

Reading Between the Lines: Human-centred Classification of Communication Patterns and Intentions

Daniela Stokar von Neuforn

Brandenburg University of Applied Sciences,
Institute Safety and Security, Germany

Katrin Franke

Norwegian Information Security Laboratory (NISlab),
Gjøvik University College, Norway

Outline

- ◆ Background & Motivation
- ◆ State of the Art

- ◆ Research Objectives
- ◆ Experimental Design
- ◆ Results

- ◆ Further Directions & Applications

Application Domain: Forensic Linguistic

Established Part of Forensic Science

Frequently used in:

- Crime Investigation, Prosecution and
- Sentencing of Criminal Offenders.

Provided evidence in:

- Identifying the author of anonymous texts (such as threat letters)
- Identifying cases of plagiarism
- Tracing the ethnic origins

Important Fields are:

- Extortion, Threat,
- Industrial espionage,
- Racism, Terrorism.



Criminal acts in Germany 2006

In Total: 4500 cases involving linguistic analysis

Crime Category	Proportion
Extortion	45%
To form a criminal association	15%
Fire raising	12%
Threat	5%
Murder and attempted murder	5%
Industrial espionage	2%
Incitement of the people	2%
Libel	2%
Falsification of documents	2%
Sexual assault, rape	2%
Computer sabotage	2%
Insult	2%

Source: German Federal Police Office, Dept. of Linguistic and Author Identification

Current Situation in Forensic Labs



- ◆ Knowledge and intuition of the human expert plays a central role in daily forensic casework.
- ◆ Courtroom forensic testimony is often criticized by defense lawyers as lacking a scientific basis.
- ◆ Initiation, Extension and Adaptation of Computer-based Investigation Methods.

Computational Forensics

Computational Forensics - Definition

(Franke, Srihari 2007)

Concerns the investigation of forensic problems using computational methods.

Works towards:

- 1) **In-depth Understanding** of a forensic discipline,
- 2) **Evaluation** of a particular scientific method basis and
- 3) **Systematic Approach** to forensic sciences by applying techniques of computer science, applied mathematics and statistics.

It involves **Modeling** and computer **Simulation (Synthesis)** and/or computer-based **Analysis** and **Recognition**

Objectives of Computational Forensics

- ◆ Study and development of computational methods to
 - Assist in basic and applied research, e.g. to establish or prove the scientific basis of a particular investigative procedure,
 - Support the forensic examiner in their daily casework.
- ◆ Modern crime investigation shall profit from the hybrid-intelligence of humans and machines.



Focus of our Studies

◆ Forensic Linguistic, Author Identification

– Text-based Communication

- Handwritten or Typed Letter,
- Blog, Website,
- Email, Chat

– Verbal Communication

- Face-2-Face,
- Telephone

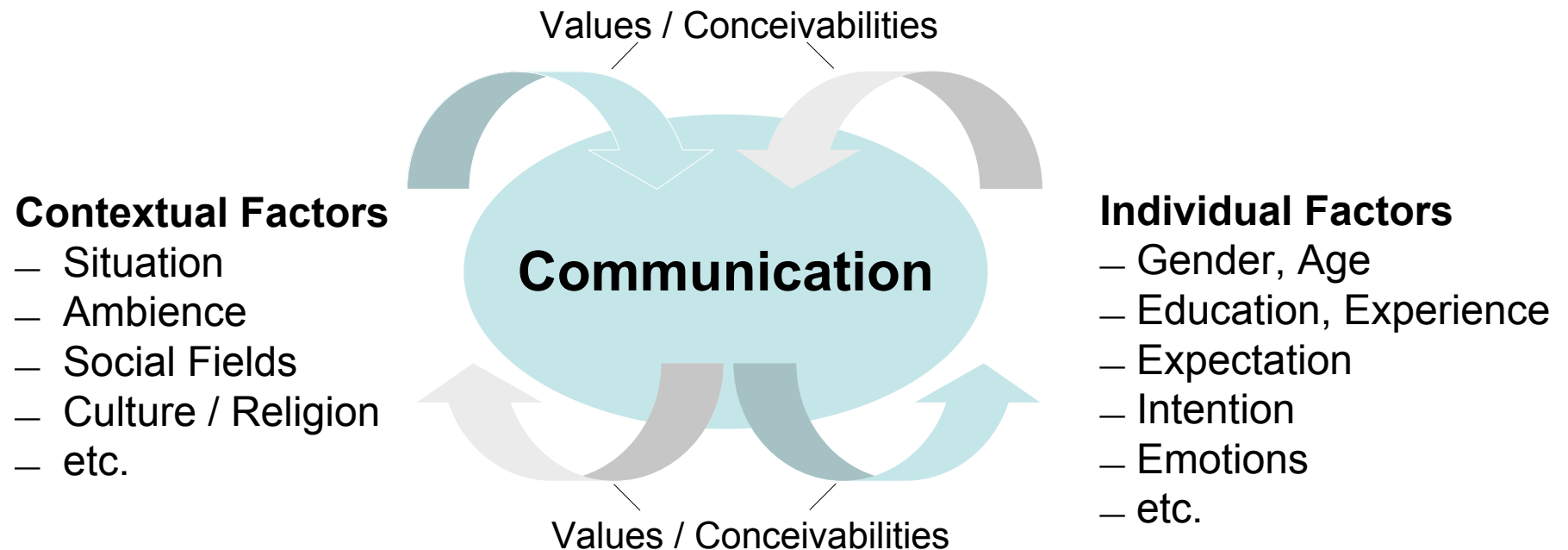
– Non-verbal Communication

- Gesture, Body language



Factors influencing the Production and Reception of communication

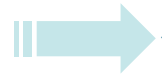
(Bourdieu 1990, Schoenthal 1998)



General Working Procedure (according to forensic experts)

- ◆ Text Analysis
 - Communication pattern / intention (Qualitative)
 - Mistake analysis (Qualitative, Quantitative)
 - Style analysis - characteristics that can be quantified: punctuation marks, orthography, syntax, lexis, structure, and text layout
- ◆ Text Comparison (in case of available reference texts produced by suspected author or target group)
- ◆ Text Collection

Related Work: Computer-based Analysis



- ◆ Ehrhardt, S.: Sprache und Verbrechen – Forensische Linguistik im Bundeskriminalamt, Ringvorlesung zum Jahr der Geisteswissenschaften, Stuttgart, 21.Mai 2007.
- ◆ De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining E-Mail content for author identification forensics. SIMOD Record, 30(4), 2001,55-64.
www.sigmod.org/record/issues/0112/SPECIAL/6.pdf
- ◆ Abbasi, A., Chen, H.: Visualizing Authorship for Identification, Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721, USA,2006
- ◆ Zheng,R., Quin, Y., Huang, Z., Chen, H.: A Framework for Authorship Analysis of Online Messages: Writing-style Features and Techniques. Journal of the American Society for Information Science and Technology 57(3), 2006,378-393.
- ◆ Dark Web Terrorism Research,
<http://ai.arizona.edu/research/terror/index.htm>, Oct. 2007.

Quantitative Analysis only :-(

Our Research

- ◆ Towards qualitative text analysis, compliment qualitative analysis
- ◆ Reading between the lines
 - Inferring communication patterns and intends
- ◆ Establish scientific basis for
 - Individual text characteristics
 - Information carrier
 - Possible interpretations

- ◆ Survey on the reception of text based communication

Survey Situation of the Target Group

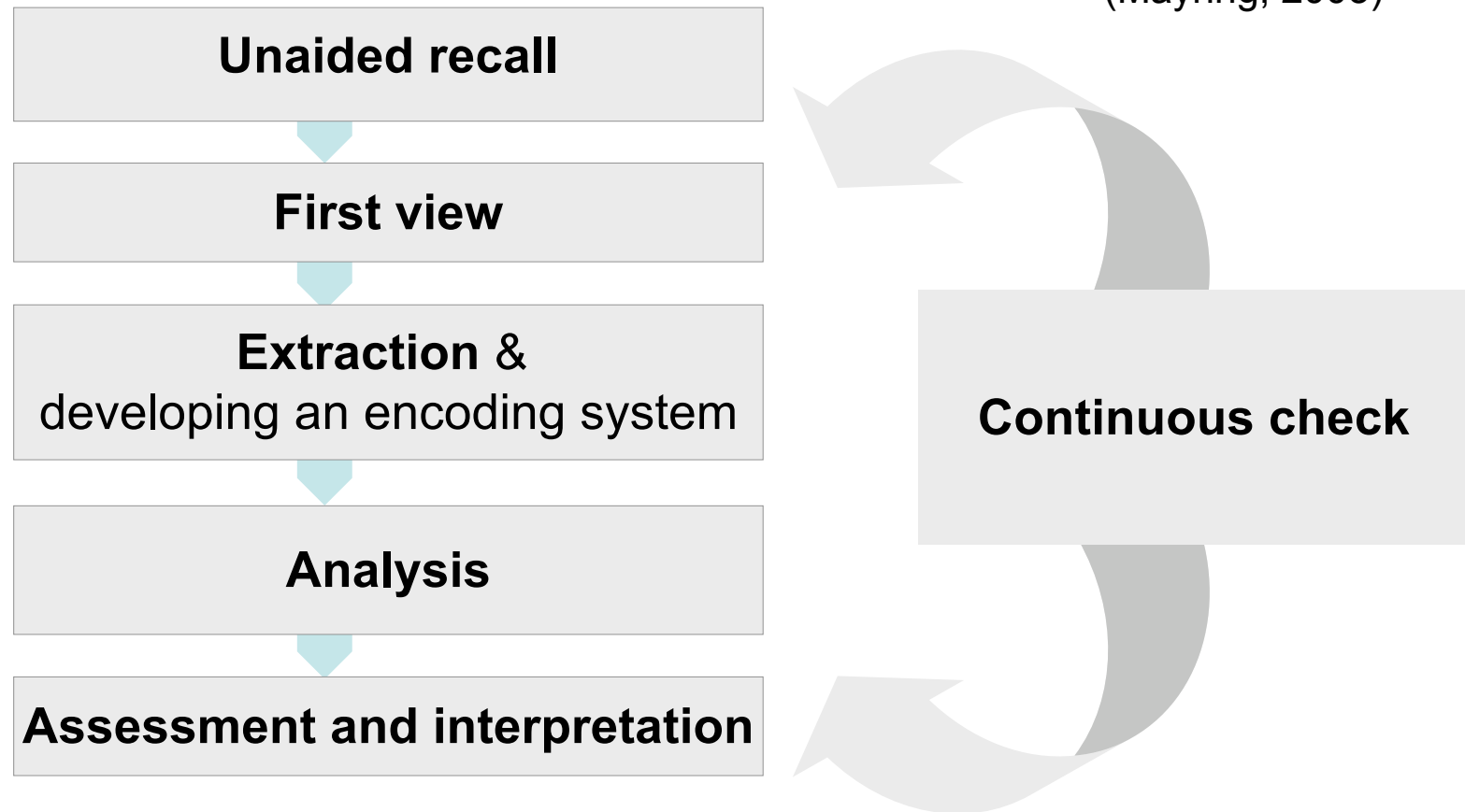
- ◆ Unaided recall at the start of present activities of online-seminars
- ◆ Media: Email
- ◆ Subjects
 - Ethnicity: German
 - 128 female & 128 male students
 - Average 26 years old
 - 5th study semester

The Questionnaire

- ◆ Unaided recall about the individual perception of text-based communication in a questionnaire
- ◆ Questioning
 - **Characteristics**, which convey information about the existential orientation or other qualities of the sender
 - **Examples**, how the characteristics were interpreted

Expiration of the context analysis

(Mayring, 2003)



Evaluation

- ◆ **Identification of characteristics** during the reception of text-based communication
- ◆ **Distribution in language areas** of text-based communication
- ◆ Evaluation of the **examples and interpretations**
- ◆ Number of entries
- ◆ Order of entries

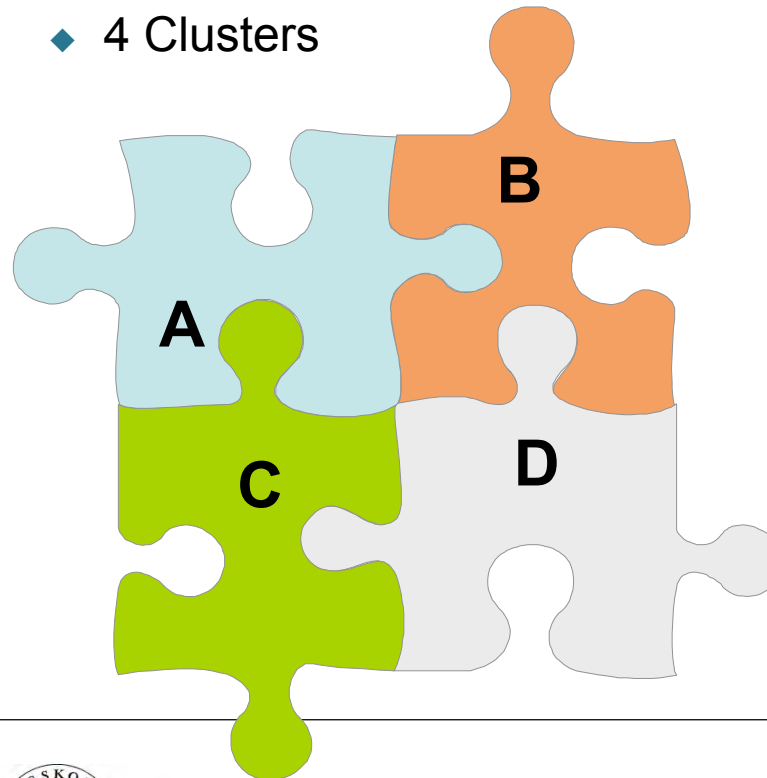
Results: Language Areas

In total

- ◆ 1535 Notations (256 Subjects)
- ◆ 47 Characteristics
- ◆ 4 Clusters

Outside appearance of the message
(14 characteristics)

Vocabulary (use of words)
(15 characteristics)



Syntax (construction of sentences)
(5 characteristics)

Empathetic communication
(13 characteristics)

Results: Identified Text-based Characteristics

Outer appearance of the message

- Address
- Length of the message
- Closure/greeting
- Capital and small writing
- Layout
- Latency of the answers
- Topic
- Pictures, colors, graphics
- Font
- Subject
- Attachments
- Signature
- Answers in the email
- Time

Syntax

- Length of the sentences
- Sentence construction
- Spelling
- Questions
- Grammar

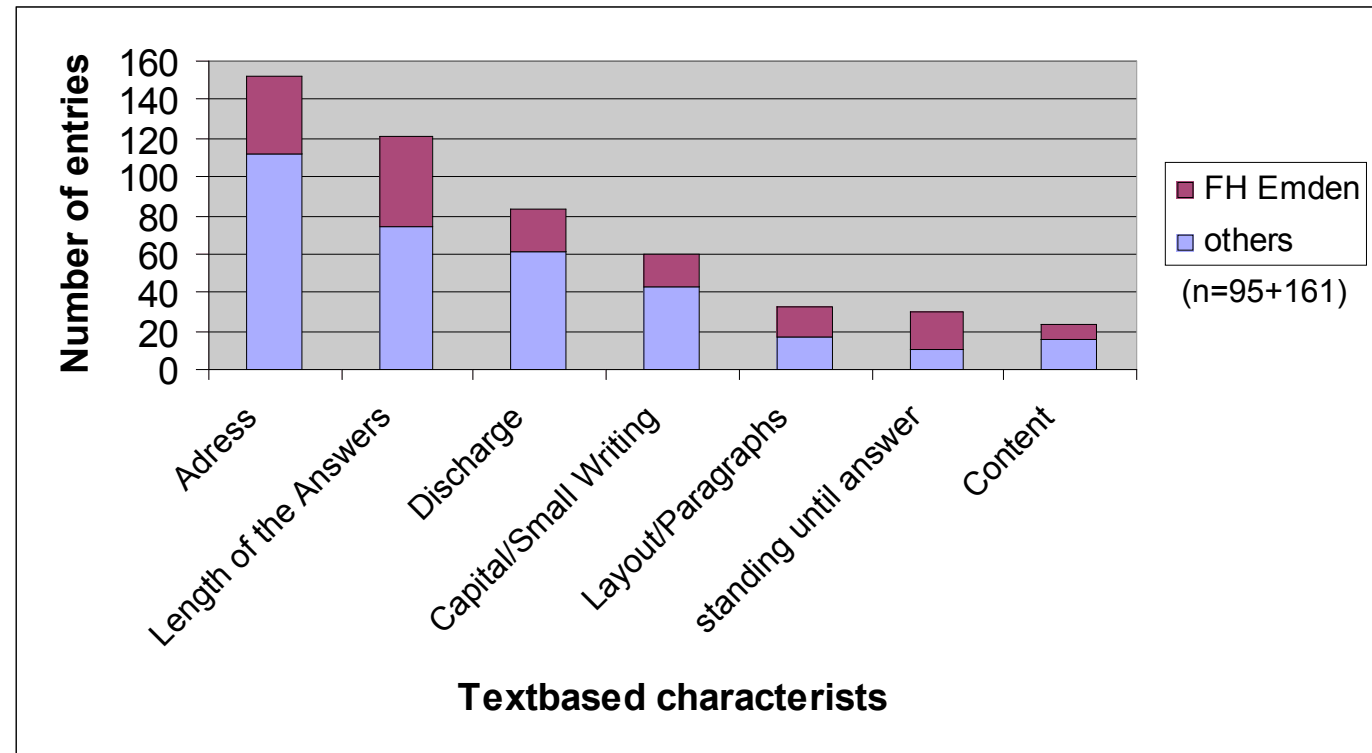
Lexik

- Writing Style
- Formal way of writing
- Colloquial language
- Abbreviations
- Adjectives, Adverbs
- Slang, Chat language
- Extensions
- Foreign words
- Filling words
- Offensive language
- Nicknames
- Vocabulary, Eloquence
- Metaphors
- Repetitions
- Conjunctive

Empathical communication

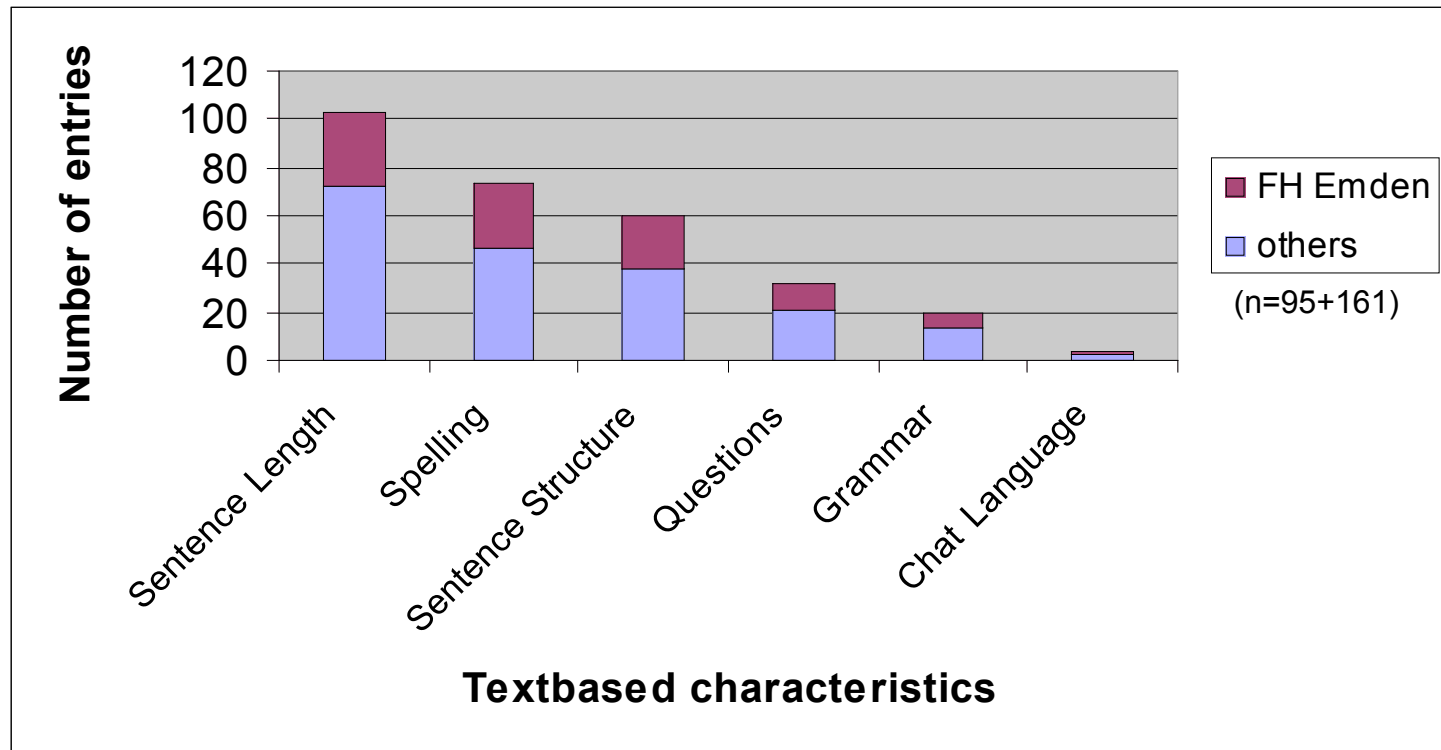
- Emoticons, symbols
- Punctuation marks
- Answers
- Content (only technical information)
- Questions about the well being
- Jokes
- Highlighting
- Compensated phonology
- Formal / informal Address
- Irony
- Personal writing
- Reasoning of the message
- Apologies

Outside Appearance of the Message



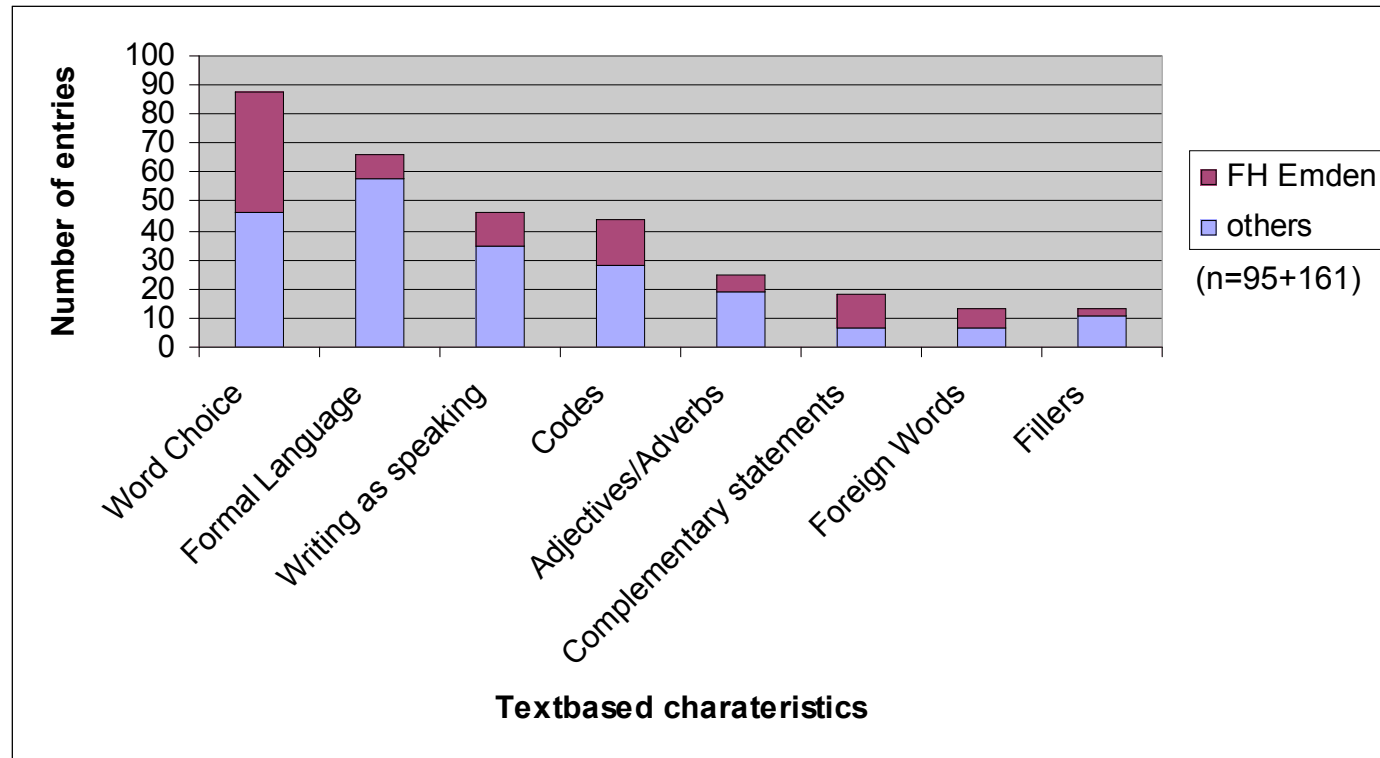


Syntax (Structure of the Sentence)



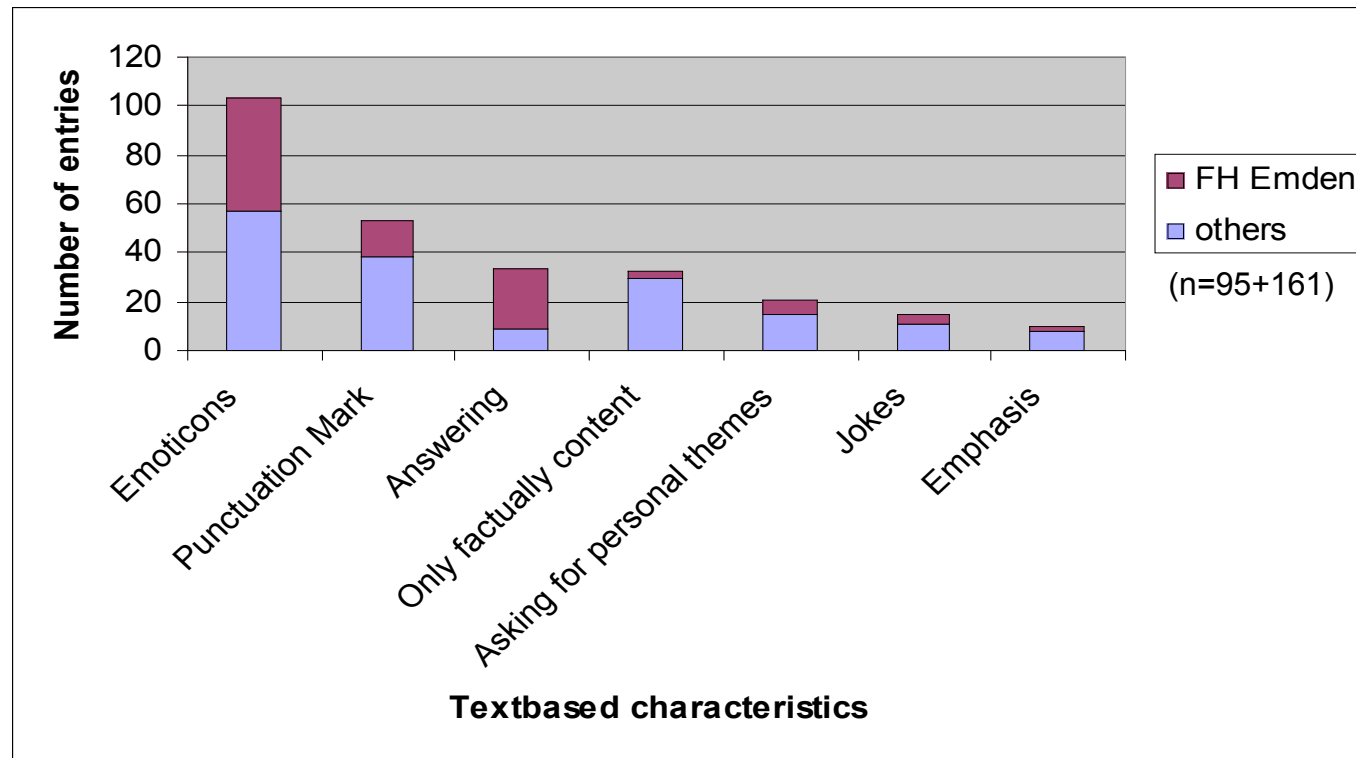


Vocabulary (Use of Words)





Empathetic Communication



Results: Specified Indicators

Atmosphere, Emotions
(63.82%)

Address, dismissal, capital and small writing, length of the message, layout, vocabulary latency, colours and graphics, font, links and attachments, length of the sentence, sentence construction, spelling, questions, grammar, writing style, slang, abbreviations, extensions, filling words, expressions of strength, naming, metaphors, repetitions, similes, punctuation marks, answering of questions, adjectives, questions about the condition, highlighting, compensated phonology, personal writing

Gender (46.80%)

Capital and small writing, length of the message/sentence, colours and graphics, subject heading, sentence construction, writing style, formal writing, abbreviations, extensions, filling words, naming, metaphors, similes, jokes, answering of questions, adjectives, questions about the condition, highlighting, reasoning of the message, personal writing, factual level

Interest (36.17%)

Address, length of the message, dismissal, layout, latency of the answers, links and attachments, answers in the email, sentence length, sentence construction, spelling, questions, extensions, answering of questions, factual level, questions about condition, jokes, reasoning of the message

Relational Aspect
(31.91%)

Address, message length, dismissal, capital and small writing, layout, writing style, formal way of writing, slang, abbreviations, naming, smilies, factual level, jokes, formal/informal address, personal writing



Competence, education
(25.53%)

Length of the message, layout, colours and graphics, links and attachments, sentence construction, spelling, grammar, writing style, slang, extensions, vocabulary

Time (23.40%)

Address, length of the message, layout, latency of the answers, font, links and attachments, answers in the email, sentence length, abbreviations, adjectives, answering of questions

Status, role, social
Background (1.27%)

Address, signature, sentence construction, writing style, formal way of writing, slang, abbreviations, foreign words, punctuation marks

Appreciation (21.27%)

Address, dismissal, layout, latency of the answers, subject heading, sentence construction, questions, formal way of writing, extensions, apologies

Degree of trust/distance
(12.76%)

Address, dismissal, capital and small writing, layout, grammar, personal writing

Taste/style/Interests
(6.38%)

Layout, Font, Slang


Identification of the person (6.38%)

Topic and reason of the message, font,

Conclusions

- ◆ Text-based language contains **information about the relationship** between communication partners
- ◆ Information hidden “between the lines”, which is sent partly consciously (emotions) and partly unconsciously (length of the sentence)
- ◆ Text-based characteristics as the
 - length of a message,
 - the presence of greetings or
 - the answering of questions in short timeare **indicators for the interest and the mood** of the communication partner, and in the same time a **factor of individual perception** of the specific communication situation

Summary & Further Direction

- ◆ Systematic study of individual text characteristics
 - ◆ Derived information carrier
 - ◆ Assigned interpretations to those information
- 
- ◆ Establish computer model for communication pattern & intends
 - ◆ Computer-based, qualitative text analysis
 - Determine communication context and target groups
 - Identify authors

2008 International Workshop on Computational Forensics

National Academy of Sciences:
Keck Center in Washington DC,
August 7-8, 2008

iwcf08.arsforensica.org



References

- [1] Ehrhardt, S.: Sprache und Verbrechen – Forensische Linguistik im Bundeskriminalamt, Ringvorlesung zum Jahr der Geisteswissenschaften, Stuttgart, 21.Mai 2007. (in German).
- [2] Mena, J.: Investigative Data Mining for Security and Criminal Detection, Butterworth Heinemann / Elsevier Science (USA), 2003, ISBN:0-7506-7613-2.
- [3] Stokar von Neuforn, D.: Geschlechtsstereotype Rezeption textbasierter Kommunikation in virtuellen Lernumgebungen, Shaker Verlag GmbH, Aachen, ISBN: 978-3-8322-5778-1, 2006. (in German).
- [4] De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining E-Mail content for author identification forensics. SIMOD Record, 30(4), 2001, pp. 55-64. www.sigmod.org/record/issues/0112/SPECIAL/6.pdf
- [5] Abbasi,A., Chen, H.: Visualizing Authorship for Identification, Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721, USA, 2006.
- [6] Zheng,R., Quin, Y., Huang, Z., Chen, H.: A Framework for Authorship Analysis of Online Messages: Writing-style Features and Techniques. Journal of the American Society for Information Science and Technology 57(3), 2006, pp. 378-393.
- [7] Dark Web Terrorism Research, <http://ai.arizona.edu/research/terror/index.htm>, Oct. 2007.
- [8] Fiske, S.T., Neuberg, S.L.: A continuum of impression formation from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In: Advances in experimental social psychology. Hg. Mark P. Zanna, pp. 1-74. New York: Academic Press, 1990.
- [9] Fiske, S. T., Neuberg, S.L.: The continuum model: Ten years later. In S. Chaiken and Y. Trope (Hrsg.), Dual process theories in social psychology, New York: Guilford, 1999, pp. 231 – 254.
- [10] Mayring, P.: Qualitative Inhaltsanalyse. Grundlagen und Techniken, Weinheim, 1988.
- [11] Lamnek, S.: Qualitative Sozialforschung, Band 2, Methoden und Techniken, p. 202 ff, Weinheim: Belz, Psychologie Verlags Union, 1993.
- [12] Bourdieu, P.: Was heißt sprechen? Die Ökonomie des sprachlichen Tausches. Wien, 1990. (in German).
- [13] Schoenthal, Gisela (1998): Geschlechtstypisches Kommunikationsverhalten: Ergebnisse, Konsequenzen, Perspektiven. In: Feministische Linguistik Linguistische Geschlechterforschung, Hildesheim, 1998 (in German).
- [14] Franke, K., Srihari, S.N.: Computational Forensics: Towards Hybrid-Intelligent Crime Investigation. In Proc. 3rd International Symposium on Information Assurance and Security / Workshop on Computational Forensics, Manchester, UK, 2007, pp. 383 - 386.