

# Was Lesende denken: Assoziationen zu Büchern in Sozialen Medien

## Beck, Jens

jens\_beck@gmx.de

Institut für Maschinelle Sprachverarbeitung, Universität  
Stuttgart, Deutschland

## Willand, Marcus

marcus.willand@ilw.uni-stuttgart.de

Institut für Literaturwissenschaft, Universität Stuttgart,  
Deutschland

## Reiter, Nils

Nils.Reiter@ims.uni-stuttgart.de

Institut für Maschinelle Sprachverarbeitung, Universität  
Stuttgart, Deutschland

## Einleitung

Im vorliegenden Abstract stellen wir eine Methode sowie erste Ergebnisse der Analyse von Entitäten-Assoziationen realer Leserinnen und Leser vor.

Literaturwissenschaftliche Rezeptions-, Lese- und Lesertheorien gehen seit ihren hermeneutischen und wirkungsästhetischen Anfängen (Schleiermacher 1838, insb. 309f.; Iser 1976) von professionellen (Dijkstra 1994), informierten (Fish 1970, 86), Modell- (Eco 1979) oder sogar idealen (Schmid 2005) Lesern aus (vgl. Willand, 2014). Diesen wird die Kompetenz zugeschrieben, idealerweise sämtliche Textmerkmale referentialisieren zu können, wobei je nach literaturtheoretischer Provenienz unterschiedliche Kontexte die Grundlage der Zuschreibungen an den Text bilden. Dazu gehören u.a. Informationen über den Autor oder über die sozialhistorischen Bedingungen der Textproduktion, über die Rezeptionsbedingungen, über Vorgänger- oder zeitgenössische Texte oder über Wissen aus dem Bereich der Literaturwissenschaftlerin bzw. des Lesers selbst.

An bestimmte Wissensbestände dieser *realen* Leserinnen und Leser literarischer Texte können wir uns durch eine computergestützte empirische Analyse von Rezeptionszeugnissen aus sozialen Medien annähern. Konkret ist unser Ziel die Rekonstruktion und Analyse der von literarischen Texten ausgelösten Assoziationen. Dabei beschränken wir uns auf die Assoziationen, die reale oder fiktive Entitäten betreffen, also etwa Personen des öffentlichen Lebens oder Figuren aus fiktionalen Werken.

Die Plattform Goodreads bietet Leserinnen und Lesern die Möglichkeit des freien schriftlichen Austauschs über literarische Texte in einer großen Community. 55 Mio.

Mitglieder haben bis 2017 über 50 Mio. Reviews geschrieben, wobei die Besprechungen die Inhalte der Bücher selbst und nicht - wie etwa bei Verkaufsplattformen wie Amazon - die Distribution, den Preis o.ä. fokussieren (Piper et al. 2015).

## Verarbeitung

Als Grundlage unserer Analysen wurden die Reviews in einer lokalen Datenbank gespeichert.

Die Datenbank enthält 1,3 Millionen englischsprachige Reviews zu 5.481 besprochenen Texten. Die Reviews umfassen insgesamt etwa 150 Mio. Tokens, d.h. uns steht eine große Datenmenge zur Extraktion der Entitäten zur Verfügung. In einem ersten Schritt wurden die Reviews bereinigt: HTML-Tags wurden entfernt und Wiederholungen von mehr als dreimal dem gleichen Zeichen oder Wort auf drei reduziert.

Zur Extraktion der Entitäten aus den Reviews haben wir den Stanford Named Entity Recognizer (Finkel et al., 2005) verwendet. Der Tagger klassifiziert die gefundenen Entitäten in mehrere Klassen. Für uns ist die Klasse „PERSON“ relevant, da diese alle gefundenen Entitäten von Personen enthält.

Im nächsten Schritt disambiguieren wir die extrahierten Entitäten, da z.B. ein Name wie „Harry“ auf viele mögliche Träger des Namens verweisen kann. Mit Hilfe von UKB (Agirre et al., 2009) und UKB-wiki (Agirre et al., 2015) können den Entitäten Wikipedia-Seiten zugeordnet werden, welche die möglichen Entitäten repräsentieren. Für diese Disambiguierung verwendet UKB den PageRank-Algorithmus (Page et al. 1999), der Dokumente nach ihrem Verlinkungsgrad bewertet. Sobald Namen wie „Ron“ und „Dumbledore“ im selben Kontext erwähnt werden, wird die Wahrscheinlichkeit größer, dass mit „Harry“ *Harry Potter*, mit „Ron“ *Ron Weasley* und mit „Dumbledore“ *Albus Dumbledore* aus der Romanreihe *Harry Potter* gemeint sind, weil diese Entitäten aus dem selben Kontext kommen und dies in der Wissensbasis Wikipedia durch Verlinkungen explizit ablesbar und quantifizierbar ist.

UKB-wiki stellt einen herunterladbaren Graphen zur Verfügung, der Wikipedia-Seiten und Links auf andere Wikipedia-Seiten repräsentiert. In einem mitgelieferten Wörterbuch sind Entitäten mit allen möglichen Entitäten (Verweise auf Wikipedia-Seiten) aufgeführt.

Die auf diese Weise gewonnenen Wikipedia-Einträge wurden anschließend hinsichtlich des ontologischen Status der referenzialisierten Entität kategorisiert, also ob es sich um eine reale Person oder fiktionale Figur handelt. Dazu wurde die Wissensbasis DBpedia<sup>1</sup> verwendet, die die Daten aus Wikipedia strukturiert und maschinenlesbar kodiert. Da die Disambiguierung Wikipedia-Einträge liefert, können wir anhand dieser die auf den zugehörigen DBpedia-Eintrag zugreifen. Über DBpedia lassen sich neben ontologischen Kategorien auch andere Eigenschaften extrahieren, die für eine Analyse

ggf. interessant sind, etwa Geschlecht oder Relationen zu anderen Figuren.

Die extrahierten Daten werden zunächst als Tabelle gespeichert und erlauben somit eine flexible weitergehende Verarbeitung, etwa in einem Netzwerk. Eine Zeile der Tabelle besteht aus dem Werktitel, der disambiguierten Entität (Verweis auf Wikipedia Seite), einer Liste der extrahierten Entitäten aus den Reviews, einer Liste von Review-IDs, um nachvollziehen zu können in welchen Reviews der Name erwähnt wird, der Anzahl der Erwähnungen und der Angabe ob es sich um eine Figur handelt oder nicht.

Titel	Disambiguierte Entität (Verweis auf Wikipedia Seite)	Extrahierte Entität	Review IDs	Anzahl der Erwähnungen	Handelt es sich um eine fiktionale Figur?
The Hound of the Baskervilles	Agatha_Christie	christie, agatha_christie, agatha_christy	470784120, ..., 188608568	20	False
The Hound of the Baskervilles	Spock	spock	42971473	1	True
The Hound of the Baskervilles	Robert_Downey_Jr.	robert_downey_jr, robert_downey	103754109, ..., 1250976986	18	False
The Hound of the Baskervilles	Ann_Radcliffe	ann_radcliffe	435380655	1	False

**Tabelle 1:** Auszug aus den extrahierten Daten. Die extrahierten Entitäten stammen aus den Reviews zu *The Hound of the Baskervilles*.

## Zwischenergebnisse

Um ein exemplarisches Resultat zu präsentieren, haben wir Reviews zu "The Hound of the Baskervilles" (deutsch: "Der Hund von Baskerville") analysiert. Unter den häufig erwähnten Entitäten finden sich erwartungsgemäß Sherlock Holmes, Dr. Watson, sowie der Autor Arthur Conan Doyle. Weitere häufig erwähnte Figuren aus der fiktionalen Welt des Sherlock Holmes' sind James Mortimer und Charles Baskerville. Aber auch Professor Moriarty wird häufig erwähnt, obwohl er in diesem Buch der Sherlock-Reihe gar nicht auftaucht. Das System erzeugt jedoch auch Fehler. Beispielsweise wird der Antagonist Stapleton zwar sehr oft erwähnt, da zu ihm aber kein eigener Wikipedia-Eintrag existiert, wird er fälschlicherweise mit dem Fußballspieler Frank Stapleton verknüpft. Henry Baskerville, der Sohn von Charles und Erbe des Anwesens, wird im Buch fast

durchgehend als Sir Henry bezeichnet, und kommt mit diesem Namen ebenfalls häufig in den Reviews vor. Da auch für ihn kein eigener Wikipedia-Eintrag existiert und der Name Henry extrem mehrdeutig ist, werden eine Reihe klar falscher Entitäten verknüpft: Henry II. von Frankreich; Henry County (Alabama); oder Henry I. von England.

Bemerkenswert sind insbesondere jedoch die referenzialisierten extra-textuellen Entitäten, also diejenigen, die nicht aus der fiktionalen Welt Sherlocks stammen. Es finden sich etwa *Hercule Poirot* und *Agatha Christie* unter den erwähnten Entitäten, was als klares Zeichen dafür gesehen werden kann, dass die Leserinnen und Leser den Text vor dem Hintergrund eines starken Gattungsbewusstseins rezipieren. Dafür spricht auch, dass mit *Benedict Cumberbatch*, *Robert Downey, Jr.*, *John Barrymore* und *Jeremy Brett* gerade die Schauspieler unter den assoziierten Referenzen vertreten sind, die in einer der vielen Verfilmungen die Rolle des Sherlock Holmes verkörpert haben.

## Fehleranalyse

Das am häufigsten auftretende Problem ist das Fehlen eines Wikipedia-Eintrages für eine Figur. In der englischsprachigen Wikipedia sind fiktionale Figuren zwar nicht per se davon ausgeschlossen -- Richtschnur hier ist deren "Notability". Viele Figuren sind jedoch nur auf den Einträgen des entsprechenden Werks erwähnt. Da der Algorithmus nicht in der Lage ist, *keinen* Eintrag zu liefern, wird in solchen Fällen eben ein anderer Eintrag verwendet, auch wenn dieser relativ weit entfernt sein mag. Eine technische Lösung wäre sicher, nur ab einem gewissen Schwellwert eine Disambiguierung vorzunehmen, und die nicht-disambiguierten Einträge zumindest als solche erkennen zu lassen. Eine andere Möglichkeit läge in der (zusätzlichen) Verwendung von Literaturlexika, die (womöglich) eine größere Abdeckung zu fiktionalen Figuren aufweisen. Beide Optionen werden wir in zukünftigen Arbeiten genauer untersuchen.

Da es sich bei den Reviews letztlich um Inhalte aus einem sozialen Medium handelt, kommt es auch vor, dass Namen falsch geschrieben werden oder gar der gesamte Text schriftsprachliche Konventionen übergeht. Prima vista sind diese Fälle im Vergleich zu Buchrezensionen zwar häufig anzutreffen, wir können das Problem aber umgehen, indem wir nur diejenigen Erwähnungen berücksichtigen, die mehr als einmal vorkommen. Festzuhalten bleibt aber ebenfalls, dass die Texte im Vergleich zu z.B. Twitter-Daten deutlich sauberer sind.

Eine weitere mögliche (jedoch noch nicht tatsächlich beobachtete) Fehlerquelle liegt in der Natur des PageRank-Algorithmus: Wenn eine Figur in einem Werk existiert, ein Leser oder eine Leserin jedoch explizit z.B. eine Person des öffentlichen Lebens mit dem gleichen Namen erwähnt, wird der Algorithmus diese Erwähnung eher der Figur zuschlagen, da diese dichter mit anderen Figuren verknüpft ist.



Taschenbücher. Band 1026. Hrsg. v. Gerhard Köpf, 201–218. München.

**Christmann, Ursula / Margrit Schreier** (2003). „Kognitionspsychologie der Textverarbeitung und Konsequenzen für die Bedeutungskonstitution literarischer Texte“. Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte. Revisionen. Hrsg. v. Fotis Jannidis, Gerhard Lauer, Matías Martínez & Simone Winko, 246–284. Berlin.

**Dijkstra, Katinka** (1994): Leseentscheidung und Lektürewahl. Empirische Untersuchungen über Einflussfaktoren auf das Leseverhalten. Berlin.

**Dimitrov, Stefan / Zamal, Faiyaz / Piper, Andrew / Ruths, Derek** (2015): “Goodreads vs Amazon: The Effect Of Decoupling Book Reviewing And Book Selling”, in: International Conference on Web and Social Media (ICWSM-14).

**Eco, Umberto** (1979): The Role of the Reader. Explorations in the Semiotics of Texts. Bloomington, IN.

**Finkel, Jenny Rose / Grenager, Trond / Manning, Christopher** (2005): “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”, in: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363–370.

**Fish, Stanley E.** (1970): „Literature in the Reader: Affective Stylistics“, in: *New Literary History* 1(2): 123–162.

**Funke, Mandy** (2003). „Das Abenteuer der Fragebögen. Aspekte zur empirischen Wirkungsforschung in der DDR“. Wissenschaft und Systemveränderung. Rezeptionsforschung in Ost und West – Eine konvergente Entwicklung? Euphorion. Band 44. Hrsg. v. Wolfgang Adam, Holger Dainat & Gunther Schandera, 119–126. Heidelberg.

**Groeben, Norbert** (1979). „Zur Relevanz empirischer Konkretisationserhebungen für die Literaturwissenschaft“. Empirie in Literatur- und Kunstwissenschaft. Grundfragen der Literaturwissenschaft. Hrsg. v. Siegfried J. Schmidt, 43–82. München.

**Iser, Wolfgang** (1976). Der Akt des Lesens. Theorie ästhetischer Wirkung. Band 636. München.

**Page, Lawrence / Brin, Sergey / Motwani, Rajeev / Winograd, Terry** (1999): “The PageRank Citation Ranking: Bringing Order to the Web”, technical Report. Stanford InfoLab.

**Schleiermacher, Friedrich** (1838): Hermeneutik und Kritik mit besonderer Beziehung auf das Neue Testament. Aus Schleiermachers handschriftlichem Nachlasse und nachgeschriebenen Vorlesungen herausgegeben von Friedrich Lücke. In: Friedrich Schleiermacher’s sämtliche Werke. Berlin: Reimer.

**Schmid, Wolf** (2005): Elemente der Narratologie. Narratologia. Band 8. Berlin.

**Veldhues, Christoph** (1995): "Gleich- und Gegenüberstellung". Intratextuelle und intertextuelle Bedeutung in der Literatur. Zeitschrift für französische Sprache und Literatur 40/3 (1995), 243–267.

**Willand, Marcus** (2014): Lesermodelle und Lesertheorien. Historische und systematische Perspektiven. Narratologia. Band 41. Berlin.

**Wübben, Yvonne** (2009). „Lesen als Mentalisieren? Neuere kognitionswissenschaftliche Ansätze in der Leseforschung“. Literatur und Kognition. Bestandsaufnahmen und Perspektiven eines Arbeitsfeldes. Poetogenesis. Band 6. Hrsg. v. Martin Huber & Simone Winko, 29–44. Paderborn.