Formalisierung von Märchen

Declerck, Thierry
declerck@dfki.de
DFKI GmbH, Deutschland

Aman, Anastasija
aamann@coli.uni-saarland.de
Universität des Saarlandes, Deutschland

Grünewald, Stefan
stefang@coli.uni-saarland.de
Universität des Saarlandes, Deutschland

Lindemann, Matthias
malinux@t-online.de
Universität des Saarlandes, Deutschland

Schäfer, Lisa
lkschae@gmail.com
Universität des Saarlandes, Deutschland

Skachkova, Natalia
s9naskac@stud.uni-saarland.de
Universität des Saarlandes, Deutschland

Introduktion

Im Rahmen eines Softwareprojektes [1], das sich mit der automatisierten Analyse von Märchen in deutscher Sprache befasst, hat sich die Notwendigkeit ergeben, eine formale Repräsentation von Märchen zu bestimmen, damit die einzelnen Komponente des Systems miteinander integriert werden können.

Wir beschreiben in diesem Beitrag zum einen, welche Informationen in dieser formalen Repräsentation enthalten sind, und zum anderen, wie diese Informationen in XML bzw. Python konkret codiert werden.

Kodierte Information

Ein Märchen besteht im Sinne unseres Projektes aus den folgenden Bestandteilen:

- Eine Menge von Orten, an denen die Handlung spielt;
- Eine Menge von Charakteren, die an der Handlung beteiligt sind;
- Eine zeitliche Abfolge von Szenen, die jeweils an einem bestimmten Ort spielen und an denen jeweils eine Teilmenge der Märchencharaktere beteiligt ist;
- Jede Szene besteht ihrerseits aus einer zeitlichen Abfolge von Dialogakten zwischen den Märchencharakteren oder vom Erzähler zum Zuhörer. Zusammengenommen bilden diese Dialogakte den Märchentext.

Im Folgenden werden die verschiedenen Bestandteile, sowie ihre Eigenschaften und Beziehungen untereinander, näher beschrieben.

Orte, an denen das Märchen spielt, werden nur über ihren **Typ** (Attribut type) charakterisiert. Mögliche Ortstypen sind dabei z.B. Wald, Schloss oder Stall. Daneben existiert außerdem der Typ „Nirgendwo“ für Szenen ohne eindeutig bestimmbaren Ort (z. B. Abschnitte des Märchens, an denen nur der Erzähler beteiligt ist). Jeder Ort erhält eine spezifische ID der Form **loc1**, **loc2** etc.

Charaktere werden über eine Reihe von Eigenschaften beschrieben, welche zum einen inhärente demographische Eigenschaften (Name, Alter, Geschlecht, Typ), sowie zum anderen externe Eigenschaften (Einstellung, Propp-Archetyp – s. (Propp 1977)) beinhalten. Beim **Namen** (name) des Charakters handelt es sich um eine Zeichenkette, z.B. „Rapunzel.“ (Wird ein Charakter auf mehrere Arten gerufen, so wird die häufigste Bezeichnung gewählt.) **Alter** (age) des Charakters wird nicht in Zahlen, sondern in Stufen angegeben, da Märchen im Allgemeinen keine genauen Altersangaben enthalten; die möglichen Werte sind dabei „toddler“, „child“, „teenager“, „young adult“, „adult“ und „senior“. Das **Geschlecht** (gender) des Charakters wird den klassischen Vorstellungen folgend entweder mit „male“ oder „female“ angegeben. Zusätzlich gibt es den Wert „none“ für geschlechtlich unterspezifizierte Charaktere wie Tiere, Monster usw. Der **Typ** des Charakters unterscheidet z. B. zwischen „human“ oder „animal/monster“. Für „animal/monster“ unterscheiden wir zusätzlich nach **Subtypen**, z.B. für Tiere nach Größe, also „small“, „medium“ oder „big“, oder „witch“ und „demon“ für einen bestimmten Monstertyp. Eine binäre Feststellung der **Einstellung bzw. Gesinnung** des Charakters verortet diesen auf der Gut-/Böse-Achse: „evil“ oder „neutral“. Außerdem wird der **Propp-Archetyp** des Charakters angegeben: „hero“, „villain“ etc. (Propp, 1977). Jeder Charakter erhält eine spezifische ID von der Form **ch1**, **ch2** usw. Außerdem gehören zu jedem Märchen zwei „Dummy“-Charaktere für Erzähler und Zuhörer, welche stets die IDs **ch0** bzw. **ch-1** und die Typen „narrator“ bzw. „listener“ zugewiesen bekommen. Dies ist nötig, um auch Passagen darstellen zu können, welche vom Erzähler gesprochen werden, der selbst ja kein eigentlicher Charakter der Handlung ist. Dies ist notwendig, um ein automatisches „Vorlesen“ des Märchens zu implementieren.

Szenen werden im Hinblick auf Zeit, Ort, beteiligte Charaktere sowie Propp Funktionen (Propp, 1977) beschrieben. Der **Zeitpunkt** (time), zu dem die Szene spielt, wird anhand einer ID der Form **t1**, **t2** usw. angegeben, wobei die IDs den linearen Ablauf der Zeit darstellen. Der **Ort** (location), an dem die Szene spielt,

wird als String in Großbuchstaben angegeben, ausgewählt aus einer Liste mit Möglichkeiten. Der **Übergang zur nächsten Szene** (transition) wird ebenfalls codiert, indem das Bewegungsverb, das den Übergang von einem Ort zum anderen beschreibt, oder die Phrase, die stattdessen den Szenenwechsel einleitet, angegeben wird. Die an der Handlung der Szene beteiligten **Charaktere** werden mit ihren IDs angegeben, also z. B. **ch2**, **ch3**, **ch5**. Dabei werden alle Charaktere berücksichtigt, die in der Szene zugegen sind, auch wenn diese bspw. nicht sprechen. Die Propp-Funktionen und -Subfunktionen der Szene werden mit ihrem Symbol (nach der englischen Ausgabe Propp (1977)) angegeben, also z. B. A4 – „theft of daylight“. Jede Szene erhält eine spezifische ID der Form **s1**, **s2** etc. Da die Märchenhandlung im Allgemeinen linear erzählt wird, ist der Index üblicherweise (aber nicht notwendigerweise) identisch mit demjenigen des Zeitpunkts der Szene, d. h. die Szene **s1** wird üblicherweise zum Zeitpunkt **t1** spielen usw. Jeder Szene sind **Dialogakte** untergeordnet, denen der zu dieser Szene gehörige Text entspricht.

Dialogakte werden im Hinblick auf ihre Sprecher und Adressaten, ihren Inhalt sowie ihren Zeitpunkt beschrieben. Der **Zeitpunkt** (time), zu dem der Dialogakt geäußert wird, wird anhand einer ID angegeben, welche eine Spezifizierung der ID des Zeitpunkts der zugehörigen Szene darstellt. Spielt z. B. Szene **s5** zum Zeitpunkt **t5**, so haben die zugehörigen Dialogakte die Zeitpunkte **t5.1**, **t5.2** usw. Der **Sprecher** (speaker) des Dialogakts wird über seine ID angegeben. Der **Adressat** bzw. die **Adressaten** (receiver) des Dialogakts werden über eine Liste von Charakter-IDs angegeben, z.B. **ch2**, **ch4**, **ch6**. Passagen des Erzählers stellen dabei einen Spezialfall dar: Sie werden als Dialogakte des Erzählers mit dem Zuhörer bzw. Leser betrachtet, d. h. der „Dummy“-Charakter des Erzählers wird als Sprecher angegeben und der Dummy-Charakters des Zuhörers als Empfänger. Abgesehen davon werden sie behandelt wie Dialogakte zwischen Charakteren. Jeder Dialogakt erhält eine spezifische ID, die – unabhängig von der Szenestruktur – linear hochgezählt wird, also **d1**, **d2** usw.

XML-Repräsentation

Die oben beschriebenen Informationen lassen sich im XML-Format darstellen. Dabei wird eine XML-Baumstruktur genutzt, um die Hierarchie der verschiedenen Objekte zu repräsentieren. Das Wurzelelement des Dokuments hat stets den Bezeichner Tale und die Attribute „title“ und „annotator“, welche Titel und den Namen des Annotators des jeweiligen Märchens enthalten:

1: Struktur des Tale-Wurzelelements (Beispiel).

```
<Tale title="Froschkönig" annotator="Lisa Schäfer"> ... </Tale>
```

Diesem Element untergeordnet sind die Elemente Characters, Locations und Text. Das Characters-Element enthält Character-Subelemente, die jeweils die gesammelten Informationen für einen Charakter speichern:

2: Struktur des Characters-Elements (Beispiel).

```
<Character id="ch1" name="Frosch" age="adult" gender="male" type="animal_monster" subtype="small" attitude="neutral" archetype="hero"> </Character>
```

Analog dazu enthält das Locations-Element untergeordnete Location-Elemente, die jeweils einen Ort codieren:

3: Struktur des Locations-Elements (Beispiel).

```
<Location id="loc1" type="WALD"> </Location>
```

Das Text-Element enthält schließlich den eigentlichen Märchentext. Dieser ist auf die verschiedenen Szenen – repräsentiert durch Szene-Elemente – aufgeteilt, welche wiederum die verschiedenen Dialogakte (Dialog-Elemente) enthalten:

4: Struktur des Text- und Szene-Elemente (Beispiel).

```
<Text><Scene id="s2" time="t2" location="loc1" characters="ch1,ch2" propp_functions="d|e" propp_subfunctions="D7|E10" transition="gehen">
  ...
  <Dialogue id="d5" time="t2.4" speaker="ch2" receiver="ch1"> Ach, du bist’s, alter Wasserpatscher, </ Dialogue>
  ...
</Scene></Text>
```

Beim Entwurf des XML-Schemas wurde besonders Wert auf Übersichtlichkeit und Leserlichkeit gelegt. Trotz der Vielzahl der kodierten Informationen sind die resultierenden XML-Dateien daher vergleichsweise kompakt; so besteht die XML-Repräsentation des (vergleichsweise langen) Märchens „ *Hänsel und Gretel“* bspw. nur aus 226 Zeilen.

Diese XML Repräsentation basiert auf und erweitert das Annotation Schema, das in (Scheidel & Declerck, 2010) beschrieben wird.

Python-Repräsentation

Auf der Grundlage der oben beschriebenen XML-Struktur kann eine Python-Klassenstruktur aufgebaut werden, die ein Märchen sowie seine einzelnen Teile als Python-Objekte repräsentiert.

Neben einer Oberklasse Tale gibt es für jeden der oben beschriebenen Teile eine eigene Python-Klasse, d. h. die Klassen Location, Character, Scene und

Dialogue. (Insgesamt bestehen die Dateien zur Märchen-Repräsentation aus 288 Zeilen Code.) Jede Klasse enthält dabei als Attribute die oben beschriebenen Eigenschaften, wobei diese auch Verweise auf andere Elemente darstellen können. So verweisen bspw. Dialogue-Objekte auf die Character-Objekte von Sprecher und Empfängern. Der Python-Code dient als Interface für drei Anwendungen. Erstens können Märchen aus bestehenden XML-Dateien eingelesen werden; zweitens können XML-Dateien anhand einer anderweitig (z. B. durch automatische Klassifizierung) erzeugten Python-Märchenstruktur generiert werden; und drittens kann anderer Python-Code auf die Märchen-Information zugreifen, was die Grundlage für Anwendungen wie Text-to-Speech oder Visualisierung bildet. Sowohl die XML Kodierung als auch die Python Objekte interagieren mit einer Märchen-Ontologie interagieren, die eine Erweiterung der in (Koleva et al., 2012) beschriebenen Ontologie ist.

Somit haben wir eine formale Repräsentation von Märchen, die in verschiedenen Anwendungen zum Tragen kommen kann.

Fußnoten

1. Mit Beiträgen von Anastasija Aman, Stefan Grünewald, Matthias Lindemann, Lisa Schäfer, Natalia Skachkova.

Bibliographie

Propp, Vladimir ; Scott, Laurence (Hrsg.): Morphology of the folktale. 2. überarbeitete Auflage. Austin, TX u.a., 1977

Antonia Scheidel and Thierry Declerck. 2010. Apftml - augmented proppian fairy tale markup language. In Sándor Dar´anyi and Piroska Lendvai, editors, First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts. Szeged University.

Nikolina Koleva, Thierry Declerck, and Hans-Ulrich Krieger. 2012. An ontology-based iterative text processing strategy for detecting and recognizing characters in folktales. In Jan Christoph Meister, editor, Digital Humanities 2012 Conference Abstracts, pages 467–470, Hamburg, 7. University of Hamburg, Hamburg University Press