

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.

$$\text{Wortlänge-Wert: } \frac{1}{|W|} \cdot \sum_i \frac{|w_i| - 2}{6}$$

Skala

| Text | min | max | avg |
|----------------------------|------|------|------|
| it_could_happen | 3.25 | 5.66 | 4.29 |
| the_halloween_house | 2.00 | 9.00 | 4.10 |
| the_little_gingerbread_man | 2.50 | 6.33 | 4.17 |
| who_did_patricks_homework | 2.00 | 6.00 | 3.84 |

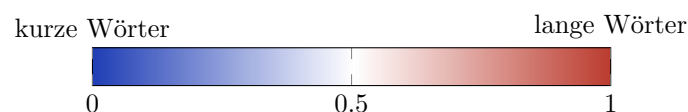
} untere Grenze: 2.00

Tabelle 1: Wortlänge: einfache Texte

| Text | min | max | avg |
|------------------------|------|------|------|
| black_and_white | 3.50 | 6.15 | 5.00 |
| fight_terrorism | 3.72 | 6.50 | 5.04 |
| jura_paper | 4.52 | 6.94 | 5.55 |
| paper_medicine | 3.06 | 7.55 | 5.13 |
| poems | 2.00 | 4.86 | 3.47 |
| political_english_text | 3.50 | 5.91 | 4.85 |

} obere Grenze: 7.55 \approx 8.00

Tabelle 2: Wortlänge: schwere Texte



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

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 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.

Das Ergebnis ist also **inhärent normiert**, die Tests dienen nur der subjektiven Überprüfung der Güte der Aussagekraft des entsprechenden Wertes.

$$\text{Vokabelkomplexitäts-Wert: } \frac{|w_i \notin L|}{|W|}$$

Skala

| Text | min | max | avg |
|----------------------------|------|------|------|
| it_could_happen | 0.20 | 0.75 | 0.41 |
| the_halloween_house | 0.00 | 1.00 | 0.42 |
| the_little_gingerbread_man | 0.00 | 0.75 | 0.37 |
| who_did_patricks_homework | 0.00 | 0.75 | 0.37 |

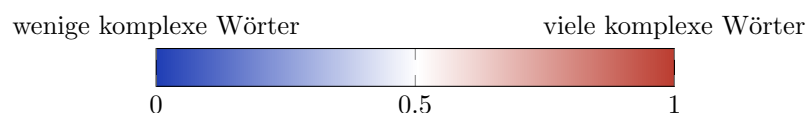
untere Grenze: 0.00

Tabelle 3: Komplexität der Vokabeln: einfache Texte

| Text | min | max | avg |
|------------------------|------|------|------|
| black_and_white | 0.08 | 0.75 | 0.54 |
| fight_terrorism | 0.14 | 0.71 | 0.49 |
| jura_paper | 0.24 | 0.68 | 0.50 |
| paper_medicine | 0.47 | 0.93 | 0.73 |
| poems | 0.00 | 0.64 | 0.32 |
| political_english_text | 0.16 | 0.64 | 0.45 |

obere Grenze: 1.00

Tabelle 4: Komplexität der Vokabeln: schwere Texte



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

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

Nominalisierungskomplexität

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. Generell sind Texte mit vielen Verben leichter verständlich als Sätze mit wenigen.

Die Berechnung der Nominalisierungskomplexität ergibt sich aus dem Verhältnis von Verben zu Nomen und der geschätzten Anzahl Nominalisierungen. Diese wird anhand der Endung des Wortes (z.B. -tion, -ity, ...) ermittelt. Das entsprechende Nomen wird doppelt gezählt (Anzahl der Nominalisierungen wird zur Anzahl der Nomen addiert), um ihm mehr Gewicht zu verleihen.

$$\text{Nominalisierungskomplexität: } \frac{|Verben|}{|Nomen| + |Nominalisierungen|}$$

Achtung! Hier ist ausnahmsweise ein höherer Wert besser als ein kleiner! Für die Visualisierung wird daher der normierte Kehrwert zurückgegeben.

$$\text{Nominalisierungskomplexitäts-Wert: } 1 - \frac{1}{5} \cdot \frac{|Verben|}{|Nomen| + |Nominalisierungen|}$$

Skala

| Text | min | max | avg |
|----------------------------|-----|-----|------|
| it_could_happen | 0 | 5 | 1.00 |
| the_halloween_house | 0 | 4 | 0.76 |
| the_little_gingerbread_man | 0 | 5 | 1.29 |
| who_did_patricks_homework | 0 | 3 | 1.10 |

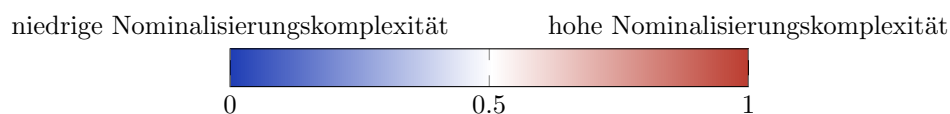
obere Grenze: 5.00

Tabelle 5: Nominalisierungen: einfache Texte

| Text | min | max | avg |
|------------------------|------|------|------|
| black_and_white | 0 | 3 | 0.45 |
| fight_terrorism | 0.14 | 2.00 | 0.58 |
| jura_paper | 0.00 | 1.40 | 0.49 |
| paper_medicine | 0.00 | 0.75 | 0.33 |
| poems | 0.00 | 3.00 | 1.08 |
| political_english_text | 0.00 | 1.33 | 0.46 |

untere Grenze: 0.00

Tabelle 6: Nominalisierungen: schwere Texte



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

Satzlänge

Hier wird die Anzahl der Wörter $|W|$ 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.

$$\text{Satzlänge-Wert: } \frac{|W|}{100}$$

Skala

| Text | min | max | avg |
|----------------------------|-----|-----|-------|
| it_could_happen | 3 | 42 | 18.00 |
| the_halloween_house | 1 | 26 | 9.68 |
| the_little_gingerbread_man | 1 | 31 | 10.91 |
| who_did_patricks_homework | 2 | 21 | 10.00 |

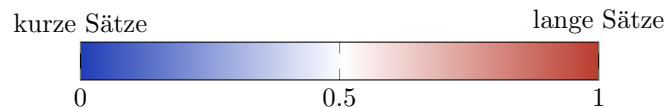
} untere Grenze: 1.00

Tabelle 7: Satzlänge: einfache Texte

| Text | min | max | avg |
|------------------------|-----|-----|-------|
| black_and_white | 4 | 81 | 26.65 |
| fight_terrorism | 11 | 52 | 25.49 |
| jura_paper | 6 | 74 | 29.60 |
| paper_medicine | 8 | 101 | 28.83 |
| poems | 2 | 54 | 17.22 |
| political_english_text | 6 | 39 | 19.15 |

} obere Grenze: 100.00 weil schöner zu rechnen

Tabelle 8: Satzlänge: schwere Texte



⁵https://de.wikipedia.org/wiki/Satzl%C3%A4nge#Durchschnittliche_Satzl.C3.A4nge

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.

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.

$$\text{Verschachtelungswert: } \frac{\text{Baumtiefe} - 3}{23}$$

Skala

| Text | min | max | avg |
|----------------------------|-----|-----|-------|
| it_could_happen | 5 | 21 | 10.00 |
| the_halloween_house | 3 | 17 | 6.74 |
| the_little_gingerbread_man | 3 | 19 | 8.08 |
| who_did_patricks_homework | 4 | 17 | 7.36 |

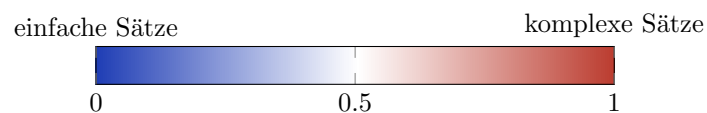
} untere Grenze: 3.00

Tabelle 9: Komplexität der Satzstruktur: einfache Texte

| Text | min | max | avg |
|------------------------|-----|-----|-------|
| black_and_white | 4 | 23 | 11.84 |
| fight_terrorism | 7 | 25 | 13.00 |
| jura_paper | 3 | 26 | 14.00 |
| paper_medicine | 3 | 11 | 5.83 |
| poems | 3 | 10 | 5.78 |
| political_english_text | 5 | 19 | 11.12 |

} obere Grenze: 26.00

Tabelle 10: Komplexität der Satzstruktur: schwere Texte



Erläuterung

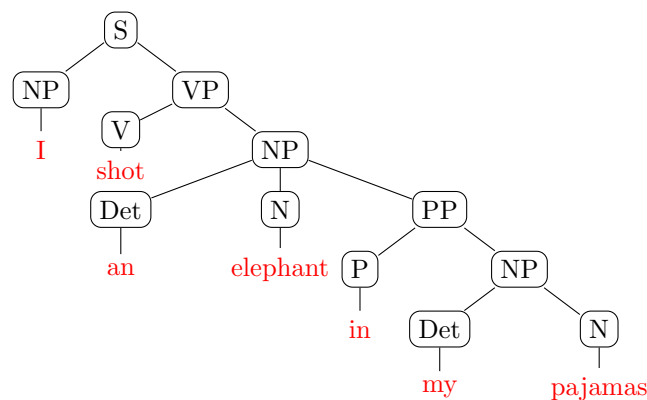


Abbildung 1: Beispiel 1 für Mehrdeutigkeit, 6 Level

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

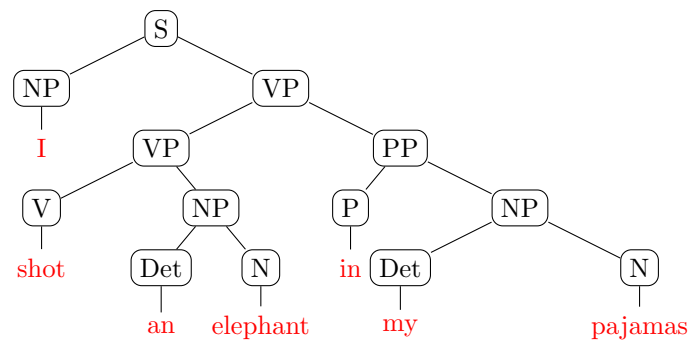


Abbildung 2: Beispiel 2 für Mehrdeutigkeit, 5 Level

| 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 |