### **Tools für die Auswertung der Daten**

Durch die inhaltliche Annotation wird es möglich, eine große Menge von Texten einheitlich und mit präziser Fokussierung auf die konkreten Forschungsfragen auszuwerten und Ergebnisse nachvollziehbar und anschaulich darzustellen. Hier betrachten wir Anwendungsbeispiel mit dem Ziel die Normalform eines Typs in Westeuropäische Kulturraum in der visualisierten Form auszuwerten.

Im Vorfeld der Auswertung muss es vorbestimmt werden (a) nach welchem Typ bzw. Typencluster wir suchen und (b) im Welchem Teil der Daten die Suche durchgeführt werden soll.

Angenommen, interessieren wir uns für die Normalform des Märchentypus *a551 – Wasser des Lebens* im europäischen Märchenrepertoire[[1]](#footnote-1) (Deutsch, Österreichisch, Französisch, Italienisch, Ungarisch, Griechisch, Russisch).

Nach der Eingabe von entsprechenden Suchparameter greift der Assistent auf die annotierten Märchentexten mit passenden Identifikatoren zu und stellt die Auswertungsergebnisse in der Form von Graphen und dabei in der einfachen Liste der Kompositionsbau beteiligten Motive dar.[[2]](#footnote-2)

Der Graph zeigt ein zusammengeführtes Bild der in den verschiedenen Texten gegenseitig geketteten Motiven mit gesuchtem ***a*** Attributwert. An den Peripherien der Kette können auch einige Motiven auftreten, die andere ***a*** Attributwerte haben. Diese stellen ihrerseits die Peripherien der Motivkette dar, die den mit dem gesuchten Typ kombinierten Typen ausmachen.[[3]](#footnote-3) Nach der Eingabe der Werten von solchen ***a*** Attribute ist zu erwarten, dass der Graph der Beziehungen von Motiven innerhalb eines Typs zum Graphen der Beziehungen der Motive innerhalb eines Typenclusters umgewandelt wird.

Die Auswertung der Motivbestand in einem Typ bzw. Typcluster, wird automatisch durch die Aggregation der Bibliothek der dabei involvierten Motive gefolgt. Diese kann, solange es keine entsprechende Aufforderung vorliegt, eine einfache Reihe von Motiven ohne Hinweise auf jegliche Verbindung darstellen. Nach der Eingabe eines konkreten Motivs wird die Liste wie folgt reorganisiert: das eingegebene Motiv bekommt die Nummer Null und wird an seine Stelle in der Reihe von gesamten Motiven platziert, die auf der Zeitachse ihm nach vorne und nach hinter liegende Motive bekommen entsprechende Nummerierung mit Minus- und Pluszeichen und werden den Platz oberhalb bzw. unterhalb des mit Null bezeichneten Motivs übernehmen. Wird die Kette der Motive von einem Typ durch eine oder mehrere Motiven mit einem anderen ***a*** Attribut getrennt so werden sie ohne Hinweis auf ihre Position in der Motivkette erfasst. Somit erhalten wir eine Liste von Motiven, die mit dem zusammengefassten Inhalt des ausgewerteten Typs übereinstimmt.

Angesichts der Unübersichtlichkeit, die durch die Wechsel der handelnden Personen hervorgerufenen Variationen der Motive mit sich bringen, ergibt sich die Empfehlung, das zusammenfassende Bild der Komposition des Märchens auf einige wenige Elemente zu verdichten. In diesem Sinne scheint es effektiv zu sein, das letzte Attribut aus dem annotierten Markupelement zu entfernen.

Beispiel: Im Typ a551 (=Attribut a) kann der Auftrag "F" (=Attribut b) zur Beschaffung des Lebenswassers (=Attribut c) verschiedenen Protagonisten (Attribut d) erteilt werden, sei es dem kranken Vater (rHH) oder dem ratgebenden alten Mann (rST). Dies führt automatisch dazu, dass bei der Zusammenfassung im Prinzip derselben Handlung zwei Markupelemente erstellt werden:

a551:F:Heilmittel\_beschaffen:rHH

a551:F:Heilmittel\_beschaffen:rST

Durch die Abschaffung des Attributs ***d*** bleibt nur ein Markupelement übrig, das trotz des Informationsverlustes in der Lage ist, ein gemeinsames Bild der Handlungskette zu vermitteln. Entsprechender Code findet man in GitHub-Repositorium.[[4]](#footnote-4)

Letztendlich hängt der wissenschaftliche Wert und die Plausibilität einer solchen Auswertung vollständig vom Grad der Repräsentativität der erfassten Forschungsdaten ab. In dieser Publikation beschränken wir uns darauf, die Funktionsweise des Codes zu demonstrieren. Die mit dem Einsatz des Codes erzielten Ergebnisse anhand des kaukasisches Erzählgutes werden separat veröffentlicht.

## **Ausblick**

Die vorliegende Arbeit untersucht allgemeine Regeln für die Organisation und Kombination von inhaltlichen Elementen in allen Ebenen des Märchens: Motiv, Episode, Vollgeschichte und gibt die formalisierte Beschreibung der erkannten Ordnungen. Auf Basis des, dadurch erfassten konzeptuellen Models des Genres entwerfe ich ein Machine\_Learning\_Modell, das uns ermöglicht die kosten und zeitintensivste Aufgaben des vergleichenden Märchens zu automatisieren.

Ein besonderer Mehrwert der erzielten Forschungsergebnisse liegt in der Bereitstellung eines künstlichen Assistenten zur Erkennung inhaltlich-struktureller Elemente innerhalb einer unüberschaubaren Menge von Textdaten. Hervorzuheben ist dabei die Zusammenführung der Leistungen von zwei unabhängigen Algorithmen in einem einzigen. Nur durch diesen Ansatz ist es möglich, die Vorhersagefunktion erfolgreich zu verwenden, selbst wenn die nach dem Modell getrennten positiven und negativen Klassifikationsmerkmale in einem untersuchten Textabschnitt zusammengehalten werden.

Ein wichtiger Bestandteil der im Zuge der vorliegenden Arbeit bereitgestellten Infrastruktur ist das Modul zur visualisierten Darstellung der ausgewerteten Normalformen von einzelnen Typen. Besonders effektiv scheint die Anwendung der Visualisierungsmethode durch Digraphen zu sein. Das daraus entwickelte Bild vermittelt eine klare und lückenlose Vorstellung über die Popularität einzelner Typen und die Häufigkeitsfälle ihrer gegenseitigen Kombinationen.

Die Ergebnisse dieser Studie werden gegenwärtig praktisch angewandt, um zur Erschließung der nordkaukasischen Folklore beizutragen. Das von der DFG geförderte Projekt verfolgt das Ziel, ein elektronisches Datenkorpus zu erstellen und durch die Annotation der inhaltlich-strukturellen Eigenschaften der in diesem Korpus gesammelten Texte diese für die international-vergleichende Märchenforschung bereitzustellen. Als besondere Herausforderungen gelten dabei:

1. sprachlichen Vielfältigkeit des Materials[[5]](#footnote-5),

2. Einschränkungen bei der Anwendung der Open-Date Prinzipien.[[6]](#footnote-6)

Nach nur anderthalb Jahren Anwendung der in diesem Buch erfassten theoretischen und praktischen Forschungsergebnisse zeichnet sich ihre Nachhaltigkeit deutlich ab. Besonders bemerkenswert ist dabei die Fähigkeit, sich neuen Entwicklungsanforderungen anzupassen. Dies wird deutlich in der Umstellung von der CSV-Datenkollektion auf ein XML-Datenkorpus sowie im Übergang von grundlegenden Segmentierungsverfahren von Episoden auf die Motivebene.

Bei fortgesetzter Anwendung dieser Arbeitsmethode lässt sich das Erreichen eines besonderen Synergieeffekts gut prognostizieren. Wie bereits angedeutet, entwickelt sich die vergleichende Märchenforschung sukzessiv, nach dem Vorbild eines selbstregulierenden Systems. Dies bedeutet unter anderem, dass jedes neu erschlossene Repertoire die vergleichende Analyse eines anderen Repertoires begünstigt; beide tragen wiederum zur Erschließung des nächsten bei, und so weiter. Unter Berücksichtigung dieses Aspekts werden die Ergänzungen der einzelnen infrastrukturellen Elemente, wie beispielsweise Markups oder synthetische Trainingsdaten, als beiläufiger Beitrag bei jedem neuen Einsatz der Arbeitsmethode betrachtet.

1. Repertoire entspricht den in der elektronischen (CD) Ausgabe der „Europäische Märchen“ zusammengetragene Texte unter dem Typnummer ATU 551. [↑](#footnote-ref-1)
2. GitHub: edadunashvili/VerMa/Auswertung\_abcd.ipynb [↑](#footnote-ref-2)
3. Stellt ein Mitglied aus den gesuchten Motiven den Anfang oder das Ende eines Textes dar, so wird dieses durch extra Zeichen „Anf-“ und „End-“ gekennzeichnet. [↑](#footnote-ref-3)
4. GitHub: edadunashvili/VerMa/Auswertung\_abc.ipynb [↑](#footnote-ref-4)
5. Auf dem heutigen Territorium der Föderationsrepublik Dagestan werden ungefähr 25 einheimische Sprachen gesprochen. [↑](#footnote-ref-5)
6. Als Ausweg aus dieser Situation bleibt die Umwandlung der Originaltexte in abgeleitete Forschungsdaten. Häufig verwenden wir dabei Texte, die von Stoppwörtern bereinigt wurden. Referenzen zu den Quelldaten sind bei der Freigabe solcher Ableitungen immer routinemäßig nachweisbar. [↑](#footnote-ref-6)