ECPAT Deutschland e.V.
Arbeitsgemeinschaft zum Schutz der Kinder vor sexueller Ausbeutung

Explorative Studie

Forensische Linguistik – Sprachanalyse in Darknet-Foren zu sexuellem Missbrauch und Ausbeutung von Kindern

Verfasser: Maximilian Fuß M.A.
Redaktion: Dr. Dorothea Czarnecki, ECPAT Deutschland e.V., info@ecpat.de , www.ecpat.de
24.09.2020

I. Forensische Linguistik, ihre Methoden und Leistungen

Forensische Linguistik kommt als sprachwissenschaftliche Untersuchungsmethode überall dort zum Einsatz, wo kriminalistische Hintergründe von Textproduktion vermutet werden. Bekannt ist ihr Einsatz in Erpresserbriefen oder Bekennerschreiben. Weniger üblich ist ihre Anwendung im Themenfeld der sexuellen Ausbeutung und sexualisierter Gewalt gegen Kinder online, wenn in Foreneinträgen oder Chats die Autorschaft entsprechender Beiträge (*Posts*) beziehungsweise deren Inhalte in Frage stehen oder Aussagen zur Autorschaft getroffen werden müssen.

Forensische Linguistik bietet Methoden und Verfahrensweisen zur Darstellung und Aufklärung unterschiedlicher Aspekte des fraglichen Textes. Sie leistet dies im Hinblick auf die Produktionssituation, den Autor[1] und dessen Charakterisierung anhand strukturell identifizierbarer Marker, also denjenigen vom Verfasser unterbewusst in die Textproduktion integrierten auffälligen linguistischen Merkmale, im Text. Im Allgemeinen spricht man diesbezüglich von der sogenannten *Autorenerkennung*, wobei Analysen mit dem Ziel einer möglichst genauen Beschreibung eines Autors gemeint sind. Die Autorenerkennung umfasst in erster Linie die Ergebnisse von Textanalyse und Textvergleich, zwei zentrale Elemente der forensischen Linguistik. Doch was umfasst der Begriff Text in diesem Kontext?

Definition: Text und Gattungsart

Der Begriff Text in seiner allgemeinen Form umfasst jegliche Art von schriftlich oder mündlich produzierter Sprache (vgl. Brinker 2014:17). Konkret geht es für die forensische Linguistik in erster Linie um die Analyse schriftlicher Textproduktion. Texte werden in unterschiedliche Kategorien, sogenannte Textgattungen, aufgeteilt. Natürlich nimmt auch der Autor stets Einfluss auf die Erscheinung seines Textes und dessen Einordnung in die unterschiedlichen Gattungen. Eine Unterscheidung zwischen Gattungsmerkmalen und individuellen Einflüssen des Autors ist jedoch nicht in jedem Text zu erwarten (vgl. Brinker 2014: 139). So ist im digitalen Raum die Frage nach einer Textsorte in der Regel nicht leicht zu beantworten und kann in den meisten Fällen mit Hinblick auf Foren oder Chats vernachlässigt werden. Dennoch ist es nicht unerheblich, sich über die grundlegenden Komponenten des Textbegriffs und dessen Umsetzung im Klaren zu sein, denn: Auffälligkeiten bei der Umset-

[1] Der Begriff Autor und Verfasser beschreibt schriftlich textproduzierende Personen im forensisch-linguistischen Kontext und stellt keine Einordnung hinsichtlich des tatsächlichen Geschlechts dar.

zung von Textproduktion, auch im digitalen Raum, sind jederzeit hinsichtlich einer möglichen Autorencharakterisierung aufschlussreich. Im forensisch-linguistischen Kontext sind fünf Punkte zu beachten: Welches Medium nutzt der Autor (*Textmedium*)? Wie drückt er sich aus (*sprachliches Mittel*)? Was ist das Thema des Textes (*Thema*)? Welche Intention verfolgt der Autor (*Funktion*)? In welchem sozialen Kontext bewegt sich der Text (*Handlungsbereich und Kommunikationssituation*)?

- *Das Medium*

Die wichtigsten Textmedien für die forensische Linguistik im Kontext sexualisierter Gewalt gegen Kinder und Jugendliche sind SMS, WhatsApp, Signal, Telegram oder ähnliche, Chatnachrichten, in den gängigen Chatmedien wie Skype oder über Dienste wie das PlayStation Network, aber auch die Chatfunktion in Videospielen, wie beispielweise Fortnite oder Minecraft. Für den Austausch über Missbrauchstaten oder -fantasien zwischen Angehörigen der hier analyserelevanten Tätergruppen sind insbesondere und Forenbeiträge relevant. Die Textkonzeption des verwendeten Mediums ist ausschlaggebend, also ob dieses tendenziell mündlich oder vermehrt schriftlich intendiert ist. Dabei ist die tatsächliche Umsetzung nicht zwangsweise die eigentliche Konzeption. Manche Medien, wie zum Beispiel Chatnachrichten, sind von der Sprachverwendung her mündlich konzipiert und bringen u.a. Dialekt oder unvollständigen Satzbau mit sich, obwohl das Medium eine schriftliche Umsetzung voraussetzt. Dies ist auch in der entgegengesetzten Richtung zu beobachten, zum Beispiel bei Reden, die zwar mündlich vorgetragen, jedoch in ihrer Sprachgestaltung schriftlich konzipiert wurden. Für eine forensische Untersuchung ist dieser Sachverhalt insofern von Belang, als dass aus dem Medium Rückschlüsse auf den Autor, nicht jedoch seinen persönlichen Einfluss auf die Textproduktion gezogen werden können (vgl. Dern 2009: 37).

- *Sprachliches Mittel*

Korrelierend zum gewählten Medium stehen die verwendeten sprachlichen Mittel, also die sprachliche Ausdrucksweise. Im Normalfall berücksichtigt der Autor Formulierungsmuster bestimmter Textsorten und Medien, um bei seinen Kommunikationspartnern auch bestmöglich verstanden zu werden. Eine (Nicht-) Einhaltung dieser sprachlichen Mittel sind für eine Analyse zu berücksichtigen (vgl. Dern 2009: 38).

- *Textthema*

Die Wortwahl und Stilebene einer Textproduktion sind oftmals vom gewählten Thema des Textes abhängig, ein Kondolenzschreiben stellt andere Erfordernisse an den Autor als es eine Urlaubskarte, oder ein Forumsbeitrag tut. Bedingt durch den thematischen Rahmen des Schreibens an sich ist die linguistische Entfaltung des Autors durch diesen Rahmen begrenzt. Dabei ist nicht nur die sprachliche Gestaltung des Themas zu berücksichtigen, sondern auch die Ausführung des resultierenden Textthemas, die sogenannte thematische Entfaltung. Diese kann auf argumentativer, deskriptiver, explikativer oder narrativer Ebene geschehen (vgl. Brinker 2014: 60ff.).

- *Textfunktion*

Jeder Text hat eine Funktion: Das Apell, die Deklaration, die Information, der Kontakt und die Obligation (vgl. Brinker 2014 : 106ff.), wobei in der Regel Mischungen zu erwarten sind. Die Funktion, verstanden als Intention eines Textes, bringt gewisse sprachliche Strukturen mit sich: Imperativkonstruktionen in Erpresserschreiben, vulgäre Ausdrücke in Schmähschreiben.

- *Handlungsbereich und Kommunikationssituation*

Der Handlungsbereich umschreibt einen bestimmten sozialen Sektor, für den Aktions-und Bewertungsnormen angewendet werden, sei es wissenschaftlich, alltäglich, rechtlich oder künstlerisch. Diese Bereiche werden weiter in offiziell, öffentlich oder private Gebiete unterschieden. Dieses Phänomen kann auch leicht an sich selbst beobachtet werden, in der Regel sprechen wir mit unseren Vorgesetzten anders, als wir dies mit unseren Freunden tun.

Definition: Autorenerkennung

Autorenerkennung bezeichnet die linguistische Bewertung in Frage stehender "schriftsprachlicher Texte in forensischen, kriminalistischen oder sonst einer Form sicherheitsrelevanten Kontexten" (Dern: 2009:19). Konkret ist dabei die Rede von einer Charakterisierung des Autors mittels Textanalyse und Textvergleich, auf deren genauen Prozeduren und Verfahrensweisen später genauer eingegangen wird. Obwohl der Begriff der Autorenerkennung eine eindeutige Identifizierung eines Autors impliziert, ist diese nicht in dieser Deutlichkeit zu erwarten. Lediglich über einen Textvergleich mit einem bekannten Autor lässt sich eine solche Charakterisierung mit einer gewissen Wahrscheinlichkeit erlangen.

Die Suche nach Auffälligkeiten

Jede Textanalyse sucht nach Auffälligkeiten, die sich auf situativ unterschiedliche Sprachverwendung zurückführen lassen. Dies kann sich darauf beziehen, in wie weit der Text von einer schriftlichen Hochsprache abweicht (*mediale Variation*), welcher öffentliche Bereich der soziale Kontext des Autors womöglich ist (*Fachlichkeit und soziale Sphäre*), welche *Gruppenzugehörigkeit* und welcher *Dialekt* sich herauslesen lässt und welcher Altersgruppe der Autor angehört (*Alter*). Dabei ist jedoch der auch der immer existente *Einfluss des Individuums* nicht zu vernachlässigen.

- *Mediale Variation*

Mediale Variation bedeutet, ähnlich der oben dargestellten Unterschiede im sprachlichen Ausdruck, die generelle Sprachnutzung in Abhängigkeit des gewählten Mediums, z.B. schriftlich vs. mündlich. Um überhaupt Variation entstehen zu lassen, muss zunächst ein Standard vorherrschen, von dem sich ein Autor sprachlich unterscheiden kann. Dieser Standard wird als Standardsprache über die gesamte Sprachproduktion verstanden und untergliedert sich weiterführend in Hochsprache für die schriftliche Norm und Umgangssprache (vgl. Jäger 1980:376). Für die Analyse ist diese Norm von Bedeutung, weil der Grad ihrer Einhaltung bereits Rückschlüsse auf die Textproduktionserfahrung des Autors zulässt.

- *Fachlichkeit und soziale Sphäre*

Auch als Funktiolekt bekannt ist die sprachliche Veränderung innerhalb einer sozialen Sphäre, also demjenigen öffentlichen Bereich, in dem sich ein Sprecher zu gegebenem Anlass bewegt. Nicht zu verwechseln ist soziale Sphäre mit gesellschaftlicher Schicht. Es gibt unterschiedliche soziale Sphären, die ein Sprecher bei Bedarf ad hoc wechseln kann: Beruf, Politik, Justiz, Journalismus, und Wissenschaft. Jeder Sphäre ist ein eigenes sprachliches Verhalten zuzuschreiben, welches einen unterscheidbaren Funktiolekt darstellt. Ein Beispiel für einen Funktiolekt ist der Gebrauch von sogenannter Fachsprache, welche die Kommunikation zwischen Angehörigen dieser sozialen Sphäre verein-

facht und präzisiert. Eine fachliche Zugehörigkeit lässt sich in der Regel einerseits durch die zugrundeliegende Bildung und andererseits anhand des Gebrauchs von fachspezifischer Terminologie determinieren.[2] Dabei ist dennoch stets zu berücksichtigen, dass manche Fachsprachen, z.B. aus dem Umfeld der Informatik, durchaus bereits in die Allgemeinsprache durchgesickert sind. Vorteilhaft ist, dass dabei bestimmte Begriffe umgangssprachlich möglicherweise nicht korrekt verwendet werden, eher ein Indiz für einen geneigten Laien. Im Rahmen dieser Studie ist unter anderem zu ergründen ob sich ein bestimmter Funktiolekt aus pädophilem Sprachgebrauch ableiten, ob sich also ein "Pädolekt" nachweisen lässt.

- *Gruppenzugehörigkeit*

Ähnlich zur sozialen Sphäre ist die soziale Gruppe, hauptsächlich zu unterscheiden sind beide durch die Flexibilität. Wo soziale Sphären dynamisch und situativ bedingt sind, demnach also auch an wechselnde Situationen angepasst werden können, sind soziale Gruppen starre Gebilde, die sich nicht ohne weiteres hinsichtlich ihrer sprachlichen Eigenheiten verändern. Insbesondere ideologische Eigenheiten sind markant für soziale Gruppen, was sich auch in der Ausdrucksweise des einzelnen Sprechers niederschlägt (vgl. Dern 2009 : 48f). An dieser Stelle stellt sich die Frage, ob und in wie weit eine pädophile Gemeinschaft ausreichend ideologische Gemeinsamkeiten aufweist, um als eine soziale Gruppe mit eigener Lexik zu fungieren.

- *Räumliche Varianzen (Dialekte)*

Bereits mehrfach ist auf den mündlichen Einfluss auf schriftliche Sprachproduktion zu sprechen gekommen. Ein wichtiges Element in diesem Zusammenhang sind räumliche Varianzen im Sprachgebrauch, auch als Dialekte bekannt. Dialekte erlauben eine recht grobe Lokalisierung des vom Autors genutzten sprachlichen Verhaltens und lassen Rückschlüsse seinen möglichen Bildungsstand zu (vgl. Dern 2009 : 46). Dies hat zumindest besonders für schriftliche Textgattungen Gültigkeit. In konzeptionell mündlichen, aber schriftlich realisierten Kommunikationsbereichen, wie es beispielsweise Chats sind, lässt sich der Rückschluss auf einen etwaigen Bildungsstand des Autors aufgrund der Dialektverwendung nicht ohne weiteres durchführen. Da jedoch Foren und Chats durchaus anfällig für das Einbeziehen von Dialekten sein können, besteht die Chance einer zumindest groben Einteilung der Sprecher, die deutschsprachige Pädophilen-Foren und Chats als Plattformen aktiv nutzen. Anhand bestimmter Lexik lässt sich zum Beispiel recht eindeutig zwischen Deutschland und Österreich unterscheiden, insbesondere was Begriffe wie Januar – Jänner, Kartoffeln – Erdäpfel etc. betrifft.

- *Alter*

Sprache unterliegt einem stetigen Wandel, begründet durch die Zeit in der sie sich bewegt und deren Stand sie zu entsprechen hat. Sprachwandel und Evolution entsteht fortlaufend durch das Bedürfnis der Sprecher, sich einerseits bequem und schnell auszudrücken (sprachliche Kompaktheit) und andererseits auch bestmöglich verstanden zu werden (sprachliche Deutlichkeit). Auch finden künstliche Einflussnahmen auf die Sprache statt, beispielsweise durch eine Rechtschreibreform. Stringentes Einhalten solcher, durch künstliche Eingriffe erzeugter Marker kann das Rück-

[2] Es sollte sich ein homogenes Bild ergeben. Höhere Bildung geht in der Regel mit qualifizierten Tätigkeiten einher.

schließen auf ein ungefähres Alter, in manchen Fällen auch auf eine gewisse ideologische Gesinnung erlauben. Eine fehlerfreie Umsetzung von neueren Rechtschreibreformen lässt eine recht präzise Aussage zu minimalem oder maximalem Autorenalter zu, wobei textproduktive Berufe auszuklammern sind. Davon ausgehend, dass die Rechtschreibreform in der Schule angeeignet und symptomatisch-intuitiv verwendet wird, lässt sich anhand des Einführungsdatums der Rechtschreibreform ein recht präzises Autorenalter deduzieren[3].

Eine grobe Einteilung des Autorenalters kann auch über den natürlichen Sprachgebrauch determiniert werden. Eine saloppe Wortwahl und informelle Textgestaltung lassen eher auf einen jugendlichen Autor schließen. Eine antiquierte Wortwahl und erhabenere Ausdrucksweise hingegen deuten auf ältere Erwachsene hin. So geben beispielsweise Foreneinträge, deren Aufmachung einem förmlichen Brief ähneln, sicherlich auch Hinweise auf das Autorenalter.

- *Einfluss des Individuums*

Nicht nur äußere Einflüsse ermöglichen eine Sprachvariation in Texten, sondern auch der Autor selbst, entweder schriftlich, mündlich, spontan gebildet oder mit klarer Intention gestaltet. Erfahrungen, Bildung, Vorlieben und kognitive Fähigkeiten des Sprachaktanten fließen in dessen Sprachgebrauch ein und bilden den Idiolekt, das individuelle, unikale sprachliche Potenzial eines Sprechers, auf dem die Theorie eines Individualstils basiert (vgl. Dittmar 1997: 181f.). Der Idiolekt wird zusätzlich durch soziale und psychologische Parameter ergänzt, beispielsweise durch den Einfluss von Mehrsprachlichkeit.

- *Muttersprachlichkeit*

Ein bestimmtes Kompetenzniveau[4] noch nicht vorausgesetzt, lassen sich die meisten Fremdsprachensprecher des Deutschen von denjenigen unterscheiden, welche als Muttersprachler aufgewachsen sind. Steht die Muttersprachlichkeit des Autors in Frage, empfiehlt sich, systematisch nach typischen Fehlern von Fremdsprachen-Nutzern zu suchen, u.a. Auslassungen, Hinzufügungen, Selektionsfehler, Anordnungsfehler und Kontaminationen (vgl. Fobbe 2011 : 155), wobei die genaueren Fehlerausprägungen sicherlich auch mit der eigentlichen Muttersprache des Textproduzentens zusammenhängen[5]. Die Feststellung der eigentlichen Muttersprache des Verfassers übersteigt allerdings die Kompetenz der Textanalyse.

- *Bildung*

Hinsichtlich der Determinierung eines Bildungsstandes sind Anzahl und Schwere von Fehlern sowie die allgemeine Ausdrucksweise des Autors ein Indikator[6]. Im Rahmen digitaler Analysen sind auch

[3] Dabei wird die Annahme Derns (2009 : 65) zugrunde gelegt, dass Sprecher, die eine Rechtschreibreform nicht von Beginn an erlernt, oder sie beruflich nachgeholt haben, diese nur in Mischformen produzieren können.

[4] Natürlich ist es denkbar, dass Fremdsprachler durch ausreichend langen Umgang mit der Sprache eine dem Muttersprachler in Nichts nachstehende Sprachkompetenz erwerben, oder dieser Unterschied aufgrund des Textumfangs einfach nicht mehr stichhaltig nachzuweisen ist.

[5] Dies ergibt sich zumindest aus der Aussage Kaufmanns (2008 : 135) wonach manche Fehler "[...] vorwiegend von türkischen Sprechern und solchen aus dem ehemaligen Jugoslawien [...]" produziert werden.

[6] In diesem Zusammenhang zu nennen ist die Frage nach der Aussagekraft von Fehlern. Um die weiterführenden Betrachtungen von Fobbe (2011 : 160) kurz zusammenzufassen: Fehler unterteilen sich in *errors* und

Type-Token Relationen aussagekräftig hinsichtlich des Bildungsstandes. Fehlende schriftsprachliche Kompetenz, wie sie sich durch erhöhte Fehlerzahl und unterdurchschnittlichem sprachlichen Niveau äußert, lässt nicht direkt den Schluss auf mangelnde allgemeine Sprachkompetenz zu. Es bedeutet zunächst lediglich, dass der Autor mit den Anforderungen, die eine schriftliche Textproduktion stellt, nicht vertraut ist. Von einer schriftlichen kann nicht direkt auf eine mündliche und somit allgemeine Sprachkompetenz geschlossen werden. Geringe oder kaum vorhandene Schreibkompetenz lässt es jedoch zu, den Grad des erreichten Schulabschlusses abzuschätzen. An weiterführenden Schulen und insbesondere nach einer universitären Laufbahn ist die Vermittlung von schriftlicher Kompetenz grundlegender Teil der Ausbildung oder Anforderung. Ist ein hohes sprachliches Niveau und eindeutige Schreibkompetenz vorhanden, kann mit relativer Sicherheit auf einen Ausbildungsabschluss geschlossen werden. Versuche, einen unvertrauten, gehobenen Stil überzeugend umzusetzen sind in der Regel nicht erfolgreich und können grundsätzlich nur in die nach unten gerichtete Sprachebene angenommen werden (vgl. Dern 2009 : 65).

Methodisches Vorgehen der Textanalyse

Für die Textanalyse ist es wichtig, sich zu jedem Zeitpunkt der Einflussfaktoren auf den Text, dessen Produktion und den produzierenden Autor bewusst zu sein, nehmen sie nicht nur entscheidenden Einfluss auf die Gestaltung des Textes, sondern sind auch grundsätzliche Voraussetzung für Rückschlüsse auf den Autor, die als Merkmalen und Charakteristika herausgearbeitet werden. Zunächst sind dies alldiejenigen Auffälligkeiten im Text, die sich von der erwartbaren Norm in irgendeiner Weise absetzen.[7] Sind alle Auffälligkeiten markiert, lassen sich diese in verschiedene Analysekategorien einordnen, die Aufschluss über Charakterzüge des Autors geben: Fremd- oder Muttersprachler, berufliche Einordnung, Alter des Autors, Bildung und regionale Zuordnung (Dern 2009: 64). Diese Kategorien *können* anhand bestimmter sprachlicher Phänomene im Text gefunden werden, *müssen* es aber nicht zwangsweise.

Marker wie diese haben gemeinsam, dass sie sich gut durch manuelle Analyse erkennen und interpretieren lassen. Sprechen wir jedoch von digitaler Analyse, bedeutet dies eine automatisierte Suche nach quantifizierbaren Mustern und Relationen mehrerer Texte und Textkomponenten. Für den Zweck der vorliegenden Analyse wurde ein vom Verfasser eigens entwickeltes Programm angewandt, das innerhalb von Chatbeiträgen aus einschlägigen Foren im Darknet eine quantitative Datenstrukturierung durchführte, die anschließend durch den Verfasser interpretiert wurde. Dabei orientierte sich die Analyse auf folgende Aspekte der Musterbildung:

- Durchschnittliche Satzlänge der einzelnen Beiträge eines Users mit Angabe der einzelnen Extremwerte (kürzester und längster Satz). Dies lässt, bei korrekter Satzkonstruktion, Rückschlüsse auf die Satzkomplexität zu. Die Minimallänge eines Satzes wurde auf drei Wörter festgelegt (Subjekt, Verb, Objekt), kürzere 'Sätze' wurden ignoriert.
- Anzahl der in einem Beitrag vorhandenen Paragraphen und deren durchschnittliche Länge, bestimmt anhand der pro Absatz enthaltenen Sätze. Getrennt wird dabei zwischen jedem Alinea (*Whitespace*), eine Mindestlänge pro Absatz ist jedoch nicht vorgesehen.

lapses. Dabei sind *errors* 'echte' Fehler, die sich auf mangelnde Autorenkompetenz zurückführen lassen, dementsprechend also an Aussagekraft gewinnen. Weniger aussagekräftig sind *lapses*, deren Entstehung sich in Unaufmerksamkeit oder Stress gründet (vgl. Fobbe 2011 : 160f.).

[7] Um an dieser Stelle bereits kurz auf die digitale Analyse vorzugreifen: Rechtschreibfehler sind bei automatischen Analysevorgängen mittels Software höchst problematisch.

- Zählung der am häufigsten im Text vorkommenden Wörter ohne sog. Stoppwörter, also nicht inhaltsrelevante Textbestandteile, die durch die Syntax gefordert werden aber keinen inhaltlichen Mehrwert bieten, z.B. Artikel und Konjunktionen. Dennoch wurden beide Zählungsarten durchgeführt, sowohl mit als auch ohne Stoppwörter, um etwaige im Userbeitrag vorzufindende Auslassungen von Stoppwörtern ebenfalls in die Analyse einfließen lassen zu können. So fallen beispielsweise Artikelauslassungen, die sich als Elemente individueller sprachlicher Charakteristika eines bestimmten Users verstehen lassen, in der Analyse auf.
- Zählung der häufigsten Wortarten sowie Muster in der Wortartnutzung, womit eine bestimmte, häufig auftretende syntaktische Struktur erkannt werden soll.[8] Die Wortarten werden dabei unter Zuhilfenahme des bekannten STTS[9] automatisch durch ein selbstlernendes Verfahren mittels eines Vergleichskorpus bestimmt. Das Verfahren als solches funktioniert mit einer Trefferquote von 96%, setzt aber orthographisch und syntaktisch korrektes Deutsch voraus.
- Type-Token-Ratio (TTR): Unter Tokens versteht man ein Zeichen oder eine Zeichenkette, wie zum Beispiel ein Wort oder Satzzeichen mit einer Mindestlänge von 1. Diese kann man in einer Liste sortieren, wobei vorkommenden Tokens quantifiziert und somit die absoluten Häufigkeiten berücksichtigt werden. Anders verhält sich dies bei einer Type-Liste. Diese setzt sich zwar ebenfalls aus Tokens zusammen, quantifiziert diese jedoch nicht. Es werden alle Tokens lediglich einmal aufgeführt, wodurch das im Korpus zur Verfügung stehende Vokabular beschrieben werden kann. Durch einen direkten Vergleich beider Listen in einer sogenannten Type-Token Relation lässt sich die lexikalische Vielfalt eines Korpus oder bestimmter Texte innerhalb eines Korpus bestimmen (vgl. Perkuhn 2012: 38f.). Die Type-Token-Ratio gilt als Indikator für das Sprachniveau eines Textes und das Bildungsniveau des erstellenden Autors und repräsentiert das Verhältnis von einzigartigen zu vorhandenen Wörtern im Text. Die TTR ist mathematisch bedingt stets kleiner oder gleich 1, da logischerweise in einem Text nicht mehr einzigartige Wörter vorkommen können als Wörter gesamt vorhanden sind. Bei der Analyse von TTR ist das Ergebnis allerdings nicht nur durch den Autor und dessen Können spezifiziert, sondern auch zu großen Teilen durch die Textlänge bedingt. Mit steigender Textlänge muss zwangsweise die TTR sinken, da es zunehmend schwierig wird, Mehrfachnennungen von Wörtern zu vermeiden. Die direkte Vergleichbarkeit von TTRs, die aus unterschiedlichen Texten generiert wurden, ist also nur mit Vorsicht zu genießen. Bisher ist kein praktikabler Lösungsansatz für dieses Problem bekannt.

Sind alle systematisch relevanten Komponenten eines Textes analysiert und die wesentlichen Marker herausgearbeitet, ist der Analyseprozess abgeschlossen. Nun bleibt der Textvergleich durchzuführen. Hierfür werden alle Befunde der einzelnen Texte hinsichtlich möglicher Ähnlichkeiten und Unterschiede auf mehreren linguistischen Ebenen verglichen und quantifiziert. Abschließend sollte eine Bewertungsskala anhand der festgestellten Befunde angewendet werden, wie wahrscheinlich oder unwahrscheinlich die verschiedenen Texte von dem gleichen Autor verfasst wurden: Mit an Sicherheit grenzender Wahrscheinlichkeit identisch, mit an hoher Wahrscheinlichkeit identisch,

[8] Im Allgemeinen wären dies bestimmte wiederkehrende Phrasen, die durch die Strukturanalyse abstrahiert dargestellt werden.

[9] Stuttgart-Tübingen Tagset. Die genaue Ausprägung des STTS ist hier zu finden: www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (23.11.2019).

wahrscheinlich identisch, kommt in Betracht / ist nicht auszuschließen, nicht entscheidbar, wahrscheinlich auszuschließen, mit hoher Wahrscheinlichkeit auszuschließen, mit an Sicherheit grenzender Wahrscheinlichkeit auszuschließen (vgl. Steinke 1990 : 336).

Limitationen

Die forensische Linguistik vermag also unterschiedliche Aussagen zu den Autoren eines in Frage stehenden Textes sowie auch zu dem Text als solchen treffen. Im Rahmen der hier durchgeführten Studie kann sie insbesondere zur Abgrenzung der einzelnen Autoren untereinander sowie zur Herausarbeitung besonderer individueller Sprachkomponenten, die spezifisch für eine Gruppe von Sprechern mit pädophilen Neigungen sein könnte, eingesetzt werden. Um dies zu gewährleisten ist allerdings ausreichend Datenmaterial essentiell. Zudem sind sprachliche Analysen zu großen Teilen auf den Textumfang angewiesen, da sich sehr kurze Texte[10] nicht aussagekräftig analysieren lassen.

Wie die Bewertungsskala verdeutlicht, handelt es sich bei der forensischen Linguistik um einen Forschungsbereich, der zwar Tendenzen und Wahrscheinlichkeiten ausdrücken kann, allerdings aufgrund der Wandelbarkeit, Individualität und feinen Nuancen von Textsprache Aussagen nicht mit absoluter Sicherheit treffen kann. Dies bezieht sich vor allem auf äußerst kurze Schrifterzeugnisse wie im Extremfall Ein-Wort-Antworten, oder auch auf Imitations- oder Verstellungsversuche des Autors. Daher ist die Vorstellung eines Linguisten, der lediglich anhand eines Textes einen Delinquenten zu überführen vermag ist ebenso realitätsfremd wie unwissenschaftlich (vgl. Kniffka 1990: 1). Dennoch vermag sie es, aufgrund der technischen Verschleierungsmethoden des Darknets (siehe Kapitel II), zu Ergebnissen zu kommen, wo technische Nachverfolgungsmethoden scheitern.

Dennoch: Der größte Vorteil der forensischen Linguistik ist, dass sich Autoren in der Regel nicht über die Aussagekraft ihrer Texte im Klaren sind und somit keinen bewussten Aufwand betreiben, um persönliche Merkmale zu verschleiern. Es ist davon auszugehen, dass ein Verfasser von Chatbeiträgen in Darknet-Foren aus zwei Gründen nicht auf Verstellungsversuche zurückgreifen wird: Einerseits erschließt sich für ihn keine Notwendigkeit dafür, er wiegt sich in Sicherheit und unter seinesgleichen, versteckt in einem zugangsbeschränkten Forum. Dort rechnet er nicht mit der Analyse seiner Sprachstrukturen durch Behörden, im Gegensatz zu beispielsweise Verfasser von Erpressungsbriefen. Andererseits stellt die konsistente Umsetzung von Verstellungsversuchen für regelmäßige und aktive Beiträge in Foren einen unverhältnismäßigen Aufwand dar.

II. Das Darknet und warum eine linguistische Analyse nötig ist

Es gibt gute Gründe für eine sprachbasierte Analyse im Darknet und teilweise auch im Clearweb, wo andere Methoden nicht zielführend sind. Zum einen ist eine Rückverfolgung des Urhebers von Beiträgen in Darknet-Foren technisch beinahe unmöglich, zum anderen bietet sie auch keinesfalls die nötigen Einsichten und Erkenntnisse, die für die Ableitung eines bestimmten 'Pädolekts', einem auffälligen Sprachgebrauch der prädominant von Pädophilen genutzt werden könnte, erforderlich sind. In strafrechtlichen Ermittlungen in Fällen von Missbrauch und Missbrauchsabbildungen von

[10] Im Allgemeinen wird von 200 textrelevanten Wörtern (textrelevant = keine Stoppwörter) ausgegangen.

Kindern stoßen die Ermittler häufig auf eine große Anzahl von IP-Adressen, die mit einem Fall zusammenhängen[11]. Doch sind 30.000 IP-Adressen gleichzusetzen mit 30.000 Tatverdächtigen? Die Frage, ob ein User mehrere Accounts betreibt, lässt sich aufgrund der technischen Beschaffenheit des Darknets und der Möglichkeiten der User zur Verschleierung ihrer digitalen Spuren nicht ohne erheblichen Aufwand beantworten. Somit kann bisher mit überschaubarem technischen und finanziellen Aufwand nur die forensisch angewandte Linguistik die Werkzeuge und Methoden zur Verfügung stellen, die sowohl für die Strafermittlung als auch für Prävention von Cyber-Grooming und Kindesmissbrauch erforderlich sind. Sie bietet außerdem mittels des Textvergleichs eine Methode, das User-Account-Verhältnis in Darknet-Foren einschätzen zu können, sprich die Frage zu klären ob einem User nur ein, oder mehrere textuell aktive Accounts zugeordnet werden können. Somit kann eine ungefähre Userzahl besser abgeschätzt werden

Begriffe wie *Darknet* oder *Deepweb* werden häufig synonym gebraucht, sind allerdings konzeptuell voneinander abzugrenzen. Wo das Darknet über ein eigenes Protokoll (*Onion Routing*) kommuniziert und funktioniert, ist das Deepweb technisch nicht vom normalen Surface-Web (auch *Clearweb*, dem normal per Browser aufrufbarem Internet) zu trennen, sie funktionieren auf derselben Basis im selben System und mit gleichen oder ähnlichen Protokollen.

Das Deepweb bezeichnet diejenigen Bereiche des Internets, die nicht für das Auffinden durch Suchmaschinen (z.B. Google) indiziert wurden und umfasst in der Regel spezifische themengebundene Datenbanken oder Webseiten, die meist aus Gründen der Performanz (Suchtiefe), der Zugriffsbeschränkung (Login, z.B. Onlinebanking), wirtschaftlicher Interessen (Online-Shops) oder technischer Unmöglichkeit außen vor gelassen werden.

Das Clearweb bezeichnet all diejenigen Bereiche des Internets, die man per üblichen Browser und mittels der Nutzung einer Suchmaschine oder direkt per URL in der Adresszeile aufrufen und ohne weitere Beschränkungen nutzen kann.

Dienste, welche über Clear- und Deepweb funktionieren, wurden nicht in die vorliegende Studie einbezogen, da in diesen Bereichen andere, nicht linguistische Methoden der Useridentifizierung schneller und eindeutiger Ergebnisse liefern können. Warum diese Methoden im Darknet nicht funktionieren wird klar, führt man sich den technischen Aufbau und die Funktionsweise genauer vor Augen:

Die technische Umsetzung des Darknets erfolgt mittels des Onion-Routing Protokolls, welches in seiner Grundkonzeption in den 1990er Jahren durch das U.S. Naval Reasearch Laboratory und der Defense Advanced Research Projects Agency (DARPA) mit dem Ziel gesicherter, nicht via sogenannter Traffic Analysis verfolgbarer, abhörgeschützter Echtzeitkommunikation entwickelt und patentiert wurde[12] (vgl. Reed, Syverson, Goldschlag, 1998: 1).

Onion Routing stellt keine direkte Verbindung zwischen zwei kommunizierenden Endgeräten her, sondern durch eine Sequenz von Geräten, welche als Onion Routers bezeichnet werden, wodurch die Verbindung zwischen Sender und Empfänger anonym bleibt. Die Geräte innerhalb des Onion-Routing-Netzwerks sind untereinander mit feststehenden Sockel-Verbindungen verbunden und

[11]Siehe ARD, 02.07.2020 „30.000 Spuren sind nicht 30.000 Täter“: www.tagesschau.de/investigativ/zapp/kindesmissbrauch-bergisch-gladbach-103.html (04.09.2020).
[12] Vgl. hierzu US Patent 6266704.

Kommunikationen werden durch das Netzwerk vervielfältigt. Die Sequenz der Onion-Router ist bei dem initialen Verbindungsaufbau vorgegeben, wobei jeder Onion-Router jeweils nur seinen Vorgänger und Nachfolger zu identifizieren vermag. Die Daten, welche entlang der Onion-Route verschickt werden, erscheinen jedem Router unterschiedlich (vgl. Reed et. Al. 1998 : 2).

Die Verbindung mit dem Onion-Netzwerk wird durch eine Reihe von Proxies (Kommunikationsschnittstellen im Netz, welche als eine Art Vermittler fungieren) hergestellt. Eine initiierende Anwendung stellt eine Socket Verbindung zu einem sogenannten Application-Proxy her, der die Kommunikation und später entsprechende Daten in ein generisches Format umwandelt, welches durch das Onion-Netzwerk weitergegeben werden kann. Anschließend wird von dem selben Application-Proxy (auch als Entry-Node bekannt) eine Verbindung mit einem Onion-Proxy hergestellt, welcher die Route durch das Onion-Netzwerk festlegt. Dies wird durch das Erzeugen einer geschichteten Datenstruktur mit dem Namen Onion (Zwiebel) bewerkstelligt. Diese Datenstruktur wird daraufhin an einen sogenannten Entry Funnel weitergegeben, welcher eine dauerhafte Verbindung zu einem Onion-Router aufrechterhält und die zu diesem Router führenden Verbindungen vervielfacht. Dieser Onion-Router ist dafür zuständig die erste Schicht des Datenpaketes (der Onion) zu verarbeiten, wobei jede Schicht des Paketes den nächsten Schritt in der Abfolge innerhalb des Netzwerkes definiert. Ein Onion-Router, welcher eine Onion erhält, interpretiert die in der äußersten Schicht des Datenpaketes enthaltenen Informationen und sendet das verbleibende Datenpaket entsprechend der in der für diesen Router bestimmten Informationen weiter. Der letzte Router der Reihe leitet das verbleibende Datenpaket an einen exit-Funnel weiter, der die Daten zwischen dem Onion-Netzwerk und dem Kommunikationsziel austauscht. (vgl. Reed et. Al. 1998 : 2).

Jede Onion Schicht enthält, neben den Informationen die zum Weiterleiten an den korrekten nächsten Onion-Router im Netzwerk notwendig sind, weiterhin alle nötigen Daten um einen Schlüssel zu erzeugen, mit dem Kommunikationspakete zwischen den einzelnen Knotenpunkten des Netzwerks verschlüsselt hin und her verschickt werden können. (vgl. Reed et. Al. 1998 : 2).

Nachdem diese Verbindung hergestellt wurde, können Daten verschickt werden. Zunächst fügt der Onion-Proxy für jeden genutzten Onion-Router innerhalb des Netzwerks eine Verschlüsselungsschicht hinzu. An jedem Router wird diese mit der Verarbeitung der äußersten Datenschicht entfernt und das verbleibende Paket, minus einer Verschlüsselungsschicht, an den nächsten Knotenpunkt weitergeschickt. Was beim Kommunikationspartner ankommt, ist dann lediglich die gewünschte Anfrage im Klartext, ohne jegliche Verschlüsselung. In die Gegenrichtung erfolgt dieser Vorgang in der exakt umgekehrten Reihenfolge. (vgl. ebd.).

Zugang zum Darknet bekommt man nicht einfach über einen gewöhnlichen Browser. Es ist ein spezieller Browser nötig, welcher die Kommunikation mit dem Onion-Routing Netzwerk erlaubt. Zu diesem Zweck gibt es viele verschiedene Browser-Optionen, exemplarisch sei hier einer der bekanntesten Vertreter vorgestellt, der Tor-Browser, abgekürzt für The Onion Router. Dieser kann von jedermann über die frei zugängliche Homepage des TOR-Projekts[13] heruntergeladen und installiert werden. Ab diesem Zeitpunkt verhält sich der Tor-Browser auf den ersten Blick wie ein gängiger Browser, mit dem man wie gewohnt im Internet surfen kann.

[13] www.torproject.org

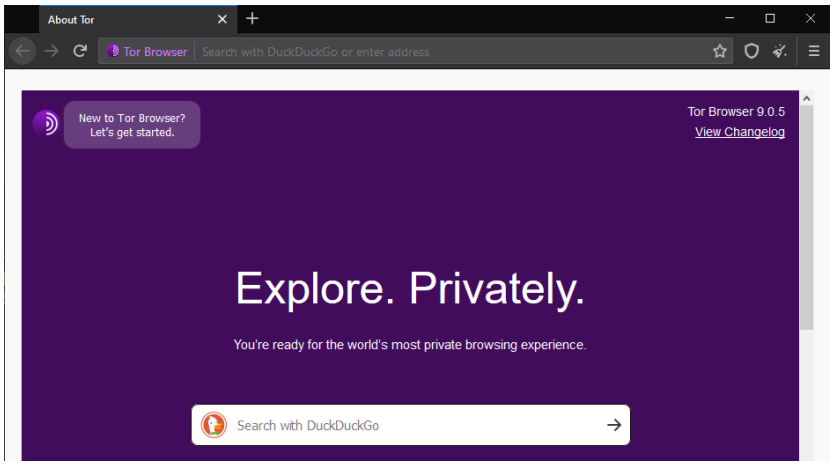

Abbildung 1: Startbildschirm des TOR-Browsers

Beim ersten Aufrufen einer beliebigen Website lässt sich bereits nachvollziehen, dass Tor Unterschiedliche Nodes verwendet um auf die Zieladresse zuzugreifen. Die zuvor illustrierte Funktionsweise mit unterschiedlichen Knotenpunkten im Netzwerk, welche den Traffic entsprechend weiterleiten, wird hier auch für den Anwender verdeutlicht.

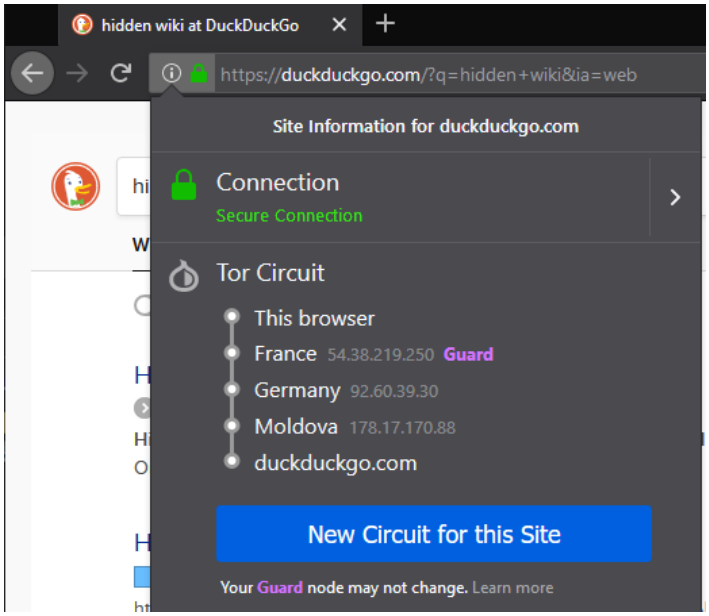

Abbildung 2: Verbindungsansicht mit Standorten der einzelnen Nodes

Ab diesem Zeitpunkt kann man Tor voll funktionsfähig mit den entsprechenden Features für das Einhalten der Privatsphäre und der Anonymität nutzen. Will man tatsächlich das Darknet erkunden, so ist für gewöhnlich der nächste Schritt das Ansurfen eines der vielen zur Verfügung stehenden sogenannten Hidden Wikis (oder dessen oft gleich benannter Klone). Bei Hidden Wikis handelt es sich um eine Art Katalog, der die unterschiedlichen Adressen des Darknets sammelt und kurz darstellt. Allen diesen Adressen ist gemein, dass sie keine gewöhnlichen namentlichen Bezeichnung haben, wie beispielsweise Google als google.com, sondern stattdessen eine kaum einprägsame alphanumerische Zeichenfolge, welche stets sie auf .onion[14] endet. Solche Adressen lassen sich ausschließlich über einen Browser ansurfen, der sich mit dem Onion-Routing Netzwerk verbinden

[14] Beispielsweise (erfunden, wahrscheinlich nicht existent): 92zghsh84mbvl09d8jsxhg83gj.onion.

kann. Die Adressen erhält man entweder über Hidden Wiki oder über eine beliebige Darknet-spezifische Suchmaschine, die sich ebenfalls auf den Hidden Wikis finden lassen.

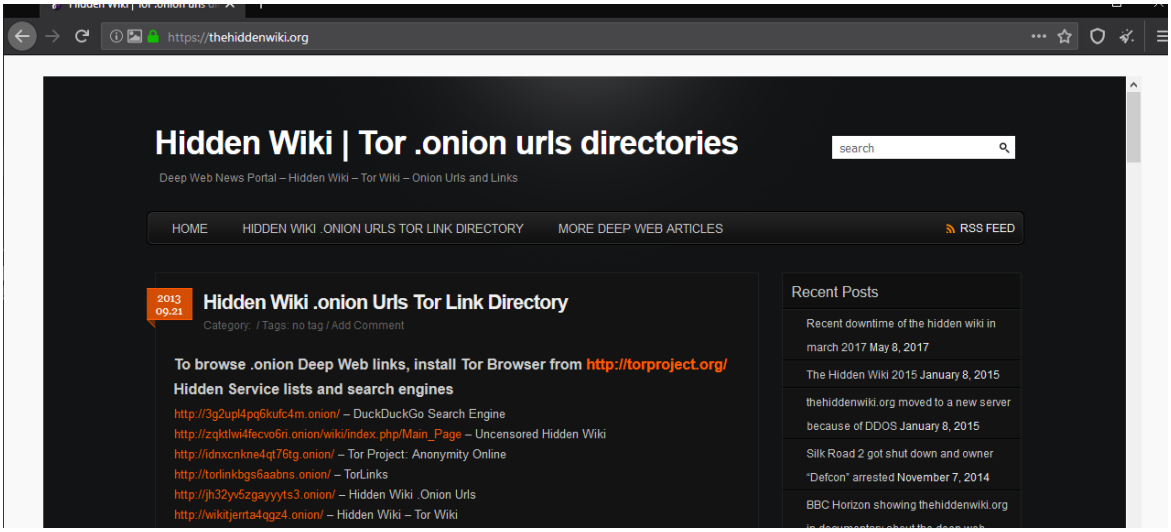

Abbildung 3: Exemplarisches Hidden-Wiki mit aufgelisteten unkritischen weiteren Darknet-Links im .onion Format

Von diesem Zeitpunkt an ist es, wenn man das Ziel der vorliegenden Studie postulierten Problematik verfolgen möchte, eine Frage entsprechend ausführlicher Recherchen und Suchen in den dargestellten Plattformen, ehe man auf entsprechende Foren stößt, die thematisch mit Missbrauchsabbildungen von Kindern und sexualisierter Gewalt zu tun haben. Der genaue Rechercheweg wird hier aus offensichtlichen Gründen nicht dargestellt.

III. Datenlage und Analyseschritte

Die Vorliegende Studie hatte mit zwei Beschränkungen zu tun: Dem Finden geeigneter Foren und Beiträgen in denselben und dem Zugriff auf die Daten in den Foren. Viele Foren und Beiträge konnten in der Auswertung aufgrund ihres unzureichenden Umfangs nicht berücksichtigt werden, das Aufspüren eines Forums mit konsistent längeren Postings von mehreren Usern gestaltete sich als problematisch. Von fünf durchsuchten Foren konnte lediglich das Forum „Rinderwahn" [Anm. d.Verf.: Name aus offensichtlichen Gründen verändert] als geeignet identifiziert werden. Es dient der vorliegenden Studie als Datengrundlage. Mit 1.143.740 registrierten Usern und 183.800 Beiträgen, 3.567 davon auf Deutsch[15] gehört es zu den aktivsten multilingualen Foren mit Pädophilie-Thematik. Da Rinderwahn bei der Registrierung keinerlei persönliche Informationen sammelt oder abfragt, ist die Registrierung vollständig anonym, die User kaum rückverfolgbar und somit für eine linguistische Analyse prädestiniert.

[15] Stand: 05.02.2020.

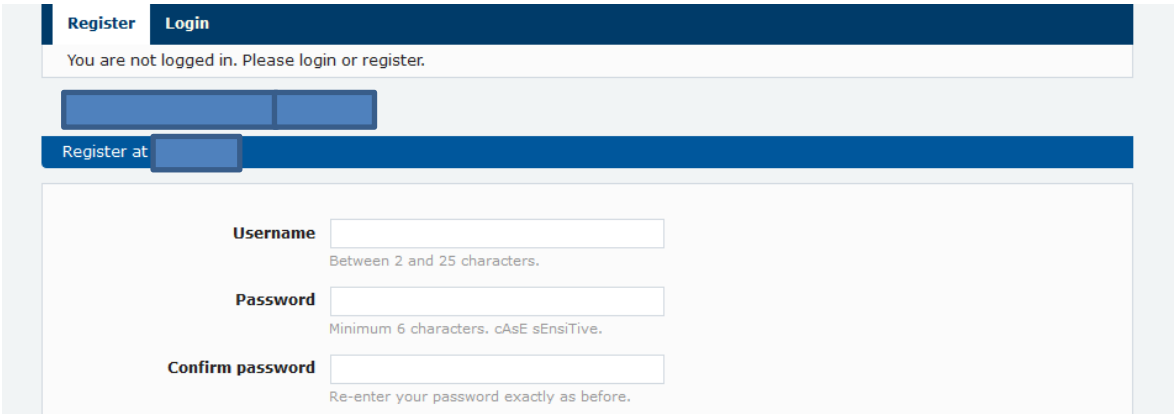

Abbildung 4: Registrierung eines neuen Users im Forum „Rinderwahn“

Für den Zugriff auf möglicherweise relevante Daten problematisch gestaltet sich insbesondere die gängige Praxis vieler derart gerichteter Foren, Zutritt erst nach dem Hochladen kinderpornographischer Inhalte zu gewähren. Eine Handlung, die nach deutschem Recht als Verbreitung von kinderpornographischen Inhalten strafbar wäre und damit für eine Verwendung im Rahmen einer Studie außer Frage stand. Daher beschränkte sich die Suche auf entweder frei oder über einfache Anmeldung zugängliche Foren. Deren Inhalte sind dabei auf die Kommunikation zwischen erwachsenen Pädophilen ausgelegt, nicht auf Kontakte mit Minderjährigen. Damit war der Zugang zu Daten bezüglich der Frage, in wie weit ein bestimmter Sprachgebrauch mit dem Ziel des Cyber-Ggroomings von Kindern angewandt wird, als versperrt. Für die Klärung dieser Frage wären Daten von echter Kommunikation mit Kindern erforderlich, eine Erhebung derselben ist allerdings mitunter aus datenschutzrechtlichen Gründen äußerst problematisch.

Ein weiteres Problem für die Datenerhebung in einem pädophilen Kontext des Darknets stellt die generell begrüßenswerte, doch für Untersuchungszwecke sehr hinderliche Ablehnung etwaiger Foren durch größere Teile der Internet-Community dar, die sich in vielen Fällen durch ausgedehnte DDoS Attacken auf Foren oder deren Hoster darstellt[16]. Häufig sind Foren nur kurz oder bereits nicht mehr erreichbar, was die Datenextraktion extrem erschwerte und die Nachvollziehbarkeit reduzierte. Auch konventionelle Datenerhebungsmethoden des Clearwebs, insbesondere Webscraping, das maschinelle "Herauskratzen" (=scraping) von Daten via automatisierter Crawler, Programme, die eine Verzeichnisstruktur einer gegebenen Website auf bestimmte Inhalte durchsuchen, funktionieren im Darknet über den Tor-Browser nicht.

Das Forum Rinderwahn gliedert mehrere Unterforen in verschiedenen Sprachen ein, wobei für die vorliegende Studie nur Daten von der deutschsprachigen Sektion[17] extrahiert wurden. Beispielhaft sollen Exzerpte von vier textuell aktivsten User vorgestellt werden, um die Funktionsweise und Aussagekraft der durchgeführten Untersuchung abschätzen sowie die Ergebnisse bewerten zu können. Zentrale Auswahlkriterien waren adäquater Umfang der einzelnen Posts einerseits, sowie der Gesamtheit an Beiträgen andererseits. Eine größtenteils sprachliche Richtigkeit der Inhalte wäre zwar,

[16] Auch die meisten Hidden Wikis haben mittlerweile aus vermutlich denselben Gründen (DDoS-Angriffe) die Referenzen auf kinderpornographische und pädophile Foren entfernt, was das Aufspüren entsprechender Plattformen erheblich erschwert und nur mit großem zeitlichen Aufwand und entsprechender Erfahrung im Umgang mit dem Darknet überhaupt möglich ist.

[17] Dazu sei weiterführend angemerkt, dass auch bei Gelegenheit Postings in Englisch von primär in den deutschen Bereichen aktiver Nutzer erhoben wurden, um deren Fremdsprachenkompetenz, subsequent den Bildungsstand und etwaige Verstellung zu verifizieren.

mit Hinblick auf die spätere digitale Analyse, wünschenswert gewesen, wurde allerdings aus Rücksicht auf etwaige Verfälschung der Testgruppe nicht mit einbezogen. Der Verfasser hat keine weiteren Kriterien für die Selektion der Textbeispiele angewandt, sondern die User vollkommen zufällig ausgewählt.

Insgesamt waren 659 Forenbeiträge Datengrundlage für die vorliegende digitale Textanalyse, die mittels eines eigens zu diesem Zweck entwickeltes Programm durchgeführt wurde. Dieses wurde in Python3[18] verfasst und erzeugt eine implizite Datenbank für den Zeitraum der Analyse. Auf eine eigenständige Datenbank, wie sie mit SQL umsetzbar wäre, wurde aus Gründen der Performanz verzichtet. Zentrale genutzte Bibliotheken sind `textblob(_de)`, `nltk`[19] (in Kombination mit TIGER[20] und STTS) und `suffix_trees`[21]. Das Programm berechnet verschiedene Sachverhalte auf Satzebene (längster, kürzester und durchschnittlicher Satz), Paragraphenebene (Anzahl und durchschnittliche Satzanzahl), Wortebene (Wortart, Stoppwort, Type-Token Relation mit und ohne Stoppwörter, häufigste Wörter) und Musterebene (repetitive Muster in der Wortartenabfolge).

Die vollständige Ausgabe der Analyse eines gegebenen Beitrages stellt sich exemplifiziert folgendermaßen dar:

```
--------------------------------------------------------------------------------

txt/Chldprn/1.txt

--------------------------------------------------------------------------------

Longest sentence: 19 words

Shortest sentence: 4 words

Average sentence: 11.166666666666666 words

Paragraph Count: 1

Avg. sentences per paragraph: 6.0

Most common words: [('ich', 3), ('habe', 3), ('eine', 2), ('und', 2), ('es', 2),
('zu', 2), ('welcher', 2), ('zeit', 2), ('noch', 2), ('mir', 1)]

Type-Token-Relation: 0.835820895522388 @ 67 tokens
```

[18] Python ist eine multiparadigmale, höhere Programmiersprache, d.h. es erlaubt einen objektorientierten, aspektorientierten oder funktionalen (Funktionen dürfen als Argumente auftreten) Programmierstil.
[19] Das Natural Language Toolkit (NLTK) ist eine Python-basierte Plattform zur natürlichen Sprachverarbeitung unter Bereitstellung von mehr als 50 Sprachkorpora sowie selbstlernender Komponenten für automatisiertes tagging und tokenizing. Es übernimmt gemäß Spezifikation Wortartenerkennung und Stoppwortentfernung. Für weiterführende Dokumentation zu NLTK siehe: www.nltk.org (15.12.19)
[20] Das TIGER Korpus besteht aus 900.000 Tokens aus deutschen Zeitungen (vgl. www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html, 15.12.19)
[21] Suffix_Trees implementiert Suffixbäume in Python, welche zur Mustererkennung in Wortarten verwendet werden. In der String-Analyse ist dieses Verfahren, aufgrund des linearen Verhaltens hinsichtlich Zeit und Speicherbedarfs, weit verbreitet. Für weiterführende Dokumentation zu Suffix_Trees siehe: www.daimi.au.dk/~mailund/suffix_tree.html (15.12.19)

```
Most common words (w/o stopwords): [('zeit', 2), ('vorgenommen', 1), ('nächsten', 1), ('wochen', 1), ('mal', 1), ('kleine', 1), ('tor-cp-board-history', 1), ('erstelle', 1), ('poste', 1), ('einfach', 1)]

Type-Token-Relation (w/o stopwords): 0.9615384615384616 @ 26 tokens

PRON ADP DET (2 Occurances)

DET ADJ NOUN (2 Occurances)

ADP DET NOUN (2 Occurances)

DET NOUN VERB (2 Occurances)

ADV DET NOUN (2 Occurances)

AUX PRON ADV (2 Occurances)

DET NOUN PUNCT (2 Occurances)

PRON ADV DET (2 Occurances)

PRON AUX PRON (1 Occurances)

AUX PRON VERB (1 Occurances)

PRON VERB PUNCT (1 Occurances)

VERB PUNCT SCONJ (1 Occurances)

PUNCT SCONJ PRON (1 Occurances)

--------------------------------------------------------------------------------
```

Das Programm kann Sätze und Wörter nur dann korrekt erkennen, wenn diese den Vorgaben entsprechen, wie ein Satz oder ein Wort auszusehen hat. Ein Satz wird durch ein beendendes Satzzeichen (. / ! / ?) eingegrenzt, ein Wort wird stets durch Leerzeichen getrennt und ist nur dann einer bekannten Wortart zuzuordnen, wenn es in den verwendeten Vergleichslisten (*Tagsets*) aufzufinden ist. Eine semantische Erkennung findet nicht statt. Insbesondere bei syntaktisch und orthographisch inkorrekten Texten sind die Ergebnisse der digitalen Analyse also stellenweise zu prüfen und entsprechend anzupassen.

Jeweils zehn Einträge werden exemplarisch von jedem User tabellarisch dargestellt. Die Auswahl wurde nicht hinsichtlich bestimmter Merkmale selektiert, sondern ergibt sich aus dem ersten Block der Berechnungen durch das Analyseprogramm. Die Sortierung findet auf String-Ebene (einer endlichen Folge von Zeichen in einer Zeichenkette) statt und ist dementsprechend nicht durchgehend numerisch fortlaufend (die Sortierung erfolgt nach dem Muster a>aa>aaa>ab>abb>b>bb etc.). In den folgenden Tabellen dargestellt sind die markantesten Merkmale der digitalen Textanalyse, dazu zählen neben den Satzkomponenten, längster, kürzester und durchschnittlicher Satz, den Absätzen und durchschnittlichen Sätzen pro Absatz auch die nicht Stoppwort-bereinigte Type-Token Relation (TTR) sowie die Stoppwortbereinigte Type-Token Relation (TTR2). Die Token-Spalte quantifiziert dieselben.

Die generelle Struktur der hier vorgestellten Beiträge zeigt Parallelen zu gewöhnlichen Forenbeiträgen auf. Es gibt zwei Arten von Beitragssträngen: Längere Initialpostings gefolgt von mehreren Usern, die daraufhin in einer textanalytisch zu kurz dimensionierten Art reaktiv in Erscheinung treten; und Initialpostings, die sich schnell zu einer längeren, ausführlicheren Diskussion zwischen zwei hauptsächlich daran beteiligten Usern entwickeln.

Zunächst sollen die gegebenen Textbeispiele aus dem Rinderwahn-Forum hinsichtlich sprachlicher Auffälligkeiten und Musterbildungen untersucht werden. Dabei liegt der spezielle Fokus darauf, die User nicht nur sprachlich voneinander abzugrenzen, sondern auch musterhafte Sprachstrukturen über alle Texterzeugnisse hinsichtlich eines Pädolektes zu analysieren und möglicherweise auftretende individuelle Stilmarker herauszuarbeiten. Mögliche Indikatoren, die auf einen Pädolekt schließen lassen, sollen anschließend mit Beiträgen aus Eltern-Foren[22] verglichen werden.

Der Analyseaufbau strukturiert sich in allgemeine syntaktische, orthographische und lexikalische Auffälligkeiten, Aussagen zu Autorencharakteristika wie Alter, regionale Zugehörigkeit, Fachlichkeit, Bildungsstand und Muttersprachlichkeit, sowie die Ergebnisse der digitalen Mustererkennung. Die genannten Kriterien sind dabei als Rahmen zu verstehen, denn nicht immer kann jeder Punkt hinlänglich am zu prüfenden Text nachgewiesen werden. Es sollen vier User vorgestellt werden, der Textanalyse am deutlichsten unterschiedliche Merkmale und Ausprägungen hervorgebracht hat.

IV. Sprachanalyse: Vier User

U*ser: Chldprn*

- *Sprachliche Charakteristik*

Syntax (Satzbau und Grammatik): Die syntaktische Struktur in den Texten von Chldprn ist stringent auf einem gehobenen Niveau, der typische Satzbau dabei zu großen Teilen von komplexer Natur, es wird mit häufigen Einschüben (Parenthesen) und Beisätzen (Appositionen) sowie mehreren untergeordneten Nebensätzen gearbeitet. Auffällig ist insbesondere, wie bestimmte Phrasen[23] teilweise in das Nachfeld des Satzes geschoben werden, obwohl sie dort typischerweise nicht zu finden sind. Beinahe vollständig durchgehend ist die konsequente Nutzung von Ausrufezeichen anstelle von Punkten. Zudem zeit der User einen korrekten Gebrauch von Kommasetzung, insbesondere bzgl. der häufigen Verwendung von Einschüben.

Lexik (Wortschatz): Chldprn nutzt auffallend häufig Anglizismen, im Speziellen solche aus dem Bereich der Informationstechnologie. Auffällig ist hierbei die teilweise inkonsistente Verwendung der Begriffe. Daneben fallen gelegentlich kontextuell falsche Komposita-Bildungen (zum Beispiel *kurzfristig* anstelle von *kurzzeitig)* auf. Diesen gelegentlichen Irregularitäten steht die gehobene Lexik gegenüber (beispielsweise *etablierte* oder *an diesem Tage*) bei gleichzeitiger Erscheinung einer umgangssprachlichen Mündlichkeit (*cool!, schon mal, rumgesprochen* oder *halt*[24]*).*

[22] Die Wahl des Vergleichsmaterials fiel auf Elternforen, weil dort eine erwachsene Gruppe von kinderfreundlichen und kinderorientierten Usern ohne pädophile Vorlieben und mit äquivalenter Sprachkompetenz zu z.B. Rinderwahn zu vermuten ist.
[23] Einzelne Wörter oder ganze Teilsätze.
[24] Kontextueller Bezug: „[…] da halt eben einfach […]“.

Orthographie: Mit Hinblick auf orthographische Auffälligkeiten fallen nur vereinzelt Verstöße auf. Diese beziehen sich insbesondere auf die Differenzierung zwischen Klein- und Großbuchstaben. Die wiederholte Falschumsetzung dieser Exempel lässt die Vermutung zu, dass es sich nicht um Flüchtigkeitsfehler handelt (Lapse), sondern um einem Error im Sinne einer mangelnden Autorenkompetenz. Größere orthographische Unregelmäßigkeiten kommen nicht vor. Einige hinsichtlich des Genus falsch determinierte Artikel und Pronomina fallen besonders mit der Verwendung von Anglizismen auf.

- *Aussagen zum Alter*

Chldprn verwendet keine offensichtlichen Marker, die sich eindeutig einer bestimmten Generation zuordnen lassen könnten. Die Rechtschreibreformen werden konsequent eingehalten, realisierte Umsetzungen von älteren Reformen lassen sich nicht erkennen. Gelegentliche Verwendung von jugendsprachlichen Phrasen (*cool!*) lassen damit verbunden eine ältere Generation als mit an Sicherheit grenzender Wahrscheinlichkeit ausschließbar erscheinen. Die sprachliche Ausdrucksweise ist strukturiert, längere Posts lassen teilweise bereits Ausblicke auf künftige Aussagen erkennen und greifen auf im späteren Verlauf erwähnte Inhalte vor, ist somit strukturell komplexer, die Verwendung von bestimmten rhetorischen Mitteln ist vorhanden und hinreichend häufig um einen jugendlichen Textproduzenten ebenfalls auszuschließen. Der Autor ist wahrscheinlich ein Erwachsener im Bereich von 25 bis 40 Jahren.

- *Aussagen zu einer regionalen Zugehörigkeit und Muttersprachlichkeit*

In Chldprns Textproduktion lassen sich keine eindeutigen Hinweise auf eine regionale Zugehörigkeit innerhalb Deutschlands feststellen. Das gelegentlich auftretende Verschieben von Phrasen in das Nachfeld eines Satzes findet sich zwar gelegentlich im mündlichen Sprachgebrauch Süddeutschlands, ist jedoch quantitativ im Verhältnis zur gesamten Textproduktion nicht ausreichend, um mit überzeugender Wahrscheinlichkeit von einem Sprecher dieses dialektalen Gebiets auszugehen. Nicht bezweifelt werden kann jedoch, dass der Autor Deutsch als Muttersprache hat oder Deutsch auf demselben Niveau beherrscht. Die Interlanguage ist gesamtheitlich inhomogen, Fehler in der Genus-Zuordnung englischer Wörter sind sowohl parallel zu korrekt produzierten anspruchsvolleren sprachlichen Ebenen produziert, als auch hinsichtlich des Englischen als einer anzunehmenden Stammsprache inkorrekt, Deutsch als Muttersprache somit anzunehmen.

- *Aussagen zum Bildungsstand und fachlicher Zugehörigkeit*

Die von Chldprn durchwegs korrekt umgesetzte komplexe und gehobene Lexik und die strukturierte Satzbildung lassen auf einen höheren Bildungshintergrund schließen, wahrscheinlich Gymnasium oder mindestens Realschule, wobei auch weiterführende Erfahrung im Umgang mit schriftlicher Textproduktion denkbar ist. Dem entgegen steht die kontinuierliche Nutzung von Ausrufezeichen anstelle von Punkten, handelt es sich dabei schlichtweg um einen, im Schulsystem wohl kaum tolerierten, stilistischen Fauxpas. Orthographische Auffälligkeiten sind in ihrer Gesamtheit zu selten und in ihrer Beschaffenheit wenig gravierend, so dass sie nicht als nennenswerter Einfluss auf die Bewertung des Bildungshintergrundes zu deuten sind. Die Kriterien zugunsten einer höheren Bildung überwiegen auf einer sprachlichen Kompetenzebene die Fehler und negativen Auffälligkeiten.

Eine fachliche Zugehörigkeit ist indes schwer zu beurteilen. Obgleich auffällig häufig Begriffe aus dem IT-Bereich kontextuell korrekt verwendet werden, wäre der Rückschluss auf einen IT-Hintergrund verfrüht. Zu berücksichtigen gilt, dass Fachbegriffe aus IT-Bereichen in den vergangenen Jahrzehnten vermehrt Einfluss auf die allgemeine Ausdrucksweise hatten und sich Begriffe sich in der Umgangssprache manifestiert haben. Zudem ist Rinderwahn ein Darknet-Forum, dessen User sich mit den grundlegenden Mechaniken und der Bedienungsweise zumindest insoweit vertraut gemacht haben müssen, um in der Lage gewesen zu sein, einen Account in genanntem Forum zu eröffnen. Eine gewisse technische Affinität lässt sich somit bei allen Usern als Grundvoraussetzung annehmen. Die weitere Sprachproduktion von Chldprn bietet keinerlei hinreichend aussagekräftige Hinweise auf eine bestimmte fachliche Zugehörigkeit, womit dieser Bereich ungeklärt bleiben muss.

- *Auswertung digitaler Analysen und Mustererkennung*

Insgesamt wurden 36 Beiträge des Users aus dem Forum Rinderwahn erhoben, von denen hier exemplarisch zehn Einträge tabellarisch dargestellt werden[25].

Text Nr.	Längster Satz	Kürzester Satz	Avg. Satz	Absätze	Avg. Satz/Abs.	TTR	TTR2	Token
1	19	4	11,16667	1	6,00000	0,835	0,961	67
10	30	11	19,66667	1	4,00000	0,745	0,866	59
11	27	13	19,25000	1	5,00000	0,792	0,945	77
12	57	7	21,53846	5	4,40000	0,614	0,716	280
13	22	13	17,50000	1	3,00000	0,971	1	35
14	36	5	14,83333	1	6,00000	0,808	0,95	89
15	97	7	26,66667	1	13,00000	0,65	0,833	240
16	38	5	18,77778	2	6,00000	0,674	0,848	169
17	18	5	11,50000	1	2,00000	0,956	0,909	23
18	34	6	17,14286	1	9,00000	0,658	0,771	120

Tab. 1: Numerische Werte des Analyseprogramms zu User "Chldprn", 10 Schreiben (Exzerpt)

Zunächst stellt sich das Bild innerhalb der Textproduktion des Users nicht als inhomogen dar. Tabelle 1 lässt sich in zwei Arten von Schreiben untergliedern, solche mit einem Umfang von unter und solche mit einem Umfang von über 120 Tokens (als vom Verfasser gesetzte Grenze), wobei 6 Beiträge darunter und 4 darüber liegen. Allgemein ist hier zu bemerken, dass die eigentlich vorgegebene Mindestlänge von etwa 200 textrelevanten Tokens nicht eingehalten wird und somit mit größeren Inkonsistenzen zu rechnen ist. Hinsichtlich der Satzlänge ergibt sich, außer Beitrag 15, dass die längsten Sätze in der Umgebung von etwa 35 Wörtern angesiedelt sind. Abweichungen sind dabei zu erwarten. Mit 97 Wörtern fällt der längste Satz in Schreiben 15 aus dieser Reihe heraus, dies ist auf eine umständliche Klammerstruktur und Schachtelung zurückzuführen, die in ihrer sprachlichen Gestaltung und allgemeinen Komposition keine auffälligen Hinweise für Fremdproduktion aufweist. Die kürzesten Sätze reihen sich homogen im Umfeld von etwa 8 Wörtern pro kürzesten Satz ein. Bezüglich der verwendeten Absätze gibt es, wie für Foren typisch, zumeist aus einem Absatz komponierte Beiträge, wobei umfänglich längere Posts (280 Tokens, Beitrag 12) in mehrere (5) Absätze untergliedert werden. In Bezug auf TTR und TTR2 ergibt sich ein Muster, wie

[25] Zur Verdeutlichung wird die Analyse des Users Chldprn ausführlich dargestellt, bei den folgenden Usern wird auf eine detaillierte Analysebeschreibung verzichtet.

es sich von einem Textproduzenten der zuvor angenommenen Bildung und Ausprägung, unter Vorbehalt des Umfangs, vermuten lässt. Kürzere Beiträge sind nach der Entfernung der Stoppwörter nahe oder gleich einer TTR2 von 1, was keine doppelt vorkommenden Lexeme (in ihrer Basis, ohne jegliche Flexion) bedeutet. Weniger kongruent zur initialen Hypothese verlaufen die TTR2 Werte von längeren Beiträgen (Beitrag 12), eine TTR2 von 0,71 bei 280 Tokens ist, verglichen mit Beiträgen anderer hier vorgestellter User, als eher unterdurchschnittlich zu bezeichnen. Die Ursache dieser Diskrepanz lässt sich inhaltlich feststellen, Beitrag 12 behandelt Fernsehserien, die wiederholt namentlich gemeinsam mit TV relatierten Begriffen genannt werden.

Die drei am häufigsten verwendeten Phrasen des Users sind, ohne Stoppwörter, *AiW*[26] und *FF*[27] (beide jeweils insgesamt 14 Nennungen), *Topic* (insgesamt 12 Nennungen) und *Serie* (insgesamt 8 Nennungen). Mit Bezug auf kindesmissbräuchliche Intentionen ist zusätzlich *baby* (5 Nennungen insgesamt) sowie *child pornography* (insgesamt 4 Nennungen, als 2 Wörter gezählt und hier begrifflich subsummiert) zu berücksichtigen. Spezifische sprachliche Muster hinsichtlich spezieller Phrasen oder Begrifflichkeiten können für diesen User anhand der gegebenen Datenlage nicht zweifelsfrei festgestellt werden.

User: DeviantDennis

Syntax: Syntaktische Auffälligkeiten lassen sich in der Textproduktion von DeviantDennis nur gelegentlich feststellen. Seine Satzstruktur bewegt sich auf einem unauffälligen Niveau, zumeist bestehend aus einem Haupt- und einem weiteren untergeordneten Nebensatz. Parenthesen und Appositionen finden nur gelegentlich und komplexere Sätze fast ausschließlich im Kontext technischer Erklärungen Verwendung. Großschreibung am Satzanfang ignoriert er ausnahmslos und hält die Interpunktionsregeln hinsichtlich der korrekten Verwendung von Kommata nur sporadisch ein. Bei manchen Gelegenheiten wird ein Genitiv-Apostroph nach englischem Vorbild umgesetzt.

Lexik: Die lexikalischen Merkmakle sind insbesondere bei der Verwendung von nicht fachspezifischen Anglizismen (beispielsweise: *easy, that said, topic*) und die Nutzung von Apostrophen auffällig. Insbesondere Letzteres wird durchgehend, wenngleich nicht stringent korrekt, in abkürzender Funktion (beispielsweise *steh'n, 'ner oder geht's*) verwendet. Auch sind häufig konzeptuell klar mündliche Phrasen wie *haste, nee* oder *sonem* zu finden. Diminutive finden stellenweise (*Beispielbildchen*) Anwendung und sollen hier nicht unerwähnt bleiben. Komplexere Lexik (*suggerieren*) kommt nur selten zum Einsatz.

Orthografie: Im Rahmen orthographischer Betrachtungen fallen gelegentliche Fehler auf, die sich allerdings, wahrscheinlich aufgrund der allgemein größtenteils schlichten Lexik, auf wenige, dann aber eher komplexere Lexeme beschränken (**Addresse, *Philipinobitches*). Die Großschreibung ist nicht nur am Satzanfang fehlerhaft, sondern auch wiederholt, jedoch nicht durchgängig, im Fließtext.

Außerhalb der hier aufgeführten Kategorien auffällig ist zudem die parallele Verwendung von einfachen und doppelten Anführungszeichen, welche in ihrer Anwendung keinem durchgehend gleichbleibenden Muster folgen. Bei technischen Beschreibungen scheinen bestimmte Termini mit einfa-

[26] Alice in Wonderland.
[27] Forbidden Fruit.

chen, Eingaben mit doppelten Anführungszeichen gekennzeichnet zu werden. Diese Parallelanwendung findet jedoch auch außerhalb von technischen Beschreibungen oder Beschreibungen im Allgemeinen ihre Anwendung und folgt dort nicht den gängigen Konventionen. So wird beispielsweise ein Titel einer Diskussion mit einfachen Anführungszeichen zitiert, im selben Satz aber eine Interpretation dieses Titels mit doppelten Anführungszeichen. Warum der Wechsel erfolgt, ist nicht direkt ersichtlich.

- *Aussagen zum Alter*

Eine eindeutige Aussage zum Alter ist im Fall von DeviantDennis nicht möglich. Unterschiedliche denkbare Szenarien lassen sich anhand des sprachlichen Ausdrucks skizzieren, so wäre ein jüngerer Sprecher aufgrund der orthographischen und syntaktischen Unregelmäßigkeiten durchaus denkbar, wogegen jedoch die Bildung von Abkürzungen unter Nutzung von Apostrophen spricht. Letzteres ließe eher auf einen älteren Sprecher oder einen Sprecher mit umfangreicher Erfahrung im Kontext von Textproduktion schließen, wobei die flapsige Verwendung von Anglizismen im Sprachgebrauch dem entgegensteht. Für einen Sprecher mittleren Alters können keine unterstützenden oder entkräftenden Argumente an den gegebenen Textbeispielen festgemacht werden. Auf einer inhaltlichen Ebene könnte angebracht werden, dass aufgrund von Rückblicken in die Telemedientechnik der 80er/90er Jahre zumindest ein Sprecher der jüngeren Altersgruppe ausgeschlossen werden kann. Aufgrund unterschiedlicher Tipps und Ratschläge seitens des Textproduzenten, was die genauere Konfiguration des TOR-Browsers anbelangt, erscheint ein Sprecher der älteren Generation zumindest unwahrscheinlich. Da es keine textuellen Anhaltspunkte für einen Sprecher der mittleren Altersgruppe gibt, bleibt lediglich das Ausschlussverfahren. In diesem Fall erscheint eine Altersgruppe zwischen 25 und 60 Jahren am wahrscheinlichsten, ist jedoch aufgrund der hohen Spannweite wenig aussagekräftig.

- *Regionale Zugehörigkeit und Muttersprache*

Obwohl DeviantDennis nur gelegentlich lexikalische Elemente verwendet, die sich zur Bestimmung eines regionalen Hintergrundes eignen, verweisen diese (*haste* oder *sonem*) unter Berücksichtigung eines Vergleichs mit dem Digitalen Wörterbuch der Deutschen Sprache(DWDS) auf eine häufige Verwendung im Norden Deutschlands, insbesondere im Raum Berlin. Den nördlichen Raum als Sprachgebiet von DeviantDennis anzunehmen erscheint hinreichend wahrscheinlich, für eine weitere Spezifizierung auf den Raum Berlin ist die Datenlage jedoch zu vage. Die Muttersprache kann als Deutsch angenommen werden. Die Verwendung von Anglizismen lässt sich, unter Berücksichtigung der vorhandenen Probleme mit Artikelbildung, mit größerer Wahrscheinlichkeit auf Internet- und Forenjargon zurückführen. Fehler, die auf fremdsprachliches Verhalten zurückzuführen wären, fallen nicht auf.

- *Bildungsstand und fachlicher Hintergrund*

Hinsichtlich eines Bildungsstandes ergibt sich ein äußerst inhomogenes Bild, was stichhaltige Aussagen erschwert. Fehler in der Großschreibung am Satzanfang und die einfache Satzstruktur lassen auf einen niedrigeren Bildungsstand schließen. Dies wird jedoch durch die Apostrophimplementierung in Abkürzungen und durch die andernfalls korrekte Verwendung der Groß- und Kleinschreibungsregeln negiert. Orthographische Unregelmäßigkeiten lassen sich teilweise auf englische Schreibweisen zurückführen (engl. *address* vs. dt. *Adresse*), die Verwendung von korrekt umgesetzten und komplexeren Wörtern (wie *suggeriert)* bleibt nicht aus. Somit erscheint es schlüssiger, von

einer unüberlegten Textproduktion seitens eines lediglich ungeübten Autors auszugehen als ein niedriges Bildungsniveau anzunehmen.

Eine fachliche Zugehörigkeit ist im Kontext der hier aufgeführten User und Foren zu betrachten. Obwohl DeviantDennis ausführliche Erklärungen zu den unterschiedlichen Konfigurationsoptionen des TOR-Browsers gibt und auch an anderer Stelle, unter Verwendung des zu erwartenden Fachvokabulars, Bezug auf die IT nimmt, soll hier lediglich von einem technisch affinem User ausgegangen werden, der nicht notwendigerweise einen professionellen Hintergrund in diesem Bereich hat.

- *Auswertung digitaler Analysen und Mustererkennung*

Insgesamt wurden 184 Beiträge des Users aus dem Forum Rinderwahn erhoben, von denen hier exeplarisch zehn Einträge tabellarisch dargestellt werden.

Text Nr.	Längster Satz	Kürzester Satz	Avg. Satz	Absätze	Avg. Satz/Abs.	TTR	TTR2	Token
1	28	20	22,6000	3	2,00000	0,814	0,904	113
10	28	11	19,3333	4	1,25000	0,810	0,84	58
100	39	20	29,5000	1	3,00000	0,915	0,937	59
101	11	11	11,0000	2	1,00000	1	1	11
102	14	14	14,0000	1	1,00000	0,928	1	14
103	10	8	9,00000	2	2,00000	0,925	1	27
104	24	9	16,5000	1	2,00000	0,969	1	33
105	14	4	11,0000	3	1,66667	0,954	1	44
106	30	15	22,5000	2	1,50000	0,955	0,970	45
107	34	6	15,5000	8	2,12500	0,653	0,827	248

Tabelle 2: Numerische Werte des Analyseprogramms zu User "DeviantDennis", 10 Schreiben (Exzerpt)

Insgesamt am häufigsten verwendete Tokens sind *Ermittler* (mit gesamt mindestens 10 Treffern), Metzelder (9 Treffer) und Staatsanwaltschaft (8 Treffer). Mit besonderem Fokus auf sexuelle Gewalt gegen Kinder sind die Tokens *Ass* (12), *Pornographie* und *Schrift* (jeweils 9 Mal), *Kinder* (6) und *Mädchen* (37) auffällig. Ein musterhafter Sprachgebrauch unter besonderer Berücksichtigung einer pädophilen Lexik lässt sich anhand der gegebenen Datenlage nicht erkennen. Als eine allgemeine Anmerkung, und im Gegensatz zu Chldprn, ist die Verwendung von Emojis anzumerken.

User: garfield66

Syntax: Die häufigen syntaktischen Auffälligkeiten der Textproduktion des Users lassen sich zum größten Teil auf inkorrekte Syntaxumsetzung zurückführen und sind fast ausschließlich, da häufig wiederholt und strukturell homogen, als Error zu betrachten. Der Satzbau stellt sich als eher unstrukturiert und gedankenflussartig dar, eine geordnete Struktur ist nur selten zu erkennen und der erwartete Kasus wird häufig nicht umgesetzt *(und dann gibts gerademal [...] lächerliche Likes eventuell* oder *Fragst dann in deinen eigenen Forum die User über einen Banner zu Mithilfe und kein Schwein meldet sich).* Parenthesen oder Appositionen werden kaum bis gar nicht umgesetzt, oder sind mangels korrekter Interpunktion nicht als solche zu erkennen. Kommata werden zur Satzstrukturierung entweder nicht verwendet oder stehen alleine zwischen Leerzeichen, häufig kommen mehrere gleiche Satzzeichen *(???)* zum Einsatz, welche ebenfalls durch ein Leerzeichen vom Satz

abgetrennt werden. Englische Fachterminologie wird, wie bei den vorangehenden Usern auch, hinsichtlich des Genus inkorrekt determiniert und somit ebenfalls fehlerhafte Artikel verwendet (z.B. *diesen Topic*, wobei Topic = Thema, das; Neutrum).

Lexik: Die lexikalische Ebene entspricht dem Niveau der syntaktischen, insbesondere was orthographische Fehler anbelangt. Der User verwendet häufig prädominant mündlich konzipierte Umgangssprache in schriftlicher Form (*einkassiert, Bullen* oder *aufgeklatscht)*, teilweise auch in abgekürzten Versionen (*gehts, heut, wünsch* oder *kauf*). Auch Anglizismen werden, nicht immer korrekt, umgesetzt. So beginnen viele Beiträge mit *Hello* obwohl es sich ansonsten um einen vollständig deutschsprachigen Beitrag handelt oder es ist von *pics* als – durchaus nicht ungeläufige – englischsprachige Abkürzung für Bilder (Pictures) die Rede. Auffällig ist ebenfalls die teilweise inkorrekte Pluralbildung wie sie konsistent beispielsweise bei Forum – **Forums* (nicht Foren) produziert wird.

Orthografie: Eindeutig am ergiebigsten ist die Analyse der zahlreichen und schwerwiegenden orthographischen Auffälligkeiten. Grundlegende deutsche Rechtschreibung kann nicht vorausgesetzt werden, so ist durchgehend **währ* anstelle von *wär(e)* in Verwendung, eine Seite **läd* anstelle von *lädt*, bestimmte Personengruppen stehen im **Focus* der Ermittler, Geräte werden mit Schadsoftware **infisziert,* es geht um **Platzersparniss* beim Wechsel von **Kasetten* oder die Gefahr von **Vieren* (Computerviren) für ein bestimmtes digitales System. Groß- und Kleinschreibungsregeln finden nicht durchgängig Anwendung, wird allerdings noch am zuverlässigsten korrekt verwendet. Eine Unterscheidung zwischen *dass* und *das* findet kaum statt, die Fehler diesbezüglich sind häufig. Auch mit Hinblick auf verwendete Anglizismen unterlaufen schwerwiegende Fehlbildungen, beispielsweise **loggin* anstelle von Login oder **unfähr* anstelle von *unfair.* In wenigen Situationen werden die Regeln der alten Rechtschreibung, so zum Beispiel bei *muss* umgesetzt, welches in der zweiten Person Singular in mindestens einer Situation als **mußt* realisiert wird.

Das Niveau und die Qualität der Textproduktion des hier untersuchten Users verbessert sich deutlich, sobald von technischen Erläuterungen gesprochen wird, insbesondere im Hinblick auf das von garfield66 angesprochene Thema Staatstrojaner. Syntaktisches Niveau und lexikalischer Ausdruck verbessern sich enorm, die orthographische Fehlerrate ist nicht mehr vorhanden, die Interpunktion wird gänzlich korrekt angewandt und zuvor erkennbare wiederkehrende Fehlermuster oder Auffälligkeiten sind verschwunden. Vor diesem Hintergrund liegt der Schluss nahe, dass bestimmte, insbesondere längere, Beiträge des Users nicht gänzlich durch ihn selbst verfasst, sondern in Teilen fremdproduziert und mittels Copy-Paste eingefügt wurden. Dieser Verdacht erhärtet sich durch das Wiederkehren bekannter Fehlermuster zum Ende des Beitrags, nach dem Abschluss der längeren Erklärung zum Thema Staatstrojaner.

- *Aussagen zum Alter*

Das Alter des Autors ist mittels der verfügbaren lexikalischen Marker eher als ein Angehöriger der mittelalten bis älteren Generation zu verorten. Die häufige Produktion von *muß* wäre als einziges lexikalisches Indiz diesbezüglich zwar zugegebenermaßen dürftig, wird aber durch weitere auffällige und in schriftlicher Textproduktion eher unübliche Elemente gestützt. So wurde *zischen,* laut DWDS*,* im Sprachgebrauch am häufigsten zwischen 1940 und 1960, *einkassiert* zwischen 1955 und 1960 sowie später erneut um 2015 herum, und *reingezogen* um das Jahr 1990 herum verwendet. Außerdem sind zahlreiche semantische Aspekte zu finden, insbesondere Aussagen im Hinblick auf die Historie von Windows, den ersten Anfängen des Internets und Kassetten, welche die initiale

Einschätzung des Alters stützen und von keinem anderen sprachlichen Merkmal direkt widerlegt werden können. Der User ist wahrscheinlich nicht unter 40 Jahre alt.

- *Regionale Zugehörigkeit und Muttersprachlichkeit*

Eine innerdeutsche regionale Zuordnung wird, aufgrund der zahlreichen orthographischen Fehler und der generellen Struktur, den teilweise fremdproduzierten Textkomponenten und der allgemeinen Sprachverwendung, deutlich erschwert. Bestimmte Worte, wie *aufgeklatscht*, sind nach einem Abgleich mit dem DWDS zwar häufiger im süddeutschen Raum anzusiedeln, jedoch nicht zwingenderweise typisch für diesen regionalen Sprachraum und in ihrer Verwendungshäufigkeit deutlich zu gering, um einen legitimen Rückschluss auf eine bestimmte Region innerhalb Deutschlands zu erlauben. Dementsprechend muss dieser Punkt unbeantwortet bleiben.

Hinsichtlich der Muttersprachlichkeit fällt insbesondere die schlechte deutsche Rechtschreibung in Kombination mit ungewöhnlichen Anglizismen wie *Focus,* sowie unübliche Phrasen wie *gibt zu denken auf* ins Auge. Die Sprachproduktion eines nicht muttersprachlich Deutschen Autors lässt sich zwar nicht eindeutig ausschließen, so kann das gegebene Fehlerbild, insbesondere hinsichtlich der lautlichen Kongruenz zwischen Aussprache und Verschriftlichung, durchaus als eine Parallele zur türkischen Verschriftlichung interpretiert werden. Zu erwartende Fehlübertragungen, wie sie von einer agglutinierenden Sprache wie dem Türkischen zu erwarten wären, finden allerdings nicht statt, eine Interlanguage ist somit inhomogen. Vor diesem Hintergrund erscheint die Wahrscheinlichkeit einer niedrigeren Bildung, wie im nächsten Punkt dargelegt werden soll, bedeutend höher. In Bezug auf die umgesetzten Anglizismen, insbesondere *k -> c* Wandlungen, kann in jedem Fall dennoch ein geübter oder gar muttersprachlicher Sprecher des Englischen aufgrund von Bildungen wie *unfähr* oder *loggin* mit ziemlicher Sicherheit ausgeschlossen werden.

- *Aussagen zu Bildungsstand und fachlicher Zugehörigkeit*

Anhand des mehrfach erwähnten Fehlerbilds ist mit an Sicherheit grenzender Wahrscheinlichkeit von einem Sprecher ohne höherer Bildung und subsequent ohne jegliche Erfahrung in textproduktiven Bereichen oder Berufen auszugehen. Obwohl der User mehrfach tiefergehendes informationstechnologisches Wissen beweist und dieses durchaus strukturiert verschriftlichen kann, lässt er sich als eigentlicher Autor dieser Passagen in Zweifel ziehen. Die generelle Verfügbarkeit dieser Informationen und der erhöhte Kontakt von Darknet-Usern mit dieser technischen Nische trifft nach wie vor zu, dennoch ist das allgemeine Bildungsniveau nicht hoch genug, um von einem beruflichen Hintergrund in der IT-Branche ausgehen zu können. Eine fachliche Zugehörigkeit kann somit nicht hinsichtlich sprachlicher Marker determiniert werden.

- *Ergebnisse der digitalen Analyse und Mustererkennung*

Insgesamt wurden 71 Beiträge des Users aus dem Forum Rinderwahn erhoben, von denen hier exeplarisch zehn Einträge tabellarisch dargestellt werden.

Text Nr.	Längster Satz	Kürzester Satz	Avg. Satz	Absätze	Avg. Satz/Abs.	TTR	TTR2	Token
1	16	16	16,00000	1	1,00000	0,937	1	16
10	21	10	14,00000	2	3,50000	0,797	0,95	84
11	26	6	15,33333	6	1,66667	0,851	0,96	94

12	17	3	8,77778	3	3,33333	0,802	0,888	81
13	33	4	14,06250	7	2,57143	0,690	0,785	226
14	36	9	20,60000	2	3,50000	0,776	0,942	103
15	20	3	11,40000	2	3,00000	0,864	0,966	59
16	19	7	11,33333	3	2,33333	0,897	0,923	68
17	25	25	25,00000	1	1,00000	0,96	1	25
18	38	4	16,30769	26	2,30769	0,508	0,741	850

Tabelle 3: Numerische Werte des Analyseprogramms zu User "garfield66", 10 Schreiben (Exzerpt)

Die am häufigsten verwendeten Begriffe des Users spiegeln dessen technische Affinität wider, die bereits in der klassisch durchgeführten Analyse zu erkennen war. *Linux* (22 Treffer), *Staatstrojaner* bzw. *Bundestrojaner* (11 + 6 Treffer) sowie *Tor* (10 Treffer) werden mit Abstand am häufigsten verwendet. Im Rahmen des sprachlichen Ausdrucks von sexueller Gewalt gegen Kinder wird *AiW* (6 Treffer) sowie *Toddler* (3 Treffer) am häufigsten gebraucht. Von den in der klassischen Analyse bereits aufgezeigten linguistischen Verhaltensmustern abgesehen ergeben sich keine non-redundanten sprachlichen Muster in der digitalen Auswertung, ein gemeinsamen Lekt über alle analysierten User ergibt sich bis dato ebenfalls noch nicht.

Als eine allgemeine Anmerkung sei auch hier die Verwendung von Emoticons in der Textproduktion erwähnt.

User: girliefriend

Syntax: Syntaktische Auffälligkeiten im Sinne von Fehlern gibt es im Falle dieses Users kaum. Die Sprache ist stets wohlgeformt, die zu erwartende Grammatik fast gänzlich korrekt, bis auf gelegentliche Fehler in der Interpunktion. Parenthesen oder Appositionen kann der Autor korrekt verwenden. Die Satzstruktur ist häufig komplex, der Einsatz von Stilmitteln, insbesondere Sarkasmus und Ironie, sind deutlich erkennbar. Konjunktive werden hinsichtlich der Differenzierung von Konjunktiv I und II grammatikalisch korrekt umgesetzt. Diminutive werden ebenfalls stellenweise korrekt angewandt. Von kleineren Verstößen gegen die Interpunktionsregeln abgesehen lassen sich keine wiederholt auftretenden Unregelmäßigkeiten oder Konventionsbrüche diesbezüglich feststellen. Die syntaktische Umsetzung befindet sich durchgehend auf einem hohen Niveau und lässt sich als anspruchsvoll bezeichnen.

Lexik: Das Bild der Syntax zeichnet sich auch auf lexikalischer Ebene ab. Der User setzt in seiner Textproduktion wiederholt Begrifflichkeiten ein, die nach DWDS als eher selten (*Karaffe, Lügenbaron* oder *Nixe*) einzuordnen oder einer fach- bzw. bildungssprachlichen Gruppe (*Nymphomanin* oder *Lethargie*) zuzuordnen sind. Diesem lexikalischen Niveau stehen zwei konträre Komponenten gegenüber. Einerseits werden häufig Worte repetitiv verwendet (*fast*), ebenfalls zu erkennen in der späteren digitalen Analyse, andererseits finden auch umgangssprachliche und deutlich mündlich konnotierte Begrifflichkeiten Einzug in die Textproduktion (zum Beispiel: *Stimmt´s*[28]*, Pussy,* oder *Muschi, äh, jaaa* oder *waaas)*.

Orthografie: Auffälligkeiten in der Orthographie lassen sich in zwei Kategorien aufgliedern, beide davon kommen im Verhältnis zur Textmenge jedoch eher selten vor. Zunächst ist hier die fehler-

[28] Die Nutzung eines inkorrekten Apostrophs sei an dieser Stelle erwähnt.

hafte Groß-/Kleinschreibung, nebenliegenden Buchstaben auf der Tastatur (**stiudieren,* herumeyperimentierte* oder **dreche)*, fehlenden Buchstaben, die wahrscheinlich auf der Tastatur nicht komplett gedrückt wurden (zum Beispiel **gbt*) oder gelegentlich ausgelassenen Leerzeichen zu nennen. Seltener beziehungsweise einmalig treten auch konkrete Errors auf (**wahr*[29] oder **Indianers)*. Nicht zwischen Error und Lapse zu unterscheiden sind gelegentlich auftretende dass/das Fehler.

- *Aussagen zum Alter*

Eine Einschätzung hinsichtlich eines wahrscheinlichen Alters lässt sich im Falle des Users hinsichtlich unterschiedlicher Punkte kongruent vornehmen. Viele Posts, v.a. solche, die auf einen zweiten reaktiv Bezug nehmen, kommen in ihrer Aufmachung und Gestaltung einem persönlichen Brief nahe, u.a. durch Anrede und Grußformel - ein Verhalten das eng mit der Produktionserfahrung von Briefen verwandt ist. Die Nutzung konzeptueller Briefe ist in den jüngeren Altersgruppen zwischen zehn und 44 Jahren, bei einem Anteil von 89 und 61 Prozent, laut einer Erhebung des statistischen Bundesamtes[30], größtenteils durch Messaging-Dienste[31] abgelöst worden. Eine Assoziation mit konzeptuellen Briefen erfolgt somit sehr wahrscheinlich tendenziell bei Sprechern des höheren mittleren bis fortgeschrittenen Alters. Dieser Altersgruppe ist der User somit zuzuordnen. Weiterhin zu untermauern ist diese Einschätzung durch die wiederholte Umsetzung der traditionellen Rechtschreibung bezüglich der Trennung von Präverbfügungen (*wiedergetroffen*, *reingesteckt* oder *wiedergefunden*) welche mit der Rechtschreibreform 1996 zu trennen war und erst mit der erweiterten, dritten Reform 2006 wieder zusammen geschrieben werden konnte. Da ein 20-Jähriger Sprecher (Einschulung mit sechs Jahren, Geburt im Jahr 2000), nicht nur aufgrund der zuvor angestellten Überlegungen mit hoher Wahrscheinlichkeit nicht in Frage kommt, sondern auch aufgrund des gesamtheitlichen Niveaus in sowohl Lexik als auch Syntax und Satzbau mit wenigen sprachfokussierten Ausnahmen - Studenten der Sprach- und Literaturwissenschaften seien hier erwähnt – mit großer Wahrscheinlichkeit auszuschließen ist, ergibt sich ein definitiv anzunehmendes Mindestalter von 30 Jahren. Bezieht man semantische Aspekte und inhaltliche Rückblicke an zurückliegende zeitliche Gegebenheiten und deren sprachlicher Darstellung ("Ich könnte heute *noch* das Alte Beta Max System reparieren") mit ein, erscheint die Einordnung des Users in eine höhere Altersgruppe von 50 Jahren oder älter immer wahrscheinlicher.

- *Regionale Zugehörigkeit und Muttersprachlichkeit*

Der hohe Sprachstandard, die zu erkennende sprachliche Wohlgeformtheit und die Komplexität der Texte lassen einen fremdsprachlichen Autor mit an Sicherheit grenzender Wahrscheinlichkeit ausschließen. Die Datenlage hinsichtlich einer regionalen Einordnung erweist sich als recht dürftig, da sich der User an die hochdeutsche Schriftsprache hält. Regional identifizierbare Begriffe sind kaum vorhanden, als eines der am ehesten zuzuordnenden Lexeme sei *Nackedei* erwähnt, welches das etymologische Wörterbuch des DWDS im 19. Jahrhundert im norddeutschen Raum verortet. Anhand dieses einmalig vorkommenden Wortes jedoch eine regionale Zugehörigkeit, insbesondere

[29] Ob die einmalige Produktion von **wahr* u.U. aufgrund einer erfolgten längeren Konversation mit dem User garfield66 und dessen repetitiver Nutzung von **währ* erfolgt ist, kann an dieser Stelle nicht holotisch von der Hand gewiesen werden.

[30] Vgl. hierzu: www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/IT-Nutzung/Tabellen/internetaktivitaeten-personen-alter-ikt.html

[31] Dass die Kommunikation eines Messaging-Dienstes konzeptuell mündlich aufgebaut ist stellt sich als trivial dar.

vor dem Hintergrund der zeitlichen Diskrepanz, vorzunehmen ist im Kontext des gesamten Textumfangs des Users nicht haltbar und muss somit unspezifiziert bleiben.

- *Aussagen zu Bildungsstand und fachlicher Zugehörigkeit*

Hinsichtlich einer adäquaten Evaluation des Bildungsstandes gilt es im Falle des Users mehrere Komponenten einzubeziehen. In seinen Texten finden bestimmte Stilmittel, wie beispielsweise Ironie oder rhetorische Fragen korrekte und situativ passende Anwendung. Die Verwendung anspruchsvollerer und seltenerer Lexik erfolgt routiniert, ohne erkennbare Fauxpas und ohne dabei gestellt zu wirken. Im Vergleich zu den ebenfalls produzierten umgangssprachlichen Begriffen ist die gehobene Lexik als bildungsrelevanteres Merkmal zu verstehen und somit stärker zu gewichten. Auf gleicher Ebene anzusiedeln ist die allgemeine Komplexität der produzierten Sätze sowie die grammatikalische Umsetzung bestimmter sprachlicher Hürden, zum Beispiel die korrekte Umsetzung des Präteritums starker Verben (*las*). Diese Merkmale lassen zum einen eindeutig auf das Erfahren von höherer Bildung des Autors schließen und eine gewissen Erfahrung mit der Produktion von Texten vermuten, ein schreibender Beruf ist dabei nicht auszuschließen.

- *Ergebnisse der digitalen Analyse und Mustererkennung*

Insgesamt wurden 368 Beiträge des Users aus dem Forum Rinderwahn erhoben, von denen hier exemplarisch zehn Einträge tabellarisch dargestellt werden.

Text Nr.	Längster Satz	Kürzester Satz	Avg. Satz	Absätze	Avg. Satz/Abs.	TTR	TTR2	Token
1	24	3	9,44444	5	5,80000	0,603	0,815	257
10	30	3	10,13846	13	5,15385	0,477	0,759	662
11	21	3	9,14286	12	2,66667	0,633	0,892	262
12	26	3	8,50000	8	3,62500	0,65	0,925	240
13	40	3	8,61061	157	5,80255	0,244	0,445	6766
14	38	3	9,60909	191	4,27225	0,263	0,474	6410
15	15	4	8,66667	3	2,33333	0,888	0,925	54
16	31	3	10,86842	6	7,33333	0,588	0,843	423
17	17	4	10,00000	2	3,00000	0,880	0,894	42
18	23	3	8,62857	10	3,90000	0,631	0,866	307

Tabelle 4: Numerische Werte des Analyseprogramms zu User "girliefriend", 10 Schreiben (Exzerpt)

Die häufigsten verwendeten Tokens sind in Bezug auf sexuelle Gewalt gegen Kinder, *Mädchen (*45 Treffer), *Kind* (34 Treffer) und *Pädophilie* (12 Treffer), allgemeine Worthäufungen sind, aufgrund der literarischen Textproduktion (*Herbert* mit 48 Treffern) verzerrend und werden dementsprechend nicht mit einbezogen.

V. Ergebnisse und Schlussfolgerungen

Die im Rahmen dieser Studie vorgestellten User wurden zufällig exemplarisch ausgewählt um den Analyseprozess zu veranschaulichen und stellen nicht den vollständigen Analyseumfang dar. Die Aussagen des gesamten Analyseprozesses stellen sich zusammenfassend wie folgt dar:

Ist jeder User eine individuelle Person? Hinsichtlich der Frage, ob die Anzahl der User mit einer gleichen Zahl legaler Personen korreliert oder zu Gunsten gefälschter Accounts, sogenannter *Sockpuppets*, mehrere Nutzerkonten die von einer legalen Person betrieben werden, verschoben ist, kann eindeutig beantwortet werden. Die Anzahl analysierbarer Accounts, das heißt Accounts mit ausreichender Aktivität in der Textproduktion, sind hinsichtlich ihres sprachlichen Verhaltens eindeutig zu unterscheiden, *Sockpuppets* lassen sich mit an Sicherheit grenzender Wahrscheinlichkeit ausschließen. Das sprachliche Verhalten der analysierten User ist zu unterschiedlich. Es gilt dabei zu berücksichtigen, dass viele Accounts deren Aktivität nicht den benötigten Textumfang zu liefern vermag, in diese Analyse nicht einbezogen werden konnten, so dass eine Aussage dazu nicht möglich ist. Auch die Zahl inaktiver Accounts, sogenannter *Lurker*, kann mit der hier vorgestellten Methode nicht zugeordnet werden.

Fremdproduktionen sind in den Beiträgen mancher Accounts anhand der Mustererkennungen deutlich zu erkennen. Dass ein Account jedoch von mehreren unterschiedlichen Personen betrieben wird, lässt sich anhand der Daten nicht bestätigen. Alle auffälligen Fremdproduktionen in Beiträgen wurden als Zitate von entweder anderen Usern verwendet oder lassen sich auf Texte von journalistischen Beiträgen zurückführen, die sich nicht nur durch Sprachanalyse vom linguistischen Verhalten der User, sondern auch durch Kennzeichnung betreffender Stellen durch dieselben abgrenzen lassen.

Gibt es einen Pädolekt? Hinsichtlich der Frage, ob sich aus den analysierten Beiträgen ein sprachliches Muster ableiten lässt, welches sich im Besonderen auf Angehörige einer pädophilen Interessensgruppe anwenden lässt, lässt sich bei der momentanen Datenlage keine definitive Aussage treffen. Obwohl bestimmte wiederkehrende Begrifflichkeiten und Phrasen auffallen, insbesondere in Bezug auf andere thematisch verwandte Foren wie AiW oder FF, sind diese noch nicht auf empirische Weise eindeutig genug unikal der betreffenden Gruppe zuzuordnen, die Datenlage zu inkonsistent und an dieser Stelle noch nicht aussagekräftig genug. Selbstverständlich ist das Gesprächsthema eines Pädophilenforums einzigartig für Angehörige dieser Gruppe, daraus lässt sich allerdings kein Sprachgebrauch ableiten, der nach außen hin erkennbar bleibt.

Literatur

Brinker, Klaus (2014):
Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden. 8., überarb. und erw. Auflage. Berlin

Dern, Christa (2009):
Autorenerkennung Theorie und Praxis der linguistischen Tatschreibeanalyse. Stuttgart

Dittmar, Norbert (1997):
Grundlagen der Soziolinguistik - Ein Arbeitsbuch mit Aufgaben. Tübingen

Fobbe, Elikia (2011):
Forensische Linguistik - Eine Einführung. Tübingen

Jäger, Siegfried (1980):
Standardsprache. In: Althaus, Hans Peter/Henne, Helmut/ Wiegand, Herbert Ernst (Hrsg.): Lexikon der Germanistischen Linguistik. 2., Tübingen

Sherman, Chris; Gary Price (2003):
"The invisible web: uncovering sources search engines can't see."

Reed, Michael G., Paul F. Syverson, David M. Goldschlag (1998):
"Anonymous Connections and Onion Routing" Naval Research Laboratory.

Kaufmann, Susan (2008):
Fortbildung für Kursleitende Deutsch als Zweitsprache, Ismaning

Kniffka, Hannes (1990):
Texte zu Theorie und Praxis forensischer Linguistik.Tübingen

Löffler, Heinrich (2005):
Germanistische Soziolinguistik. Berlin

Olsson, John (2008):
Forensic Linguistics: An Introduciton To Language, Crime and the Law, London, Bloomsbury

Seiffert, Jan (2010):
Verstellungs- und Imitationsstrategien in Erpresserschreiben: Empirische Studien zu einem Desiderat der forensisch-linguistischen Textanalyse. Bonn

Selinker, Larry (2001):
Analysing interlanguage: how do we know what learners know? in: Second Lan-guage Research 17,4

Steinke, Wolfgang (1990):
Die linguistische Textanalyse aus kriminalistischer Sicht. In: Kniffka, Hannes (Hrsg.): Texte zu Theorie und Praxis forensischer Linguistik. Tübingen, S. 321-338