Benutzerdokumentation

Technical Report IDS-KL-2018-01

zum Produkt

Korpusbasierte Zeichenhäufigkeitslisten

DeReChar
v-uni-XXX-2018-02-28-1.0

Institut für Deutsche Sprache, Mannheim
Februar 2018

Inhaltsverzeichnis

Vorwort

Bei Spielereien mit Wörtern, bei denen es um die Zusammensetzung aus einzelnen Zeichen geht, etwa bei Scrabble, dem Glücksrad oder Galgenmännchen, geht man mit den unterschiedlichen Zeichen unterschiedlich um. Beim Scrabble werden für verschiedene Zeichen unterschiedlich viele Punkte gutgeschrieben, bei den Ratespielen versucht man, zunächst die Zeichen zu nennen, die man am ehesten erwarten kann, auch wenn das berühmte „ERNSTL“ des Glücksrads durch Einschränkung auf einen Vokal und womöglich auf Annahmen über Kombinationsrestriktionen bedingt ist. Hintergrund ist die zumindest vermutete Häufigkeit des Zeichens. Genaugenommen müsste man an dieser Stelle hinterfragen, ob sich dieser gefühlte Wert auf eine Vorstellung eines Vokabulars oder auf den tatsächlichen Gebrauch mit allen Wiederholungen bezieht – wir werden uns im Folgenden stets auf die zweite Variante berufen und dazu die Menge der verschiedenen Zeichen und ihre jeweilige Anzahl in einer umfangreichen Sammlung deutschsprachiger Texte, dem Deutschen Referenzkorpus DeReKo, auswerten (s. Abschn. 2.3).

Das Wissen um die Häufigkeitsverteilung von Zeichen ist aber nicht nur interessant für Spielereien, sondern durchaus auch praxisnah relevant: Das Anwendungsspektrum reicht von der Anordnung der Zeichen auf einer Tastatur über die Optimierung von Kodierung (vgl. Morse-Code), Ver-/Entschlüsselungsstrategien (vgl. Caesar-Kode), aber auch zum Erkennen von Auffälligkeiten, der verwendeten Sprache oder technischen Artefakten.

Download

Die Originale dieser DeReChar-Zeichenhäufigkeitslisten können unter http://www.ids-mannheim.de/derewo zusammen mit der Dokumentation in der jeweils aktuellen Version abgerufen werden.

Bitte beachten Sie die Lizenzbestimmungen am Ende dieses Dokuments.

1 Grundsätzliches zu DeReChar v-uni-XXX-2018-02-28-1.0

1.1 Was ist DeReChar v-uni-XXX-2018-02-28-1.0?

DeReChar v-uni-XXX-2018-02-28-1.0 ist eine Sammlung von korpusbasierten Zeichenhäufigkeitslisten.

1.2 Wie wurde DeReChar v-uni-XXX-2018-02-28-1.0 erstellt?

DeReChar v-uni-XXX-2018-02-28-1.0 wurde – nach eingehenden Explorationsphasen – weitestgehend über automatische Verfahren erstellt.

2 Ressourcen und Kategorien

2.1 Korpusbasiertheit

Den Zeichenhäufigkeitslisten liegen die Korpora des DeReKo-Archivs (Stand 31.12.2017, entspricht: DeReKo 2017) zugrunde. Es wurden nur die sprachlichen Primärdaten genutzt, d.h. Annotationen und weiteres Markup wurde ignoriert. In dieser Ausgangsfassung umfasst die Datengrundlage insgesamt 238.376.881.328 (mehr als 238 Milliarden) Zeichen.

2.2 Fokussierung

In der o.g. DeReKo-Version kommen **29.944** verschiedene Zeichen vor. Ein Großteil dieser Zeichen würde man aber nicht i.e.S. als Bestandteil des Zeicheninventars der deutschen Sprache bezeichnen. Selbst wenn ein Text sich auf den ersten Blick als „deutschsprachig“ klassifizieren lässt, können nicht-deutsche Zeichen vorkommen, sowohl in lateinischer Schrift wie auch in anderen Schriftsystemen, z.B. in Eigennamen (Piëch, Толстóй), Ortsnamen (Øresund, 東京), Zitaten (لا شيء على ما يرام, πάντα ῥεῖ), in fachsprachlichen Passagen (α, β, γ, ɸβ, dʒ, ...) usw.

Die allermeisten der 29.944 verschiedenen Zeichen sind für die wissenschaftliche Dokumentation des öffentlichen Deutsch kaum von Bedeutung, zumal sie in typischen Texten des öffentlichen Deutsch extrem selten sind. Die Gesamtübersicht kann auf Anfrage angefordert werden, für die Veröffentlichung an dieser Stelle wurde aber eine kompaktere Übersicht angestrebt.

Um sich der Aufgabenstellung, „typisch deutsche“ Zeichen in „typisch deutschen“ Texten zu zählen, anzunähern, kann einerseits die Datenauswahl eingegrenzt werden (auf möglichst wenig Fremdsprachliches oder Idiosynkratisches) oder der Zählweise

Kategorien zugrundegelegt werden (relevant vs. irrelevant), die dann auch beim Bestimmen der relativen Häufigkeiten optional herausgenommen werden können.

2.3 Definition eines Teilkorpus

Zur Bildung eines Teilkorpus, das überwiegend „typisch deutsche" Texte enthält, wurden die Kategorien „in deutschen Texten erwartbare Zeichen" (alle, die in (Basic) Latin-Zeichensätzen, diversen Ergänzungen und in „allgemeine Zeichensetzungen" enthalten sind[1]) und „nicht-deutsche Zeichen" (alle übrigen) definiert.

Für alle Texte in DeReKo wurde der Anteil nicht-deutscher Zeichen bestimmt. Die Texte wurden nach diesem Anteil sortiert und das Viertel der Texte mit dem höchsten Anteil an nicht-deutschen Zeichen aus der weiteren Betrachtung herausgenommen.

Als Resultat liegt somit ein Teilkorpus vor, das 179.617.554.197 (knapp 180 Milliarden) Zeichen umfasst und nur noch **365** verschiedene Zeichen enthält. Die Liste der verwendeten Texte (in Form von Archiv-Siglen) kann bei Bedarf angefordert werden.

2.4 Zeichen-Kategorien

Zu allen Zeichen wurde ihre Zugehörigkeit zu der Kategorie nach den Unicode General Categories[2] nachgeschlagen. Eine Teilmenge dieser Kategorien (grafische, mathematische oder z.B. auch chinesische Zeichen) wurde für die weitere Auswertung mit einem Status eingestuft, dass alle damit markierten Zeichen zusammengefasst als ein Zeichentyp „andere Zeichen" betrachtet werden. Unabhängig von ihrer UGC-Kategorie wurden auch sehr seltene Zeichen[3] dieser Gruppe zugeschlagen. Insgesamt wurden 161 Zeichen auf diese Weise herausgefiltert, die im gewählten Korpus eine Gesamtfrequenz von 11.813.717 aufweisen, was einem relativen Anteil von ca. 0,006% entspricht.

Es verbleiben also **204** Zeichen, deren Vorkommen ca. 99,994% des gewählten Korpus abdecken und deren Häufigkeit einzeln ermittelt werden soll.

3 Varianten

Da viele Fragestellungen ein besonderes Interesse vor allem an der Verteilung der Zeichen des deutschen Alphabets haben, werden weiter vereinfacht zusammengefasste Versionen angeboten, zu denen jeweils auch aus zwei Varianten zur Berechnung der relativen Häufigkeiten ausgewählt werden kann.

3.1 Deutsches Alphabet

Das deutsche Alphabet umfasst 59 Zeichen, sofern zwischen Groß- und Kleinschreibung unterschieden wird (a-zäöüßA-ZÄÖÜ[4]). In einer Zählung wurde nur zwischen diesen Zeichen (und allen anderen als ein Zeichentyp „andere Zeichen") unterschieden und ihre Häufigkeit getrennt ausgewiesen.

1 ftp://www.unicode.org/Public/UNIDATA/Blocks.txt: Basic Latin, General Punctuation, Latin-1 Supplement, Latin Extended-A, Latin Extended-B, Latin Extended Additional

2 http://www.unicode.org/reports/tr44/#General_Category_Values: Cf Ll Lu Nd No Pd Pe Pf Pi Po Ps Sk Sm Zs

3 relative Häufigkeit < 10^{-9}, höchstens ein Vorkommen pro eine Milliarde Zeichen

3.2 Groß-/Kleinschreibung ignorieren

Sofern zwischen Groß- und Kleinschreibung nicht unterschieden wird, reduziert sich das Repertoire auf 30 Zeichen (a-zäöüß), die allerdings – abgesehen vom ß[4] – nicht mehr nur für sich alleine stehen, sondern jeweils Groß- und Kleinschreibweise zusammenfasst. In der kompaktesten Darstellung sind nunmehr nur noch 30 verschiedene Zeichentypen ausgewiesen, zu der die Häufigkeit des kleingeschriebenen Zeichens plus ggf. die Häufigkeit des großgeschriebenen Zeichens (h(„a“) = h(‚a‘) + h(‚A‘)) angegeben ist.

3.3 Berechnung relativer Häufigkeiten

Zu diesen beiden kompakten Formen der Zeichenhäufigkeitslisten wurde auf zweierlei Arten die relative Häufigkeit ermittelt. In der ersten Fassung wurde der Anteil der Häufigkeit eines Zeichens (bzw. Zeichentyps) an der Gesamthäufigkeit aller betrachteten Zeichen berechnet (obere Formel), in einer zweiten Fassung nur in Relation zu der Gesamthäufigkeit aller Zeichen des deutschen Alphabets (untere Formel).

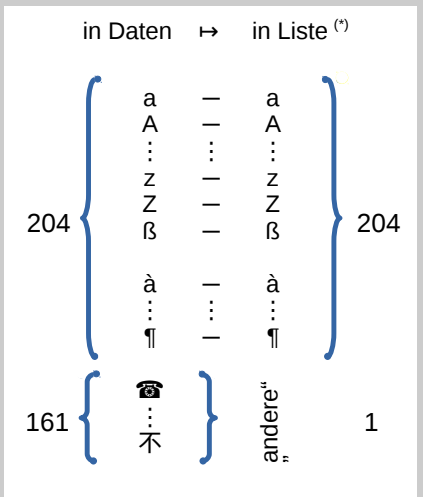
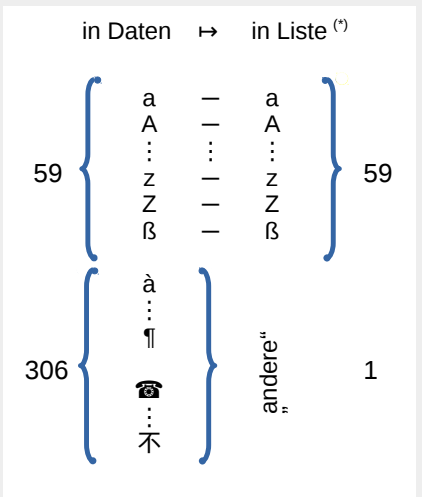
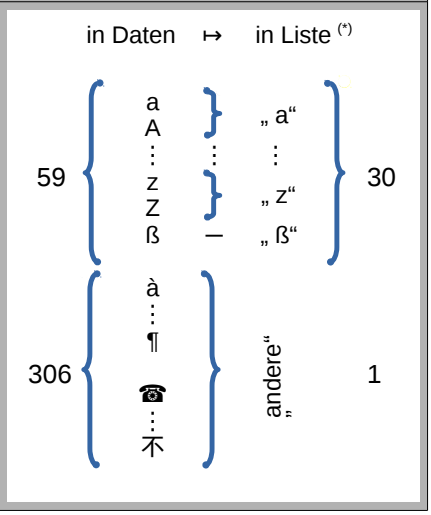
$$(a)\quad rh(zeichen_x) = \frac{ah(zeichen_x)}{\sum_{i=1}^{365} ah(zeichen_i)}$$

$$(b)\quad rh(zeichen_x) = \frac{ah(zeichen_x)}{\sum_{i=1}^{59} ah(zeichen_i)}$$

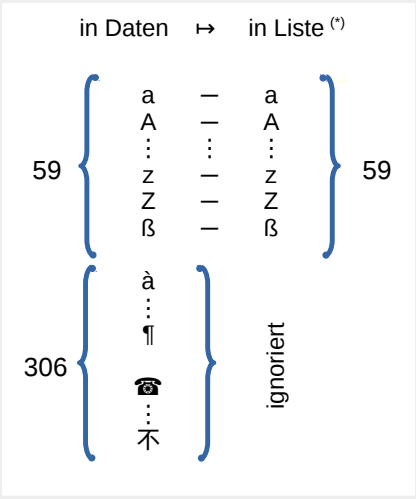
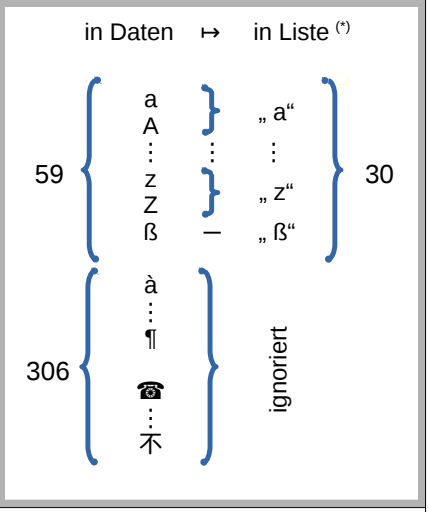
4 Zusammenfassung

Korpus	Anzahl Zeichen(-typen) gezählt	
DeReKo 2017-II	29.994	(vgl. 2.2)
Teilkorpus „typisch deutsch“	365	(vgl. 2.3)

Abbildung der in dem Teilkorpus beobachteten 365 Zeichen auf die in den Listen dargestellten Einheiten:

Version 1a: „markante Zeichen“ (vgl. 2.4)	Version 2a: deutsches Alphabet, große/kleine Zeichen (vgl. 3.1)	Version 3a: deutsches Alphabet, groß/klein zusammengefasst (vgl. 3.2)
		

[4] Auch wenn die großgeschriebene Variante der Ligatur ß inzwischen definiert ist, ist sie im Teilkorpus nicht belegt und wird deshalb nicht berücksichtigt (im Gesamtbestand 906 Mal).

	Nichtberücksichtigung der „anderen Zeichen“ (Variante für andere Berechnung der relativen Häufigkeiten, vgl. 3.3)	
	Version 2b:	Version 3b:
		
(*) in den Spalten sind (jeweils links:) die in den Daten beobachteten unterschiedlichen Zeichen angedeutet, (jeweils rechts:) die in den Listen dargestellten Einheiten, die z.T. 1:1 die beobachteten Zeichen repräsentieren, ansonsten n:1 als nicht relevante Zeichen zur Kategorie „andere Zeichen“ oder von groß- oder kleingeschriebenen Varianten 2:1 auf eine Nennform reduziert werden.		

5 Übersicht

Folgende Varianten werden zum Download angeboten:

- Eine Version mit allen markanten Zeichen und der Zusammenfassung aller anderen Zeichen zur Kategorie „andere Zeichen“ wie in Kapitel 2 hergeleitet.
- Jeweils zwei Versionen, die auf das deutsche Alphabet mit bzw. ohne Unterscheidung von Groß- und Kleinschreibung reduziert wurden, zu denen wiederum in zwei Varianten die relativen Häufigkeiten mit Bezug zur Gesamtheit (dabei mit Kategorie „andere Zeichen“) oder zur Summe der dadurch abgedeckten Zeichen (dabei ohne Kategorie „andere Zeichen“) berechnet wurden.
-

uniXXX=	**alle markanten Zeichen**	**nur deutsches Alphabet**	
Berechnung relative Häufigkeit (vgl. 3.3)		**Groß-/Kleinschreibung unterscheiden**	**Groß-/Kleinschreibung ignorieren**
mit „andere Zeichen“ (Varianten a, s.o.)	...uni-204-a-c...	...uni-059-a-c...	...uni-030-a-l...
ohne „andere Zeichen“ (Varianten b, s.o.)		...uni-059-b-c...	...uni-030-b-l...

6 Formate

Alle Zeichenhäufigkeitslisten werden auf einer eigenen Webseite angeboten. Auf dieser ist jeweils die Liste selber als Tabelle im html-Format eingebettet. Durch Klicken auf die Tabellenköpfe lassen sich die Spalten jeweils umsortieren. Angegeben ist

- die relative Häufigkeit **RF**
- die absolute Häufigkeit **AF**
- der hexadezimale Unicode-Codepunkt **U**(**)
- der dem Code entsprechende Dezimalwert **D**(**)
- die Unicode General Category **GC**(**)
- die Glyphe (Darstellung des Zeichens) **G**
- der Name des Zeichens **N**(**)
- der Unicode-Block-Bezeichnung **B**

In den Versionen, in denen Groß- und Kleinbuchstaben zu einer Einheit zusammenfasst werden, sind die mit (**) versehenen Angaben angepasst worden (im Falle des Namens N) bzw. fehlen.

Auf den Seiten wird ebenfalls diese Dokumentation sowie die jeweilige Liste als Dateien in drei Formaten zum Download angeboten.

- rtf: Rich Text Format
- txt: „einfache“ Textdatei
- csv: eigentlich „comma-separated values“

Das Rich Text-Format kann von allen gängigen Textverarbeitungsprogrammen verarbeitet werden und eignet sich somit für das Öffnen mit bzw. Einlesen in Word, Open/Libre Office oder ähnliche Systeme. Die Liste wird als Tabelle dargestellt, ähnlich wie auf der Webseite.

Einfache txt-Textdateien enthalten keine Formatierungshinweise. Sie lassen sich mit den allereinfachsten Programmen (wie etwa Texteditoren) öffnen und bearbeiten. Um den Eindruck der spaltenartigen Anordnung einer Tabelle zu simulieren, werden die Lücken zwischen den Werten mit der jeweils erforderlichen Anzahl Leerzeichen aufgefüllt.

Eine csv-Datei ist im Grunde genommen auch eine einfache Textdatei, bei der die spaltenartige Aufteilung aber durch ein dafür ausgewähltes Zeichen kodiert ist. Statt eines Kommas (von der sich die Bezeichnung ableitet) sind auch andere Zeichen möglich. In diesem Fall hat sich das Tabulator-Zeichen angeboten, da es im Gegensatz zum Komma sonst nicht im Dokument verwendet wird.

Falls Sie mit einer der beiden Textdateitypen arbeiten, beachten Sie bitte, dass diese im Zeichensatz utf-8 kodiert sind. Falls Sie eine Zeichenhäufigkeitsliste in ein Tabellenkalkulationsprogramm (wie Excel oder Open/Libre Calc) importieren wollen, bietet sich die CSV-Datei an. Stellen Sie bitte ebenfalls diese Kodierung ein und beachten Sie als Spaltentrenner nur das Tabulatorzeichen.

Die Kodierung der html- und der rtf-Varianten ist selbstdokumentierend, d.h., dass idealerweise auch alle (druck-/darstellbaren) Glyphen auf dem Bildschirm und beim

Ausdruck angezeigt werden. Je nach Konfiguration Ihrer lokalen Arbeitsumgebung (Betriebssystem, Webbrowser, installierte Zeichensätze) kann es jedoch zu Abweichungen kommen.

Referenzen

DeReKo (2017): *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-II* (Release vom 01.10.2017). Mannheim: Institut für Deutsche Sprache. http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html.

Kontakt

Falls Sie speziellere Informationen benötigen, als dieses Werk bereithält, wenden Sie sich bitte an
derewo@ids-mannheim.de.

Bei Veröffentlichung auf diesem Werk aufbauender Forschungsergebnisse bitten wir um eine kollegiale Mitteilung an derewo@ids-mannheim.de.

Lizenzbestimmungen

(zu zitieren als, XXX durch jeweilige Instanzen zu ersetzen:)

> DeReChar v-uni-XXX-2018-02-28-1.0, Korpusbasierte Zeichenhäufigkeitslisten zum Deutschen Referenzkorpus DeReKo, http://www.ids-mannheim.de/derewo, © Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2018.

Die Zeichenhäufigkeitslisten und die Dokumentation bilden eine Einheit. Diese Lizenzbestimmung darf aus keinem der Dokumente entfernt werden.

Dieses Werk ist unter die Creative Commons-Lizenz (by) gestellt (http://creativecommons.org/licenses/by/4.0/deed.de).

Namensnennung 4.0 International (CC BY 4.0)

Dies ist eine allgemeinverständliche Zusammenfassung der Lizenz (die diese nicht ersetzt). Haftungsbeschränkung.

Sie dürfen:

- **Teilen** – das Material in jedwedem Format oder Medium vervielfältigen und weiterverbreiten
- **Bearbeiten** – das Material remixen, verändern und darauf aufbauen
- und zwar für beliebige Zwecke, sogar kommerziell.

- Der Lizenzgeber kann diese Freiheiten nicht widerrufen, solange Sie sich an die Lizenzbedingungen halten.

Unter folgenden Bedingungen:

- **Namensnennung –** Sie müssen angemessene Urheber- und Rechteangaben machen, einen Link zur Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden. Diese Angaben dürfen in jeder angemessenen Art und Weise gemacht werden, allerdings nicht so, dass der Eindruck entsteht, der Lizenzgeber unterstütze gerade Sie oder Ihre Nutzung besonders.
- **Keine weiteren Einschränkungen –** Sie dürfen keine zusätzlichen Klauseln oder technische Verfahren einsetzen, die anderen rechtlich irgendetwas untersagen, was die Lizenz erlaubt.

Hinweise:

- Sie müssen sich nicht an diese Lizenz halten hinsichtlich solcher Teile des Materials, die gemeinfrei sind, oder soweit Ihre Nutzungshandlungen durch Ausnahmen und Schranken des Urheberrechts gedeckt sind.
- Es werden keine Garantien gegeben und auch keine Gewähr geleistet. Die Lizenz verschafft Ihnen möglicherweise nicht alle Erlaubnisse, die Sie für die jeweilige Nutzung brauchen. Es können beispielsweise andere Rechte wie Persönlichkeits- und Datenschutzrechte zu beachten sein, die Ihre Nutzung des Materials entsprechend beschränken.