Kriterien für die Lesbarkeitsanalyse

Die folgenden 5 Kriterien sind in der Literatur[1] durch empirische Analyse als für die Lesbarkeit bedeutsam befunden worden. Die Beschreibung enthält jeweils einen Vorschlag zur Berechnung des Kriteriums aus den zu erhebenden Features.

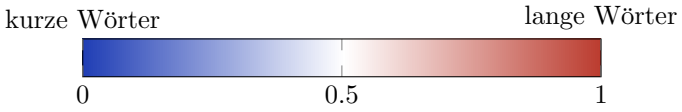
Alternativ zur Ad-hoc-Normalisierung könnte analog zu den *Grundwahrheiten* im Paper ein Korpus von etwa 5 besonders leicht bzw. schwer lesbaren Texten analysiert und das resultierende Minimum / Maximum als Grundlage für die Normalisierung genutzt werden.

Wird die Farbkodierung adaptiv in Bezug auf den zugrundeliegenden Maßstab für die Normalisierung (vgl. Figure 2 im Paper) implementiert, überlappen sich vermutlich die Werte des längsten Wörter in den leicht lesbaren Texten mit den Werten der kürzesten Wörter in den schwer lesbaren Texten. Es ist generell zu diskutieren, ob dieser Umstand in der Farbkodierung reflektiert werden sollte oder ob dies die Interpretation der Analyseergebnisse nicht sogar erschwert.

Wortlänge

Hierfür wird zunächst die durchschnittliche Wortlänge analysiert und normiert. Sei W die Menge aller Wörter w_i im zu analysierenden Text mit Wortlänge $|w_i|$. Die minimale Wortlänge ist 1 (bzw. 2 im Deutschen), die maximale ist $|w_i|_{max} = max(|w_i|)$ bzgl. aller Wörter $w_i \in W$.

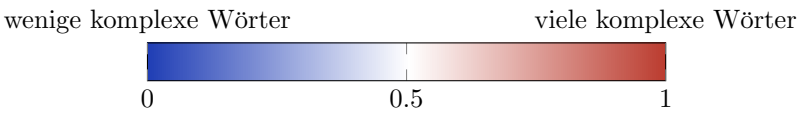
Der Lesbarkeitswert jedes Wortes wird normiert durch $\frac{|w_i|}{|w_i|_{max}}$ und der summierte Wert der Wörter des entsprechenden Satzes durch die Anzahl der Wörter $|W|$ geteilt. Anschließend wird der Wert z.B. auf Farbwerte zwischen blau $(32, 62, 181)$, weiß und rot $(186, 57, 44)$ abgetragen.



Komplexität der Vokabeln

Hier wird der Prozentanteil eines Absatzes/Satzes gemessen, der nicht in einer Liste häufig verwendeter Wörter vorkommt. Dazu kann entweder Wikipedia[2] (deutsch/englisch), ein Korpus aus Zeitungsartikeln[3] oder evtl. eine fachspezifische Textsammlung ausgewertet werden. Der Anteil der Wörter w_i, die nicht in der Liste L sind, wird dann durch die Anzahl $|W|$ der Wörter im zu analysierenden Text W geteilt, also

$$Komplexität_W = \frac{|w_i \notin L|}{|W|}.$$



[1] http://bib.dbvis.de/uploadedFiles/305.pdf
[2] https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#German
[3] http://wortschatz.uni-leipzig.de/html/wliste.html

Nominalisierungen

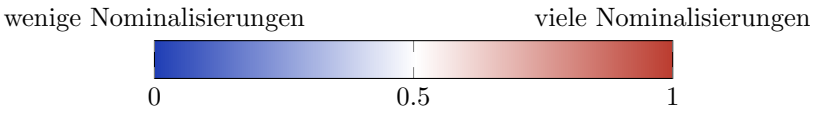
Die Nominalisierung ist die Bildung eines Substantivs aus einer anderen Wortart, vor allem aus Verben und Adjektiven (z.B. *das Böse, etwas Hübsches; the evil, something pretty*). Ein Gerundium ist ein substantivierter Infinitiv eines Verbs (z.b. ***climbing** is dangerous; **das Klettern** ist gefährlich*).

Da Nominalisierungen schwer grundsätzlich vermeidbar sind, die Lesbarkeit des Textes aber auch nicht zwingend schwer unter ihrer Verwendung leidet (z.B. *Es geschah aus **Versehen**; The **use** of drugs is dangerous*), muss die Bewertungsskala kontextsensitiv angelegt werden[4]. Bei einem wissenschaftlichen Fachartikel wird die Lesbarkeit bzgl. dieses Kriteriums evtl. zugunsten einer präzisen Formulierung vernachlässigt, in der Unterhaltungsliteratur wiederum als Stilmittel, etwa um eine Gesinnung über eine bestimmte Ausdrucksweise zu transportieren.

Ein mögliches Maß ist die Anzahl der Nominalisierungen $|W_{Nominalisierung}|$ geteilt durch die Anzahl der Substantive $|W_{Substantive}|$ im zu analysierenden Text, also

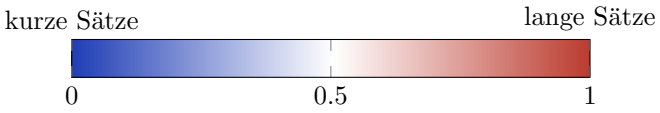
$$Nominalisierungen_W = \frac{|W_{Nominalisierungen}|}{|W_{Substantive}|},$$

wobei $W_{Nominalisierungen} \subseteq W_{Substantive}$. Der resultierende Wert kann wie beim Kriterium Wortlänge normalisiert und farbkodiert werden.



Satzlänge

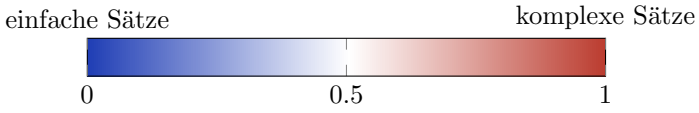
Hier wird die Anzahl der Wörter in einem Satz gemessen. Sollte kein Katalog an *Grundwahrheiten* (vgl. Einleitung) gebildet werden, könnten entsprechende Werte aus anderer Literatur[5] für die Maßstabsfindung übernommen werden, was jedoch in je nach Kontext (nicht berücksichtige Textarten, Sprachwandel) zu weniger aussagekräftigen Ergebnissen führen könnte.



Komplexität der Satzstruktur

Dieses Kriterium basiert auf der Annahme, dass der für das Verständnis eines Satzes erforderliche mentale Arbeitsaufwand mit dem Grad an Verschachtelung und der Verwendung von Klammern steigt.

[4] `https://ps.ipd.kit.edu/backend/index.php/veroeffentlichungen-details/items/3801.html`
[5] `https://de.wikipedia.org/wiki/Satzl%C3%A4nge#Durchschnittliche_Satzl.C3.A4nge`



Der dem Maßstab zugrundeliegende Verzweigungsfaktor des Satzstruktur-Baums muss zunächst experimentell ermittelt werden. Um Mehrdeutigkeiten aufzulösen, wird der Stanford Parser[6] verwendet. Dieser hat auch eine eingebaute Visualisierungsmöglichkeit für den entstehenden Syntaxbaum.

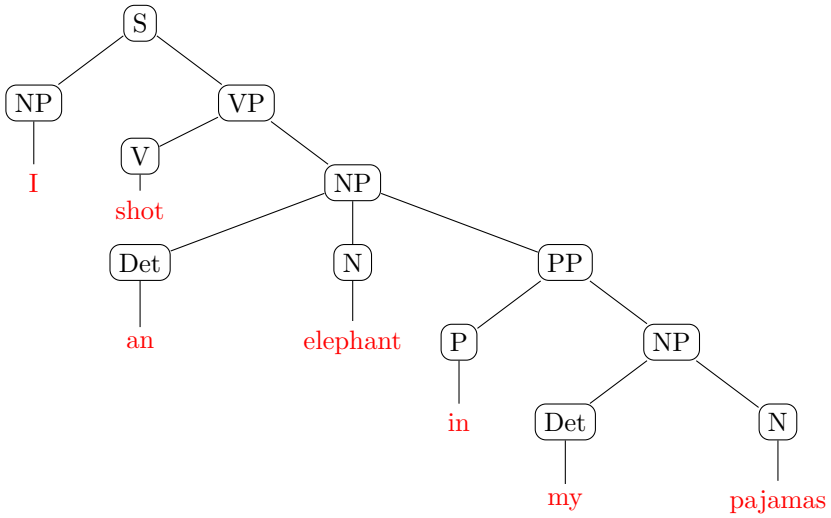


Abbildung 1: Beispiel 1 für Mehrdeutigkeit, 6 Level

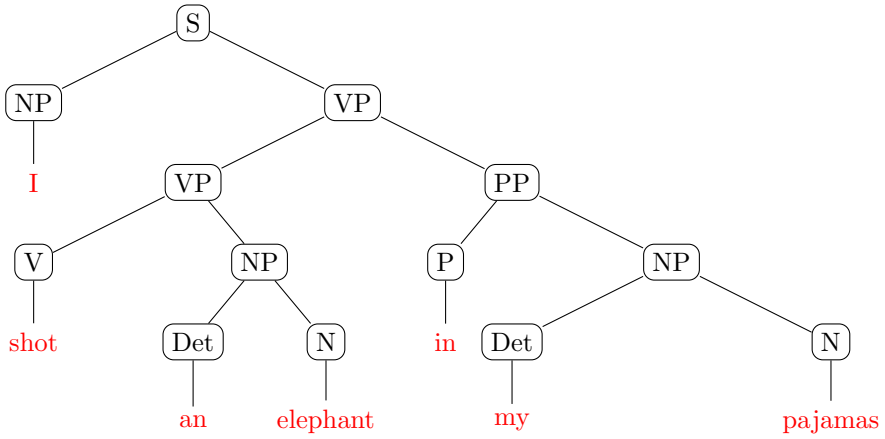


Abbildung 2: Beispiel 2 für Mehrdeutigkeit, 5 Level

[6] `http://nlp.stanford.edu/software/lex-parser.shtml`

Symbol	Bedeutung	Beispiel
S	sentence	the man walked
NP	noun phrase	a dog
VP	verb phrase	saw a park
PP	prepositional phrase	with a telescope
Det	determiner	the
N	noun	dog
V	verb	walked
P	preposition	in