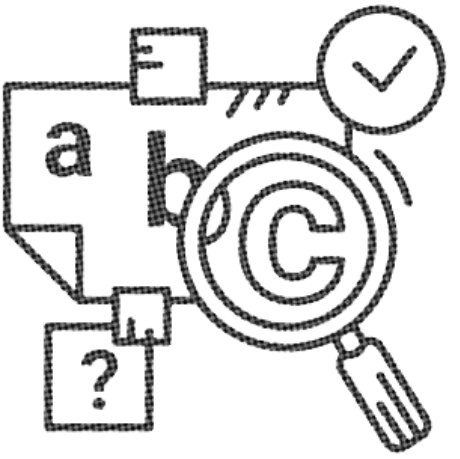
Who wrote that?



Who wrote that?

Original text in German
Wer schreibt denn da?
Zündlumpen #76
2020
zuendlappen.noblogs.org/post/2020/10/07/wer-schreibt-denn-da

Translation and layout
No Trace Project
notrace.how/resources/#who-wrote

Contents

Introduction 3
Author Identification at the BKA 4
Methods of author recognition and author profiling 7
And now what? 9

Introduction

A brief overview of modern forensic linguistics methods for determining authorship.

The following article tries to give an overview from a non-technical perspective and to make a corresponding evaluation. There are some academic publications on this topic that could be evaluated for a better assessment. However, my main purpose here is just to raise the issue, not to provide a sound and conclusive view so if you know anything more, publish it!

Avoiding traces that could be your undoing down the road—perhaps even after years or decades—is probably of interest to most people who occasionally commit a crime and come into conflict with the law. Avoiding fingerprints, avoiding DNA traces, avoiding shoe prints and textile fiber traces or at least disposing of clothing afterwards, avoiding surveillance cameras, avoiding tool traces, avoiding recordings of any kind, recognizing surveillance, etc.—all this should be a concern for anyone who commits crimes from time to time and wants to protect themselves from identification. But what about those traces that often arise only after a crime has been committed, out of the urge to explain one's deed anonymously or even by using a recurring pseudonym? When writing and publishing a communiqué?

My impression is that in many cases no special attention is paid to these traces despite a rapid technological development of analytical capacities.

This may be intentional, negligent, or a compromise of competing needs. Without wishing to make a general suggestion here on how to deal with these traces—after all, everyone must determine that for themselves—I would like to outline the methods the investigative authorities in Germany and elsewhere are currently (probably) working with, what seems possible in theory, and what could become possible in the future.

Perhaps I should note in advance that everything or at least most of what I present here is scientifically as well as legally controversial. I am also less interested in the legal validity of linguistic analyses—and not in the scientific one either—than in whether it seems plausible that these investigations could guide a surveillance effort, because even if a trail is not useful in court by itself, it could still lead to other, useful trails.

Author Identification at the BKA

According to its own information, the Federal Criminal Police Office (BKA) maintains a department dedicated to identifying the authors of texts. The focus is on texts related to criminal acts, such as responsibility claims, but also "position papers" from the "left-wing extremist spectrum," among others. All collected texts are processed by linguistic studies in a so-called collection of communiques and can be compared and searched with the Criminal Information System for Texts (KISTE). According to the BKA, the texts are

classified according to the following biographical characteristics of their (alleged) authors: origin, age, education and occupation.

All incoming texts are also compared with previously saved texts to determine whether several texts may have been written by the same author.

In the context of case-specific investigations, the stored texts can also be compared with texts whose authorship is known in order to determine whether they were written by the same author or whether this can be ruled out.

This is the official information from the BKA about this department. What does this mean in practice?

I think that one can assume that at least all responsibility claims are recorded in this database and analyzed to see whether there are other responsibility claims by the same author(s). The finding that they also record "position papers" allows us to draw further conclusions: at the very least, it seems possible that in addition to texts with criminal relevance, they also store other texts that are thought to come from a particular scene. For example, texts from newspapers, statements from political groups/organizations, calls, blog posts, etc. In the worst case, I would assume that all published texts on known "left-wing extremist" websites (after all, it is quite easy to get hold of them), as well as texts from print publications that appear interesting to the investigating authorities, would be fed into this database.

This would mean that for each responsibility claim, the BKA would have a cluster of texts that they presume to have the same author. These can consist of other claims as well as texts that have been fed into the database. In addition to series of crimes, further clues to perpetrators can be obtained, such as pseudonyms, group names—or, in the worst case, names—under which an author of a claim may have written other texts, but also, depending on the text, all kinds of other information that it provides, often including clues to a person's place of residence and activity, thematic focus, biographical characteristics, educational background, etc. All of this information can at the very least be used to narrow down the circle of suspects.

What remains unclear in all of this is what other comparison samples the BKA might obtain. For most people, there is certainly a whole series of texts to which investigating authorities (could) have access and which could be fed into the database in the event of suspicion or possibly also partly as a precaution—if a person is on file with an entry such as "violent left-wing extremist", etc. This could be anything with your name under it, from a letter to an authority to a letter to the editor in the newspaper. I will intentionally name only the most obvious sources here, so as not to inadvertently provide the investigating authorities with decisive inspiration, but I'm sure you can answer for yourself which texts of yours might be accessible. If the profilers of the BKA succeed in narrowing down the circle of suspects to a specific characteristic, which allows the comparison with masses of available

text samples (for example, if it is assumed that a scientist of a certain discipline is responsible for a letter, all publications in this field could be used as comparison samples). This would, for example, be a possible (partial) explanation for how it might have gone with Andrej Holm in the case against the militante gruppe (mg), at least if one assumes that the BKA did not just Google "gentrification", so I think it is quite possible that such analyses are also carried out.

Methods of author recognition and author profiling

All this, however, only considers what the BKA claims to be able to do and takes these considerations to some logical conclusions. But how does author recognition or author profiling actually work?

Who hasn't felt the fear that maybe the German teacher will expose you after a mocking poem about a teacher appeared in the washrooms and the whole school is making fun of how only you could have written "vacuum" [*Leerer*] instead of "teacher" [*Lehrer*]. Fortunately, the entire German faculty fell for it, adopting the narrative of a spelling mistake and turning a blind eye to the all-too-accurate pun. Forensic linguistics does seem to require a bit of practice, or at least a criminological motivation, who knows. In any case, error analysis, which most have probably heard of, was one of the BKA's most important analysis tools around 2002

along with style analysis, according to a promotional article by language cop Christa Baldauf. Spelling mistakes, grammatical errors, punctuation, but also typos, new or old spelling, hints on keyboard peculiarities, etc., all this serves the language cops to collect clues about the author. For example, if I write "muß" instead of "muss", that could be a clue that I missed some of the more recent spelling reforms when I was in school. If, on the other hand, I constantly write terms that, according to spelling rules, use "ß" and not "ss", it could mean that there is no "ß" on my keyboard. For example, if I speak of "dem Butter" [*rather than "die Butter"*], it could be a reference to the fact that I grew up in Bavaria, etc. But I could also be faking all these things just to mislead the language cops. The plausibility of my error profile, is also part of such an analysis. Similarly, stylistic analysis examines peculiarities of my writing style. What kind of terms do I use, does my sentence structure show specific patterns, are there repeated constellations of terms that may even appear in different texts, etc.? I think everyone who takes a closer look at his or her texts will recognize some stylistic characteristics of their own.

Such qualitative analyses primarily serves to profile the authors. While it is certainly possible to match different texts in this way, the real value of such analyses lies in being able to determine things like age, "level of education", "scene affiliation", regional origins, and sometimes perhaps even indications of occupation/training, etc. Attempts to determine things like gender are also heard of, but generally do not seem to be quite as straightforward.

In contrast, there are also more quantitative and statistical analyses that examine everything from word frequencies to word constellations to syntax sentence structure that can be measured in this way. These methods, known as stylometry, are sometimes very controversial because it is not possible to say exactly what they are meant to measure, but they sometimes deliver astonishing results, especially in combination with machine learning approaches. I think that these approaches are therefore likely to be used primarily to cluster different texts according to their similarities.

The clear advantage of such quantitative analyses is that they can be performed en masse. All digitally available or digitizable texts can be analyzed in this way. From social media posts to books, texts can be captured using these methods. Although the success of these methods is currently still relatively modest, and it has often turned out that supposedly similar texts are often more similar in their genre than in their authorship, if one assumes that individual writing styles could certainly leave behind quantitative patterns, this means that once these patterns are known, a mass assignment of texts to certain authors will be possible.

And now what?

There were and are, of course, various approaches to dealing with this knowledge, one not better or worse than another. Those who do not write communiqués anyway largely avoid this problem, but are still

affected by the problem of participation in publications and authorship of other texts. Whoever obscures texts before publication, for example, by having several people successively rewrite and rephrase passages from them, etc., runs the risk of also developing exploitable linguistic and stylistic characteristics in repeatedly similar constellations or also of failing to successfully conceal characteristics. Whoever thinks that they can dismiss the whole thing because none of their text samples are available or also because they are convinced that the legal value of author recognition is too shaky, risks that in the future text samples might somehow be available (for example because they are successfully convicted of authorship) or the legal assessment of the procedure changes. Those who trust that technology is not (yet) good enough may be surprised by future developments. Those who use technical solutions to obscure their authorship run the risk of leaving new characteristics and traces, and also of producing poorly written communiqués that no one wants to read anyway. If you never write any texts regardless, you just don't write any texts.

So do whatever appeals to you most, but do it from now on—if you haven't already—keeping these traces in mind and the queasy feeling in your stomach, which is said to have saved many a person from making a careless mistake at the crucial moment.

A brief overview of modern forensic linguistics methods for determining authorship. The following article tries to give an overview from a non-technical perspective and to make a corresponding evaluation.

No Trace Project / No trace, no case. A collection of tools to help anarchists and other rebels **understand** the capabilities of their enemies, **undermine** surveillance efforts, and ultimately **act** without getting caught.

Depending on your context, possession of certain documents may be criminalized or attract unwanted attention. Be careful about what zines you print and where you store them.