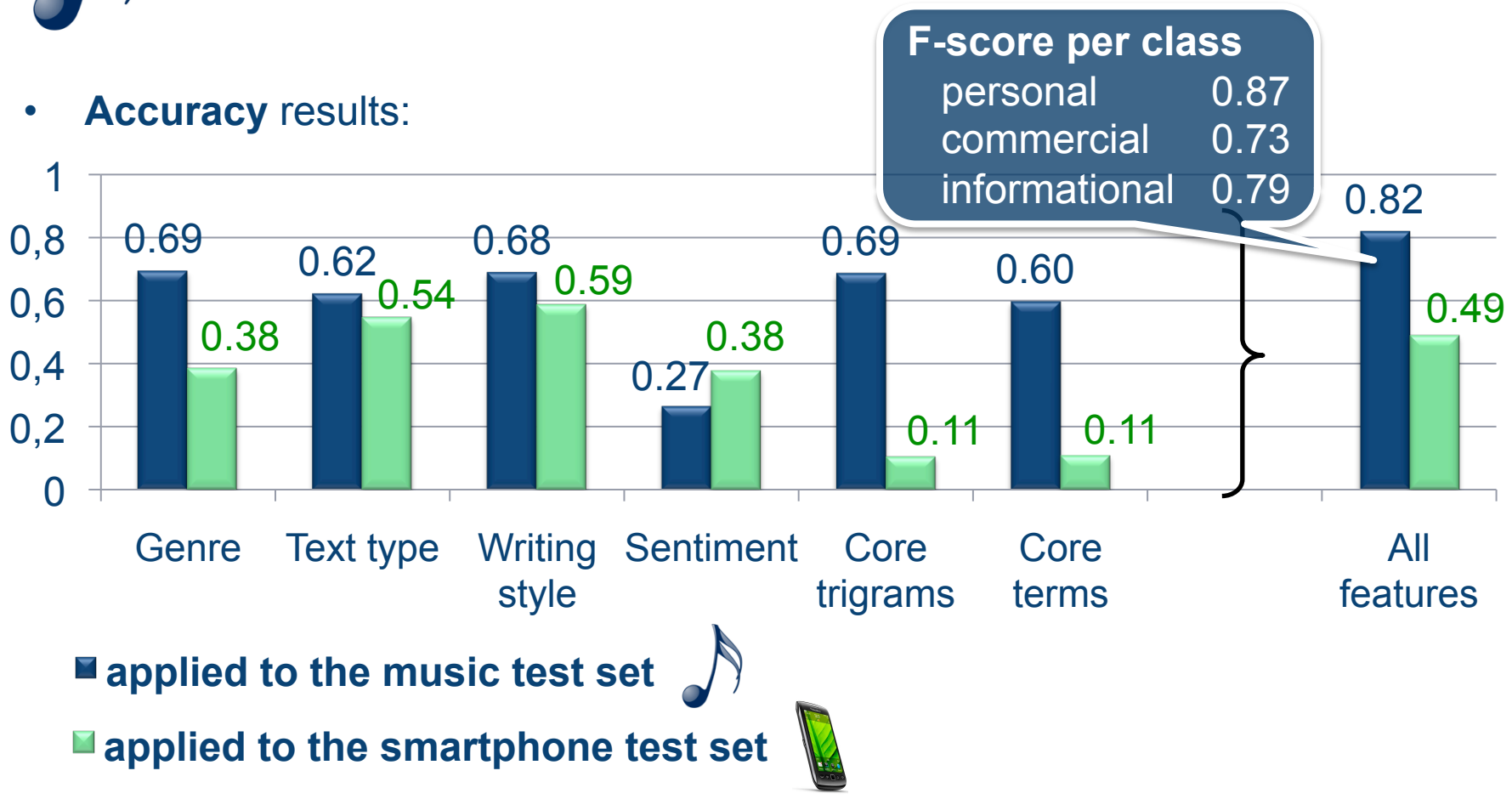


  - Smartphone sets even more imbalanced

Results: From music to smartphones

- We first trained a classifier **on the music training set** for each feature group as well as for all features

- **Accuracy** results:



Results: From smartphones to music

- Next, we retrained the classifiers **on the smartphone training set**

- **Accuracy** results:

F-score per class

personal	0.75
commercial	0.31
informational	0.68

applied to the smartphone test set

applied to the music test set

Key observations

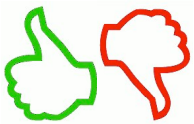
- **Machine learning** appears to work well for LFA on homogeneous collections, such as the music texts

- Classification of very heterogeneous collections as well as of out-of-domain data remain **open problems**

- The **best-performing features** are common in authorship attribution

- **Writing style and text type features** appear to be only weakly domain-dependent in LFA

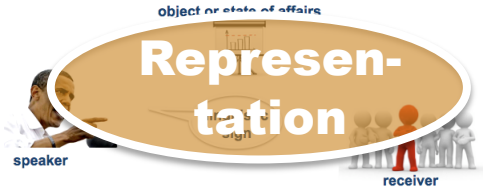
- Language functions and **sentiment polarities** seem to have few correlation



Take away messages

- In our view, we need to go **back to the roots of genres** in order to achieve progress in the field

- We introduced **Language Function Analysis (LFA),** a very general classification task that addresses one single aspect



- It is possible to **determine the predominant language function** of a text using machine learning

- There‘s much room for doing better than us in LFA, so **start working on it** ☺

Thank you for your attention.

s-lab – Software Quality Lab
University of Paderborn

Zukunftsmeile 1, 33102 Paderborn
Germany

http://is.upb.de/?id=wachsmuth
hwachsmuth@s-lab.upb.de