

# Web Mining

## Übung 1

Gruppe 15

Patrick Bogdan, Christian Krebs, Rene Wilmes

# Aufgabe 1



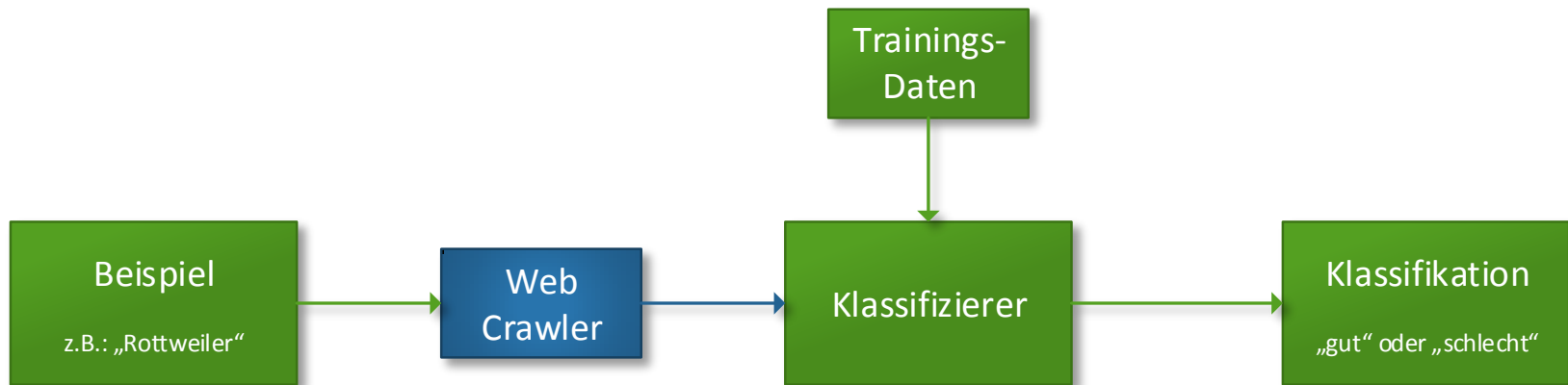
**Idee:** Web-Mining Anwendung, die eine Hunderasse entgegen nimmt und diese nach der Meinung des Webs klassifiziert.

**Umsetzung:**

1. Klassifizierer wird mit Trainingsdaten gefüttert die positive als auch negative Beispiele enthalten. (z.B. *unfreundlich, beißen, gefährlich...*)
2. User übergibt die gewünschte Hunderasse als Eingabe.
3. Web Crawler durchsucht das Web nach Seiten auf denen die angegebene Hunderasse erwähnt wird.
4. Klassifizierer klassifiziert die Erwähnungen jeweils in *gut* bzw. *schlecht*.
5. Nach dem alle Daten klassifiziert wurden wird die Anzahl der positiven und negativen Einträge gegenüber gestellt. Ausgegeben wird was überwiegt.

# Aufgabe 1

## Schematischer Ablauf:



# Aufgabe 2 - Programm

Programm wurde in Python (v. 3.4) geschrieben. (Link zur Datei: [word\\_count.py](#),  
Manual: [word\\_count.man](#))

## Workflow:

1. Liest Text-Datei ein welche als Parameter übergeben wird.
2. Normalisiert den Text
  - Ersetzt vorkommenden Satz und Sonderzeichen (z.B.: . , - \_ „“ etc.) durch Leerzeichen.
  - Ersetzt alle Großbuchstaben durch Kleinbuchstaben.
3. Falls Stopword Flag gesetzt ist:
  - Löscht alle Stopwords aus dem Text.
4. Zählt alle Wörter die durch Leerzeichen getrennt sind, sortiert diese anhand ihrer Häufigkeit und gibt die **x** häufigsten aus. (Ausgabe sowohl im Terminal als auch in Output-Datei möglich, **x** kann als Parameter übergeben werden.)

(Anmerkung: Alle Texte stammen aus dem [Project Gutenberg](#). Copyright- und Lizenzhinweise wurden vor der Analyse händig aus den Dateien entfernt.)

## Aufgabe 2 - a)



Verglichen wurden die Texte:

[Die Geschwister](#) – Johann Wolfgang von Goethe

[Jedermann](#) – Hugo von Hofmannsthal

Die 30 am häufigsten vorkommenden Worte sind bei den Texten ähnlich, sie haben 21 Worte gemeinsam (**rot**).

Das heißt bei einer Klassifizierung über die 30 häufigsten Worte werden unterschiedliche Texte trotzdem ähnlich klassifiziert und wahrscheinlich gleich zugeordnet.

Viele dieser Wörter sind sogenannten Stoppwörter, welche in allen Texten sehr häufig auftreten aber i.d.R. irrelevant für die Erfassung des Inhalts sind.

**Ergebnisse:** [../results/word\\_with\\_stopwords/geschw](#)

[../results/word\\_with\\_stopwords/jeder](#)

| Text Die Geschwister | Text Jedermann |
|----------------------|----------------|
| ich                  | und            |
| und                  | ich            |
| nicht                | jedermann      |
| sie                  | ist            |
| du                   | die            |
| marianne             | nit            |
| mir                  | der            |
| fabrice              | ein            |
| wilhelm              | das            |
| ist                  | mir            |
| s                    | in             |
| er                   | du             |
| es                   | zu             |
| so                   | auf            |
| der                  | mich           |
| wenn                 | mit            |
| mich                 | dir            |
| die                  | so             |
| das                  | mein           |
| zu                   | von            |
| ein                  | was            |
| was                  | daß            |
| den                  | dich           |
| daß                  | an             |
| wie                  | den            |
| in                   | wie            |
| sich                 | sein           |
| ihn                  | er             |
| mit                  | ihr            |
| dir                  | sie            |

## Aufgabe 2 - b)



Verglichen wurden die selben Texte, allerdings wurden nun alle Stoppwörter anhand einer [Blacklist](#) aus den Rankings entfernt.

**Ergebnisse:** [../results/word/geschw](#)  
[../results/word/jeder](#)

Es gibt nur noch 6 Übereinstimmungen in den 30 häufigsten Wörtern.

Eine Klassifizierung der beiden Texte wäre nun wesentlich ausschlaggebender als zuvor.

| Text Die Geschwister | Text Jedermann |
|----------------------|----------------|
| marianne             | jedermann      |
| fabrice              | nit            |
| wilhelm              | vetter         |
| s                    | gesell         |
| bruder               | werke          |
| immer                | recht          |
| wohl                 | geht           |
| liebe                | ja             |
| muß                  | tod            |
| sagen                | gott           |
| manchmal             | mutter         |
| herz                 | glaube         |
| ganz                 | gar            |
| gut                  | o              |
| ab                   | wohl           |
| gar                  | geld           |
| nein                 | muß            |
| glücklich            | kommt          |
| weiß                 | jedermanns     |
| hast                 | mann           |
| recht                | weiß           |
| geht                 | wär            |
| wäre                 | hast           |
| nie                  | schon          |
| gern                 | buhlschaft     |
| lieb                 | dicker         |
| leben                | halt           |
| schon                | weh            |
| mehr                 | alls           |
| lieben               | stund          |

# Aufgabe 3



Verglichen wurden die folgenden Texte:

Deutsch-sprachig:

[Die Geschwister](#) – Johann Wolfgang von Goethe

[Jedermann](#) – Hugo von Hofmannsthal

Englisch-sprachig:

[Ten Acres Enough](#) – Edmund Morris

[Liliom](#) – Ferenc Molnar

## Aufgabe 3 – Anmerkungen



1. Die Ergebnisse auf den folgenden Folien schließen Stopp-Wörter mit ein, sie wurden also nicht wie in Aufgabe 2 – b) herausgefiltert, da in der Aufgabenstellung nicht weiter spezifiziert. Wir haben die Untersuchungen jedoch für beide Fälle analog vorgenommen. Die Ergebnisse ohne Stopp-Wörter finden sich im Verzeichnis: [../results/word/](#) (analoge Dateinamen)
2. Da in der Aufgabe nicht genau angegeben, wurde die Analyse im Folgenden auf die 30 häufigsten Wörter beschränkt. Analoge Ergebnisse, die alle Wörter miteinbeziehen, finden sich im Verzeichnis: [../results/big/](#)
3. Alle Plots wurden mit Gnuplot durch generische Skripte erzeugt. Diese finden sich in den jeweiligen „*plotscript*“ Unter-Verzeichnissen. Skripte für normale Plots heißen „plot“ und die für die Anzahl der Worte mit Häufigkeit über die Häufigkeit „plot\_occ“.



## Aufgabe 3 – a)

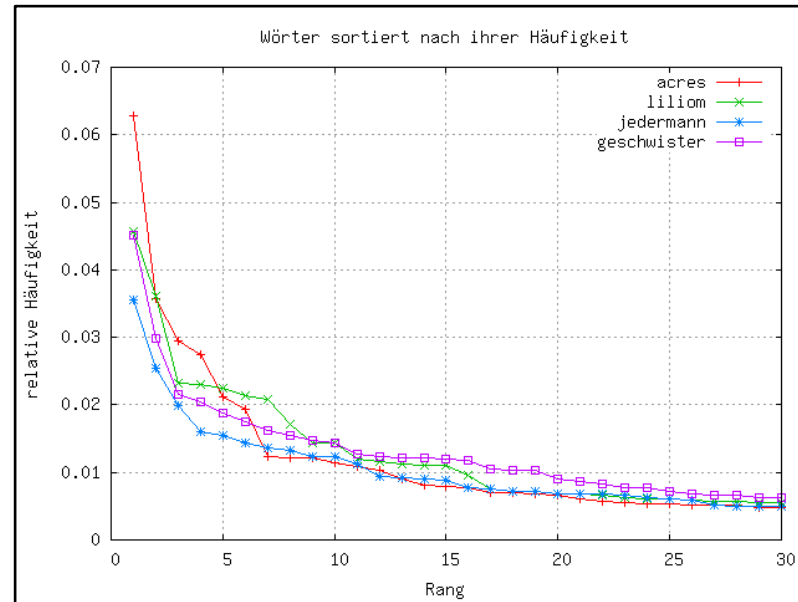


Abbildung 3.1: [../results/word\\_with\\_stopwords/plots/plot\\_rel.png](http://results/word_with_stopwords/plots/plot_rel.png)

Bei der **absoluten** Achsen-Skalierung kann man gut erkennen, dass die relative Häufigkeit der Wörter eine logarithmische Stagnation zeigt je mehr man sich vom niedrigsten Rang entfernt. Interessant ist weiterhin, dass die relativen Häufigkeiten über alle Texte ähnlich verlaufen, obwohl diese teilweise sogar in verschiedenen Sprachen verfasst wurden.

## Aufgabe 3 – a)

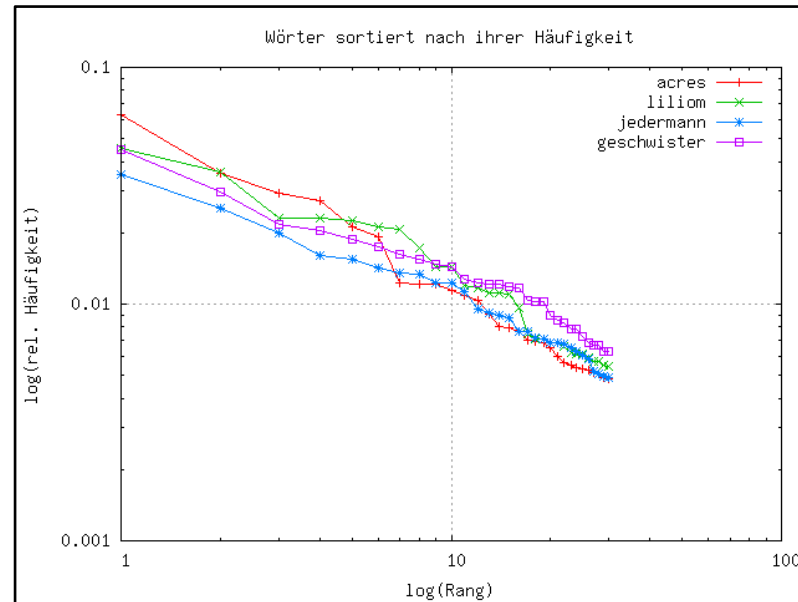


Abbildung 3.2: [../results/word\\_with\\_stopwords/plots/plot\\_rel\\_log\\_xy.png](http://results/word_with_stopwords/plots/plot_rel_log_xy.png)

In dieser Abbildung weisen nun sowohl die x- als auch die y-Achse **logarithmische** Skalierung auf. Auffällig ist, dass der vorherige logarithmische Rückgang der Kurven nicht mehr sichtbar ist. Stattdessen verlaufen die Kurven nun linear. Es lässt sich folgern, dass die Auftrittswahrscheinlichkeit tatsächlich einer doppelt logarithmischen Verteilung folgt.

## Aufgabe 3 – b)



Für diesen Teil der Aufgabe haben wir unseren Programm-Code aus Aufgabe 2 zusätzlich erweitert um uns auch eine Liste der Anzahl der Worte, die mit einer bestimmten Häufigkeit vorkommen, ausgeben zu lassen.

Die untersuchten Texte bleiben die selben, auch hier wurden Stopp-Wörter nicht herausgefiltert und dementsprechend mitberücksichtigt.

Die Ergebnisse finden sich in den folgenden Dateien:

- Die Geschwister: [../results/word\\_with\\_stopwords/geschw\\_OCC](http://../results/word_with_stopwords/geschw_OCC)
- Jedermann: [../results/word\\_with\\_stopwords/jeder\\_OCC](http://../results/word_with_stopwords/jeder_OCC)
- Ten Acres Enough: [../results/word\\_with\\_stopwords/acres\\_OCC](http://../results/word_with_stopwords/acres_OCC)
- Liliom: [../results/word\\_with\\_stopwords/lili\\_OCC](http://../results/word_with_stopwords/lili_OCC)

## Aufgabe 3 – b)

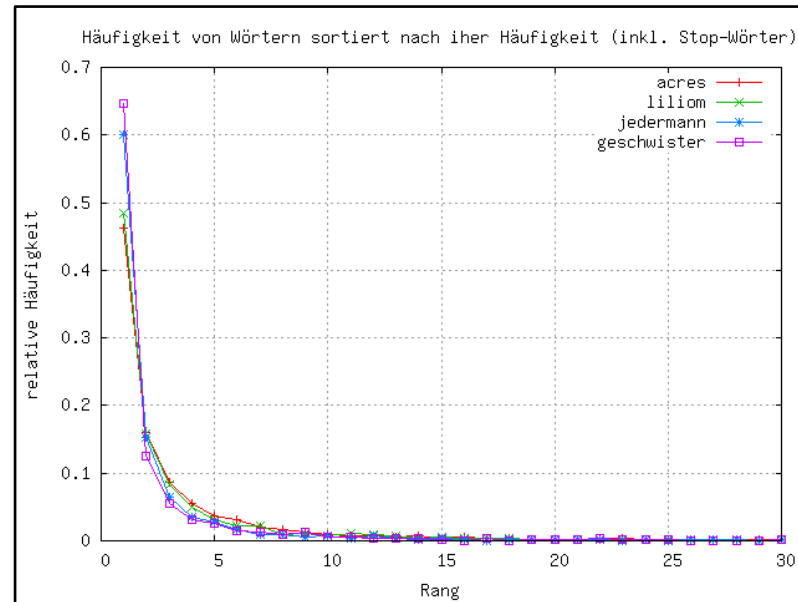


Abbildung 3.3: [../results/word\\_with\\_stopwords/plots/occ\\_plot\\_rel.png](#)

Hier zeigen sich ähnliche Ergebnisse wie in der vorherigen Teilaufgabe. Man erkennt, dass Wörter die mit einer niedrigen Häufigkeit auftreten, trotzdem einen großen Anteil an der Gesamtanzahl der Wörter ausmachen. Z.B. besteht der Text *Die Geschwister* zu mehr als 6% aus verschiedenen und nur ein mal vorkommenden Wörtern.

## Aufgabe 3 – b)

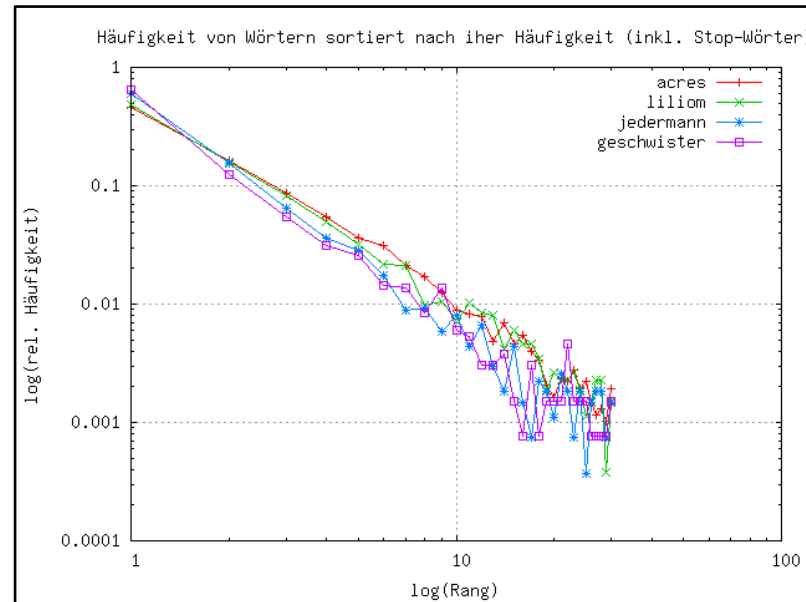


Abbildung 3.4: [../results/word\\_with\\_stopwords/plots/occ\\_plot\\_rel\\_log\\_xy.png](http://results/word_with_stopwords/plots/occ_plot_rel_log_xy.png)

Eine logarithmische Skalierung beider Achsen zeigt auch hier, dass die Anzahl der Worte mit einer gewissen Häufigkeit über die Häufigkeit ebenfalls einer doppelt logarithmischen Verteilung folgt.

# Aufgabe 4

Für diesen Teil der Aufgabe haben wir unseren Programm-Code aus Aufgabe 2 bzw. 3 b) zusätzlich erweitert, sodass es nicht nur Wörter sondern auch Buchstaben bzw. Buchstaben-Paare zählen kann.

Die untersuchten Texte bleiben die selben, auch hier wurden Stopp-Wörter mit berücksichtigt und nicht herausgefiltert.

Die Ergebnisse finden sich in den folgenden Verzeichnissen:

- Buchstaben: [../results/char/](http://localhost:8080/..../results/char/)
- Buchstaben-Paare: [../results/pair/](http://localhost:8080/..../results/pair/)

# Aufgabe 4 – Buchstaben

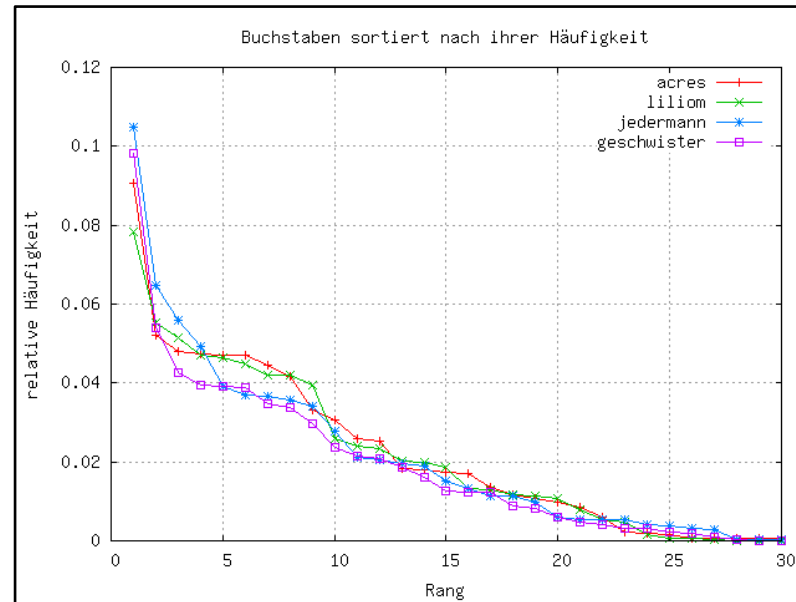


Abbildung 4.1: [../results/char/plots/plot\\_rel.png](http://results/char/plots/plot_rel.png)

In dieser Abbildung wird die relative Häufigkeit bestimmter Buchstaben in den einzelnen Texten verglichen. Die Kurven liegen alle relativ nah beieinander, was eine direkte Unterscheidung schwierig macht. Die Häufigkeitsverteilungen scheinen ähnlich zu sein.

## Aufgabe 4 – Buchstaben

| Rang | Jedermann |       | Geschwister |       | Ten Acres |       | Liliom |       |
|------|-----------|-------|-------------|-------|-----------|-------|--------|-------|
| 1    | e         | 0.105 | e           | 0.098 | e         | 0.091 | e      | 0.078 |
| 2    | n         | 0.065 | n           | 0.054 | r         | 0.052 | i      | 0.055 |
| 3    | t         | 0.055 | r           | 0.043 | t         | 0.048 | l      | 0.051 |
| 4    | r         | 0.049 | i           | 0.040 | s         | 0.047 | o      | 0.047 |
| 5    | h         | 0.039 | a           | 0.039 | a         | 0.047 | s      | 0.047 |

**Tabelle 4.1:** Die fünf häufigsten Buchstaben und ihre relative Häufigkeit in den verschiedenen Texten.

Im direkten Vergleich der jeweiligen Buchstaben lassen sich interessante Beobachtungen machen. Der Buchstabe *e* kommt in allen Texten am häufigsten vor. In beiden deutschen Texten sind außerdem die Buchstaben *n* und *r* oft vertreten, während die englischen Texte nur noch den Buchstaben *s* gemeinsam unter den Top 5 haben.



## Aufgabe 4 – Buchstaben-Paare

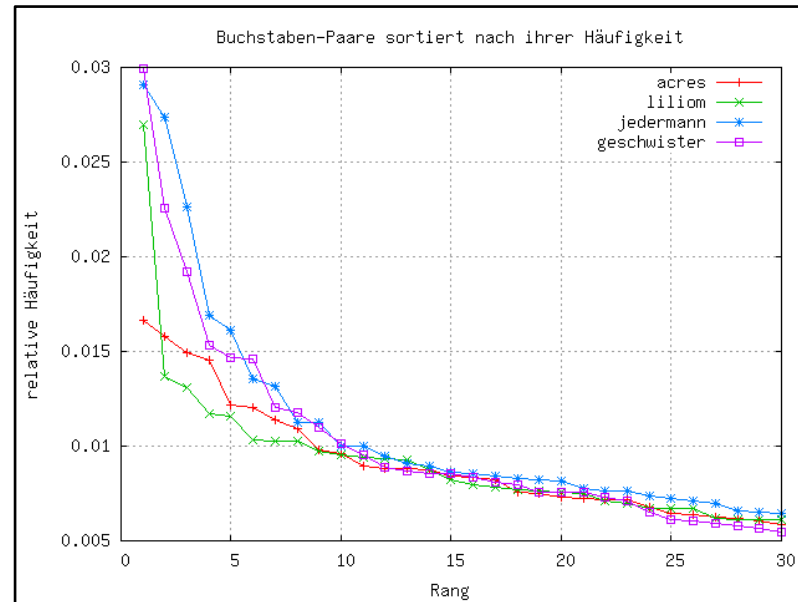


Abbildung 4.2: [../results/pair/plots/plot\\_rel.png](http://results/pair/plots/plot_rel.png)

In dieser Abbildung wird die relative Häufigkeit bestimmter Buchstaben-Paare in den einzelnen Texten verglichen. Auffällig ist, dass, besonders in den Bereichen zwischen Rang 1 und 10, die deutschen Texte relativ nah beieinander liegen, während die englischen Texte besonders in den niedrigeren Rängen größere Varianzen aufweisen.

## Aufgabe 4 – Buchstaben-Paare

| Rang | Jedermann |       | Geschwister |       | Ten Acres |       | Liliom |       |
|------|-----------|-------|-------------|-------|-----------|-------|--------|-------|
| 1    | er        | 0.029 | en          | 0.030 | er        | 0.017 | li     | 0.027 |
| 2    | en        | 0.027 | er          | 0.023 | in        | 0.016 | in     | 0.014 |
| 3    | ch        | 0.023 | ch          | 0.019 | re        | 0.015 | er     | 0.013 |
| 4    | ge        | 0.017 | ge          | 0.015 | ed        | 0.015 | ar     | 0.012 |
| 5    | ei        | 0.016 | he          | 0.015 | es        | 0.012 | om     | 0.012 |

**Tabelle 4.2:** Die fünf häufigsten Buchstaben-Paare und ihre relative Häufigkeit in den verschiedenen Texten.

In der obigen Tabelle sind die fünf häufigsten Buchstaben-Paare der jeweiligen Texte sowie ihre relative Häufigkeit aufgelistet. Auffällig ist, dass die ersten vier Buchstaben-Paare in den deutschen Texten die gleichen sind und mit ähnlicher Häufigkeit auftreten. Das Paar *er* findet sich in allen Texten sehr häufig. Bei den englischen Texten ist sonst nur das Paar *in* in den Top 5 beider Texte vertreten.