Korpusbereinigung für größere Textmengen. Eine (kurze) Problematisierung und ein Lösungsansatz für Duplikate

Adelmann, Benedikt
adelmann@informatik.uni-hamburg.de
Universität Hamburg, Deutschland

Gius, Evelyn
gius@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Korpuserstellung in der Literaturwissenschaft

Eine neue Praxis

Die Zusammenstellung von Primärtexten als Forschungsgrundlage ist in der Literaturwissenschaft Alltagsgeschäft. Die beiden Standardfälle der (nicht-digitalen) Korpusanalyse gelten in Bezug auf die Korpuszusammensetzung als wenig problematisch:

1. Wenn die Forschungsfrage eine spezifische Textgrundlage erfordert, steht das Korpus von vornherein fest (z. B. alle Romane von Thomas Mann für eine Studie zur Repräsentation von Krankheit in Manns Romanwerk).
2. In Korpora kanonischer Texte werden weitergehende Fragen untersucht (z. B. ausgewählte Texte aus dem Realismus zur Untersuchung des Ausdrucks von Idylle im Realismus).

Mit der Verfügbarkeit digitaler Texte wurde die Praxis der Korpuserstellung jedoch erweitert und es wurde offensichtlich, dass sie nicht ohne weiteres in die bestehende literaturwissenschaftliche Disziplinarmatrix (Kuhn 1970) integriert werden kann. Viele der digital verfügbaren Texte sind nämlich weder kanonisch noch repräsentativ, die Qualität einzelner Texte ist aus philologischer Sicht oft fragwürdig und ein Korpus enthält trotz der Vielzahl der verfügbaren digitalen Texte selten die gesamte Population relevanter Texte, sondern nur eine Teilmenge.[1]

Über die philologisch einwandfreie Auswahl von Texten hinaus birgt die Kuration von Korpora weitere Herausforderungen.[2] Die vermutlich größte ist, sich einen Überblick über ein Korpus zu verschaffen, das mehr Texte enthält, als man lesen kann. Bei einer nicht von einem Einzelnen erfassbaren Textmenge kann selbst eine scheinbar einfache Aufgabe wie die Erkennung von Duplikaten unlösbare Probleme bereiten.

Ein exemplarisches Korpus

Der vorgestellte Ansatz wurde für ein Korpus entwickelt, das in einem Forschungsprojekt zur geschlechtsspezifischen Darstellung von Krankheit in literarischen Texten im Rahmen der Forschungskooperation hermA[3] erstellt wurde. Ausgangspunkt war das Kolimo-Korpus, das mehr als 42.000 literarische und nicht-literarische deutsche Texte vor allem von 1880 bis 1930 aus drei großen Repositorien deutscher Texte enthält: dem Deutschen Textarchiv[4], dem TextGrid Repository[5] und Projekt Gutenberg-DE[6] (vgl. Herrmann & Lauer 2017). Wir haben alle Prosatexte von 1870 bis 1920 ausgewählt, die ursprünglich auf Deutsch verfasst waren (vgl. Gius et al., 2019).

Im daraus resultierenden Korpus von mehr als 2.500 Texten mussten Artefakte behandelt werden, die durch unterschiedliche Digitalisierungsstrategien verursacht wurden. Nicht nur die Erhaltung von Sonderzeichen wie das recht häufige lange s (ſ) zwischen oder innerhalb der Repositorien war inkonsistent, sondern auch die Verwendung von Bindestrichen (Wortverbindung, Worttrennung an Zeilenumbruch, andere Bindestriche, Gedankenstriche) oder die Kodierung von Zeilenumbrüchen und Absätzen. Diese Probleme konnten mit einer relativ einfachen Heuristik angegangen werden.

Die Identifizierung von Duplikaten

Ein schwierigeres Problem ist die Frage der Duplikate: Insbesondere bei der Zusammenstellung eines Korpus aus verschiedenen Quellen kann es vorkommen, dass der gleiche Text mehrfach vorhanden ist. In der Regel ist es nicht erwünscht, mehr als eine Instanz desselben Textes im Korpus zu haben, da die Überrepräsentation einzelner Werke bei statistischen Analysen zu verzerrten Ergebnissen führen kann. Daher sollte die Identifizierung von Duplikaten ein wesentlicher Bestandteil der Korpuserstellung sein.

Dabei gibt es zwei Probleme: Erstens wächst die Anzahl der ungeordneten Werkpaare, die alle potenziell Duplikate sein könnten, quadratisch mit der Anzahl der Werke. In unserem Korpus mit gut 2.500 Texten müssten deshalb 3,1 Millionen Werkpaare überprüft werden. Zweitens ist auch für jedes einzelne Textpaar die Feststellung, ob es sich um Duplikate handelt, aufgrund von Metadaten- und Textinkonsistenzen eine nicht-triviale Aufgabe. Ansätze zur *text reuse detection* (z. B. Bär et al. 2012) blenden den kombinatorischen Aspekt oft aus.

Wir haben mit zwei Methoden zur automatischen Duplikatidentifizierung experimentiert. Beide sind

Heuristiken für die Suche nach Werkpaaren, die Duplikate sind; sie lösen aber nicht das daran anschließende Problem, zu entscheiden, welche von mehreren Instanzen tatsächlich in das Korpus aufgenommen werden sollen.

Für die Evaluation wurden alle Duplikatkandidaten, die von mindestens einer der beiden Methoden gefunden wurden, manuell auf ihre Richtigkeit überprüft. Wir berichten über den Prozentsatz der automatisch als Duplikate identifizierten Paare, die tatsächlich Duplikate sind (Precision), und den Prozentsatz der tatsächlichen Duplikate, die automatisch identifiziert werden (Recall). Allerdings lässt sich der Recall nur exakt bestimmen, wenn alle 3,1 Millionen Werkpaare manuell untersucht werden. Als Annäherung verwenden wir daher stattdessen den Prozentsatz der bei der manuellen Prüfung identifizierten tatsächlichen Duplikate (Gesamtzahl: 355), die ebenfalls automatisch gefunden werden.[7]

Die erste Methode basiert auf Metadaten und ist deshalb schnell genug, um alle ungeordneten Werkpaare im Korpus zu testen. Für jedes Werkpaar werden Autor*innen- und Titelinformationen verglichen. Was die Autor*innen-Informationen betrifft, so wird die sogenannte Edit-Distanz (Levenshtein 1965) der vollständigen Namen der Autor*innen berechnet; die Edit-Distanz ist die kleinste Anzahl von Zeicheneinfügungen, Zeichenlöschungen und Zeichenersetzungen („Edits“), mit der der erste Autor*innen-Name in den zweiten umgewandelt werden kann.

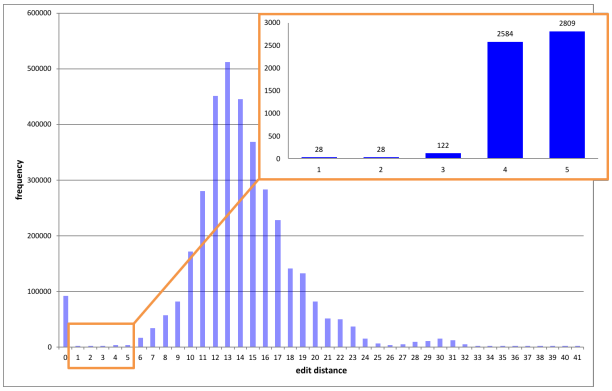

*Abbildung 1: Edit-Distanzen Autor*innen-Namen (Schwellenwert bei 2)*

Für die Titelinformationen haben wir das Maß leicht modifiziert: Wir verwenden die kleinste Anzahl von Edits, die einen der Titel in einen *Teil* (d. h. Teilzeichenkette) des anderen verwandeln können, mit der zusätzlichen Einschränkung, dass ganze Wörter vollständig abgeglichen werden müssen, um unangemessen kurze Entfernungswerte zu vermeiden (sonst ließe sich beispielsweise „Tot“ nach nur zwei Edits als Teilzeichenkette von „Das Jüngste Gericht“ finden, hätte also Abstand 2).[8] Durch dieses modifizierte Maß sollen hohe Titeldistanzen vermieden werden, wenn Untertitel in nur einem der beiden Titel enthalten sind.[9]

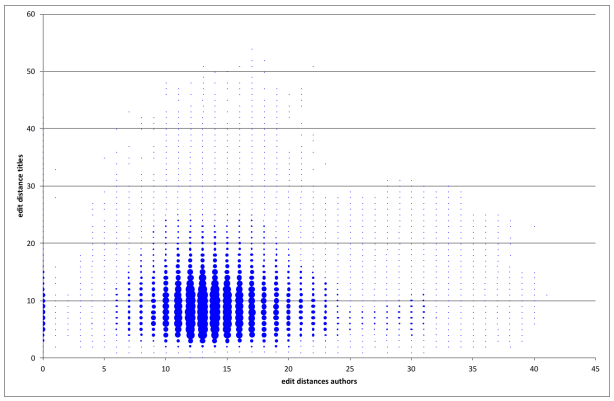

*Abbildung 2: Kombinierte Edit-Distanzen für Autor*innen und Titel*

Zwei Texte werden als Duplikate betrachtet, wenn der Abstand sowohl beim Autor*innen-Namen als auch beim Titel höchstens so hoch wie der Schwellenwert ist. Bei einem Schwellenwert von 2 wurden 672 Duplikatpaare mit einer Precision von 51,9 % und einem Recall von 98,3 % identifiziert. Unter den sechs nicht identifizierten Duplikaten finden sich beispielsweise zwei Werke von Karl May mit abweichender Behandlung von Sammelband-/Einzelwerktitel („Ardistan und Dschinnistan. 1. Band“ vs. „Der Mir von Dschinnistan“; „Satan und Ischariot III“ vs. „Im Todesthale“) und ein Fall, in dem bei einem der beiden Werke fälschlich der Name des Autors als Titel eingetragen war (Ferdinand von Saar, „Vae victis!“).

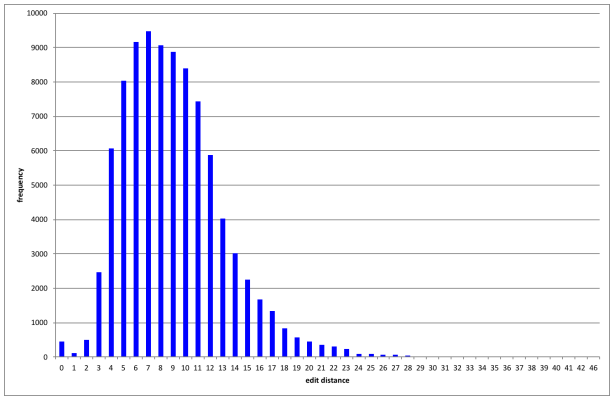

*Abbildung 3: Edit-Distanzen in Titeln, wo Autor*innen-Namen max. Edit-Distanz 2 aufweisen*

Das zweite Verfahren berechnet die Edit-Distanzen von Volltexten. Da dies eine zeitaufwändige Operation ist, beschränken wir uns auf den Vergleich von Texten, bei denen der Autor*innen-Name eine Edit-Distanz von maximal zwei hat. Manuelle Überprüfungen zeigten, dass diese Schwelle alle Rechtschreibfehler und Varianten

in unseren Daten erfasst, verschiedene Autor*innen mit ähnlichen Namen jedoch ausnimmt. Für Volltexte verwenden wir wieder Teilzeichenketten-Edit-Distanzen, da ein Text mehr oder weniger vollständig in einem anderen enthalten sein kann (z. B. bei Anthologien). Zugunsten der Rechenzeiten verwenden wir wortbezogene Distanzen mit Insertionskosten gleich Deletionskosten gleich Substitutionskosten gleich eins.[10]

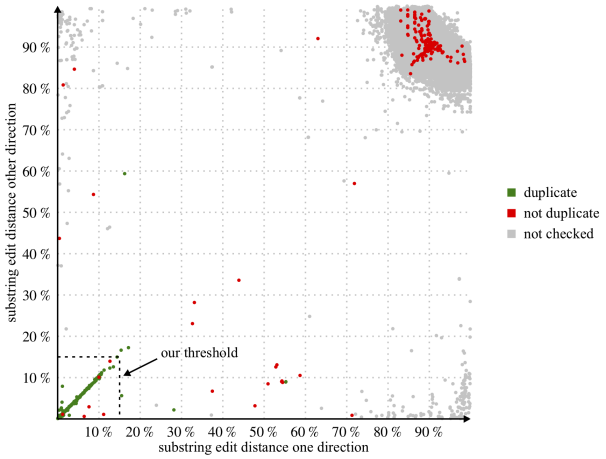

Abbildung 4: Edit-Distanzen der Volltexte

Zwei Texte gelten als Duplikate, wenn die Teilzeichenketten-Edit-Distanzen für beide Richtungen (1 als Teilzeichenkette von 2 oder umgekehrt), dividiert durch die Länge des jeweils als Teilzeichenkette einzubettenden Texts, unter 15 % liegt. Auf diese Weise können wir 307 Duplikate mit einer Precision von 98 % und einem Recall von 84,8 % bestimmen.

678 Paare wurden so durch mindestens ein Verfahren als Duplikate identifiziert (Precision: 52,4 %),[11] 301 durch beide (Precision: 98 %, Recall: 83,1 %). Letzteres ist bedeutsam für praktische Anwendungen, bei denen man mit Methode 1 kostengünstig Werkpaare auswählen möchte, die anschließend mit der teuren Methode 2 getestet werden sollen.

Fazit

Die Digitalisierung hat die Arbeit mit literarischen Korpora erheblich gefördert. Die schiere Menge an Texten in einem Korpus muss sowohl technisch als auch konzeptionell unterstützt werden. Für die Erreichung dieser Ziele ist es umso wichtiger, Qualitätskriterien für die Zusammenstellung von Korpora im Hinblick auf die verfügbaren Daten, d. h. für Texte aus heterogenen Quellen unterschiedlicher Qualität, zu entwickeln und umzusetzen. Zusätzlich zu diesen noch zu entwickelnden Kriterien für die wissenschaftliche Qualitätssicherung können einige pragmatische Entscheidungen die Qualität eines Korpus und seiner Texte in Fällen mit geringer Daten- und insbesondere Metadatenqualität erheblich verbessern.

Der vorgestellte Ansatz zum Umgang mit Duplikaten kann zusammen mit den genannten Vorverarbeitungsschritten ein wichtiger erster Schritt in diesem Prozess sein.

Fußnoten

1. Für verschiedene Typen von Datensammlungen oder Korpora vgl. Schöch (2017).
2. Einige der Probleme und Lösungsmöglichkeiten wurden in Gius et al. (2019) vorgestellt, für einen breiteren Überblick vgl. Lauer & Herrmann (2018).
3. Vgl. https://www.herma.uni-hamburg.de und Gaidys et al. (2017) [alle Links in diesem Beitrag wurden am 15.09.2019 das letzte Mal abgerufen].
4. Vgl. http://www.deutschestextarchiv.de/
5. Vgl. https://textgridrep.org/
6. Vgl. https://www.gutenberg.org/wiki/DE_Hauptseite (Achtung: seit 2018-02 nicht mehr über deutsche IP-Adressen abrufbar).
7. Unter den manuell geprüften Werkpaaren finden sich keine Duplikate innerhalb DTA, zwei innerhalb TextGrid und 46 innerhalb Gutenberg sowie 23 zwischen DTA und TextGrid, 27 zwischen DTA und Gutenberg und 257 zwischen TextGrid und Gutenberg.
8. Formal: Edit-Distanzen auf Wortebene mit folgenden Kosten: Einfügungen, Löschungen: an Anfang und Ende der Wortsequenzen 0, sonst Anzahl Zeichen im eingefügten/gelöschten Wort Substitutionskosten: zeichenbasierte Edit-Distanzen
9. Wir haben bewusst auf spezialisierte Heuristiken für Untertitel, Schreibweisealternativen und andere für dieses Korpus spezifische Titelabweichungsphänomene verzichtet.
10. Ob die Edit-Distanz über einem vorgegebenen Schwellwert liegen würde, lässt sich außerdem in einigen Fällen auch ohne deren tatsächliche Berechnung (und damit erheblich effizienter) feststellen (Tateishi & Kusui 2008), allerdings waren wir für diesen Beitrag an exakten Maßen interessiert. Außerdem ist die dortige Formel auf vollständige Edit-Distanzen und nicht auf Teilzeichenfolgen-Distanzen ausgelegt.
11. Eine Recall-Angabe ist hier nicht sinnvoll, da die manuelle Auswertung sich genau auf diese Menge beschränkt.

Bibliographie

Bär, Daniel / Zesch, Torsten / Gurevych, Iryna (2012): Text reuse detection using a composition of text similarity measures. *Proceedings of COLING 2012*, S. 167–184.

Gaidys, Uta / Gius, Evelyn / Jarchow, Margarete / Koch, Gertraud / Menzel, Wolfgang / Orth, Dominik / Zinsmeister, Heike (2017): Project Description. HermA:

Automated Modelling of Hermeneutic Processes. In *Hamburger Journal für Kulturanthropologie* 7. S. 119–123.

Gius, Evelyn / Krüger, Katharina / Sökefeld, Carla (2019): Korpuserstellung als literaturwissenschaftliche Aufgabe. In *DHd2019 Book of Abstracts*.

Herrmann, Berenike / Lauer, Gerhard (2017): Das „Was-bisher-geschah" von KOLIMO. Ein Update zum Korpus der literarischen Moderne. In *DHd 2017 Digitale Nachhaltigkeit Book of Abstracts*. S. 107–111.

Kuhn, T. S. (1970). *The Structure of scientific revolutions*. 2nd ed., enlarged. Chicago: Chicago Univ. Press.

Lauer, Gerhard / Herrmann, Berenike (2018): Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne. *Osnabrücker Beiträge zur Sprachtheorie* 92. 7–35

Levenshtein, Vladimir (1965): Binary codes capable of correcting deletions, insertions, and reversals. Englische Übersetzung in: *Soviet Physics Doklady*, Bd. 10, Nr. 8, S. 707–710, 1966.

Schöch, Christof (2017): Aufbau von Datensammlungen. In: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (Hrsg.): *Digital Humanities: eine Einführung.* Stuttgart: J. B. Metzler Verlag. S. 223–233.

Tateishi, Kenji / Kusui, Dai (2008): Fast Duplicated Documents Detection using Multi-level Prefix-filter. *Proceedings of the Third International Joint Conference on Natural Language Processing*, Bd. II.