

Kriterien für die Lesbarkeitsanalyse

Die folgenden 5 Kriterien sind in der Literatur¹ 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.

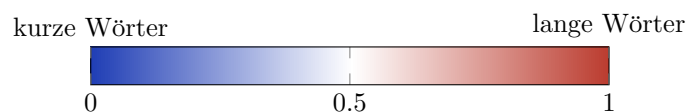
Alternativ zur Ad-hoc-Normalisierung könnte analog zu den *Grundwahrheiten* im Paper ein Korpus von etwa 5 besonders leicht bzw. schwer lesbaren Texten analysiert und das resultierende Minimum / Maximum als Grundlage für die Normalisierung genutzt werden.

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.



Skala

	min	max	avg
it_could_happen	3.25	5.66	4.46
the_halloween_house	2.0	9.0	5.5
the_little_gingerbread_man	2.5	6.33	4.42
who_did_patricks_homework	2	6	4

Tabelle 1: Wortlänge: einfache Texte

Komplexität der Vokabeln

Hier wird der Prozentanteil eines Absatzes/Satzes gemessen, der nicht in einer Liste häufig verwendeter Wörter vorkommt. Dazu kann entweder Wikipedia² (deutsch/englisch), ein Korpus aus Zeitungsartikeln³ oder evtl. eine fachspezifische Textsammlung ausgewertet werden. Der Anteil

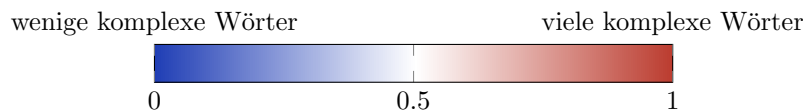
¹<http://bib.dbvis.de/uploadedFiles/305.pdf>

²https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#German

³<http://wortschatz.uni-leipzig.de/html/wliste.html>

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

$$\text{Komplexität}_W = \frac{|w_i \notin L|}{|W|}.$$



Skala

	min	max	avg
it_could_happen	0.2	0.75	0.475
the_halloween_house	0.0	1.0	0.5
the_little_gingerbread_man	0.0	0.75	0.375
who_did_patricks_homework	0.0	0.75	0.375

Tabelle 2: Komplexität der Vokabeln: einfache Texte

Nominalisierungen

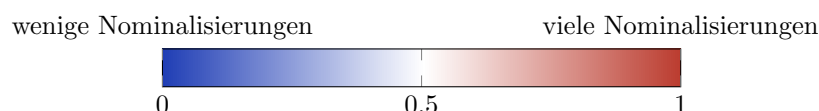
Die Nominalisierung ist die Bildung eines Substantivs aus einer anderen Wortart, vor allem aus Verben und Adjektiven (z.B. *das Böse, etwas Hübsches; the evil, something pretty*). Ein Gerundium ist ein substantivierter Infinitiv eines Verbs (z.b. *climbing is dangerous; das Klettern ist gefährlich*).

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 werden⁴. 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 $|W_{\text{Nominalisierung}}|$ geteilt durch die Anzahl der Substantive $|W_{\text{Substantive}}|$ im zu analysierenden Text, also

$$\text{Nominalisierungen}_W = \frac{|W_{\text{Nominalisierungen}}|}{|W_{\text{Substantive}}|},$$

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



⁴<https://ps.ipd.kit.edu/backend/index.php/veroeffentlichungen-details/items/3801.html>

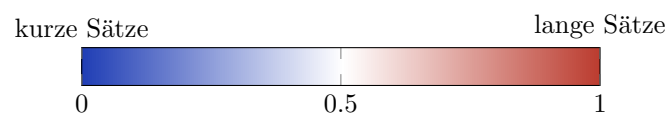
Skala

	min	max	avg
it_could_happen			
the_halloween_house			
the_little_gingerbread_man			
who_did_patricks_homework			

Tabelle 3: Einfache Texte

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 Literatur⁵ für die Maßstabsfindung übernommen werden, was jedoch in je nach Kontext (nicht berücksichtigte Textarten, Sprachwandel) zu weniger aussagekräftigen Ergebnissen führen könnte.



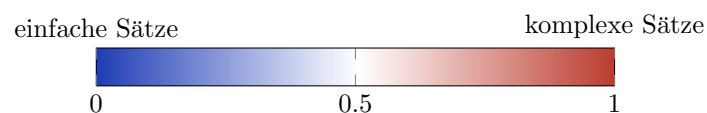
Skala

	min	max	avg
it_could_happen	3	42	22.5
the_halloween_house	1	26	13.5
the_little_gingerbread_man	1	31	16
who_did_patricks_homework	2	21	11.5

Tabelle 4: Satzlänge: einfache Texte

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.



⁵https://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 Parser⁶ verwendet. Dieser hat auch eine eingebaute Visualisierungsmöglichkeit für den entstehenden Syntaxbaum.

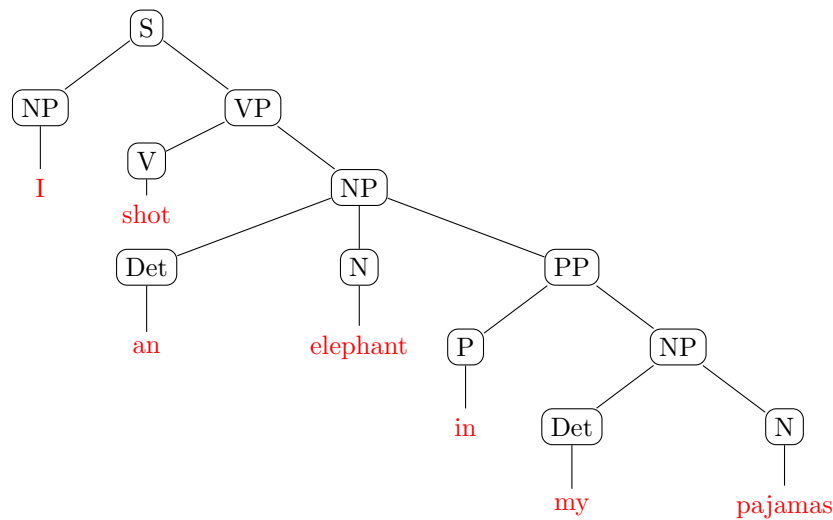


Abbildung 1: Beispiel 1 für Mehrdeutigkeit, 6 Level

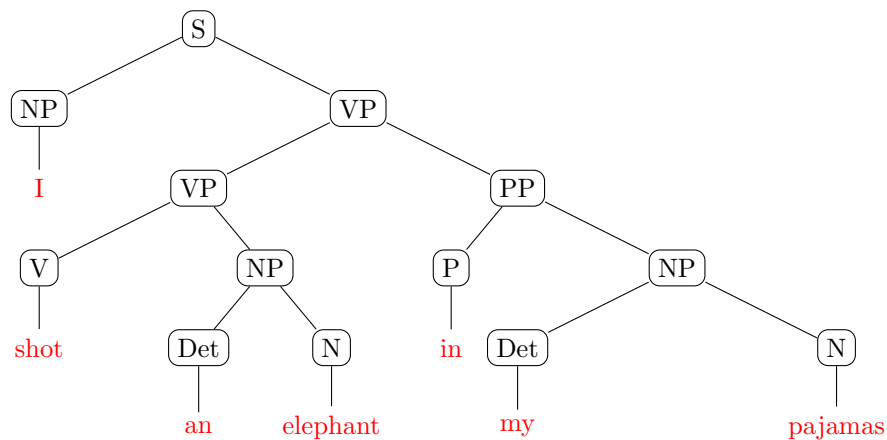


Abbildung 2: Beispiel 2 für Mehrdeutigkeit, 5 Level

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

⁶<http://nlp.stanford.edu/software/lex-parser.shtml>

Skala

	min	max	avg
it_could_happen	5	21	13
the_halloween_house	3	17	10
the_little_gingerbread_man	3	19	11
who.did.patricks_homework	4	17	10.5

Tabelle 5: Komplexität der Satzstruktur: einfache Texte