Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Medieninformatik

# Authorship Identification with Phonological Features

## Bachelor's Thesis

David Reinartz
Born Feb. 2, 1998 in Ostercappeln

Matriculation Number 3706997

1. Referee: Prof. Dr. Martin Potthast
2. Referee: Prof. Dr. Benno Stein

Submission date: July 31, 2021

# Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, July 31, 2021

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
David Reinartz

### Abstract

# Contents

# Chapter 1

# Introduction

Idea: Using phonetic features for Authorship Verification could be useful. Why? -> Assumption: authors produce different texts based on their phonetic preferences (explain reasoning, course p277) Contributions:
- Unite authorship identification methods with phonetic methods (non-refutable)
- investigate the effect of using phonetic features for authorship identificaiton

We show that there is no effect. Why? -> Either thesis is wrong or an error somewhere else (e.g. phonetic preferences too diverse, etc. look at big open questions / roadblocks)

# Chapter 2

# Theory

### 2.0.1 Authorship Verification

Humans are, by nature, beings deriving the biggest part of their information from eyesight. It comes as no surprise that a large amount of communication and information transfer is done through a visual medium: text. The people, composing their thoughts into texts are called authors of these texts. Over time it has become increasingly simple for the public to produce and distribute content in text form. Smt like: From XXXX onwards, the amount of literate people doubled every X years, until in XXXX XX percent of the population where literate. Writing supplies became cheaper. Letterpress allowed for much faster and wider distribution of text content. More literate people means text can become more important. With text gaining traction as a medium, the concept of authorship of texts also becomes more important. Often, the author of a given text is unknown (intentionally or unintentionally). Example for unknown authorship. List some cases of authorship verification: ... Author Identification is an area in broader field of stylometry, the Stylometry was done by hand. With the adoption of computers by the linguistic community, stylometry was increasingly done with computers. The wide availability of training data and the speed of computers allow for more involved and complex stylometric methods [holmes1998].

Bevendorff et al. [2019] says: "Authorship verification is a young task in the fieldof authorship analysis. Proposed by Koppel andSchler (2004)...".

Authorship Verification falls into the larger area of Author Identification, which aims at determining authors of given texts. The task of Authorship Verification is defined in Bevendorff et al. [2020] as: Given a pair of documents, determine whether they a written by the same author. For example, XXX. Note that we do not consider the actual authors but only whether they are different or not. This also means we cannot use methods that profile individual authors, because the test set might exclusively consist of texts by authors that were not seen before.

XXXX ROUGHLY based on the notation in Bevendorff et al. [2020] the task can be formalized as follows. Given a text pair $(d_1, d_2)$, classify it to $True, False$, i.e., approximate the target function $\phi : (d_1, d_2) \rightarrow \{True, False\}$ where $\phi(d_1, d_2) = True$ iff $d_1$ and $d_2$ have the same author.

### 2.0.2 Phonetic Features

According to O'Grady et al. [2017], phonetics is a branch of linguistics concerned with the inventory and structure of sounds in a given language. As such, we define phonetic features of a given text as those attributes that carry phonetic meaning. => O'Grady p.91 has different definition of this. Phonology is the sounds function in a language.

The input of the algorithms described later is plain text. Therefore, we have to use methods to extract the phonetic features from the text. One possible way of achieving this is with phonetic transcriptions. These are transformations assigning a symbol to each phoneme of a text. Phonetic transcriptions can be seen as data reduction methods. By applying them, we anticipate that the phonetic features stay apparent while other, less relevant features stand out less(?). In total, we use X different transcription systems of different granularity. The narrower a transcription, the more closely it follows the phonetic details of an utterance. This often leads to the system having a bigger alphabet, such as the IPA described below. The broader a transcription, the more it generalizes phonetic features. Table 2.1 shows more information on the properties of these systems, from most narrow to most broad. The most widely used phonetic transcription system is the International Phonetic Alphabet. It was developed by the International Phonetic Association founded in 1886 Association et al. [1999]. With 155 symbols, its alphabet is the largest of the transcription systems considered in this thesis. Therefore, the produced transcriptions are usually the narrowest. For our analysis, we use a slightly broader version of the IPA omitting prosodic markers and diacritics.

Another way to transcribe phonetic texts is by using sound classes. These group certain phonemes together to

CV -> broadest transcriptions

We define the set of "phonetic transcriptions systems" as follows: $t_{phonetic} =$

,And the other one as $t_{other} =$

**Table 2.1:** Statistical properties of the transcription systems used. "Verbatim" represents original English text.

| System | Alphabet size | Vocabulary reduction | Attribute3 | Attribute4 |
|---|---|---|---|---|
| Verbatim | | | | |
| IPA | | | | |
| Dolgo | | | | |
| ASJP | | | | |
| CV | | | | |
| Soundex | | | | |
| RefSoundex | | | | |

Integrating phonetics into stylometric methods is not an entirely new endeavour...

# Chapter 3

# Related Work

Bevendorff et al. [2019] gives related work in Authorship Verification. -> Also list possible biases here! XXX Only keep if used later!
Model Bias
B1: Corpus-relative features, e.g. document frequency -> overfitting
B2: Feature scaling -> Overfitting towards corpus specifics
Data Bias
B3: Plain text heterogeneity, artifacts that are unlikely to signal authorial style,but rather originate from other sources, e.g. features like white spaces which vary across authors but were not necessarily introduced by them. Data sets should be fully homogenized.
B4: Population homogeneity, reusing chunks when creating the corpus might over- / underrepresent certain authors' styles.
B5: Accidental text overlap, named entities / topic words / repetitions / unique character sequences might give same author pairs away, so that algorithms learn these things instead of the wanted patterns.
Evaluation Bias
B6: Test conflation, verifiers can usually access the entire test dataset, this is not how a forensic linguist would do things -> test one case at a time.

# Chapter 4

# Experiments

### 4.0.1 Methods using sub-segmental features

We will start of with a naive approach using sub-segmental features. In a preprocessing step, one of the datasets is standardized. Then, the data is transcribed using the transcription and transformation methods $t_{phonetic}$ and $t_{other}$ defined earlier. The resulting data (together with the original data) are then used as the inputs to two already existing Authorship Verification algorithms. This way we can examine the effect of said transformation methods to the results.

#### Datasets

We use learning-based classification algorithms. This means, given a set of rules, they try to induce the underlying patterns of a training set. The resulting patterns are then used to classify unseen entities of a test set. Each of the two data sets used consists of labeled text pairs. A pair has the label $True$ if both texts were authored by the same person and $False$ if not.

First, we will use the small official dataset from the PAN2020 task on Authorship Verification. This allows us to compare our results to the other methods submitted. It consists of 52.601 text pairs collected from the fanfiction website `fanfiction.net`. The dataset file is formatted in the PAN20 format with each line containing a text pair, an ID, and optionally some additional information such as the corresponding fandoms[1]. With 256.000 samples, the large dataset contains roughly five (4.86) times as many samples as the small data set. Efforts have been made to maximally optimize the methods used, but due to

[1] The franchise a fanfiction text belongs to. Can be seen as the topical domain of the text.

some (out of control) implementation details(?), the utilization of this dataset is infeasible for now. (Also, 53k samples is quite a lot already. And unmasking is reaaaally slow, won't even work on 53k samples.)
We source the second dataset from Bevendorff et al. [2019]. It presents a dataset containing science fiction and adventure texts from the 19th and 20th century, compiled from books from Project Gutenberg [2]. The aim of this dataset is to reduce common biases in data sets for Authorship Verification. This makes it a good candidate for evaluating new authorship verifiers. As it contains only 182(?) text pairs, it is well suited for the much slower Unmasking algorithm, but might lead to overfitting. To mitigate this, we use out-of-fold cross-validation to evaluate the models instead of a standard train-test-split method. This dataset is in the old PAN format[3] and is converted to the new PAN20 format for standardization [4].
To use these transcription systems in our experiments, we first transcribe a given text to IPA. This is done using g2pen because CLTS needs segmented IPA. Then CLTS is used to transcribe to sound classes. For Soundex and RefSoundex the package X is used.

### Compression Approach

The second approach, also introduced in PAN2020 and based on Teahan and Harper [2003], uses a text compression method to determine the chance that two texts were written by the same author. The compression of a text can be seen as encoding said text with a given encoding. Thus, text compression can be used to estimate an upper bound to the entropy, i.e. the amount of information, of characters in English text **??**. More specific, by using the compression model of some text A, the cross-entropy of encoding a text B with this model can be calculated. This approach uses the Prediction by Partical Matching (PPM) model, a standard model for lossless text compression, first introduced by Cleary and Witten [1984]. During training, for each pair, the PPM of the first text is used to encode the second text and vice-versa. In this process, the cross-entropy of the first to the second text can be calculated and vice-versa. In other words, if the compression of one text with the compression model (i.e. the "encoding") of the second text works well, the chance that both are written by the same author can be considered high. The source code used is based on a reimplementation of the Authorship Attribution approach from Teahan and Harper [2003] as part of a reproducibility study in Potthast et al.

[2] https://www.gutenberg.org/

[3] XXX Enter PANXX-PANXX years

[4] The code for the conversion can be accessed at XXX

[2016]. An adaption for Authorship Verification stems from PAN20[5]. The source code extending the algorithm to use phonetic features is available on GitHub[6]. (Code sources are a bit more complicated.)

**Unmasking Approach**

### 4.0.2 Methods using supra-segmental features

Manual annotation of long / short syllables and stress / no-stress syllables (p.86). ARIMA method for time series modelling.
Rhythmical series are easier to model -> goodness of fit   rhythmicity

"All the suprasegmental features are characterized by the fact that they must be described in relation to other items in the same utterance. It is the relative values of pitch, length, or degree of stress of an item that are significant. ... The absolute values are never linguistically important. But they do, of course, convey information about the speaker's age, sex, emotional state, and attitude toward the topic under discussion."Ladefoged and Johnson [2014]

[5] https://github.com/pan-webis-de/pan-code/tree/master/clef20/authorship-verification

[6] ...

# Chapter 5

# Results

As indicated earlier, through cross validation, we can use the entire data set for training and for testing, thus milking it in the most effective way.

# Chapter 6

# Conclusion

Problems:
-

Outlook:
- Improve usability of unmasking, cite DH paper (Juola 2007), got a request for use of unmasking

# Appendix A

# My First Appendix

This was just missing.

# Bibliography

International Phonetic Association, International Phonetic Association Staff, et al. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Cambridge University Press, 1999. 2.0.2

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Bias Analysis and Mitigation in the Evaluation of Authorship Verification. In *Proceedings of ACL 2019*, 2019. 2.0.1, 3, 4.0.1

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. Shared tasks on authorship analysis at pan 2020. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 508–516, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45442-5. 2.0.1

John Cleary and Ian Witten. Data compression using adaptive coding and partial string matching. *IEEE transactions on Communications*, 32(4):396–402, 1984. 4.0.1

Peter Ladefoged and Keith Johnson. *A course in phonetics.* Cengage learning, 2014. 4.0.2

William O'Grady, John Archibald, Mark Aronoff, and Janie Rees-Miller. *Contemporary Linguistics: An Introduction.* Bedford/St. Martin's, 7 edition, 2017. ISBN 1319040896,9781319040895. 2.0.2

Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Benno Stein, Efstathios Stamatatos, and Matthias Hagen. Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016)*, volume 9626 of *Lecture Notes in Computer Science*, pages 393–407, Berlin Heidelberg New York, March 2016. Springer. doi: 10.1007/978-3-319-30671-1_29. 4.0.1

William J Teahan and David J Harper. Using compression-based language models for text categorization. In *Language modeling for information retrieval*, pages 141–165. Springer, 2003. 4.0.1