## **Maschinelle Analyse des Märchens**

### **Einführung**

Angesichts ihrer einfachen Formen und Strukturen sind Genres der mündlichen Überlieferung beliebte Objekte zur Demonstration der Effektivität analytischer und generativer Modelle der künstlichen Intelligenz. Viele dieser Versuche basieren auf dem wissenschaftlichen Erbe der Finnischen Schule, insbesondere auf den ATU- und MIT-Klassifikationssystemen. Sie beschränken sich jedoch darauf, diese Instrumentarien zu automatisieren, ohne ihre Grundmethoden nach den Anforderungen des neuen Ansatzes, den Methoden der digitalen Geisteswissenschaften, grundlegend zu reformieren (DECLERCK u. a., o. D.; D'Huy 2019: passim).[[1]](#footnote-1) Es ist weiterhin populär, die durch die strukturalistische Analyse des Märchens erfassten Textbausteine, die sogenannten Propp'schen Funktionen, zu formalisieren und ihre maschinelle Erkennbarkeit zu prüfen.[[2]](#footnote-2)

Im Gegensatz zu diesen Versuchen, die als Retrodigitalisierung der traditionellen Methoden rekapituliert werden können, konzentriert sich die vorliegende Arbeit auf die Entwicklung einer qualitativ neuen, auf digitalen Ansätzen basierenden Methode zur Lösung traditioneller Fragestellungen der vergleichenden Märchenforschung.

Die wichtigste Aufgabe dieses neuen Ansatzes besteht in der formalisierten Erschließung des Forschungsobjekts und nicht in der Anpassung vorhandener Erkenntnisse und Systeme an digitale Anwendungen oder umgekehrt.

Um den artikulierten Unterschied zwischen der auf dieser Strategie beruhenden Forschung und den Ergebnissen der bereits veröffentlichten Abhandlungen im Bereich der elektronischen Märchenforschung zu verdeutlichen, versuchen wir hier beispielhaft, eine der jüngsten Publikationen zu diesem Thema zu besprechen. Es handelt sich um den Aufsatz von J. Eklund und anderen unter dem Titel: *"Teaching Tale Types to a Computer: A First Experiment with the Annotated Folktales Collection"* (Eklund et al. 2023: passim)[[3]](#footnote-3).

Wie aus dem genannten Artikel hervorgeht, setzen sich die Autoren zum Ziel, den Computer dazu zu bringen, die Zugehörigkeit der Texte zu den bekannten Märchentypen vorherzusagen. Als Basis für die Klassifikation dient der internationale Märchenkatalog (ATU). Die maschinelle Erkennung des Typs beruht auf den inhaltlichen Eigenschaften des Textes, die wiederum durch die im Motivkatalog von Thompson (MIT) erfassten inhaltlichen Einheiten identifiziert werden. Laut dem Aufsatz haben die Autoren den Computer in die Lage versetzt, die MIT-Elemente im Text zu erkennen und je nach deren Kombination im Text vorherzusagen, welchem Typ der Text zuzuordnen ist.

Zunächst etwas Näheres zu den Katalogen ATU und MIT. Der erste Versuch einer Katalogisierung von inhaltlichen Übereinstimmungen in den Märchenrepertoires geht auf die Vertreter der mythologischen Schule zurück. Es handelt sich um die handschriftlichen Sagenkonkordanzen der Brüder Jacob und Wilhelm Grimm. Leider haben die Arbeiten der Brüder nicht den Stand der Veröffentlichungsreife erreicht.

Die von den Brüdern Grimm erkannte Notwendigkeit der Katalogisierung von Übereinstimmungen verschiedener Märchenrepertoires wurde in relativ kleinem Maßstab von ihrem Anhänger J. G. von Hahn umgesetzt. Der von ihm in seiner griechischen und albanischen Märchensammlung angefügte Formelkatalog gilt als erstes Verzeichnis der Märchentypen (Hahn 1864: 74-85; vgl. Aarne 1913: 3, Fn. 3).

Die Finnische Schule legte großen Wert auf die Katalogisierung des Märchens. A. Aarne erklärte die Notwendigkeit, einen "systematischen Typenkatalog auszuarbeiten", aufgrund des Mangels an einem System, das in der Vielfalt der Volksmärchen die einzelnen Märchentypen feststellen und zu einem geordneten Ganzen vereinen könnte (Aarne 1910: iii). Damit begann die Finnische Schule bzw. die vergleichende Märchenforschung, über stabile Inhaltselemente zu sprechen und dabei mit entsprechenden wissenschaftlichen Begriffen zu operieren.

Die Klassifikation nach Typen gewährleistet, dass ähnliche Geschichten durch gemeinsame Identifikationsmerkmale gekennzeichnet und somit miteinander vergleichbar sind. Von Anfang an wurde der Katalog ausschließlich für angewandte Zwecke eingeführt. Die Implementierung und die kontinuierliche Pflege des Klassifikationssystems führten jedoch dazu, dass der Katalog allmählich von einem rein angewandten zu einem wissenschaftlichen System wurde. der Vorgang und die Konsequenzen dieser Entwicklung lassen sich wie folgt beschreiben:

1. Durch empirische Erkenntnisse über die stabil wiederholbare Inhaltselemente in Märchen aus unterschiedlichen Kulturen entstand der Bedarf, deren Vergleichbarkeit zu katalogisieren.
2. Die Katalogisierung erfüllt eine angewandte Funktion und zielt darauf ab, die Aufgabe eines Nachschlagewerks zu erfüllen.
3. Nationale Märchenkataloge identifizieren und klassifizieren Texte nach diesem System in ihren nationalen Repertoires.
4. Nach dieser Phase bietet der Katalog zusätzliche und erkenntnisrelevante Informationen für die Forschung, um folgende Fragen zu beantworten:

(a) Wie sieht die Rangliste der Typen im nationalen Märchenrepertoire aus?

(b) Aus welchen Variationen bestehen einzelne Typen in diesen Repertoires?

1. Später wird versucht, die Typenbeschreibungen aus verschiedenen lokalen Katalogen zusammenzuführen, um verallgemeinerte Zusammenfassungen einzelner Typen darzustellen.

Der Versuch, sowohl spezifische als auch allgemeine Informationen in einem Internationalen Katalog des Märchens zusammenzuführen, ist an der Komplexität der dafür kodierten Informationen gescheitert. Der Katalog identifizierte zum Beispiel neue Fälle von Kombinationen oder Kontaminationen, konnte jedoch nicht gleichzeitig angeben, aus welchen Repertoires diese Fälle stammen oder ob sie dort typische oder eher zufällige Einzelfälle darstellen.

Dieser Versuch geht auf den amerikanischen Erzählforscher Thompson zurück. Er entwickelte außerdem den Index von internationalen Motiven (MIT) und integrierte diesen in den Typenkatalog, um ihn seiner Meinung nach in ein noch präziseres Informationssystem umzuwandeln. Ab jetzt sind neben den traditionellen Zusammenfassungen des Märchens auch die Indizien zu finden, die auf die festen Inhaltselemente, die sogenannten Motive, hinweisen. Zum Beispiel sind bei der Beschreibung des Typs *ATU 300* *The Dragon-Slayer* folgende Indizes zu finden:

*B421 = Helpful dog.*

*B312.2. Helpful animals obtained by exchange.*

*B11.2.3. 1. Seven-headed dragon.*

*B1 1.10. Sacrifice of human being to dragon.*

*S262. Periodic sacrifices to a monster.*

*T68.1. Princess offered as prize to rescuer.*

*D1975. Dragon-fighter's magic sleep.*

*D1978.2. Waking from magic sleep by letting tear fall on sleeper.*

Diese stammen aus dem internationalen Motivindex (MIT), wo sie nach dem Prinzip der Taxonomie in verschiedene Themen und Unterthemen eingeteilt sind. Zum Beispiel:

*B = Animal Motifs.*

*B11= Dragon.*

*B11.2 = Form of dragon.*

*B11.2.3 = Many-headed dragon.*

Die danach kommende Zahl weist auf die Menge der Drachenköpfe hin:

*B11.2.3. 1 = Seven-headed dragon.*

*B11.2.3.2 = Three-headed dragon.*

*B11.2.3.3 = Six-headed dragon.*

usw.

Nach der Integration der MIT-Systemelemente in das ATU-System wurden die traditionellen Typenzusammenfassungen um zusätzliche MIT-Indizes ergänzt. Ein Beispiel dafür ist die folgende Zusammenfassung des oben genannten Typs: ***ATU 300*** ***The Dragon-Slayer.*** *A youth (e.g. by exchange three wonderful dogs [B421, B312.2]. He comes to a town where people are mourning and learns that once a year a (seven-headed) dragon [B11.2.3.1] demands a virgin as a sacrifice [B11.10, S262]. In the current year, the king’s daughter has been chosen to be sacrificed, and the king offers her as a prize to her rescuer [T68.1]. The youth goes to the appointed place. While waiting to fight with the dragon, he falls into a magic sleep [D1975], during which the princess twists a ring (ribbons) into his hair; only one of her falling tears can awaken him [D1978.2]. usw.*

Nach Ansicht der Autoren des betrachteten Aufsatzes gilt folgendes: Wenn diese MIT-Elemente im Text in einer mehr oder weniger ähnlichen Reihenfolge vorkommen, kann man daraus schließen, dass es sich um den Typ ATU 300 handelt. Durch das Training des Modells mit entsprechend gelabelten Trainingsdaten kann es dann die Fähigkeit erlangen, die Zugehörigkeit des Textes zum passenden Typ vorherzusagen.

Die Effizienz eines solchen Unternehmens ist nicht zu unterschätzen. Dennoch stellt sich die Frage, ob wir dadurch tatsächlich mehr über die Unterschiede und Ähnlichkeiten zwischen den einzelnen Texten oder Repertoires erfahren können. Dies hängt vollständig davon ab, wie adäquat wir das Objekt formalisieren können und nicht davon, wie geschickt wir den Computer bei der Anwendung der kompromissartigen Arbeitsmethode einsetzen.

Wenn wir einen Text erschließen müssen und dabei über keinen Computer, sondern nur ein Zettelsystem verfügen, müssen wir so viele Zettel erstellen, wie es Lemmata im Text gibt, multipliziert mit der Anzahl der Sätze, in denen diese Lemmata ein- oder mehrmals vorkommen. Wenn wir jedoch nach der Erledigung dieser Aufgabe einen Computer zur Verfügung haben, müssen wir nicht die handschriftlich ausgefüllten Zettel als einzelne Datensätze in den Computer eintragen, sondern können den fließenden Text eingeben und diesen später je nach den Fragestellungen auswerten.

Genau dasselbe gilt für die Wiedergeburt der Arbeitsmethode der vergleichenden Märchenforschung. Die Idee, Märchen durch eine gemeinsame Instanz (tertium comparationis) zu vergleichen, bleibt dabei nach wie vor relevant. Welches Modell wir für die Konstruktion des Typs als Vergleichsobjekt verwenden, muss allerdings aus der Perspektive computergestützter Lösungsstrategien bestimmt werden.

Eine automatisierte Erkennung des Textes ohne kritische Auseinandersetzung mit dem Erbe der vergleichenden Märchenforschung aus der Zeit vor den digitalen Geisteswissenschaften kann nichts anderes sein als eine Retrodigitalisierung einer Methode, die nicht nach den Anforderungen der formalisierten Erschließung des Forschungsobjektes erarbeitet wurde. Wie bereits angedeutet, distanzieren wir uns von dieser Herangehensweise und versuchen, die Schwächen der bisherigen Methode der vergleichenden Märchenforschung durch eine neue Lösung des traditionell erkannten Forschungsproblems mithilfe computergestützter Methoden zu bewältigen.

Der Grund für die Entstehung des vorliegenden Vorhabens liegt im Bedarf an einem auf philologischer Analyse basierenden Modell, das für die maschinelle Analyse von Märchentexten verwendet werden kann. Die maschinelle Analyse des Forschungsobjektes anhand der formalisierte Beschreibung der Struktur des Märchens beinhaltet sowohl die Automatisierung der Segmentierung des Märchens in Strukturelemente unterschiedlicher Größe als auch das Erzeugen eines zusammenfassenden Bildes aufgrund des Vergleichs der je nach Fragestellung ausgewählten Forschungsobjekte. Beispiele hierfür sind:

* Die vergleichende Analyse des Typs X anhand der Repertoires A, B, C usw.
* Die Kombination von Motiven innerhalb der Episode X.
* Der der Status der handlungstragenden Figuren im Motiv Y.
* Und vieles mehr.

Die Lösung der maschinellen Analyse des Märchens als nächster Schritt nach der Formalisierung des Forschungsobjektes ist mit der Entwicklung der entsprechenden digitalen Infrastruktur verbunden. Im Folgenden werden wir einige der wichtigsten Elemente dieser Infrastruktur betrachten.

1. zu einem der jüngster [↑](#footnote-ref-1)
2. Als jüngster Versuch sind die Publikationen von Finlayson (2012, 2016) sowie Lendvai u.a. (2010a, 2010b) zu nennen. [↑](#footnote-ref-2)
3. Eklund, Johan, Hagedorn, Josh and Darányi, Sándor. "Teaching Tale Types to a Computer: A First Experiment with the Annotated Folktales Collection" *Fabula*, vol. 64, no. 1-2, 2023, pp. 92-106. <https://doi.org/10.1515/fabula-2023-0005> [↑](#footnote-ref-3)