Vorschlag für ein IT-Projekt für den Sommer 2024
Betreuer: Prof. Jörg Roth

Erweiterung und Verbesserung der Strukturerkennung einer Autoren-Umgebung

Die Ausgangssituation

Ziel des Projekts *StoryChecker* [1, 2] ist, Schriftsteller von fiktiven Geschichten bei ihrer Autorentätigkeit zu unterstützen, indem Rechschreibung und Stilregeln geprüft werden. Diese Prüfungen sind nicht trivial: nicht alleine eine isoliert betrachtete Schreibweise, sondern die Kombination von Wortformen in einem Satz ist ausschlaggebend dafür, ob ein Satz richtig aufgebaut worden ist und ob bestimmte Regeln eingehalten wurden.
Aktuelle Werkzeuge (z.B. MS Word, Papyrus Autor) machen in dieser Richtung schon sehr viel, prüfen aber teilweise nur, ob einzelne Wortformen in Wörterbüchern vorkommen. Bei komplexen Regeln scheitern sie oft, da sie nicht imstande sind, eine Satzstruktur in ihrer Gesamtheit zu verstehen.
StoryChecker kann schon sehr viel:

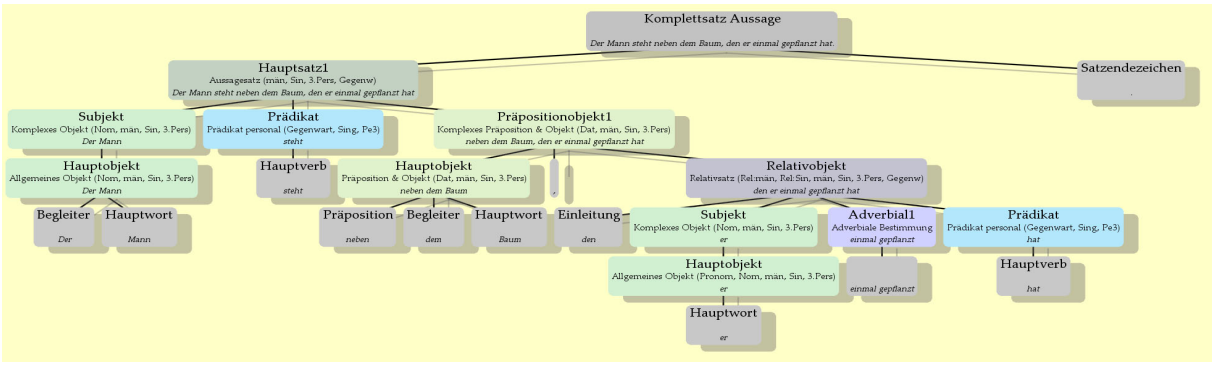
- Eine Software-Komponente liest Plain-Texts ein und zerlegt Zeichenketten in Token (z.B. *Wort*, *Satzzeichen*, *Zahl*) und in Satzstrukturen (z.B. *wörtliche Rede*, *ganze Sätze*).
- Eine Wortdatenbank ermittelt für jede *Wortform* alle Varianten mit ihren grammatikalischen Eigenschaften (Worttyp, Geschlecht, Zeitform, Kasus, Numerus etc.). Diese Wortformendatenbank kann erweitert werden.
- Eine Software-Bibliothek erlaubt den Zugriff auf diese Wortdatenbank, kann sie durchsuchen und modelliert Wortformen als Instanzen mit grammatikalischen Eigenschaften.
- Eine Musterbeschreibungssprache erlaubt die Definition von Wortkombinationen, die innerhalb eines Satzes zusammen betrachtet eine bestimmte Bedeutung haben (z.B. Objekt, Prädikat). Eine Software-Komponente kann diese Muster in Texten suchen. Viele grammatikalische Strukturen werden schon treffsicher erkannt.
- Stilregeln werden auf der Basis der Strukturen und Wortformen geprüft und die Ergebnisse über eine grafische Schnittstelle dargestellt (Abbildung).

Die Aufgabe

In der Musterbeschreibungssprache wurden schon zahlreiche Grammatikkonstruktionen formal beschrieben. So werden derzeit zwischen 50 und 70% aller Sätze aus typischen Romantexten korrekt erkannt. Bei erkannten Sätzen können alle Bestandteile (z.B. was ist das Prädikat? was ist das Objekt?) richtig zugeordnet werden. Darüber hinaus kann in einer Baumstruktur auf Details zugegriffen werden. Als Beispiel, aus dem Satz

Der Mann steht neben dem Baum, den er einmal gepflanzt hat.

wird folgende Struktur herausgerechnet:



Das ist eine geeignete Darstellung, um verschiedene Regeln des Schreibens zu prüfen.
Derzeit wird ein gemischter Ansatz eingesetzt. Der Großteil der Erkennung wird durch die formale Musterbeschreibung geleistet. Hierzu steht ein Interpreter im Quellcode zur Verfügung. Die Resultate werden durch Entwickler-Code weiterverarbeitet, da einige Eigenschaften nur schwer durch eine Musterbeschreibung geprüft werden können.
In diesem Projekt soll der generelle Ansatz verbessert und erweitert werden. Das Ziel ist, die Erkennungsquote von derzeit max. 70% signifikant zu erhöhnen. Wünschenswert wäre ein Wert zwischen 80% und 90%.
Das Projekt umfasst unter anderem folgende Arbeitspunkte:

- Einarbeitung in die aktuelle Musterbeschreibungssprache, Identifikation von Schwachstellen.
- Es fehlen einige grundlegende Muster, insb. einige Nebensatz-Konstruktionen. Diese sollten entwickelt werden.
- Die Musterbeschreibungssprache sollte überarbeitet werden, und zwar unter dem Gesichtspunkt der Ausdrucksfähigkeit und Interpretierbarkeit. Ggfs. müssen neue Elemente entwickelt, unnötige Elemente entfernt werden.
- Überarbeitung der Zuständigkeiten zwischen Musterbeschreibung und Code-Elementen. Idealerweise kann alles über die Beschreibung dargestellt werden. Wünschenswert wäre zumindest, den bisherigen Code-Anteil zu reduzieren.

Bestandteil der Arbeit ist die Integration der Ansätze in die bisherige Autoren-Umgebung (in Java).

[1] Jörg Roth: Story Checker, Internes technisches Papier
[2] Jörg Roth: Support for Fictional Story Writing and Copy Editing, 23rd International Conference on Innovations for Community Services (I4CS), Bamberg, 11.-13. Sept. 2023, Springer CCIS 1876, 151-168, ISBN 978-3-031-40851-9, DOI: 10.1007/978-3-031-40852-6_8