Projektarbeit

im Fach

Kryptographie

„Europäische Alphabete“

Ronald Wick,
in Anlehnung an
Gorąca Katarzyna
(Gaststudentin)

Inhaltsverzeichnis

1 Einführung

„Alle Geheimnisse liegen in vollkommener Offenheit vor uns. Nur wir stufen uns gegen sie ab, vom Stein bis zum Seher. Es gibt kein Geheimnis an sich, es gibt nur Uneingeweihte aller Grade.“(Christian Morgenstern)

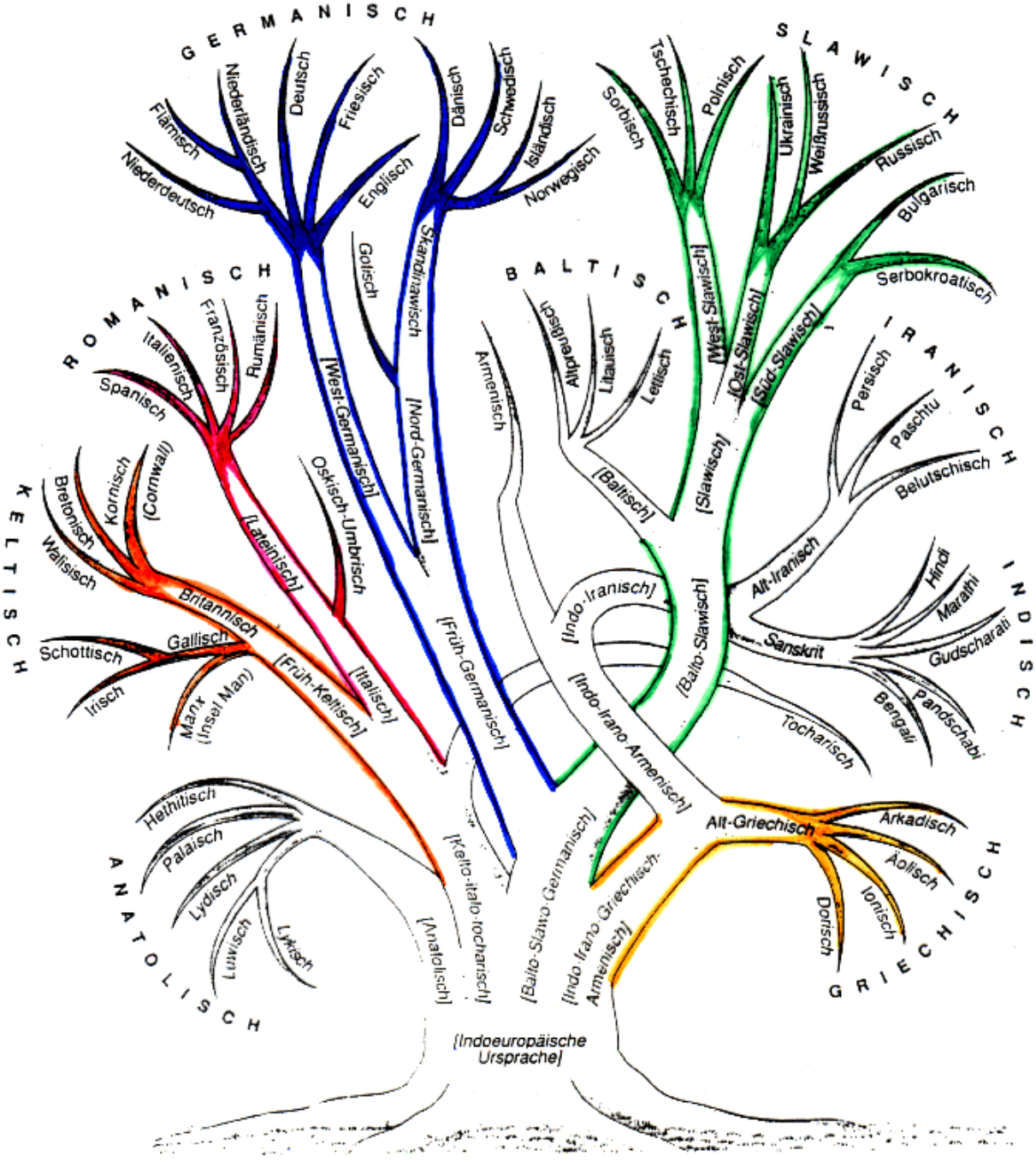
Diese Vers beschreibt, wie die Geheimnisse oder verschlüsselten Botschaften eigentlich offen vor uns und den Betrachtern liegen, aber da wie nicht den Schlüssel kennen, erschließt sich uns auch nicht die in ihr verborgene Nachricht. Um diesen Geheimnissen auf den Grund zu gehen, braucht man aber fundiertes Wissen und Anwendungen, die einem eine Basis zur Untersuchung liefern.
In diesem Projekt werden verschiedene Sprachen unterschiedlichster Herkunft aus dem europäischen Raum untersucht, um der Frage „Kann man aus den sprachlichen Eigenarten und der Anatomie von Sprachen nützliche Information zur Entschlüsselung von Chiffren generieren?“ nach zu gehen.

Für die Durchführung der Analysen werden verschiedene statistische Methoden genutzt.
Mit Hilfe der analysierten Daten sollen Zusammenhänge bzw. Gesetzmäßigkeiten erkannt werden, welche später eine Grundlage bildet um einfach verschlüsselte Chiffren zu entschlüsseln.

2 Herkunft und Verwandtschaft der heutigen Sprachen in Europa

Die Sprachen haben entsprechende Morphologien; Morphologien, die sich im Lauf der Zeit wandeln und sich von Sprache zu Sprache unterscheiden. In Zeiten politischer, ökonomischer oder gesellschaftlicher Veränderungen in Europa kommt es zu Wandlungsprozessen in Bezug auf die Sprache. Solche Veränderungen fanden immer statt. Ich werde kurz einen geschichtlichen Faden der Entwicklung der europäischen Sprachen darstellen.
Nachdem im 19. Jahrhundert eine zuverlässige Methodik von dem Vergleich von Einzelsprachen entwickelt worden war, konnte der wissenschaftliche Nachweis erbracht werden, dass verschiedenste Sprachen auf dem europäischen und asiatischen Kontinent auf eine gemeinsame Ausgangssprache zurückgehen. Zusammen bilden sie die indogermanische Sprachfamilie. Die Bezeichnung indogermanisch wurde im deutschen Sprachraum beibehalten, in anderen Sprachen wird fast nur die Bezeichnung indoeuropäisch verwendet. Etwa die Hälfte der Weltbevölkerung spricht heute eine indogermanische Sprache.
Die Schwierigkeit beim der allgemeinen Betrachtung der indogermanischen Sprachen ist, dass man schwer die gemeinsamen Wurzeln nach verschiedenen Sprachstämmen ordnen kann. Selbst in der heutigen Zeit gibt es keine plausiblen Entwürfe für die sogenannten „Supersprachstämme“. Damit ist die Rückführung der Sprachen auf einen gemeinsamen oder ähnlichen Ursprung nicht erklärbar.
Linguisten gehen davon aus, dass die Entstehung vor den einzelnen Sprachstämmen nicht nachzuvollziehen ist, da man die Entstehung neuer Sprachen und Umformung zu den jetzigen Sprachen nicht mehr rekonstruieren kann.Versuche eine mögliche Verbindung herzustellen scheiterte am Beispiel eines hebräischen Dialektes, welcher fälschlicher Weise durch eine jüngere Form als rekonstruiert galt, aber durch neue archäologische Funde wiederlegt wurde.
Um diese Abhängigkeit der heutigen Sprachstämme untereinander abzugrenzen entwickelte August Schleicher (1821-1868) im Mitte des 19. Jahrhunderts die Stammbaumtheorie. Er ging davon aus, dass sich Sprachen analog der Evolution biologischer Arten aus Ursprachen entwickeln. Danach verhalten sich die Beziehungen und Verwandtschaftsverhältnisse zwischen Sprachen genau so wie die Relationen der Arten in der Biologie, die sich in Form von Stammbäumen darstellen lassen. Ausgehend von seinen evolutionstheoretischen Überlegungen entwickelte August Schleicher u.a. das Stammbaummodell der indogermanischen Sprachfamilie (vgl. Quelle 3)Da die heutigen europäischen Sprachen aus einer Sprachfamilie stammen, weisen sie weitreichende Übereinstimmungen beim Vokabular, in der Flexion, in den grammatischen

Kategorien wie Numerus und Genus und im Ablaut auf. Inwieweit sie in statistischer Hinsicht übereinstimmen, werden wir uns nach der durchgeführten Analysen erfahren.



Drawing 1 Sprachenbaum - Der „Stammbaum“ der europäischen Sprachen

In dieser Arbeit werden folgende Zweige der indogermanischen Sprachfamilien in Betracht gezogen: germanischen, uralischen, romanischen, slawischen, baltischen, keltische und afroasiatischen.

Aus germanischen Zweig werden Deutsch, Englisch, Niederländisch, Dänisch, Norwegisch, Schwedisch; aus romanischen Französisch, Spanisch, Italienisch, Portugiesisch, Rumänisch; aus slawischen Zweig Polnisch, Tschechisch, Bulgarisch, Slowakisch; aus uralischen Finnisch, Estnisch, Ungarisch; aus dem baltischen Lettisch und Litauisch analysiert.

Zur Vergleichszwecken wurde auch Malti als Vertreter der afroasiatischen Sprache als Verbindungsglied zwischen dem europäischen Sprachraum und dem asiatisch/arabischen Sprachraum ausgewählt. Außerdem wird die Analyse ergänzt durch eine Sprache, welche eher zu den Minderheiten der europäischen Sprachen zählt und nur noch wenig gesprochen wird, dem Irischen.

3 Grundlagen der Analyse - Brechen durch statistische Methoden

Ziel der Kryptoanalyse ist es, Verschlüsselungsverfahren zu brechen, um damit eine geheime Nachricht lesen zu können. Beim Entschlüsseln lassen sich Schwachstellen eines Verschlüsselungsverfahrens ausnutzen. Es existieren dafür verschiedene Ansatzpunkte. Wir werden hier für uns interessant einen besprechen, nämlich Eigenheiten der Sprache. Sprache enthält ein schwer ausrottbares inneres Gerüst von Gesetzmässigkeiten: In jeder Sprache treten die einzelnen Buchstaben in bestimmten Häufigkeiten aus, die in längeren Texten meist recht eindeutig sind. Außerdem gibt es unmögliche oder zumindest unwahrscheinliche Buchstabenfolgen, wie "mrx", die in keinem Wort der betreffenden Sprache vorkommen. Andererseits existieren bevorzugte Buchstabenfolgen wie "ch" oder "ck", die recht häufig sind. Dieser Aspekt wird jetzt näher gebracht. (vgl. Quelle 1)

3.1 Häufigkeitsanalyse

3.1.1 Historischer Faden der Häufigkeitsanalyse

Die Idee der Häufigkeitsanalyse hat zum ersten Mal im 8. Jahrhundert das Licht der Welt erblickt. In dieser Zeit war Abu Bakr der 1. Kalif des Islams. Er fasste die Offenbarungen Mohammeds zu einer Schrift zusammen, die dann zu den 114 Kapiteln des Korans wurden. Schon im 7. Jahrhundert n. Chr. verbreitete sich der Islam rasant und ein Jahrhundert später im Jahre 750 begann das "Goldene Zeitalter". In dieser Zeit besaßen die Araber nicht nur das Wissen von Verschlüsselungsverfahren (z.B. die Beamten haben Steuerunterlagen durch eine monoalphabetische Verschlüsselung geschützt), sondern legten auch die Grundlagen für die Kryptoanalyse. Damals wurde auch Häufigkeitsanalyse nur per Zufall entdeckt. Im 8. Jahrhundert n.C. gab es in Arabien zahlreiche Koranschulen, in denen sehr viel Textforschung betrieben wurde, um zum Beispiel anhand von Analysen das Alters der einzelnen Wörter festzustellen.
Dabei fiel den Gelehrten auf, dass im Arabischen die Buchstaben mit unterschiedlichen Häufigkeiten vorkamen. Mit dieser Entdeckung war es nur noch ein kleiner Schritt bis man darauf kam, dass man dies für die Entschlüsselung der monoalphabetischen Verschlüsselung Buchstabenhäufigkeiten verwenden kann. Dies war der erste Ansatzpunkt, um die

monoalphabetische Verschlüsselung zu "knacken". Hieraus entstand dann die Häufigkeitsanalyse. Diese geschichtlichen Details wurden erst vor wenigen Jahren, 1987, entdeckt. (vgl. Quelle 4,5).

3.1.2 Technik der Häufigkeitsanalyse

Da eine Nachricht im Allgemeinen in einer Sprache formuliert wird und da in keiner Sprache jeder Buchstabe gleich häufig vorkommt, ist möglich geworden, auf diesem Grund die Chiffren zu entschlüsseln.
In einfachster Näherung kommt jedem Einzelzeichen x_i eine Wahrscheinlichkeit p_i des Auftretens (stochastische Quelle Q) zu, derart dass die Häufigkeit $m_i = Q(x_i)$ seines Vorkommens in einem Text der Länge M nahe bei M* pi liegt (vgl. Quelle 1, Seite 203).
Die Häufigkeitsanalyse lässt sich in Analyse der Einzelzeichen und n-Gramme unterteilen.
• Häufigkeitsanalyse der Einzelzeichen: Man teilt die Buchstaben des Klartextalphabets in mehrere Gruppen von „sehr häufig“ bis „unwahrscheinlich“ ein.
• n-Gramme (Bi- und Trigramme): Man untersucht die Häufigkeit von Wortpaaren und -tripeln, oder anderen Buchstabengruppen im Geheimtext
Vorgehensweise ist bei Häufigkeitsanalyse folgend: man zählt die einzelnen Buchstaben in dem Geheimtext und schreibt nach Häufigkeit auf, meist in Prozent, also relativ zur Gesamtzahl der Buchstaben (Buchstabenhäufigkeit). Die Häufigkeiten der einzelnen Buchstaben, die Häufigkeiten der Bigramme und der Trigramme kann man der vorher für jede Sprache erstellten Häufigkeitstabellen gegenüberstellen. Mit etwas Logik und Raten ist es möglich, die Buchstabenzuordnung zu finden und den Text vollkommen zu erschließen.

3.1.3 Darstellungstechniken

Zur intuitiven Häufigkeitserkennung ist es zweckmäßig, sich das `Häufigkeitsgebirge´ einer Sprache optisch einzuprägen. Manchmal reicht es einfach nicht. Im Fall der monoalphabetichen Substitution, die nicht linear mit q=1 ist, bedeutet das Häufigkeitsgebirge nichts mehr - die einzelnen Buchstabennachbarschaften sind zerrissen.
Dieses Problem kann man durch Häufigkeitsreihenfolge lösen: Das häufigste Zeichen im Geheimtext sollte dem häufigsten Buchstaben der betreffenden Sprache entsprechen; nach Entfernen des so bestimmten Paares wiederholt sich das Verfahren für den Rest, bis alle Zeichen erschöpft sind (vgl. Quelle 1, Seite 203-205).

Auf der Achse nach oben wird die Häufigkeit des entsprechenden Zeichens (in Prozent) abgetragen.

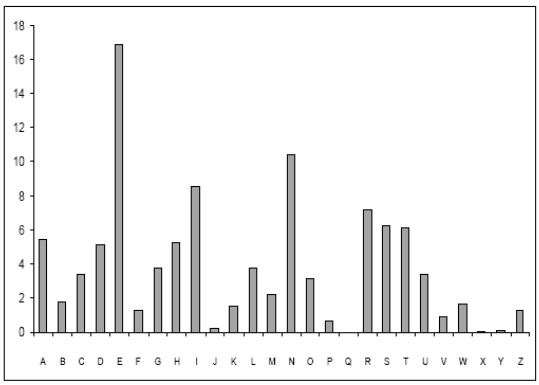


Diagramm 1 Häufigkeitsgebirge

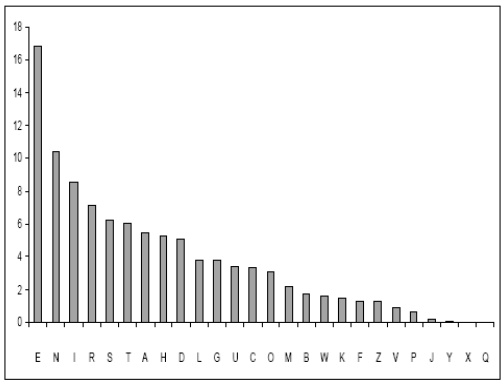


Diagramm 2 Häufigkeitsreihenfolge

3.1.4 Anwendung

Da die Genauigkeit der Häufigkeit mit der Länge einer Nachricht steigt, ist eine lange Nachricht deutlich einfacher zu entschlüsseln, als eine kurze. Dies gilt übrigens für die meisten kryptoanalytischen Verfahren.
Wieso sollte der Text mindestens 200 Zeichen umfassen? Theoretisch sollte das Verfahren zumindest für genügend lange Texte zum Ziel führen. Die Angaben über die Häufigkeiten der einzelnen Buchstaben in verschiedenen Sprachen sehr schwanken. Schon für die Häufigkeitsreihenfolge kann man verschieden Angaben und zwar in jedem Werk andere, die nur in groben Zügen übereinstimmen. Das kann durch Genre der Texte, über sich die Auszählung erstreckt hervorgerufen. Es wurde hier ein deutscher Text mit unterschiedlicher Anzahl der Zeichen analysiert. Auf Grund dieser Forschung zeigte sich, dass die Überlappung der Schwankungsbereiche geringer wird, je länger der Text ist. Wieso klappt die Häufigkeitsanalyse nicht immer? Wann findet diese Methode keine Anwendung. Dies folgt aus der letzten Frage: Ohne klare Häufigkeitsverteilung kann man auch keine Häufigkeitsanalyse machen. Die Häufigkeitsanalyse ist bei polyalphabetischen Verfahren machtlos.

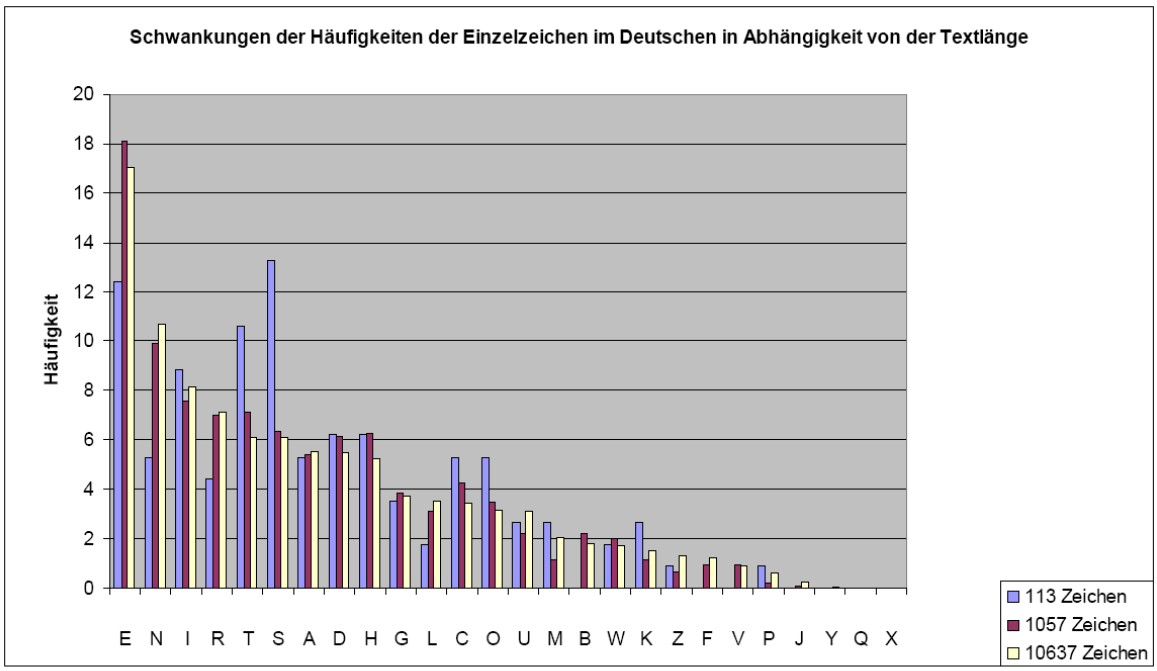


Diagramm 3 Schwankungen der Häufigkeiten der Einzelzeichen im Deutschen in Abhängigkeit von der Textlänge

3.1.5 Tools

Zur Analyse der Buchstabenhäufigkeiten wurde eine Software, das sich CrypTool nennt und ein Internettool - Kryptographiespielplatz v.1.40.2 benutzt.

3.2 Entropie

Die Entropie ist eine Kennzahl für Informationsgehalt. Der Informationsgehalt einer Nachricht M[i] ist definiert durch: *Informationsgehalt (M[i]):=log(1/ p[i])=−log(p [i])*.
Dabei ist p[i] die Wahrscheinlichkeit, mit der die Nachricht M[i] ausgestrahlt wird. Mit log ist der Logarithmus zur Basis 2 gemeint. Der Informationsgehalt hängt damit ausschließlich von der Wahrscheinlichkeitsverteilung ab, mit der die Quelle die Nachrichten erzeugt. Der semantische Inhalt der Nachricht geht nicht in die Berechnung ein. Da der Informationsgehalt einer seltenen Nachricht höher als der einer häufigen Nachricht ist, wird in der Definition der Kehrwert der Wahrscheinlichkeit verwendet.
Die Entropie gibt die Unsicherheit als Anzahl der notwendigen Ja / Nein-Fragen zur Klärung einer Nachricht oder eines Zeichens an. Hat ein Zeichen eine sehr hohe Auftrittswahrscheinlichkeit, so hat es einen geringen Informationsgehalt. Antworten, die sehr selten auftreten, haben einen hohen Informationsgehalt.
Für Dokumente, die ausschließlich Großbuchstaben enthalten, ist die Entropie mindestens 0 bit/char (bei einem Dokument, das nur aus einem Zeichen besteht) und höchstens log(26) bit/char = 4,700440 bit/char (bei einem Dokument, in dem alle 26 Zeichen gleich oft vorkommen) (vgl. Quelle 2, 6)

3.3 KI-Index

Der Koinzidenzindex ist eine statistische Methode, mit der verschlüsselte oder unverständliche Texte auf sprachliche Eigenschaften untersucht werden können. Der Koinzidenzindex ermöglicht die Unterscheidung eines Texts von anderem Text mit rein zufällig ausgewählten Buchstaben. Dieser Wert gibt an, wie zufällig ein gegebener Text ist.

$$I = \sum_{i-1}^{26} \frac{n_i * (n_i *-1)}{n * (n-1)}$$

I - Koinzidenzindex von x, symbolisch I(x); n_i - die Anzahl des Buchstabens in x
Der Koinzidenzindex beschreibt die Wahrscheinlichkeit aus einem Text zwei gleiche Buchstaben zu 'ziehen'. Er wird in der Kryptoanalyse und bei der Entschlüsselung historischer Schriftdokumente eingesetzt. Die Methode wurde von William Friedman für kryptologische

Zwecke entwickelt und im Jahr 1920 in seiner bahnbrechenden Arbeit „The Index of Coincidence and its Applications in Cryptography“ publiziert.
Geht man von Texten aus gleichverteilten Zufallszeichen über zu in einer Sprache verfassten Texten, so ändert sich der erwartete Wert erheblich. Eine Faustregel besagt, dass er etwa doppelt so groß ist. In verwandten Sprachen ähneln sich oft die Erwartungswerte, so dass bei unbekannten Texten anhand des Koinzidenzindex Vermutungen auf den zugehörigen Sprachraum angestellt werden können (vgl. Quelle 3).

4 Empirische Analyse der europäischen Alphabete

Bevor wir mit der Analyse beginnen, erklären wir, was man unter Alphabet versteht. Ein Alphabet ist eine Menge von Zeichen zur schriftlichen Darstellung von Lauten einer Sprache (vgl. Quelle 3). Die von mir analysierten Sprachen basieren auf dem lateinischen Alphabet. Falls eine Sprache nicht dem Lateinischen entspricht, wurde sie mit Hilfe der ISO Normen auf eine einheitliche und adäquate lateinische Schreibweise angepasst. Die lateinische Sprache wurde auf viele Sprachen übertragen und ist das am weitesten verbreitete Alphabet der Welt. Das archaische lateinische Alphabet bestand aus 21 Buchstaben:

A B C D E F Z H I K L M N O P Q R S T V X.

Heute unter dem Begriff lateinisches Alphabet versteht man solchen Zeichenvorrat, der sich aus folgenden 26 Buchstaben besteht:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Da in hier analysierten Sprachen noch diakritische Zeichen vorkommen, also zu Buchstaben gehörige kleine Zeichen wie Punkte, Striche, Häkchen oder Kringel, die eine besondere Aussprache oder Betonung markieren und unter oder über dem Buchstaben angebracht sind, wurden diese für den Zweck der Analyse zu einem Stammzeichen zusammengefasst. (bsp. wie in Polnischen ą -> a). Aber um sich dieser Varianten des lateinischen Alphabets bewusst zu sein, wurde bei der Darstellung der Sprache jeweils ihr Alphabet angezeigt.
Eine Besonderheit bildet die kyrillische Lautschrift, welche teilweise bei unklaren ISO-Normen oder nicht gebräuchlichen Umsetzungen in der Analyse sich stark an die Betonung und dessen betonte Silben richtet.
Um eine Analyse der Sprache durchzuführen, wurde jeweils ein Text in dieser Sprache gewählt. Da die Textbasis von unterschiedlichem Niveau ist (wegen der Sprachbarriere war es schwer das einheitliche Niveau der Sprache von Texten festzustellen), kann es zu Abweichungen kommen, d.h. die Ergebnisse können mit dieser in Literatur dargestellten nicht übereinstimmen.

4.1 Germanische Spachen

Die germanischen Sprachen als ein Zweig der indogermanischen Sprachfamilie gliedern sich noch in westgermanische, ostgermanische und nordgermanische Sprachen. Hier werden Deutsch, Englisch, Niederländisch, Dänisch, Norwegisch und Schwedisch untersucht.

4.1.1 Deutsch

Das deutsche Alphabet besteht aus 30 Buchstaben.

A Ä BC DE F GH I J K LM N OÖ PQ RS T U Ü V W X Y Z

a ä b c d e f g hi j k l m n o ö p q r s ß t u ü v w x y z

Hier wurde sowohl Alphabet der großen als auch kleinen Buchstaben angegeben. Merkwürdig ist hier das, dass diese beiden Alphabete nicht miteinander übereinstimmen. Der Unterschied besteht in Auftreten von Zeichen „β“ im Alphabet von kleinen Buchstaben.

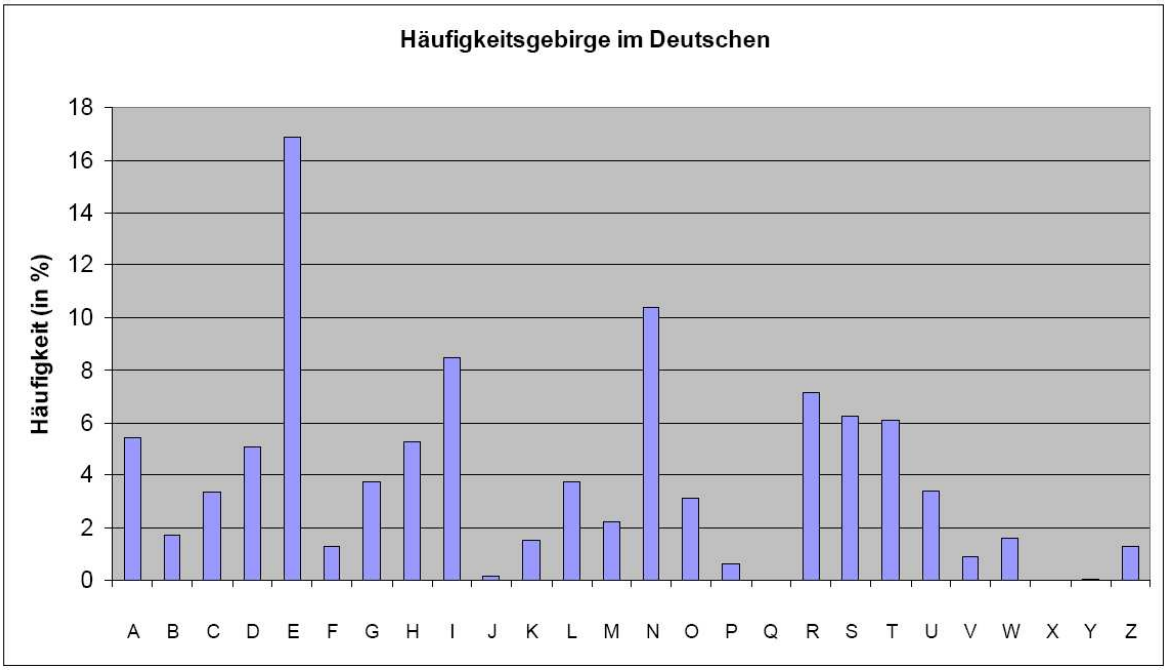


Diagramm 4 Die Buchstaben-Häufigkeitsverteilung eines deutschen Textes

Nach der Häufigkeitsanalyse eines deutschen Textes von Umfang 28462 Zeichen lassen sich sofort eine charakteristische Verteilung der Buchstaben in der deutschen Sprache erkennen: die „e“-Spitze und der „n“-Gipfel, der „b-c-d“-Anstieg, die „f-g-h-i“-Flanke mit anschließender „j-k“-Senke, die „o-p-q“-Senke mit anschließendem „r-s-t-u“-Kamm. Die Ergebnisse der n-Grammanalyse wurde unten in Tabellen abgebildet.

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	EN	5.0331	1172
2	ER	4.2214	983
3	CH	3.9552	921
4	DE	2.8515	664
5	EI	2.4349	567
6	TE	2.1214	494
7	GE	1.8466	430
8	IN	1.8466	430
9	IE	1.8122	422
10	HE	1.7951	418
11	ES	1.5675	365
12	IC	1.5073	351
13	UN	1.4687	342
14	RE	1.43	333
15	ND	1.3914	324
16	AN	1.3356	311
17	ST	1.3356	311
18	BE	1.2368	288
19	IS	1.2325	287
20	SC	1.2325	287
21	NE	1.2024	280
22	SE	1.065	248
23	DI	1.035	241
24	IG	0.9877	230
25	IT	0.9705	226
26	LI	0.9448	220

Table 1 Bigramme der deutschen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	ICH	1.857	340
2	SCH	1.5675	287
3	CHE	1.4911	273
4	DER	1.4692	269
5	EIN	1.3054	239
6	DIE	1.1634	213
7	TEN	0.9613	176
8	DEN	0.9394	172
9	ISC	0.9067	166
10	HEN	0.8411	154
11	CHT	0.8302	152
12	GEN	0.7756	142
13	INE	0.7428	136
14	UNG	0.7319	134
15	UND	0.721	132
16	NDE	0.6936	127
17	EIT	0.6063	111
18	REI	0.5899	108
19	BER	0.5298	97
20	LIC	0.5243	96
21	EIC	0.5134	94
22	ACH	0.5025	92
23	HER	0.497	91
24	NEN	0.4861	89
25	EST	0.4806	88
26	DES	0.4533	83

Table 2 Trigramme der deutschen Sprache

4- Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	ISCH	1.2061	166
2	SCHE	1.1553	159
3	CHEN	0.9954	137
4	EINE	0.9155	126
5	EICH	0.683	94
6	LICH	0.6685	92
7	ICHT	0.6612	91

Table 3 4-Gramme der deutschen Sprache

Die 5 häufigsten Buchstaben „e-n-i-s-r“ decken bereits 50 % der vorkommenden Buchstaben ab, die häufigsten 10 dann über 75 %. Nicht uninteressant ist auch die Häufigkeit des Vorkommens von Wörtern. Die zehn häufigsten Worte sind: die, der, und, den, am, in, zu, ist, dass, es (vgl. Quelle 1, Seite 221) Die statistischen Größen der deutschen Sprache lauten:

Entropie: 4.07 (*maxmögliche :*4.70)
KI = 0.0754239268745251 (0.0384615384615385)

0.07542 entspricht der Wahrscheinlichkeit der deutschen Sprache ob zwei beliebige Buchstaben gleich sind. Bei Gleichverteilung der Buchstaben ist der Erwartungwert 1/26 also etwa 0,0385. Der wesentlich höhere Wert für die deutsche Sprache gegenüber der englischen Sprache spiegelt vor allem die wesentlich größere Häufigkeit des dominanten Buchstabens E im Deutschen (etwa 17,5%) gegenüber dem Englischen (etwa 13%) wider.

4.1.2 Englisch

Das englische Alphabet besteht aus 26 Buchstaben:
ABC D E F G H I J K LM N OPQ RS T U V W X Y Z

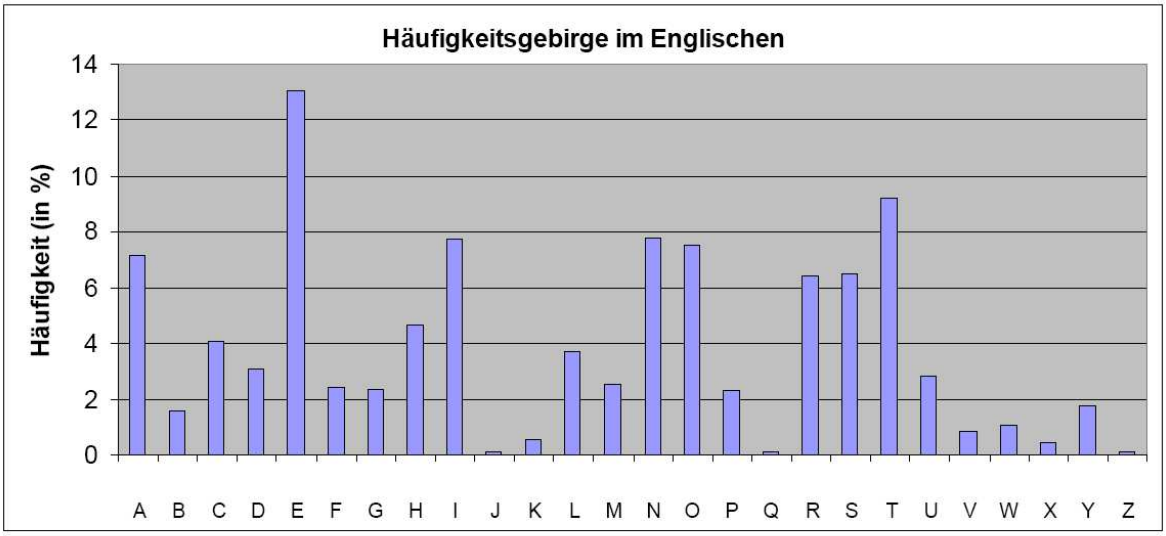


Diagramm 5 Die Buchstaben-Häufigkeitsverteilung eines englischen Textes

Hier wurde ein Text von Umfang 15956 Zeichen analysiert. Auf Grund der Abbildung lässt sich schon auf den ersten Blick e mit relativ hoher Häufigkeit, ca. 13%. Das Häufigkeitsgebirge kann man wie folgt beschreiben : ein ausgeprägter „a“-Gipfel, es besteht ein h-i-Kamm und ein „l-m-n-o“-Kamm, der „r-s-t-u“-Kamm hat einen „t“-Gipfel; jedoch finden sich „b-c-d“-Flanke, „j-k“-Senke und „v-w-x-y-z“-Niederung wieder. Die 5 häufigsten Buchstaben decken 45 % der vorkommenden Buchstaben ab, die häufigsten 10, dann 74 %. (vgl. Quelle 1)
Am häufigsten vorkommenden Wörter im Englischen sind: the, of, and, to, a, in, that, it, is und I. Als Merkwürdigkeit kann man erwähnen, dass die häufigsten Worte in den indogermanischen Sprachen weit überwiegend Formwörter, nämlich Artikel, Präpositionen, Konjuktionen und andere Hilfpartikel sind, im Gegensatz zu Begriffswörtern wie Substantiven, Adjektiven und Verben. Unter den 70 häufigsten Worten der englischen Sprache sind keine Begriffswörter, unter den 100 häufigsten sind nur 10 Begriffswörter. (vgl. Quelle 1, Seite 221)

Die statistischen Grössen der englischen Sprache lauten:

Entropie: 4.16 (*maxmögliche :*4.70)
KI = 0.0667762028303953 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	TH	3.5935	459
2	IN	3.3352	426
3	HE	3.1394	401
4	ER	2.0434	261
5	EN	1.8711	239
6	ON	1.832	234
7	AN	1.7067	218
8	ES	1.6911	216
9	ST	1.6049	205
10	NE	1.5893	203
11	NG	1.5267	195
12	TI	1.3701	175
13	RE	1.3388	171
14	AT	1.2996	166
15	RO	1.2996	166
16	CO	1.2292	157
17	ND	1.2292	157
18	TE	1.2057	154
19	TO	1.1978	153
20	IS	1.1822	151
21	ED	1.1744	150
22	OF	1.143	146
23	AR	1.1195	143
24	IO	1.0413	133
25	OM	1.0413	133
26	AL	0.9943	127

Table 4 Bigramme der englischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	THE	3.636	352
2	INE	1.5081	146
3	ION	1.2809	124
4	ENG	1.1053	107
5	GIN	1.1053	107
6	NGI	1.0949	106
7	TIO	1.0846	105
8	AND	0.9606	93
9	COM	0.8573	83
10	ING	0.8367	81
11	UST	0.6404	62
12	TER	0.5991	58
13	NES	0.5785	56
14	TRO	0.5475	53
15	TUR	0.5475	53
16	ATI	0.4958	48
17	STR	0.4855	47
18	STI	0.4752	46
19	OMB	0.4648	45
20	ATE	0.4545	44
21	RAT	0.4338	42
22	CON	0.4235	41
23	ARE	0.4132	40
24	BUS	0.4132	40
25	DER	0.4132	40
26	ENT	0.4132	40

Table 5 Trigramme der englischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	ENGI	1.4932	106
2	GINE	1.4932	106
3	NGIN	1.4932	106
4	TION	1.4086	100
5	INES	0.7888	56
6	COMB	0.6339	45
7	BUST	0.5635	40

Table 6 4-Gramme der englischen Sprache

4.1.3 Niederländisch

Niederländische Sprache steht dem Deutschen nahe. Deswegen wird sie wegen der vielen Gemeinsamkeiten von deutschen Muttersprachlern oft als eine Art Plattdeutsch angesehen. Wie falsch dieses Vorurteil ist, wird bei näherer Betrachtung deutlich, denn das Niederländische hat eine auch vom Niederdeutschen getrennte Entwicklung durchlaufen.

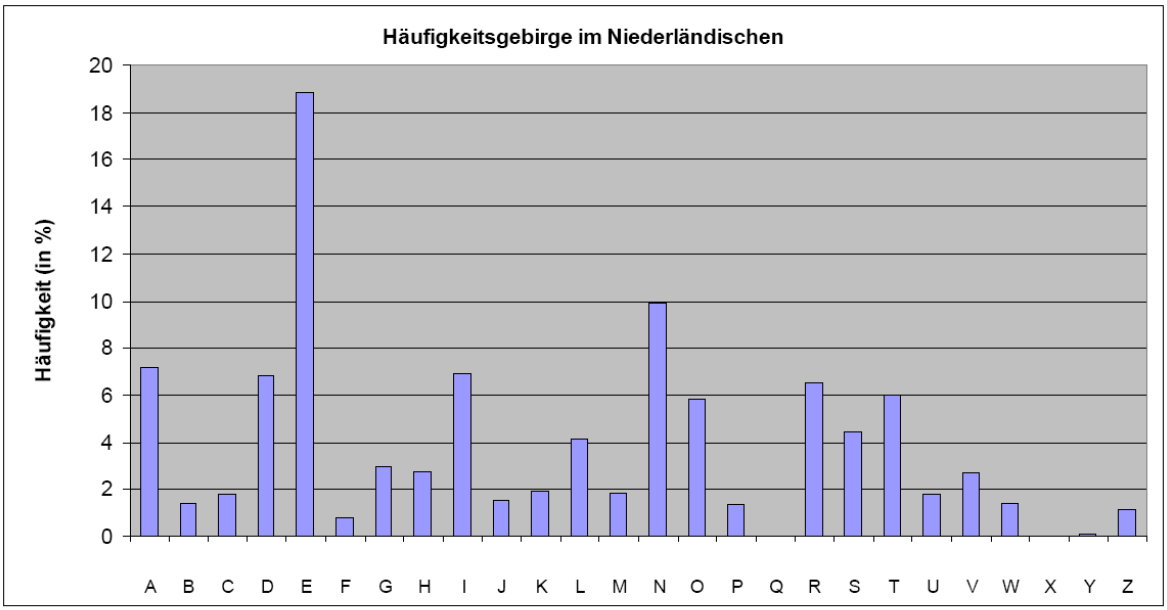


Diagramm 6 Die Buchstaben-Häufigkeitsverteilung eines niederländischen Textes

Wie im Deutschen lässt sich auf den ersten Blick e mit relativ hoher Häufigkeit, sogar höher als im Deutschen ca. 18,8 %.

Die statistischen Größen von Niederländisch lauten:
Entropie: 4.06 (*maxmögliche :*4.70)
KI = 0.0794593543153402 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	EI	6.1224	6
2	IG	5.102	5
3	TE	4.0816	4
4	ER	3.0612	3
5	FI	3.0612	3
6	GK	3.0612	3
7	IT	3.0612	3
8	KE	3.0612	3
9	ST	3.0612	3
10	UF	3.0612	3
11	CH	2.0408	2
12	DE	2.0408	2
13	EN	2.0408	2
14	GR	2.0408	2
15	IE	2.0408	2
16	IN	2.0408	2
17	IS	2.0408	2
18	NA	2.0408	2
19	ND	2.0408	2
20	RT	2.0408	2
21	AB	1.0204	1
22	AC	1.0204	1
23	AE	1.0204	1
24	AL	1.0204	1
25	AM	1.0204	1
26	AN	1.0204	1

Table 7 Bigramme der niederländischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	VAN	1.2475	132
2	AND	1.1908	126
3	SCH	1.1341	120
4	DEN	1.0301	109
5	OOR	0.9923	105
6	DER	0.8789	93
7	VER	0.8789	93
8	NDE	0.86	91
9	HET	0.8222	87
10	ING	0.7939	84
11	LAN	0.7372	78
12	ERD	0.7277	77
13	TEN	0.7183	76
14	IJK	0.6616	70
15	EDE	0.6427	68
16	EEN	0.6143	65
17	GEN	0.5954	63
18	RDE	0.5954	63
19	LIJ	0.5671	60
20	TER	0.5671	60
21	AAR	0.5576	59
22	ELI	0.5482	58
23	WER	0.5293	56
24	EER	0.5009	53
25	STE	0.5009	53
26	UIT	0.4914	52

Table 8 Trigramme der niederländischen Sprache

4-Gramme		
Nr.	Teilstring	Häufigkeit (in %)
1	FIGK	4.8387
2	GKEI	4.8387
3	IGKE	4.8387
4	KEIT	4.8387
5	UFIG	4.8387
6	TEIG	3.2258

Table 9 4-Gramme der niederländischen Sprache

4.1.4 Dänisch:

Die dänische Sprache zählt zu der indogermanischen Sprache mit der weiteren Untergliederung germanische Sprache, skandinavische Sprache und Dänisch.
Der Ursprung der dänischen Sprache als eigenständige Sprache liegt ca. 1000 Jahren zurück. Das Nordgermanische, welches sich etwa 200 n. Chr. aus dem Germanischen bildete wird als Vorläufer dieser Sprache gesehen. Weitere Ausprägungen der nordischen Sprache sind in der Zeit von 750 –1100 n. Chr. historisch nicht belegt, da es zu dieser Zeit eher um unterschiedliche Dialekte handelte als verschiedene Sprachmuster. Deutliche Einflüsse die sich in der dänischen Sprache wiederfinden gehen auf die Zeit der Wikinger zurück.
Die Ursache hierfür waren die Wikingerzüge, wobei Sprachfragmente in den englischen Raum gelangten, und somit das Altenglische beeinflußten. Beispiele dafür sind die Wörter sky oder window, welche ursprüngliche aus dem nordischen Wortschatz kamen.
Im weiteren geschichtlichen Verlauf wurde das Dänisch eher durch Deutsch bzw. Englisch bestimmt.
Einen starken Einfluss nahm die deutsche Sprache zur Zeit der Hanse auf das Dänisch, wodurch Vokabeln die in Handel und Handwerk gebräuchlich waren in den Grundwortschatz der Sprache übergingen. (Beispiele hierfür sind magt – Macht oder straks – stracks)
Des Weiteren vermehrte der französische Adel (17. –18 Jhdt) und das Englisch/ Amerikanische (20. Jhdt) den Grundwortschatz der dänischen Sprache.
Das dänische Alphabet besteht aus den typischen 26 lateinischen Buchstaben, welche durch die nordischen Umlaute ergänzt worden sind. Sie reihen sich am Ende des Alphabets ein.

A,B,C [...] X,Y,Z, Æ, Ø, AA

Bei der folgenden Erhebung würde eine Text von 13.276 Zeichen analysiert.
Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.
Die nachfolgenden nordischen Sonderzeichen wurden zum analytischen Zwecken wie folgt umbenannt.

V v	ersetzt durch W oder w	Æ	ersetzt durch AE oder ae
Ø	ersetzt durch OE oder oe	AA	bleibt so bestehen

Besonderheit der dänischen Sprachbildes ist das es keine Umlaute (ä, ö, ü) bzw. ß besitzt und, dass das V im Dänischen fast immer anstatt des W’s im Deutschen geschrieben wird.

Auffällig ist das im Dänischen kennt man außer bei AA keine Vokalverdopplung. (im Gegensatz zu den Konsonanten)
Die aufgeführten Sonderzeichen æ, ø und aa nehmen hauptsächlich den Platz für die nicht vorhandenen Umlaute ein und haben somit eine spezielle Charakteristik und treten deshalb ausschließlich (mit kleinen Ausnahmen) im nordischen Raum auf. Sie definieren die Aussprache der Wörter.
Aus *Aalborg* wird durch die Aussprache/Betonung des Sonderzeichens ein „*Ollborg*“.
„Charakteristisch für das Dänische und die übrigen nordischen Sprachen ist, dass der bestimmte Artikel nicht vor dem Substantiv steht, sondern an dieses angehängt wird. Während im Englischen, Deutschen oder Französischen Bestimmtheit durch ein vorangestelltes Element gebildet wird, hängt man in den nordischen Sprachen dem Substantiv eine Bestimmtheitsendung an. *The house, das Haus, la maison* entsprechen im Dänischen *huset*.“ (vgl. Quelle 7)

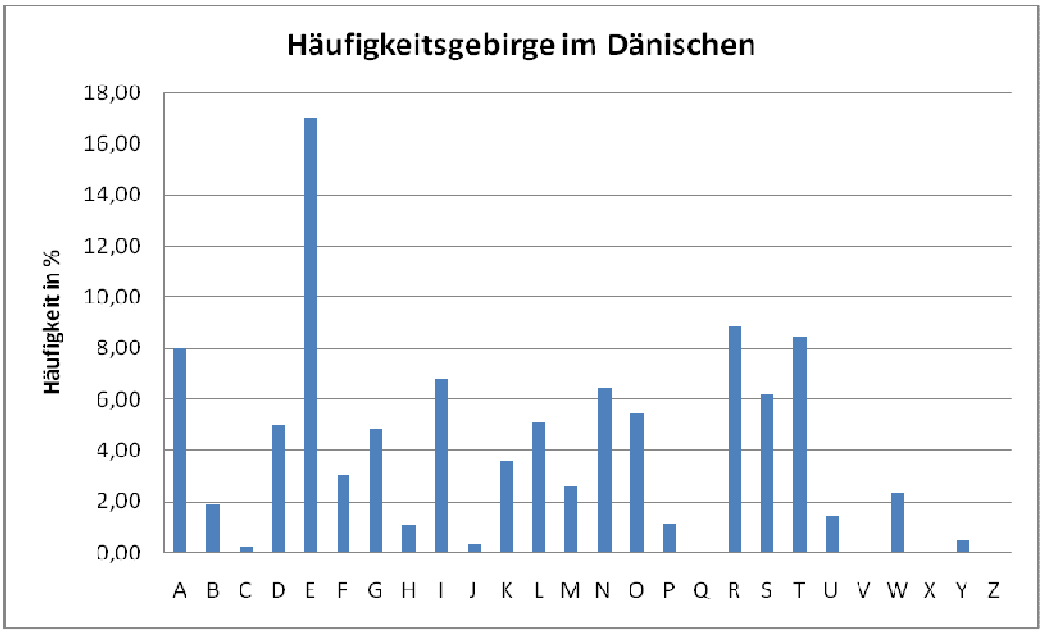


Diagramm 6.1 Die Buchstaben-Häufigkeitsverteilung eines dänischen Textes

Aus der Häufigkeitsanalyse lässt sich erkennen, dass auch wieder der Buchstabe E hier mit 16,97 % dominiert gegenüber den anderen. Zudem kann man sagen dass der Zusammenschluss von den Buchstaben A-E-I-N-R-S-T die Häufigkeit zu 61% abdecken.
Deutlich ist auch erkennbar das die Buchstaben Q und X außer in Fremdwörtern, Stadtnamen oder Personennamen kaum Anwendung in der dänischen Sprache finden.

Die statistischen Größen von Dänisch lauten:
Entropie: 4,028 (*maxmögliche :*4.70); *KI* = 0.072812406203102 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**ER**	4,60	610
2	**TE**	2,74	363
3	**EN**	2,51	333
4	**DE**	2,36	313
5	**ET**	1,96	260
6	**RE**	1,62	215
7	**OR**	1,48	197
8	**GE**	1,46	194
9	**LI**	1,27	168
10	**AN**	1,25	166
11	**EL**	1,22	162
12	**TI**	1,20	159
13	**LE**	1,16	154
14	**ST**	1,15	152
15	**FO**	1,12	148
16	**ES**	1,09	145
17	**IN**	1,09	145
18	**WE**	1,08	143
19	**ND**	1,06	141
20	**AT**	1,06	141
21	**OG**	1,05	139
22	**NG**	0,99	131
23	**RA**	0,99	131
24	**SK**	0,97	129
25	**ED**	0,96	127
26	**EK**	0,94	125
27	**IL**	0,94	125
28	**TT**	0,93	123
29	**NE**	0,92	122
30	**IG**	0,91	121

Table 9.1 Bigramme der dänischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**TER**	1,00	132
2	**FOR**	0,96	127
3	**ERE**	0,81	108
4	**TTE**	0,79	105
5	**DER**	0,72	96
6	**ERS**	0,63	84
7	**TIL**	0,63	84
8	**WER**	0,63	83
9	**ING**	0,61	81
10	**TEK**	0,58	77
11	**NDE**	0,56	74
12	**EKS**	0,46	61
13	**LIG**	0,45	60
14	**NGE**	0,45	60
15	**DET**	0,42	56
16	**STE**	0,41	54
17	**DEN**	0,41	54
18	**GEN**	0,40	53
19	**ERA**	0,39	52
20	**LER**	0,38	51
21	**KST**	0,38	51
22	**SKE**	0,38	51
23	**ERI**	0,36	48
24	**ENS**	0,36	48
25	**END**	0,36	48
26	**IND**	0,35	47
27	**GER**	0,35	47
28	**EDE**	0,35	47
29	**SAM**	0,35	47
30	**RET**	0,35	47

Table 9.2 Trigramme der dänischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**TTER**	0,45	60
2	**TEKS**	0,36	48
3	**EKST**	0,35	47
4	**INGE**	0,32	43
5	**NING**	0,29	38
6	**VERS**	0,28	37
7	**ERNE**	0,27	36
8	**ATTE**	0,27	36
9	**FATT**	0,26	34
10	**ERSI**	0,24	32

Table 9.3 4-Gramme der dänischen Sprache

4.1.5. Norwegisch:

Die norwegische Sprache zählt zu den Indogermanischen Sprachen, mit der weiteren Untergliederung Germanische Sprache, Skandinavische Sprache und Norwegisch.
Der Ursprung dieser Sprache sind auch wie bei den verwandten skandinavischen Sprachen Ruinenzeichen, welche auf das Jahr 300 v. Chr. zurückgehen.
Da aus dem ursprünglichen Dänisch erst Norwegisch[1] entstanden ist, unterscheidet sich die geschichtliche Entwicklung der norwegischen Sprache erst ab dem 19. Jh..
Zu diesem Zeitpunkt bildete sich eine Mischform aus Dänisch/Norwegisch, die Riksmal (Reichssprache), welche heute unter dem Namen Bokmal (Buchsprache) die vorherrschende Sprache im östlichen Teil Norwegens ist. Kurz nach der Bildung der Riksmal entwickelte Ivar Aasen in den 50er Jahren des 19. Jahrhunderts aus einer Zusammenstellung verschiedener westnorwegischer Dialekte eine neue nationale und unabhängigere Sprache, die Landsmal. Diese neue nationale Schriftsprache ist heute eher unter dem Begriff Nynorsk (Neunorwegisch) bekannt und wird vornehmlich in Westnorwegen gesprochen.
Beide Sprachen sind rechtlich gleichgestellt und werden an den Schulen unterrichtet.

Das norwegische Alphabet setzt sich aus 29 Buchstaben zusammen:

A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, Æ, Ø, Å

Bei der folgenden Erhebung würde ein Text von 12.103 Zeichen analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen. Bei dieser Zusammensetzung sind die Sonderzeichen wie folgt zu interpretieren bzw. zu sprechen: Æ ist dem deutschen Ä nahe, Ø dem Ö, Å dem O und Y wird oft als Ü Laut gesprochen.

Daraus ergibt sich die Ersetzungsvorschrift:

Æ	ersetzt durch AE	Ø	ersetzt durch OE
Å	ersetzt durch A		

Die Besonderheit der norwegischen Sprache ist, dass Buchstaben wie C, Y, W, X, Z nur in wenigen Fremdwörtern vorkommen. Statt CK schreibt man KK, für QU tritt KY ein, für PH/TH/KH treten F/T/K ein, Z (im Deutschen) wird meist durch S ersetzt: *sentrum / senter* entspricht "Zentrum", *sukker* entspricht "Zucker".

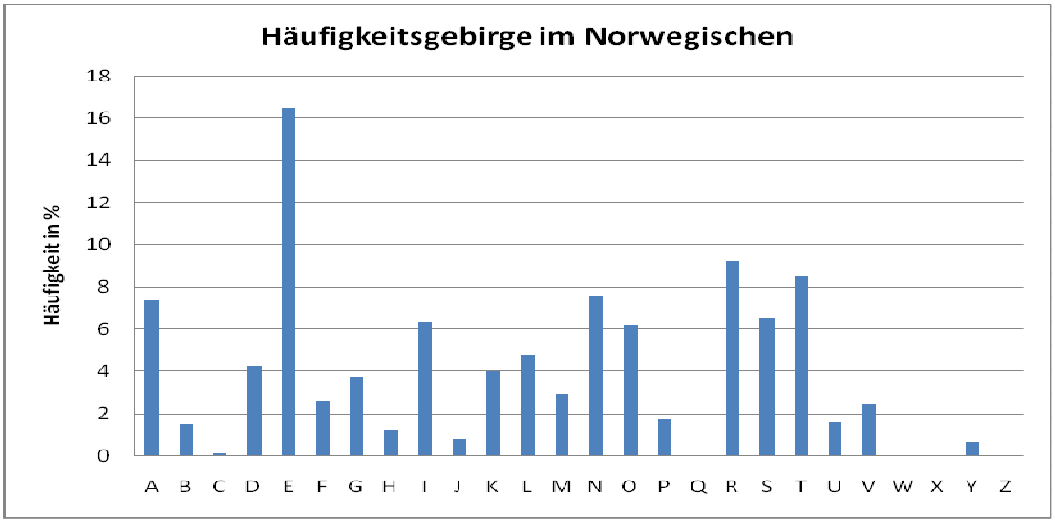


Diagramm 6.2 Die Buchstaben-Häufigkeitsverteilung eines norwegischen Textes

Das Häufigkeitsgebirge bestätigt die vorher aufgeführten Besonderheiten der Sprache, dass einige Buchstaben eigentlich nur in Fremdwörter vorkommen. Desweiteren ist zu beachten dass das E mit einer Häufigkeit von 16,47 % deutlich der am meisten vorkommende Buchstabe im Norwegischen ist. Danach folgen dann das R mit 9,21 % und das T mit 8,45 %.
Die weiteren Vokale verteilen sich statistisch gesehen mit einer normalen Streuung im Verhältnis zu den Konsonanten.
Der oft vorkommende Buchstabe E hat auch einen direkten Einfluss auf die n-Gramme, wobei sehr viele Strings mindestens ein E beinhalten.
Ausgesprochen häufig trat die Kombination ER und EN auf, oder auch FOR, FRE und TER.

Die statistischen Größen der norwegischen Sprache lauten:

Entropie: 4.01 (*maxmögliche :* 4.70)
KI = 0.075914454445111 (0.0384615384615385)

[1] Norwegen war bis 1814 Dänisch

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**ER**	3,85	466
2	**EN**	2,66	322
3	**TE**	2,15	260
4	**RE**	2,15	260
5	**ET**	1,93	234
6	**DE**	1,78	215
7	**OR**	1,62	196
8	**IN**	1,30	157
9	**AN**	1,28	155
10	**AR**	1,21	147
11	**NE**	1,20	145
12	**TI**	1,18	143
13	**RA**	1,18	143
14	**FO**	1,09	132
15	**ES**	1,08	131
16	**SE**	1,06	128
17	**ST**	1,02	124
18	**EL**	1,00	121
19	**LE**	0,95	115
20	**KE**	0,94	114
21	**LI**	0,93	113
22	**GE**	0,91	110
23	**OM**	0,91	110
24	**EK**	0,90	109
25	**TA**	0,89	108
26	**AT**	0,89	108
27	**TT**	0,88	106
28	**NG**	0,87	105
29	**SK**	0,86	104
30	**NT**	0,78	95

Table 9.4 Bigramme der norwegischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**FOR**	0,97	117
2	**ERE**	0,94	114
3	**TER**	0,71	86
4	**TIL**	0,59	72
5	**DET**	0,56	68
6	**ING**	0,52	63
7	**TEN**	0,50	60
8	**TTE**	0,42	51
9	**LIG**	0,41	50
10	**ENT**	0,41	50
11	**SOM**	0,40	48
12	**AND**	0,38	46
13	**ETT**	0,38	46
14	**VER**	0,36	43
15	**EKT**	0,36	43
16	**ERA**	0,35	42
17	**ENE**	0,34	41
18	**ENS**	0,34	41
19	**DER**	0,33	40
20	**NGE**	0,33	40
21	**RBE**	0,32	39
22	**FFE**	0,31	37
23	**ARB**	0,31	37
24	**SEN**	0,31	37
25	**RES**	0,30	36
26	**NNE**	0,30	36
27	**EID**	0,30	36
28	**ERS**	0,30	36
29	**AER**	0,30	36
30	**BEI**	0,29	35

Table 9.5 Trigramme der norwegischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**RBEI**	0,29	35
2	**BEID**	0,29	35
3	**ARBE**	0,29	35
4	**SJON**	0,26	31
5	**ASJO**	0,21	25
6	**ETTE**	0,21	25
7	**NING**	0,21	25
8	**FORS**	0,21	25
9	**OFFE**	0,19	23
10	**ETER**	0,19	23

Table 9.6 4-Gramme der norwegischen Sprache

4.1.6 Schwedisch:

Die schwedische Sprache zählt zu der indogermanischen Sprache mit der weiteren Untergliederung germanische Sprache, skandinavische Sprache (Ostskandinavisch) und Schwedisch.
Der Ursprung der schwedischen Sprache liegt auch wie bei den beiden skandinavischen Sprachen vorher in den Ruinenschriften. Aus der gemeinsamen Entwicklung mit dem Dänischen, bildete sich zwischen 800 und 1600 das Altschwedisch heraus.
Einige lateinische Wörter fanden vor der Einführung des lateinischen Alphabets im 13. Jh. Anwendung als abgelehnte Wörter im Syntax des Schwedischen. Neben den Abweichungen im Wortschatz unterscheidet sich Schwedisch vom Dänischen vor allem dadurch, dass es nach den Vokalen die alten stimmlosen Konsonanten k, t und p beibehalten hat, die im Dänischen zu g, d und b wurden, und in unbetonten Silben noch die Vokale a und o besitzt, wo im Dänischen e oder gar kein Vokal steht. Die Mehrzahl der Wörter, die den schwedischen Wortschatz umfassen, sind germanischen Ursprungs und viele wurden aus dem Lateinischen und Griechischen abgelehnt. Zur Zeit der Hanse (13. – 16 Jh.) und im 18. Jh. wurden viele Wörter aus dem Deutschen und dem Französischen Bestsandteil des Neuschwedisch.
Eine Vereinheitlichung gibt es aufgrund der vielen gesprochenen Akzenten nicht.
Das führt dazu, dass es in ganz Schweden auch heute noch große dialektale Unterschiede in der gesprochenen Sprache gibt. Die schwedischen Dialekte lassen sich in sechs Gruppen einteilen: die Sveadialekte, das Norrländische, die Götadialekte, das Südschwedische, das Gutnische und das Ostschwedische.
Mit diesem Problem und der Entwicklung der schwedischen Sprache beschäftigt sich die 1944 gegründete Schwedischen Akademie „Nämnden för svensk språkvård“.

Das schwedische Alphabet besteht aus den typischen 29 Buchstaben. Das W kommt in Lehnwörtern vor und galt bis 2006 nicht als eigener Buchstabe, sondern als Schreibvariante des V. Nach dem Z folgen noch Å, Ä, Ö, die als eigenständige Buchstaben gezählt werden und nicht wie im Deutschen als Varianten von A und O.

A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, Å, Ä, Ö

Bei der folgenden Erhebung würde eine Text von 17594 Zeichen analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.

Die nachfolgenden nordischen Sonderzeichen wurden zum analytischen Zwecken wie folgt umbenannt.

Ä ersetzt durch AE Ö ersetzt durch OE

Å ersetzt durch A

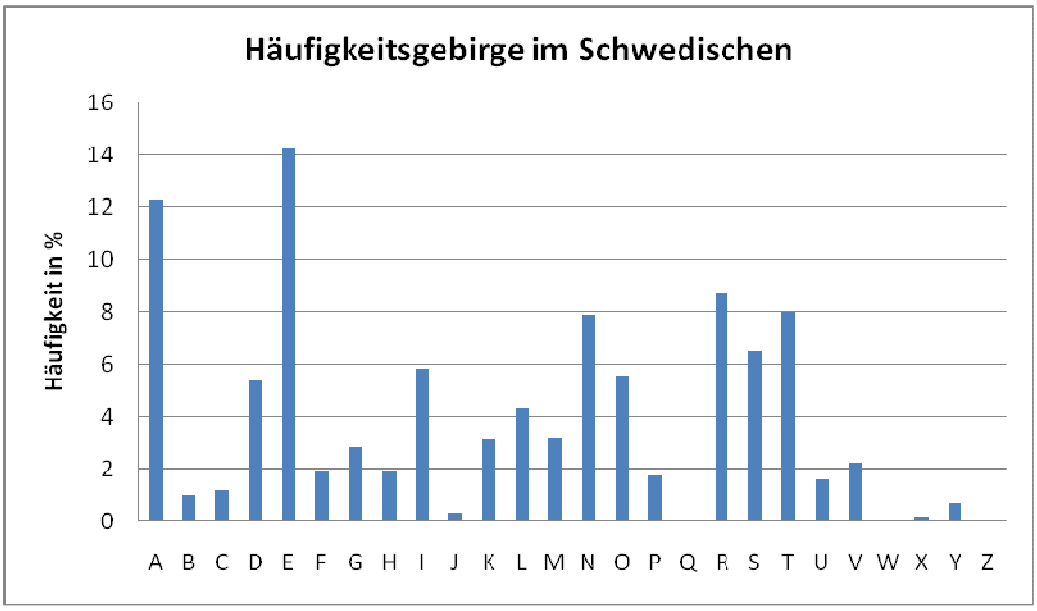


Diagramm 6.3 Die Buchstaben-Häufigkeitsverteilung eines schwedischen Textes

Aus der Häufigkeitsanalyse lässt sich erkennen, dass der Buchstabe A und E mit 14,17 % und 12,23 % sehr oft im Schwedischen vorkommen. Wenig Verwendung finden die Buchstaben B, C, J, Q, W, X, und Z in der schwedischen Sprache.

Die statistischen Größen der schwedischen Sprache lauten:

Entropie: 4.02 (*maxmögliche :* 4.70)

KI =0.075184452602578 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**ER**	4,07	716
2	**DE**	3,15	554
3	**EN**	2,51	442
4	**RA**	1,80	317
5	**AN**	1,71	300
6	**NA**	1,70	298
7	**AE**	1,62	284
8	**AR**	1,43	252
9	**TE**	1,39	244
10	**OE**	1,33	233
11	**AT**	1,31	230
12	**ET**	1,29	227
13	**TA**	1,19	210
14	**ST**	1,17	206
15	**AD**	1,17	206
16	**TT**	1,13	198
17	**RE**	1,11	196
18	**TI**	1,09	191
19	**ES**	1,07	188
20	**LL**	1,06	186
21	**ND**	1,06	186
22	**OM**	1,02	180
23	**IN**	0,99	174
24	**ED**	0,96	168
25	**RI**	0,92	162
26	**AS**	0,90	158
27	**HA**	0,88	155
28	**RS**	0,87	153
29	**KA**	0,86	152
30	**SI**	0,86	152

Table 9.7 Bigramme der schwedischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**ADE**	0,99	174
2	**OCH**	0,73	128
3	**OER**	0,71	125
4	**ERS**	0,70	123
5	**RNA**	0,69	121
6	**FOE**	0,65	115
7	**DER**	0,63	111
8	**TER**	0,62	109
9	**ERA**	0,60	106
10	**ERN**	0,59	103
11	**NDE**	0,57	101
12	**ATT**	0,56	98
13	**SOM**	0,54	95
14	**DES**	0,52	91
15	**PER**	0,48	85
16	**SKA**	0,47	83
17	**ISK**	0,46	81
18	**DET**	0,46	80
19	**ILL**	0,45	79
20	**TIL**	0,44	77
21	**HAN**	0,39	69
22	**AND**	0,39	68
23	**DEN**	0,38	67
24	**EDE**	0,36	63
25	**ARE**	0,34	60
26	**ETT**	0,34	59
27	**ENA**	0,33	58
28	**AER**	0,33	58
29	**RAD**	0,33	58
30	**VAR**	0,32	57

Table 9.8 Trigramme der schwedischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**FOER**	0,57	100
2	**ERNA**	0,56	98
3	**TILL**	0,44	78
4	**ISKA**	0,36	63
5	**PERS**	0,34	60
6	**RADE**	0,30	53
7	**KUNG**	0,26	45
8	**ADES**	0,26	45
9	**FTER**	0,23	40
10	**NDER**	0,23	40

Table 9.9 4-Gramme der schwedischen Sprache

4.1.7 Vergleich der germanischen Sprachen

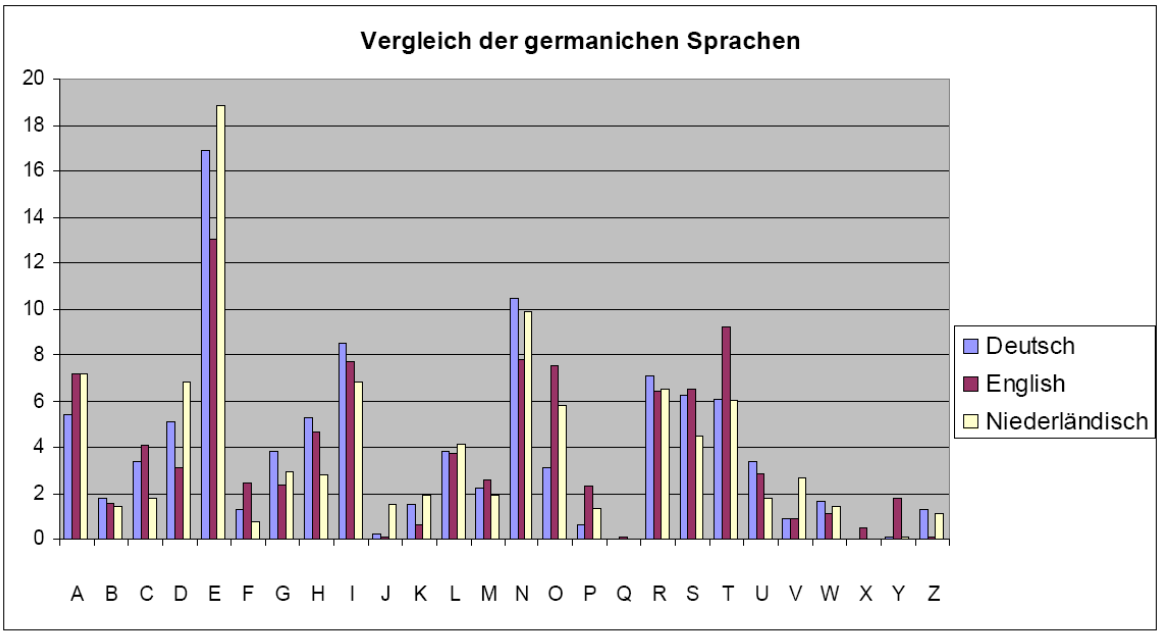


Diagramm 7 Vergleich der germanischen Sprachen nach Buchstabenhäufigkeiten

Zum Vergleichszwecken wurde gemeinsame Abbildung der Buchstabenhäufigkeiten von 3 ausgewählten Sprachen erstellt. Anhand deren lässt sich feststellen, dass die Häufigkeiten einzelner Buchstaben sich in ähnlichem Zahlenbereich verteilen. Es gibt ein paar Einzelfälle, wo diese Unterschiede relativ groβ sind (bsp. d,e, t).

4.2 Finnugrische Sprachen

Während die meisten in Europa gesprochenen Sprachen der indogermanischen Sprachfamilie angehören, gehören zu den finno-ugrischen Sprachen neben dem Finnischen nur die estnische, samische und die ungarische Sprache sowie eine Reihe von im europäischen Russland und in Nordsibirien gesprochenen Sprachen. Die finno-ugrischen Sprachen bilden zusammen mit dem samojedischen Zweig die uralische Sprachfamilie. Ähnlich wie bei den indogermanischen Sprachen sind die heutigen finno-ugrischen Sprachen das Ergebnis mehrerer Sprachspaltungen und lassen sich auf eine hypothetische Ursprache zurückführen, das Urfinnougrische (vgl. Quelle 3)

4.2.1 Finnisch

Das finnische Alphabet besteht aus 29 Buchstaben:

a ,b , c ,d ,e , f , g ,h ,i , j , k , l ,m,n ,o , p ,q , r , s ,t ,u , v ,w , x , y , z ,å , ä ,ö

Es wurde der Text von 8767 Zeichenumfang analysiert. Auf den ersten Blick lassen sich grosse Unterschiede von Buchstabenverteilungen bemerken. Entweder kommt Buchstabe oft vor, oder relativ selten.

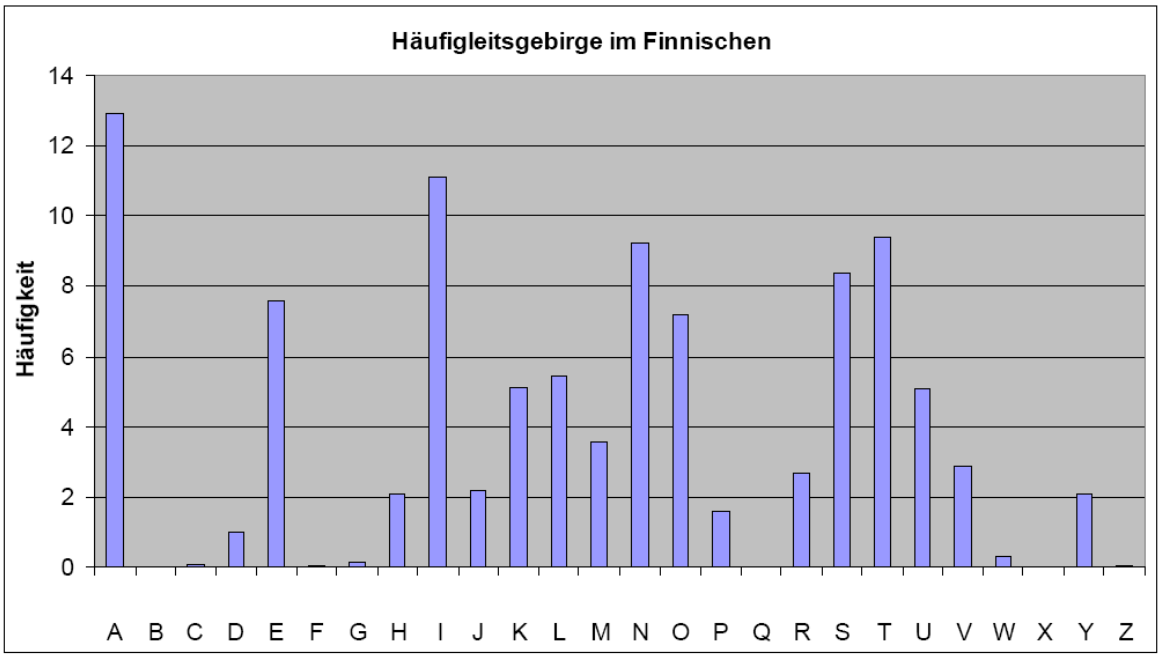


Diagramm 8 Die Buchstaben-Häufigkeitsverteilung eines finnischen Textes

Zum Vergleich mit schon oben analysierten Sprachen lassen sich ein seltenes Auftreten solcher Buchstaben wie „B“, „C“, „F“, „G“ erkennen. Die 5 häufigsten Buchstaben decken 50 % der vorkommenden Buchstaben ab, die häufigsten 10 dann 81 %.

Die statistischen Größen der finnischen Sprache lauten:

Entropie: 3.92 (*maxmögliche:* 4.70)
KI = 0.0818741527777994 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	TA	2.7621	205
2	EN	2.6543	197
3	IN	2.1423	159
4	ST	1.9671	146
5	AN	1.9132	142
6	IS	1.765	131
7	SE	1.765	131
8	SI	1.5764	117
9	SA	1.4956	111
10	TT	1.4956	111
11	ON	1.4282	106
12	AA	1.3608	101
13	MA	1.3473	100
14	IT	1.3339	99
15	VA	1.3339	99
16	LI	1.3204	98
17	TI	1.3069	97
18	LL	1.2396	92
19	OI	1.2396	92
20	AI	1.1857	88
21	JA	1.1857	88
22	KA	1.1722	87
23	MI	1.1722	87
24	II	1.1318	84
25	TO	1.0914	81
26	US	1.0914	81

Table 10 Bigramme der finnischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	STA	1.2358	77
2	SEN	0.7543	47
3	IST	0.7382	46
4	OMA	0.6901	43
5	SAN	0.658	41
6	LLA	0.642	40
7	ANO	0.6259	39
8	TTA	0.5778	36
9	DEN	0.5617	35
10	IKE	0.5617	35
11	NOM	0.5617	35
12	SSA	0.5617	35
13	VUO	0.5136	32
14	TOI	0.4975	31
15	TTI	0.4975	31
16	AAN	0.4815	30
17	ISE	0.4815	30
18	ITT	0.4815	30
19	ENT	0.4654	29
20	EUR	0.4494	28
21	LII	0.4494	28
22	OIT	0.4494	28
23	EEN	0.4173	26
24	IIN	0.4173	26
25	IIK	0.4012	25
26	IMI	0.4012	25

Table 11 Trigramme der finnischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	SANO	0.6974	36
2	ANOM	0.6587	34
3	ISTA	0.6199	32
4	NOMA	0.6199	32
5	EURO	0.4649	24
6	IIKE	0.4649	24
7	LIIK	0.4649	24
8	WSOY	0.4456	23

Table 12 4-Gramme der finnischen Sprache

4.2.2 Estnisch:

Die estnische Sprache zählt zu den Uralsprachen, mit der weiteren Untergliederung Finno-ugrische Sprache, Ostseefinnische Sprache und Estnisch.
Der Ursprung der Sprache liegt geht auf das 16. Jh. zurück, aus der die ersten schriftlichen Zeugnisse entstanden. Bei diesen Schriftstücken ist der deutliche Einfluss der Niederdeutschen Sprache zu erkennen. Ab dem 16. Jh. bewegte sich die Sprache eher in die Richtung des Hochdeutschen.
Als Folge des 2. Weltkrieges fiel Estland an Russland (Sowjetunion) und wurde eine sowjetische Teilrepublik. Damit gab es einen rückläufigen Trend, welcher die Estnische Kultur zurück-drängte. Diese Entwicklung kehrte sich nach der Eigenständigkeit Estlands nach dem Zusammenbruch der Sowjetunion um.

Das estnische Alphabet verwendet die folgenden Buchstaben:

A, B, C, [...] S, Š, Z, Ž, T, U, V, W, Õ, Ä, Ö, Ü, X, Y

Die Besonderheit dieser Sprache ist, dass einige Buchstaben so wie C, F, Š, Z, Ž, Q, W, X und Y nur selten vorkommen, und wenn dann nur in Fremdwörtern und Namensgebungen.
Die Vokale A, E, I, O, U, Ü, Ä, Ö und Õ können alle in der ersten Silbe des Wortes vorkommen, in der zweiten sind aber nur noch die Vokale A, E, I und U möglich. Wörter, die mit den Buchstaben G, B oder D beginnen, sind Fremdwörter. Auffällig für die estnische Sprache ist, dass sie die erste Silbe betonen.

Bei der folgenden Erhebung würde eine Text von 14.176 Zeichen analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.

Die nachfolgenden Sonderzeichen wurden zum analytischen Zwecken wie folgt umbenannt.

Š š	ersetzt durch S oder s	Ž ž	ersetzt durch Z oder z
Õ õ	ersetzt durch o	Ö ö	ersetzt durch oe
Ü ü	ersetzt durch ue	Ä ä	ersetzt durch ae

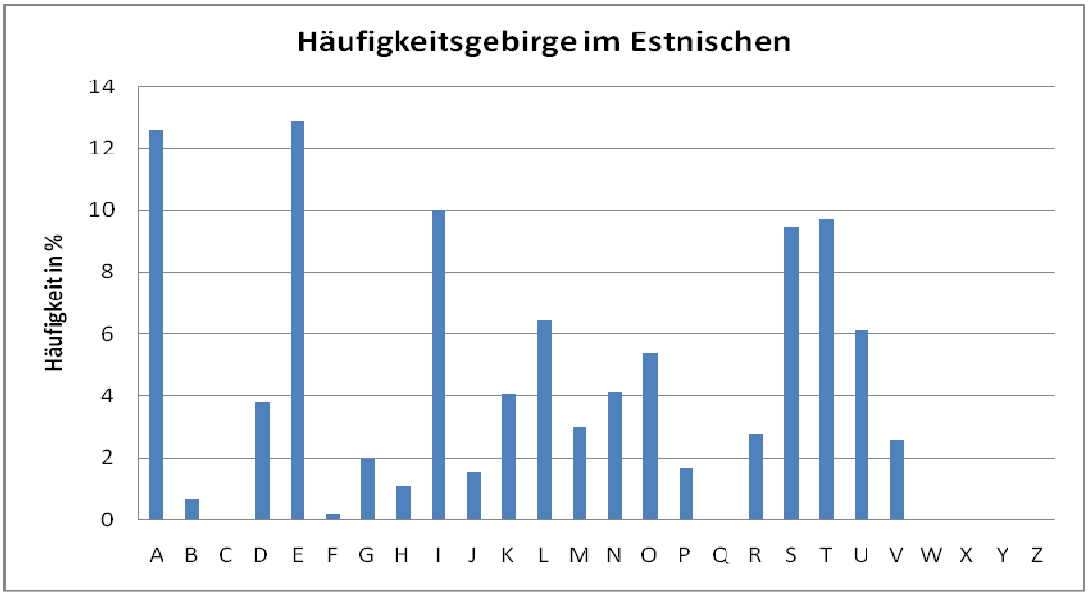


Diagramm 8.1 Die Buchstaben-Häufigkeitsverteilung eines estnischen Textes

Aus der Häufigkeitsanalyse lässt sich erkennen, dass wie auch bei der finnischen Sprache die Buchstaben A-E-I-S-T markant häufig auftreten. Das A und das E treten hierbei mit den Prozentwerten von 12,57 % und 12,88 % am häufigsten auf.
Zu beachten ist, dass das L mit 6,46 % überdurchschnittlich häufig auftaucht, hingegen wie schon vorher beschrieben die Buchstaben C, W, X, Y, Z so gut wie gar nicht auftreten.

Die statistischen Größen der estnischen Sprache lauten:

Entropie: 3.89 (*maxmögliche:* 4.70)
KI = 0.079708072531961 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**SE**	2,77	392
2	**TA**	2,41	341
3	**LE**	2,16	306
4	**ST**	2,09	296
5	**TE**	2,02	287
6	**IS**	2,01	285
7	**EL**	1,66	235
8	**IT**	1,57	223
9	**AS**	1,55	219
10	**VA**	1,50	213
11	**SI**	1,48	210
12	**US**	1,46	207
13	**AT**	1,42	201
14	**TS**	1,41	200
15	**ES**	1,35	192
16	**AL**	1,33	188
17	**AE**	1,31	186
18	**MI**	1,30	184
19	**TU**	1,26	179
20	**JA**	1,22	173
21	**IN**	1,22	173
22	**LI**	1,04	147
23	**EE**	1,02	145
24	**MA**	0,95	135
25	**AK**	0,92	131
26	**EN**	0,92	131
27	**KO**	0,92	130
28	**EK**	0,91	129
29	**OL**	0,90	127
30	**UD**	0,87	124

Table 12.1 Bigramme der estnischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**MIS**	0,82	116
2	**STA**	0,73	104
3	**ISE**	0,73	104
4	**NTS**	0,72	102
5	**ING**	0,70	99
6	**TSE**	0,65	92
7	**TSI**	0,64	90
8	**SEL**	0,61	87
9	**LEP**	0,59	84
10	**AVA**	0,59	83
11	**ITA**	0,59	83
12	**AST**	0,58	82
13	**TUD**	0,57	81
14	**USE**	0,55	78
15	**PIN**	0,50	71
16	**STE**	0,49	70
17	**ELE**	0,49	69
18	**ENT**	0,47	67
19	**EPI**	0,47	67
20	**SEN**	0,47	66
21	**ATE**	0,45	64
22	**NGU**	0,44	63
23	**IST**	0,44	63
24	**EVA**	0,44	63
25	**AMI**	0,44	63
26	**LIT**	0,44	62
27	**ITS**	0,44	62
28	**ASU**	0,44	62
29	**AJA**	0,42	60
30	**ESI**	0,42	59

Table 12.2 Trigramme der estnischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**NTSI**	0,52	74
2	**MISE**	0,51	72
3	**LEPI**	0,49	69
4	**PING**	0,49	69
5	**EPIN**	0,49	69
6	**ASTA**	0,47	66
7	**ENTS**	0,45	64
8	**INGU**	0,44	62
9	**LITS**	0,43	61
10	**ITSE**	0,43	61

Table 12.3 4- Gramme der estnischen Sprache

4.2.3 Ungarisch:

Die ungarische Sprache zählt zu den Uralsprachen, mit der weiteren Untergliederung Finno-ugrische Sprache, Ugrische Sprache und Ungarisch.
Erste schriftliche Überlieferungen des Ungarischen stammen aus dem 9. Jahrhundert, als sich die Magyaren noch der ungarischen Runenschrift bedienten. Die Ungarn bilden zusammen mit den Chanten (früher auch Ostjaken) und den Mansen[2] (früher auch Wogulen), die ugrische Untergruppe. Mit der Zeit der Christianisierung kamen neue lateinische Wörter in den Sprachgebrauch der Ungaren, welche entlehnt worden sind. Erste Zeugnisse der ungarischen Schrift tauchen im Jahre 1055 mit einem Schriftstück auf, welches zwar in Latein geschrieben war, aber vermehrt ungarische Wortverbindungen aufwies. Nach der Zeit bis 1918 stand Ungarn stark unter deutschem Einfluss, parallel zu dieser Zeit wurde eine starke Magyariesierung[3] verfolgt, was den Unmut der nichtmagyarischen Bevölkerung schürte und zum Zerfall des Königreiches Ungarn führte.
Nach dem 1. Weltkrieg verteilte sich die ungarische Bevölkerung in den Grenzgebieten und Nachbarstaaten. Diese Tendenz war bis 1956 rückläufig. Danach wanderten viele Ungaren nach Amerika und Australien aus.

Das ungarische Alphabet verwendet die folgenden Buchstaben:

A Á B C Cs D Dz Dzs E É F G Gy H I Í J K L Ly M
N Ny O Ó Ö Ő P(Q) R S Sz T Ty U Ú Ü Ű V(W) (X) (Y) Z Zs

Im Ungarischen zählen auch die Buchstaben Ö, Ő, Ü und Ű sowie die Digraphen wie cs, dz, gy, ly, ny, sz, ty, zs und der Trigraph dzs als eigener Buchstabe. Man spricht hier vom großen und kleinen ungarischen Alphabet, je nachdem, ob die nur in Fremdwörtern und historischen Schreibweisen (von z. B. Familiennamen) vorkommenden vier Buchstaben Q, W, X, Y hinzugezählt werden oder nicht. Im ersteren Fall hat das ungarische Alphabet somit 44, im zweiteren 40 Buchstaben.

Bei der folgenden Erhebung würde eine Text von 12.520 Zeichen analysiert.

[2] zwei östlich des Ural lebenden Völkern
[3] Überfremdung mit ungarisch sprechenden Bevölkerung

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.

Die nachfolgenden Sonderzeichen wurden zum analytischen Zwecken wie folgt umbenannt.

Á	ersetzt durch a	É	ersetzt durch e	Õ	ersetzt durch o
Ó	ersetzt durch o	Ö	ersetzt durch oe	Ú	ersetzt durch u
Ü	ersetzt durch ue	Û	ersetzt durch u	Í	ersetzt durch i

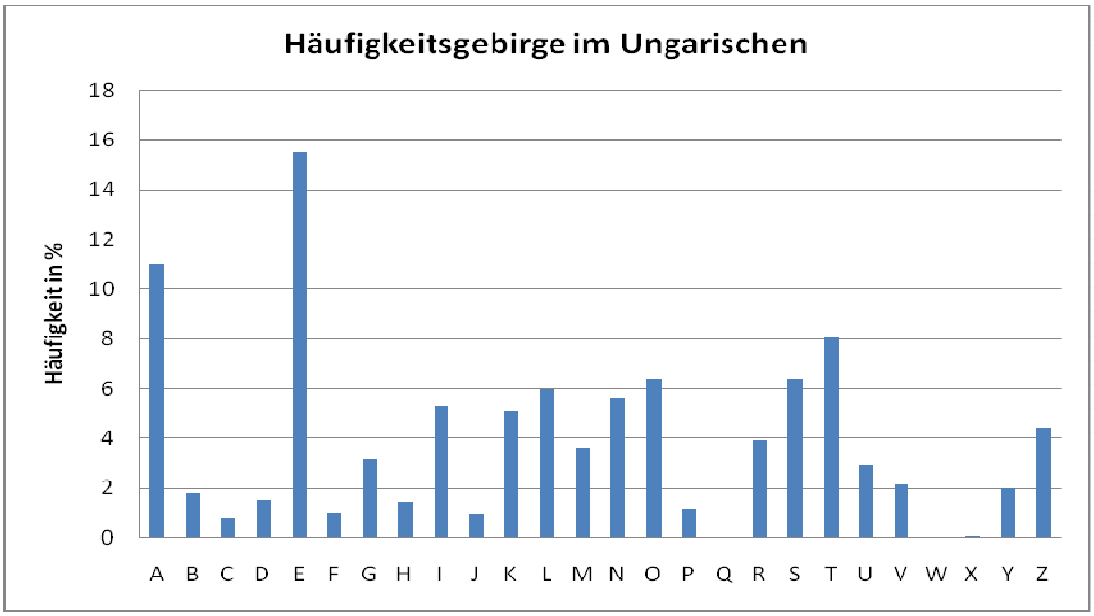


Diagramm 8.2 Die Buchstaben-Häufigkeitsverteilung eines ungarischen Textes

Aus der Häufigkeitsanalyse lässt sich erkennen, dass die Buchstaben A und E im Ungarischen sehr oft vorkommen im Vergleich zu dem Rest (26,49 %).
Andere Buchstaben treten eher weniger auf oder werden aufgrund der schon erwähnten Verkürzung des Alphabets nicht genutzt.

Die statistischen Größen der ungarischen Sprache lauten:

Entropie: 4.12 (*maxmögliche :* 4.70)
KI = 0.071344772977445 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**EN**	2,07	259
2	**EL**	2,05	256
3	**TE**	2,01	251
4	**SZ**	1,97	246
5	**ES**	1,93	241
6	**ET**	1,70	213
7	**TA**	1,49	186
8	**ER**	1,41	176
9	**AT**	1,37	172
10	**AL**	1,34	168
11	**KE**	1,31	164
12	**LE**	1,28	160
13	**AN**	1,27	159
14	**EG**	1,23	154
15	**ZE**	1,21	151
16	**LA**	1,20	150
17	**EK**	1,10	138
18	**OE**	1,03	129
19	**SE**	1,01	126
20	**ME**	0,96	120
21	**AK**	0,95	119
22	**GY**	0,93	116
23	**AS**	0,93	116
24	**VE**	0,92	115
25	**NE**	0,90	113
26	**OL**	0,88	110
27	**TO**	0,86	108
28	**KO**	0,85	106
29	**RA**	0,84	105
30	**TT**	0,78	97

Table 12.4 Bigramme der ungarischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**SZE**	0,70	87
2	**ELE**	0,56	70
3	**ESZ**	0,52	65
4	**ETE**	0,45	56
5	**SEG**	0,42	53
6	**KEP**	0,40	50
7	**ERE**	0,39	49
8	**ETT**	0,38	48
9	**ENT**	0,38	47
10	**LEN**	0,36	45
11	**TER**	0,36	45
12	**ZET**	0,35	44
13	**ESE**	0,34	43
14	**ATA**	0,33	41
15	**MEG**	0,32	40
16	**UEL**	0,32	40
17	**HAT**	0,32	40
18	**UVE**	0,31	39
19	**EGY**	0,31	39
20	**MUV**	0,30	38
21	**ASZ**	0,30	38
22	**ALA**	0,30	38
23	**TES**	0,30	38
24	**ENY**	0,30	37
25	**SSZ**	0,30	37
26	**EVE**	0,30	37
27	**KOE**	0,30	37
28	**VES**	0,28	35
29	**ZER**	0,28	35
30	**LET**	0,27	34

Table 12.5 Trigramme der ungarischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**MUVE**	0,30	38
2	**ESZE**	0,24	30
3	**ELEN**	0,23	29
4	**SZER**	0,23	29
5	**SZET**	0,21	26
6	**VESZ**	0,21	26
7	**UVES**	0,19	24
8	**ENYE**	0,19	24
9	**ESSZ**	0,17	21
10	**ILAG**	0,17	21

Table 12.6 4- Gramme der ungarischen Sprache

4.3 Romanische Sprachen

"Ursprache" die romanischen Sprachen im Gegensatz zu den meisten anderen Sprachgruppen ist gut bezeugt: es handelt sich nämlich hier um das gesprochene Latein (Volkslatein oder Vulgärlatein). Das Lateinische selbst gilt nicht als romanische Sprache, sondern wird zu den italischen Sprachen gezählt.

4.3.1 Spanisch

Das spanische Alphabet besteht aus 27 Buchstaben:

a b c ch d e f g h i j k l m n ñ o p q s t u v w x y z

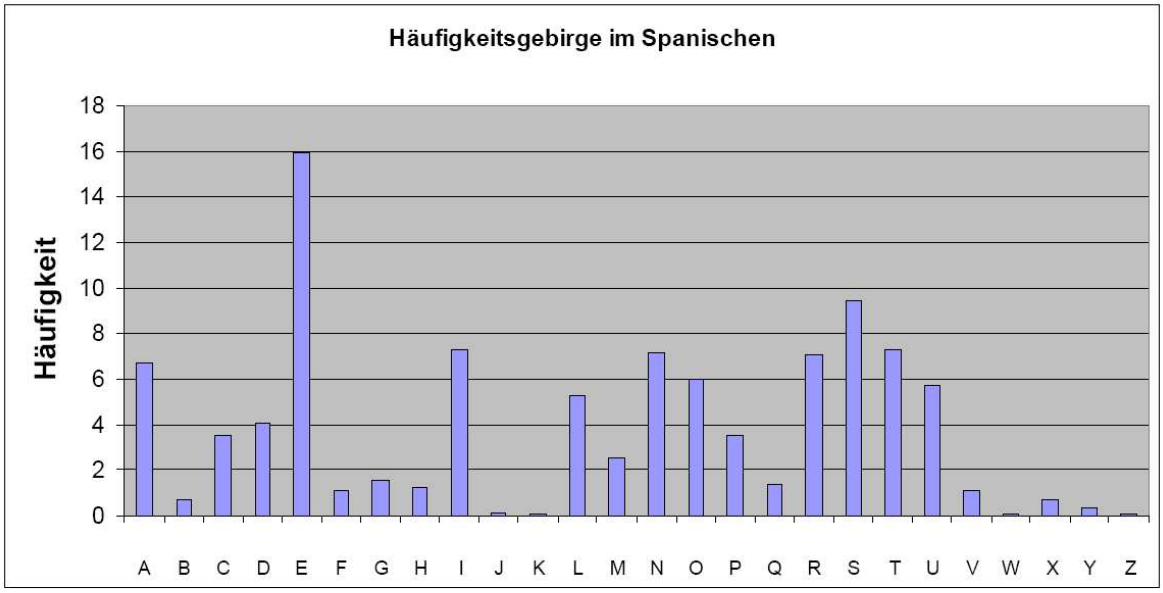


Diagramm 9 Die Buchstaben-Häufigkeitsverteilung eines spanischen Textes

Der analysierte Text hat Umfang von 14589 Zeichen. Am häufigsten auftretender Buchstabe ist e, der von anderen Buchstaben überwiegend auffällt. Die Buchstaben von der geringsten Auftretensrate sind „B“, „J“, „K“, „W“, „Z“. Die 5 häufigsten Buchstaben decken 47% der vorkommenden Buchstaben ab, die häufigsten 10 dann 77 %. Die zehn häufigsten Worte im Spanischen sind: de, la, el, que, en, no, con, un, se, sa (vgl.Quelle 1, Seite 221)
Die statistischen Grössen der finnischen Sprache lauten:
Entropie: 4.03 (*maxmögliche :*4.70)
KI = 0.0748292446185148 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	ES	4.5135	507
2	EN	2.9556	332
3	DE	2.9467	331
4	NT	2.6974	303
5	ON	2.5728	289
6	LE	2.3502	264
7	RE	2.0921	235
8	QU	1.7538	197
9	TE	1.5668	176
10	NS	1.4956	168
11	TI	1.46	164
12	UE	1.4244	160
13	ER	1.4066	158
14	UR	1.3799	155
15	ME	1.3532	152
16	OU	1.3532	152
17	RA	1.3443	151
18	LA	1.3264	149
19	IE	1.273	143
20	IS	1.2196	137
21	AN	1.184	133
22	CE	1.184	133
23	CO	1.1128	125
24	AI	1.0772	121
25	IT	1.0683	120
26	IO	1.0149	114

Table 13 Bigramme der spanischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	TION	1.4227	87
2	MENT	0.9975	61
3	IQUE	0.8504	52
4	ATIO	0.785	48
5	OGRA	0.785	48
6	GRAP	0.7522	46
7	RAPH	0.7522	46
8	IONS	0.6705	41
9	DANS	0.6378	39
10	ENCE	0.6214	38
11	QUES	0.6051	37
12	EMEN	0.5887	36

Table 14 Trigramme der spanischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	TION	1.4227	87
2	MENT	0.9975	61
3	IQUE	0.8504	52
4	ATIO	0.785	48
5	OGRA	0.785	48
6	GRAP	0.7522	46
7	RAPH	0.7522	46
8	IONS	0.6705	41
9	DANS	0.6378	39
10	ENCE	0.6214	38
11	QUES	0.6051	37
12	EMEN	0.5887	36

Table 15 4-Gramme der spanischen Sprache

4.3.2 Italienisch

Das italienische Alphabet besteht aus 21 Buchstaben:

a b c d e f g h i l m n o p q r s t u v z

Die Buchstaben wie j, k, w, x und y treten nur in Fremdwörter auf, was sich auch mittels Häufigkeitsgebirge ablesen lässt. Die 5 häufigsten Buchstaben decken 50 % der vorkommenden Buchstaben ab, die häufigsten 10 dann 80%. Die zehn häufigsten Worte im Italienischen sind: la, di, che, il, non, si, le, una, lo, in (vgl.Quelle 1, Seite 221)

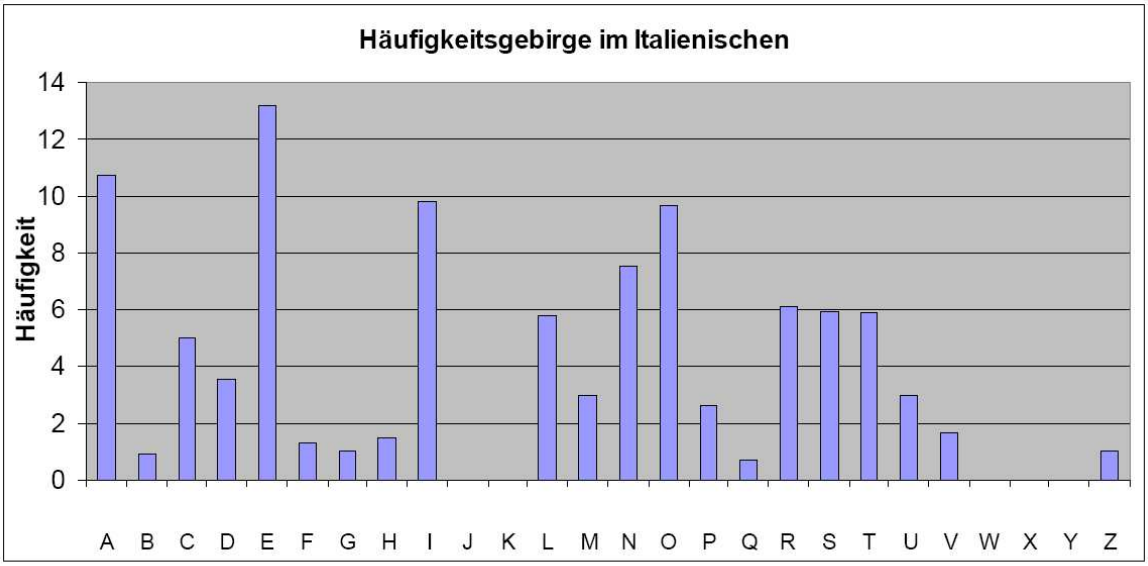


Diagramm 10 Die Buchstaben-Häufigkeitsverteilung eines italienischen Textes

Die statistischen Größen der italienischen Sprache lauten:

Entropie: 3.98 (*maxmögliche:* 4.70)
KI = 0.0748816394378137 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	ON	2.3886	622
2	RE	2.2657	590
3	ER	1.9547	509
4	LA	1.8011	469
5	CO	1.6628	433
6	NO	1.6551	431
7	CH	1.6398	427
8	ES	1.6244	423
9	AN	1.6206	422
10	EN	1.6052	418
11	DI	1.586	413
12	TE	1.563	407
13	TO	1.5246	397
14	DE	1.49	388
15	TA	1.4785	385
16	ST	1.4324	373
17	EL	1.3518	352
18	AL	1.3479	351
19	NT	1.3441	350
20	SI	1.3364	348
21	RA	1.3095	341
22	IO	1.2865	335
23	LL	1.2135	316
24	IN	1.1943	311
25	NE	1.179	307
26	HE	1.1598	302

Table 17 Bigramme der italienischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	CHE	1.5125	302
2	ELL	1.1319	226
3	ENT	1.0467	209
4	DEL	0.9466	189
5	LLA	0.9466	189
6	CON	0.8264	165
7	ION	0.8264	165
8	NTE	0.8063	161
9	MEN	0.7813	156
10	PER	0.7713	154
11	EST	0.7462	149
12	QUE	0.7212	144
13	ERE	0.6411	128
14	ESS	0.636	127
15	UES	0.636	127
16	ANC	0.586	117
17	ONE	0.586	117
18	ARE	0.5709	114
19	ZIO	0.5609	112
20	ONO	0.5459	109
21	ITA	0.5409	108
22	STO	0.5409	108
23	NCH	0.4908	98
24	PAR	0.4908	98
25	NON	0.4858	97
26	CHI	0.4758	95

Table 16 Trigramme der italienischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	ELLA	0.9717	146
2	MENT	0.8453	127
3	QUES	0.8453	127
4	UEST	0.8453	127
5	DELL	0.8319	125
6	ZION	0.7121	107
7	ENTE	0.6922	104
8	IONE	0.6789	102

Table 18 4-Gramme der italienischen Sprache

4.3.3 Französisch

Das französische Alphabet besteht aus 42 Buchstaben:

a à â ä b c ch d e è é ê ë f g hi î ï j k l m n ń o ô oe p q s t u ù û ü v w x y ÿ z

Nach der Häufigkeitsanalyse ist es zu erkennen, dass am häufigsten auftretetenden Buchstaben „E“ ist. Buchstaben wie „K“ und „W“ kommen sehr selten in dieser Sprache vor.

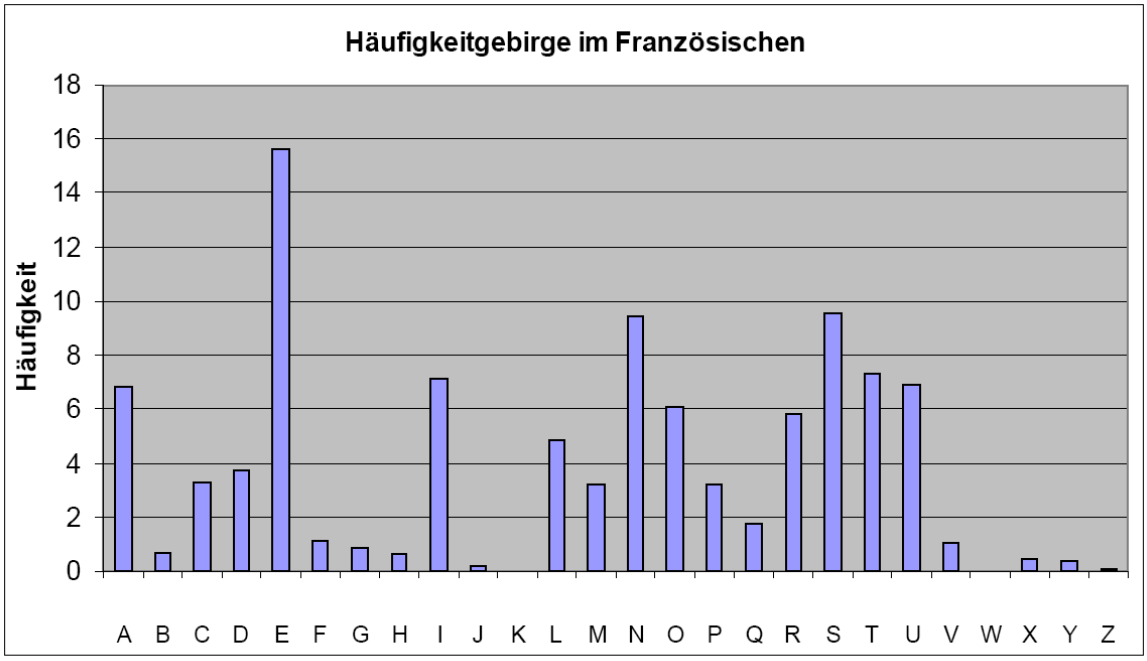


Diagramm 11 Die Buchstaben-Häufigkeitsverteilung eines französischen Textes

Die zehn häufigsten Worte im Französischen sind: de, il, le, et, que, je, la, ne, on, les.

Die statistischen Grössen der französischen Sprache lauten:

Entropie: 3.97 (*maxmögliche:* 4.70)
K I= 0.0768714949917831 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	EN	3.4734	386
2	ES	3.0595	340
3	NT	2.8435	316
4	ON	2.6186	291
5	NS	2.2766	253
6	QU	2.2766	253
7	DE	2.1416	238
8	LE	2.0696	230
9	RE	2.0067	223
10	OU	1.9617	218
11	ME	1.9167	213
12	IN	1.6017	178
13	AN	1.5387	171
14	TE	1.5297	170
15	SE	1.5207	169
16	US	1.5207	169
17	UE	1.3948	155
18	SI	1.3318	148
19	AI	1.2598	140
20	UN	1.2508	139
21	CE	1.2238	136
22	IT	1.1968	133
23	TI	1.1698	130

Table 19 Bigramme der französischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	ENT	1.9734	159
2	QUE	1.5639	126
3	ENS	1.1543	93
4	ION	1.0426	84
5	OUS	1.0426	84
6	TEN	1.0302	83
7	EME	1.0053	81
8	ONS	1.0053	81
9	NTE	0.9929	80
10	LES	0.9433	76
11	NOU	0.9309	75
12	INT	0.8936	72
13	UNE	0.8688	70
14	MEN	0.8564	69
15	NSI	0.8316	67
16	TIO	0.8316	67
17	TRE	0.8192	66
18	EST	0.7571	61
19	LUS	0.7571	61
20	PLU	0.7571	61
21	SEN	0.7571	61
22	CON	0.6578	53
23	PAR	0.6578	53

Table 20 Trigramme der französischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	NOUS	1.2104	70
2	TION	1.1586	67
3	MENT	1.1067	64
4	PLUS	1.0375	60
5	NTEN	1.0029	58
6	INTE	0.9684	56
7	TENS	0.8819	51
8	ELLE	0.8127	47
9	ENSI	0.7781	45

Table 21 4-Gramme der französischen Sprache

4.3.4. Portugiesisch:

Die portugiesische Sprache zählt zu den Indogermanischen Sprachen, mit der weiteren Untergliederung Italienische Sprache, Romanische Sprache, Iberomanische und Portugiesisch. Das Portugiesisch hat seinen Ursprung im Westen der iberischen Halbinsel, dort entwickelte sich ein sogenanntes Vulgärlatein, welches sich durch die dort stationierten römische Soldaten und Siedler in diese Region verbreitete. Die Verbreitung des Vulgärlateins ging einher mit der römischen Kolonisierung, durch welche sich alle romanischen Sprachen entwickelten. Dies spiegelt sich auch darin wieder, dass 90 % des portugiesischen Wortschatzes vom Latein abstammen. Im Jahre 409 brach dann das weströmische Reich durch Germanische Invasionen zusammen, was dazu führte das sich aus der einheitliche Sprache Latein verschiedene Abwandlungen mit unterschiedlichen Dialekten entwickelten. Einen wichtigen Punkt in der Geschichte der portugiesischen Sprache spielt der germanische Stamm der Sueben, die durch ihren Einfluss eine Art Zusammenführung ihrer eigenen Sprache mit der Sprache der dort ansässigen Bewohner herbeigeführt haben. Dadurch grenzten sie diese Sprache auch zu anderen ab. (z.B. Spanisch) Zur Zeit der Maurischen Invasionen (gegen 711) wurde die Entwicklung der Sprache durch gesprochenes Arabisch nur sehr wenig beeinflusst. Um 1100 fand man die ersten Beweise für die geschriebene portugiesische Sprache, welche sich im Norden des heutigen Portugals entwickelte. Nachdem sich das als Vulgärlatein verspottete Portugiesisch gegenüber dem „alten“
Latein durchsetzt, kam die Zeit der Eroberungen. Damit trugen die portugiesischen Seefahrer und Entdecker die Sprache bis in weite Regionen der Erde. (Asien, Afrika und Amerika). Zur Zeit der Renaissance wurden etliche Lehnwörter aus dem Französischen und dem Italienischen in die Portugiesische Sprache aufgenommen. Heute wird Portugiesisch von fast 220 Million Menschen gesprochen.
Das portugiesische Alphabet setzt sich aus 23 Buchstaben aus dem lateinischen Alphabet zusammen, wobei außerhalb von Namen kein Gebrauch von K, W und Y gemacht wird.

Zudem wird in der portugiesischen Sprache folgende Diakritika verwendet:

Á, Â, Ã, À, Ç, É, Ê, Í, Ó, Ô, Õ, Ú, Ü

Bei der folgenden Erhebung würde eine Text von 10.106 Zeichen analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.

Ersetzungsvorschrift:

Á ersetzt durch a	Â ersetzt durch a	Ã ersetzt durch a
À ersetzt durch a	Ç ersetzt durch c	É ersetzt durch e
Ê ersetzt durch e	Í ersetzt durch i	Ó ersetzt durch o
Ô ersetzt durch o	Õ ersetzt durch o	Ú ersetzt durch u
Ü ersetzt durch u		

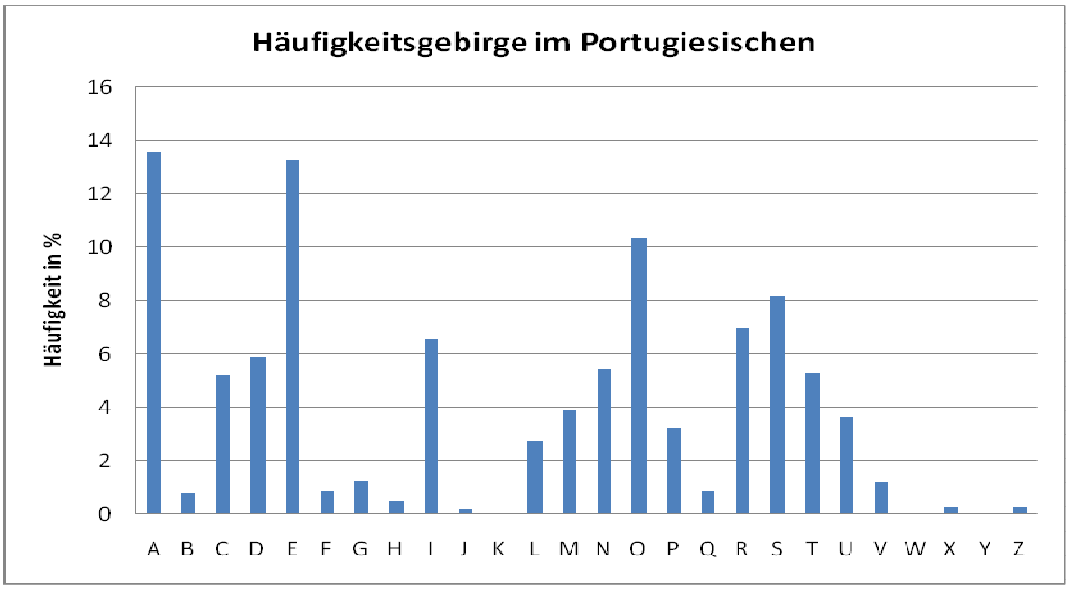


Diagramm 11.1 Die Buchstaben-Häufigkeitsverteilung eines portugiesischen Textes

Aus der Häufigkeitsanalyse lässt sich erkennen, dass das A und E in dieser Sprache eine wichtige Bedeutung haben. Wenn man die Buchstaben A E O S zusammen betrachtet so machen alleine sie schon einen Anteil von 45,24 % in der portugiesischen Sprache aus.
Hingegen einige Konsonanten wie das C und D gegenüber dem F G H K L sehr oft auftauchen, wie aus dem Diagramm ersichtlich wird.

Die statistischen Größen der portugiesischen Sprache lauten:

Entropie: 3.93 (*maxmögliche :* 4.70)
KI = 0.079143072545319 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**DE**	2,47	250
2	**OS**	2,17	219
3	**ES**	2,15	217
4	**DO**	1,84	186
5	**EN**	1,76	178
6	**RE**	1,74	176
7	**AD**	1,71	173
8	**CO**	1,61	163
9	**RA**	1,59	161
10	**CA**	1,54	156
11	**NT**	1,51	153
12	**TE**	1,51	153
13	**AS**	1,48	150
14	**SE**	1,38	139
15	**OR**	1,35	136
16	**ER**	1,34	135
17	**AN**	1,34	135
18	**AC**	1,28	129
19	**AR**	1,21	122
20	**ST**	1,10	111
21	**OD**	1,06	107
22	**TO**	1,05	106
23	**AL**	1,05	106
24	**AO**	1,02	103
25	**MA**	0,98	99
26	**RI**	0,98	99
27	**ON**	0,97	98
28	**EC**	0,97	98
29	**PO**	0,96	97
30	**DA**	0,96	97

Table 21.1 Bigramme der portugieschen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**ENT**	0,84	85
2	**ADO**	0,83	84
3	**QUE**	0,73	74
4	**NTE**	0,66	67
5	**DOS**	0,62	63
6	**CON**	0,57	58
7	**CAO**	0,55	56
8	**SDE**	0,53	54
9	**POR**	0,52	53
10	**ADE**	0,51	52
11	**ODE**	0,50	51
12	**ACO**	0,46	46
13	**RES**	0,46	46
14	**COM**	0,43	43
15	**EST**	0,43	43
16	**DES**	0,42	42
17	**ACA**	0,42	42
18	**RIA**	0,41	41
19	**MEN**	0,40	40
20	**OSE**	0,39	39
21	**ARA**	0,39	39
22	**CRI**	0,39	39
23	**PAR**	0,38	38
24	**PRE**	0,37	37
25	**ONT**	0,34	34
26	**NTO**	0,34	34
27	**ICA**	0,34	34
28	**STA**	0,33	33
29	**REC**	0,30	30
30	**ESS**	0,29	29

Table 21.2 Trigramme der portugieschen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**ENTE**	0,34	34
2	**ACAO**	0,34	34
3	**MENT**	0,34	34
4	**ADOS**	0,31	31
5	**ACON**	0,28	28
6	**CONT**	0,25	25
7	**ENTO**	0,23	23
8	**ANCA**	0,23	23
9	**CAOD**	0,22	22
10	**RIAN**	0,21	21

Table 21.3 4-Gramme der portugieschen Sprache

4.3.5 Rumänisch:

Die rumänische Sprache zählt zu den Indogermanischen Sprachen, mit der weiteren Untergliederung Italienische Sprache, Romanische Sprache, Balkanromanisch und Rumänisch. Der Ursprung der rumänischen Sprache ist ähnlich dem Portugiesischen.
Dort wurde auch das Vulgärlatein gesprochen und abgelöst nach dem Abzug der römischen Truppen. Die sprachliche Entwicklung wurde später zwischen dem 7. Jh. Und dem 10 Jh. maßgeblich durch den slawisch/bulgarischen Einfluss in diesem Gebiet geprägt. Bis 1860 wurde das Rumänisch in Kyrillisch geschrieben, danach entwickelte die Siebenbürgische Schule ein lateinisches Alphabet mit Sonderzeichen für die einzelnen benötigten Betonungen.
Diese Schreibweise des Rumänischen hat bis heute Bestand.

Das rumänische Alphabet setzt sich aus 31 Buchstaben zusammen, welche schon die 5 Sonderzeichen beinhalten.

Demzufolge setzt sich das rumänische Alphabet aus folgenden Buchstaben zusammen:

a, ă, â, b, c, d, e, f, g, h, i, î, j, k, l, m, n, o, p, q, r, s, ş, t, ţ, u, v, w, x, y, z

Bei der folgende n Erhebung würde eine Text von 12.163 Zeichen analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.

Ersetzungsvorschrift:

ă	wird ersetzt durch a	î	wird ersetzt durch i
â	wird ersetzt durch a	ş	wird ersetzt durch s
ţ	wird ersetzt durch t		

Anstelle unserer bekannten Umlaute werden in der rumänischen Sprache die 5 Sonderzeichen gesetzt, um die jeweilige Betonung deutlich zu machen.

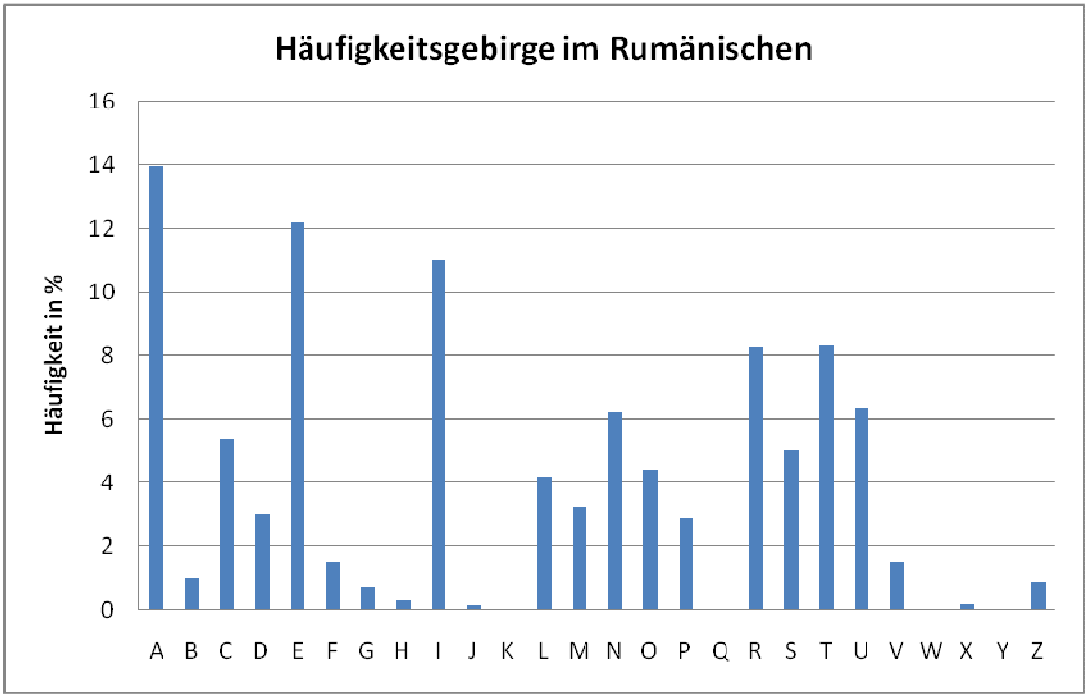


Diagramm 11.2 Die Buchstaben-Häufigkeitsverteilung eines rumänischen Textes

Das Diagramm zeigt die starke Ausprägung der 4 Vokale (A, E, I, und U), zudem werden die Buchstaben R und T oft verwendet. Diese 4 Vokale machen auf die Gesamtverteilung einen Anteil von insgesamt 43,36 % aus. Hingegen eine Menge von Konsonanten kaum eine prägnante Häufigkeit von 4,5 % aufzeigen (Zusammen B, F, G, H, J, K, Q, W, X, Y, Z).
Auffallend erscheint in dem Tri-Gramm der String ARE sehr häufig mit 1,20 %.

Die statistischen Größen der rumänischen Sprache lauten:

Entropie: 3.90 (*maxmögliche :* 4.70)
KI = 0.079940629396485(0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**RE**	2,56	311
2	**IN**	2,43	295
3	**AR**	2,06	251
4	**TI**	1,92	234
5	**AT**	1,75	213
6	**CA**	1,73	211
7	**EA**	1,73	210
8	**RI**	1,61	196
9	**TE**	1,58	192
10	**DE**	1,51	184
11	**ES**	1,51	184
12	**TA**	1,48	180
13	**NT**	1,48	180
14	**ST**	1,44	175
15	**AS**	1,25	152
16	**AC**	1,23	150
17	**TR**	1,22	148
18	**EN**	1,15	140
19	**UL**	1,15	140
20	**RA**	1,13	138
21	**SA**	1,12	136
22	**EC**	1,06	129
23	**ER**	1,00	122
24	**CE**	1,00	122
25	**IT**	0,98	119
26	**AN**	0,97	118
27	**IC**	0,93	113
28	**OR**	0,90	109
29	**UR**	0,90	109
30	**PR**	0,89	108

Table 21.4 Bigramme der rumänischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**ARE**	1,20	146
2	**ENT**	0,73	89
3	**REA**	0,72	87
4	**EST**	0,62	76
5	**CAR**	0,58	71
6	**NTR**	0,56	68
7	**TRU**	0,51	62
8	**ACE**	0,50	61
9	**ATI**	0,48	58
10	**TAT**	0,44	53
11	**STE**	0,42	51
12	**STI**	0,39	47
13	**PEN**	0,38	46
14	**URI**	0,36	44
15	**AIN**	0,35	43
16	**ECA**	0,35	43
17	**RES**	0,35	42
18	**INT**	0,35	42
19	**ITA**	0,34	41
20	**EDE**	0,33	40
21	**TRE**	0,33	40
22	**TIN**	0,32	39
23	**TUR**	0,32	39
24	**PRI**	0,31	38
25	**ULT**	0,31	38
26	**ATE**	0,31	38
27	**ICA**	0,31	38
28	**ATA**	0,30	37
29	**ERE**	0,30	37
30	**RIC**	0,30	37

Table 21.5 Trigramme der rumänischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**CARE**	0,53	65
2	**AREA**	0,41	50
3	**NTRU**	0,37	45
4	**ESTE**	0,37	45
5	**ENTR**	0,36	44
6	**PENT**	0,35	42
7	**ITAT**	0,28	34
8	**ECAR**	0,26	32
9	**TRUC**	0,25	31
10	**ULUI**	0,21	26

Table 21.6 4- Gramme der rumänischen Sprache

4.3.6 Vergleich der romanischen Sprachen

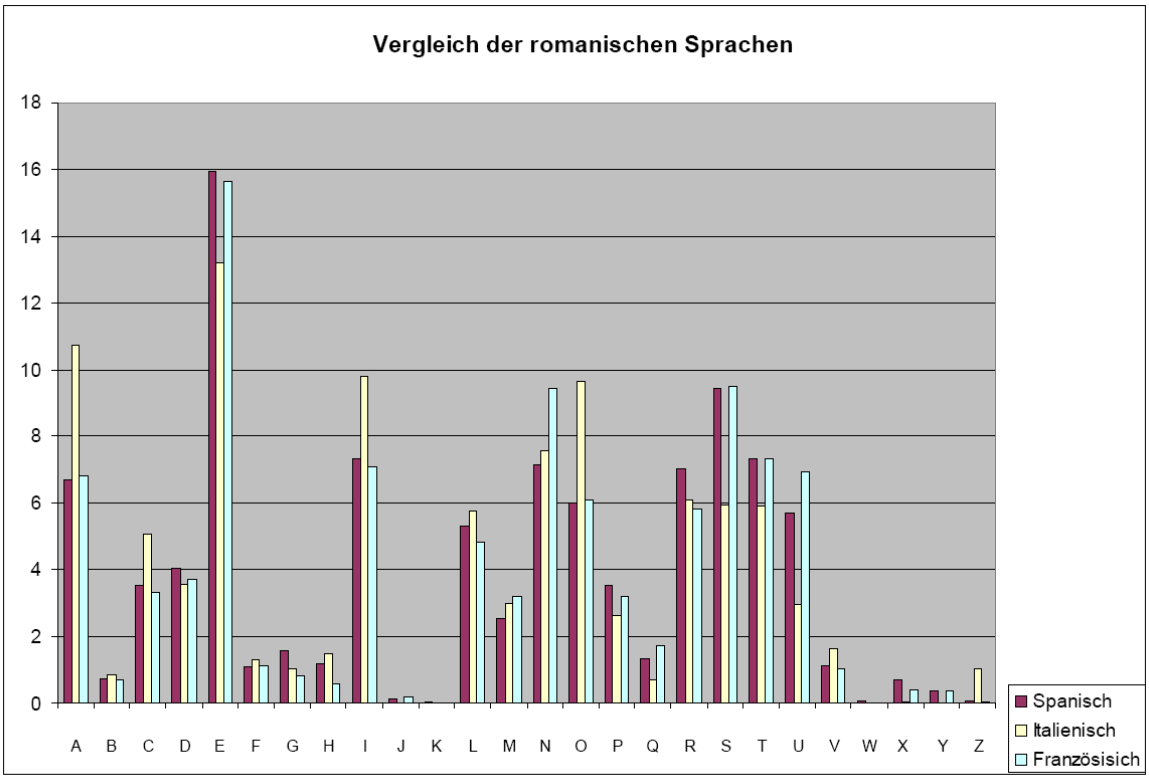


Diagramm 12 Vergleich der romanischen Sprachen nach Buchstabenhäufigkeiten

Trotz der Zugehörigkeit zu einer Sprachgruppe lassen sich die Unterschiede in Häufigkeiten der Einzelbuchstaben erkennen. Besonders stark sieht das mit A, E, K, U-Verteilungen aus. Auffällig zudem ist noch, das die Buchstaben B, J, K, W kaum auftreten, was als ein Merkmal für diese Sprachgruppe zu deuten ist.

4.4 Slawische Sprachen

Zu den slawischen Sprachen gehören die folgenden Gruppen der Sprachen:
westslawische, ostslawische und südslawische Sprachen. Untersucht wurden Polnisch, Tschechisch, Bulgarisch und Slowakisch.

4.4.1 Polnisch

Das polnische Alphabet besteht aus 35 Buchstaben und lautet vollständig:
A, Ą ,B ,C ,Ć , D ,E , Ę , F ,G ,H , I , J , K ,L , Ł ,M , N ,
Ń ,O ,Ó, P ,Q, R, S ,Ś ,T ,U ,V ,W , X ,Y ,Z , Ź , Ż.
Die Häufigkeitsanalyse des polnischen Textes von Umfang 8016 Zeichen macht das häufige Auftreten der Vokalen sichtbar. Es kommt a, e, i, o -Gipfel vor. Die 5 häufigsten Buchstaben decken 41 % der vorkommenden Buchstaben ab, die häufigsten 10 dann 79%.

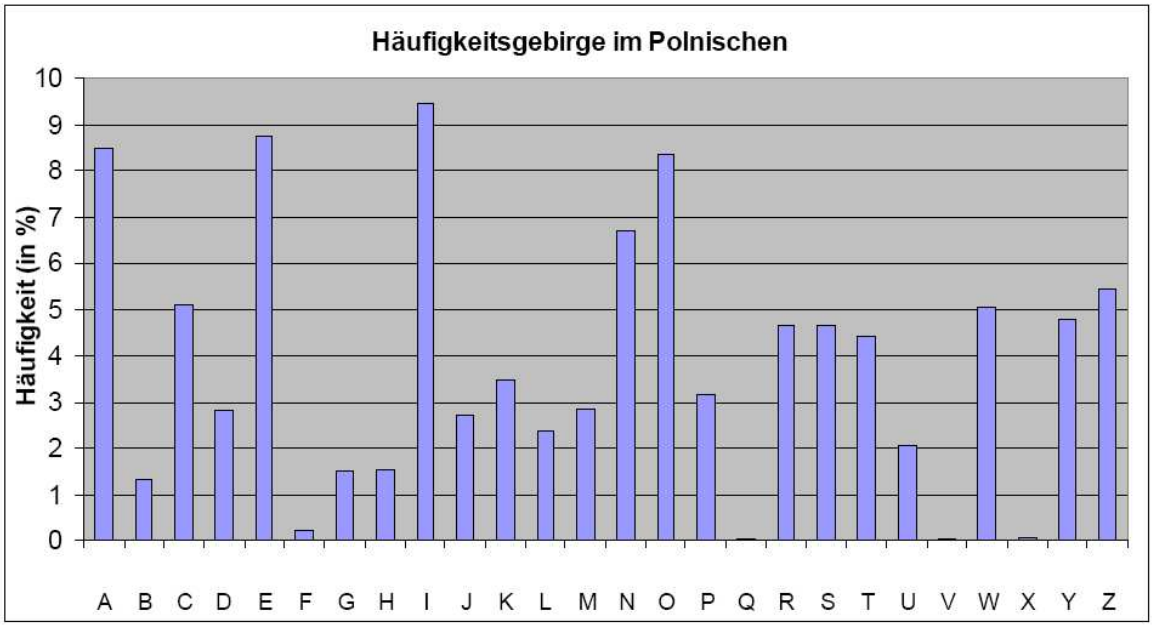


Diagramm 13 Die Buchstaben-Häufigkeitsverteilung eines polnischen Textes

Die statistischen Grössen der polnischen Sprache lauten:
Entropie: 4.28 (*maxmögliche :* 4.70)
KI = 0.0579696502192122 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	IE	4.0188	256
2	NI	2.2606	144
3	CZ	2.0565	131
4	NA	2.0408	130
5	WI	1.7425	111
6	PO	1.7111	109
7	CH	1.6484	105
8	KI	1.4914	95
9	ST	1.4286	91
10	RO	1.3344	85
11	YC	1.2873	82
12	OW	1.1303	72
13	ZN	1.1303	72
14	ZY	1.1146	71
15	RZ	1.0832	69
16	JE	1.0518	67
17	ZA	1.0518	67
18	AN	1.0361	66
19	EJ	1.0361	66
20	LI	1.0047	64
21	ER	0.989	63
22	MI	0.9419	60
23	OD	0.9419	60
24	NY	0.9262	59
25	PR	0.9262	59
26	NE	0.9105	58

Table 22 Bigramme der polnischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	NIE	1.6097	80
2	KIE	1.0262	51
3	WIE	1.006	50
4	CZN	0.9658	48
5	EGO	0.9054	45
6	YCH	0.8048	40
7	ICZ	0.7445	37
8	PRZ	0.7445	37
9	SKI	0.6439	32
10	IEJ	0.6036	30
11	OWI	0.6036	30
12	RZY	0.6036	30
13	YCZ	0.6036	30
14	POL	0.5835	29
15	IST	0.5634	28
16	TYC	0.5634	28
17	IEN	0.5433	27
18	ZNA	0.5433	27
19	IEW	0.5231	26
20	LIT	0.5231	26
21	ROD	0.503	25
22	TER	0.503	25

Table 23 Trigramme der polnischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	YCZN	0.7778	30
2	PRZY	0.5963	23
3	CZNE	0.5704	22
4	TYCZ	0.5704	22
5	IEWI	0.5445	21
6	AGRO	0.5185	20
7	ENKI	0.5185	20
8	EWIC	0.5185	20
9	IENK	0.5185	20
10	KIEW	0.5185	20
11	NKIE	0.5185	20
12	SIEN	0.5185	20

Table 24 4-Gramme der polnischen Sprache

4.4.2 Tschechisch

Das tschechische Alphabet besteht aus den folgenden Buchstaben:

a á b c č d ď e é ě f g h ch i í j k l m n ň o ó p q r ř s š t ť u ú ů v w x y ý z ž

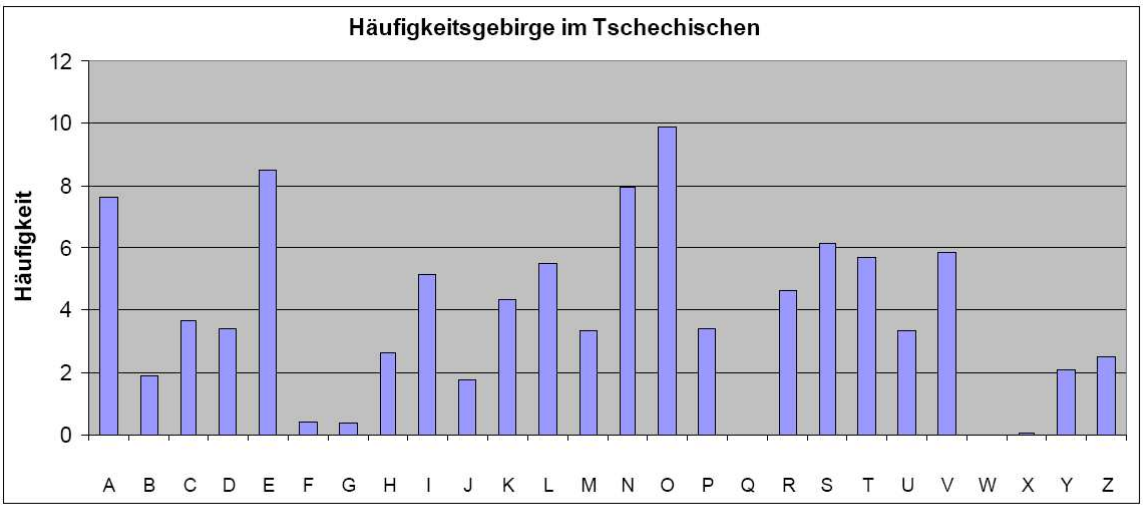


Diagramm 14 Die Buchstaben-Häufigkeitsverteilung eines tschechischen Textes

Die statistischen Größen der tschechischen Sprache lauten:

Entropie: 4.27 (*maxmögliche :* 4.70)

KI = 0.0574109394235558 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	ST	3.0671	222
2	CH	2.0309	147
3	PO	1.8513	134
4	EN	1.7823	129
5	OV	1.727	125
6	RO	1.5198	110
7	NA	1.423	103
8	LA	1.3401	97
9	RA	1.3263	96
10	LE	1.2849	93
11	OS	1.2711	92
12	AL	1.2296	89
13	JE	1.1744	85
14	LI	1.1467	83
15	KO	1.1329	82
16	NO	1.1053	80
17	AN	1.0776	78
18	TA	1.0638	77
19	OD	1.05	76
20	OL	1.05	76
21	HO	1.0362	75
22	SK	1.0086	73
23	LO	0.9947	72
24	PR	0.9671	70
25	CE	0.9533	69
26	TI	0.9533	69

Table 25 Bigramme der tschechischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	OST	1.0785	54
2	BYL	0.8588	43
3	STA	0.8388	42
4	ICK	0.699	35
5	OVA	0.6391	32
6	STV	0.5992	30
7	PRO	0.5792	29
8	POD	0.5592	28
9	STR	0.5592	28
10	NOS	0.5392	27
11	STI	0.5193	26
12	KON	0.4993	25
13	SPO	0.4993	25
14	TRA	0.4594	23
15	KTE	0.4394	22
16	RAN	0.4394	22
17	STO	0.4394	22
18	ALI	0.4194	21
19	ENS	0.4194	21
20	POL	0.4194	21
21	YLA	0.3994	20
22	ECH	0.3795	19
23	NSK	0.3795	19
24	TER	0.3795	19
25	TIC	0.3795	19
26	ESK	0.3595	18

Table 26 Trigramme der tschechischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	NOST	0.748	26
2	BYLA	0.5754	20
3	KTER	0.5466	19
4	STRA	0.4891	17
5	TICK	0.4891	17
6	TRAN	0.4891	17
7	OSTI	0.4028	14
8	ROCE	0.4028	14

Table 27 4-Gramme der tschechischen Sprache

4.4.3 Bulgarisch:

Die bulgarische Sprache ist in die indogermanische Sprache mit der Untergliederung Slawische Sprache, Südslawische Sprache und Bulgarisch einzuordnen. Sie wird in Bulgarien, der Ukraine und Moldawien gesprochen und ist dialektisch dem Mazedonischen ähnlich.
Der Ursprung der Sprache wird bis zur Zeit Alexanders des Großen (3.Jh. v. Chr.) zurückgeführt. Nach seiner Erstkolonisation durchläuft das Land verschiedene Zyklen u.a. auch römische Besatzungen.
Die dort lebenden Thraker nahmen sich der lateinischen Sprache, welche im Lande weit verbreitete war, an. Nach einer kurzen Periode der Entwicklung ließen sie sich die Slawen in dieser Region nieder, (5. Jh) und vertrieben diese romanische Sprache. Anstelle der lateinischen Sprache trat jetzt die slawische Sprache für die Entfaltung der Sprache in den Vordergrund.
Das bulgarische Reich wurde im 9. Jh. christianisiert, daraus entstand eine einheitliche Schriftsprache, die sich in 2 Schriften, Glagolitisch und Kyrillisch festigte, wobei Kyrillisch sich als spätere Standardsprache entwickelte. Die wesentlichste Veränderung in dieser Zeit ist, dass das Bulgarisch keine Nominalflexion besitzt.
In der weiteren geschichtlichen Entwicklung des Landes spielt die Eingliederung von Bulgarien in das Osmanische Reich (im 14. Jh.) eine entscheidende Rolle in der Weiterentwicklung der Sprache. In dieser Zeit wurde die Sprache stark von der islamischen und türkischen Kultur geprägt und teilweise übernommen. Bulgariens weiterer geschichtlicher Verlauf spiegelt die Zerrissenheit in diesem Land wieder, welche schließlich auch in dem Balkankrieg 1912 seinen Höhepunkt fand. Erst nach dem 2. Weltkrieg wurde diese Region, welche aus vielen einzelnen teils unabhängigen, teils eingegliederten Teilen Bulgariens bestand, geeinigt. Die Region wurde „bulgarisiert“.

<u>Fakten:</u>
Das bulgarische Alphabet setzt sich aus 30 Buchstaben zusammen.

Bei der folgenden Erhebung würde eine Text von 19.703 Zeichen analysiert.

Ersetzung im Bulgarischen:

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.
Die nachfolgenden kyrillischen Buchstaben wurden auf die ISO 9 Form zu analytischen Zwecken umbenannt. Bei den letzten 2 Gruppen könnte keine genaue Umformung vorgenommen werden, da es sich bei diesen Zeichen um Betonungszeichen oder Lautbetonungen handelt. Deshalb wurde bei dieser Betrachtung zu der normalen empirischen Analyse die Lautgebung miteinbezogen.

В в	ersetzt durch V oder v	Б б	ersetzt durch B oder b
Г г	ersetzt durch G oder g	Д д	ersetzt durch D oder d
Ж ж	ersetzt durch Z oder z	З з	ersetzt durch Z oder z
И и	ersetzt durch I oder i	Й й	ersetzt durch J oder j
К к	ersetzt durch K oder k	Л л	ersetzt durch L oder l
М м	ersetzt durch M oder m	Н н	ersetzt durch N oder n
О о	ersetzt durch O oder o	Р р	ersetzt durch R oder r
П п	ersetzt durch P oder p	С с	ersetzt durch S oder s
Т т	ersetzt durch T oder t	У у	ersetzt durch U oder u
Ф ф	ersetzt durch F oder f	Х х	ersetzt durch H oder h
Ц ц	ersetzt durch C oder c	Ч ч	ersetzt durch C oder c
Ш ш	ersetzt durch S oder s	Щ щ	ersetzt durch S oder s
Ъ ъ	ersetzt durch A oder a	Ь ь	ersetzt durch J oder j
Ю ю	ersetzt durch U oder u	Я я	ersetzt durch A oder a

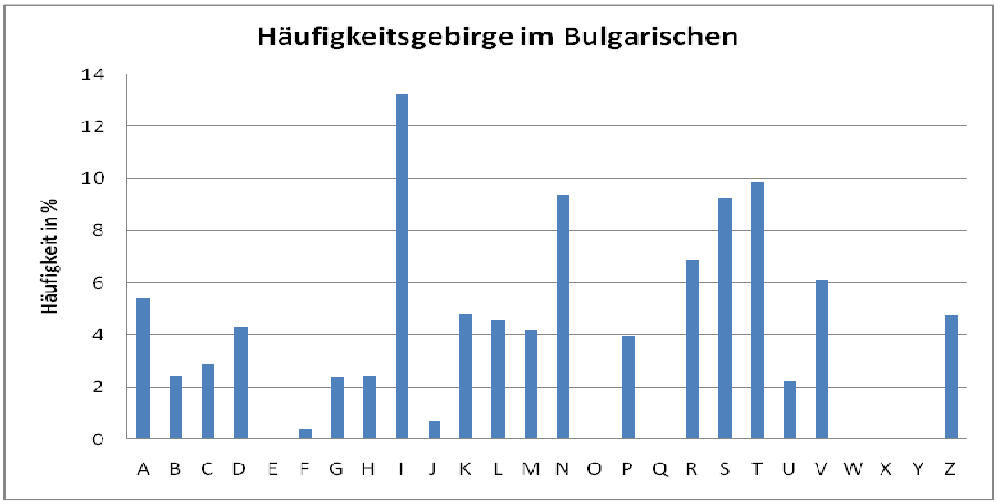


Diagramm 14.1 Die Buchstaben-Häufigkeitsverteilung eines bulgarischen Textes

Aus der Häufigkeitsanalyse lässt sich erkennen, dass im Vergleich zu z.B. den germanischen Sprachen hier das E mit 0,02 % eine untergeordnete Rolle einnimmt. Im Vergleich dazu ist der Buchstabe I mit 13,24 % sehr häufig in Wörter zu finden. Ansonsten werden die Buchstaben W, Y und Q in der bulgarischen Sprache eher nicht verwendet, außer in Eigennamen oder Fremdwörtern.
Die Buchstabenkombination von I-N-S-T deckt statistisch 41,66% der Gesamtheit der Buchstabenhäufigkeiten ab.

Eine Besonderheit in der bulgarischen Sprache ist, dass sie keinen Infinitiv haben und das Futur durch eine Partikel ausdrücken.

Aus den weiteren N-Grammen kann man deutlich die Verkettungen der einzelnen Buchstaben in Bezug auf ihre Häufigkeit in den Strings sehen.

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**NI**	2,40	330
2	**IT**	2,13	294
3	**ST**	2,03	280
4	**TI**	1,36	188
5	**PR**	1,34	184
6	**IN**	1,33	183
7	**IS**	1,29	178
8	**TN**	1,25	172
9	**RI**	1,20	165
10	**LI**	1,12	154
11	**IZ**	1,08	149
12	**NS**	1,08	149
13	**SI**	1,02	141
14	**VS**	0,94	130
15	**TS**	0,90	124
16	**AT**	0,89	122
17	**IC**	0,88	121
18	**TV**	0,88	121
19	**VI**	0,85	117
20	**CI**	0,81	112
21	**SK**	0,81	112
22	**IA**	0,81	111
23	**TR**	0,81	111
24	**NT**	0,79	109
25	**SV**	0,75	103
26	**MI**	0,75	103
27	**IV**	0,74	102
28	**RZ**	0,73	100
29	**SN**	0,70	97
30	**IP**	0,67	92

Table 27.1 Bigramme der bulgarischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**NIT**	0,53	73
2	**PRI**	0,42	58
3	**STV**	0,41	57
4	**IST**	0,33	46
5	**STI**	0,33	46
6	**TNI**	0,30	42
7	**STR**	0,30	42
8	**ITI**	0,30	42
9	**IPR**	0,29	40
10	**ITS**	0,29	40
11	**ITN**	0,29	40
12	**NST**	0,28	39
13	**IAT**	0,27	37
14	**TSI**	0,27	37
15	**NIA**	0,25	34
16	**ICI**	0,24	33
17	**ILI**	0,23	32
18	**STN**	0,23	32
19	**KIT**	0,23	31
20	**NIK**	0,23	31
21	**NIS**	0,22	30
22	**PRD**	0,22	30
23	**LNI**	0,21	29
24	**DIN**	0,21	29
25	**DNI**	0,20	28
26	**CIT**	0,20	28
27	**RNI**	0,20	28
28	**VNI**	0,20	28
29	**NIC**	0,19	26
30	**LIC**	0,19	26

Table 27.2 Trigramme der bulgarischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**ITSI**	0,12	17
2	**NICI**	0,12	16
3	**STVN**	0,11	15
4	**DRUG**	0,11	15
5	**NITS**	0,10	14
6	**ISTI**	0,10	14
7	**IPRI**	0,10	14
8	**RIST**	0,10	14
9	**STRN**	0,10	14
10	**RIAT**	0,09	13

Table 27.3 4-Gramme der bulgarischen Sprache

Die statistischen Größen der bulgarischen Sprache lauten:

Entropie: 4.026 (*maxmögliche :* 4.70)

KI = 0.070715867792682 (0.0384615384615385)

4.4.4 Slowakisch:

Die slowakische Sprache ist in die indogermanische Sprache mit der Untergliederung Slawische Sprache, Westslawische Sprache und Slowakisch einzuordnen.
Den Ursprung der Sprache liegt im 6. Jh., wo die Slawen in dem Gebiet der heutigen Slowakei siedelten und ihr Slawisch in Dialektgruppen weiter formten.
Im Jahre 833 entwickelte sich die slowakische Sprache im Zuge der Christianisierung und der Bildung des Grossmährischen Reiches weiter. Also Folge dieser Entwicklung entstand ein Dekret zu Sprachnutzung der sogenannten liturgischen Sprache.
Durch die Zerschlagung des Grossmährischen Reich gegen Anfang des 10. Jh. entwickelte sich ein echter slowakischer Dialekt und slowakische Unabhängigkeit. Zwischen dem 15. Jh. Und dem 18. Jh. wurde die Sprache etlichen Einflüssen und Veränderungen durch Feldzüge oder Eindringen des Tschechischen ausgesetzt (Lehnwörter und Fremdwörterübernahme).
Aber zum Ende des 18. Jh. Bildete sich in den Städten der Westslowakei und der Mittelslowakei eine überdialektale Koine[4], welche von den Schichten des Bürgertums und den Gebildeten gesprochen worden ist. Diese Sprachart hat bis heute Bestand.

Das slowakische Alphabet setzt sich aus 46 Buchstaben wie folgt zusammen.

a á ä b c č d ď dz dž e é f g h ch i í j k l Ĺ ľ m n ň o ó ô p q r ŕ s š t ť u ú v w x y ý z ž

Bei der folgenden Erhebung würde eine Text von 11.799 analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.

Ersetzung im Slowakischen:

á erstetzt durch a	é erstetzt durch e	í erstetzt durch i
ó erstetzt durch o	ú erstetzt durch u	ý erstetzt durch y
ä erstetzt durch ae	ô ersetzt durch o	č ersetzt durch c
š ersetzt durch s	ž ersetzt durch z	ď ersetzt durch d
ť ersetzt durch t	ň ersetzt durch n	ľ ersetzt durch l
ŕ ersetzt durch r		

Besonderheit der slowakischen Sprache ist, dass die erste Silbe immer betont wird.

Zusätzlich haben die Sonderzeichen zwei Bedeutungen:

1. Als Verlängerung der Betonung, dazu zählen á, é, í, ó, ú, ý, Ĺ, ŕ
2. Als Weichzeichnung der Betonung, dazu zählen ď, ň, ľ, ť

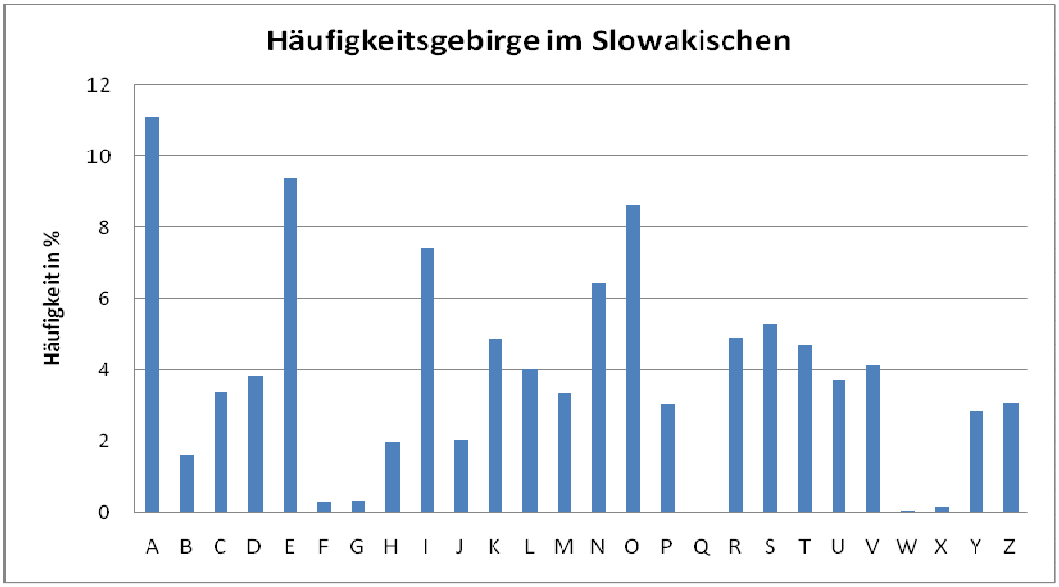


Diagramm 14.2 Die Buchstaben-Häufigkeitsverteilung eines slowakischen Textes

Das Diagramm zeigt ein vermehrtes Auftreten von Vokalen, die zusammen einen Anteil von 36,44 % ausmachen. Die weitere Verteilung der Buchstaben zeigt nur dass die Buchstaben F und G eher weniger genutzt werden.

Die n-Gramme weisen keine auffälligen oder herausstechenden Werte auf.

Die statistischen Größen der slowakischen Sprache lauten:

Entropie: 4.26 (*maxmögliche :* 4.70)

KI = 0.059667627894848 (0.0384615384615385)

[4] Übermundartliche Gemeinsprache

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**NE**	1,53	181
2	**NA**	1,31	154
3	**OV**	1,27	150
4	**EN**	1,22	144
5	**NI**	1,20	142
6	**LI**	1,20	141
7	**IE**	1,12	132
8	**RO**	1,06	125
9	**ST**	1,05	124
10	**AN**	1,03	121
11	**CH**	1,03	121
12	**PR**	0,98	116
13	**IN**	0,98	116
14	**TO**	0,97	114
15	**RA**	0,96	113
16	**ES**	0,92	108
17	**KO**	0,90	106
18	**AT**	0,86	102
19	**VA**	0,86	102
20	**RI**	0,86	101
21	**LA**	0,84	99
22	**IA**	0,84	99
23	**AK**	0,83	98
24	**PO**	0,83	98
25	**AV**	0,81	96
26	**OD**	0,81	95
27	**AJ**	0,81	95
28	**SK**	0,81	95
29	**ZA**	0,80	94
30	**RE**	0,80	94

Table 27.4 Bigramme der slowakischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**NIE**	0,42	50
2	**YCH**	0,42	49
3	**OVA**	0,40	47
4	**PRI**	0,36	42
5	**LIN**	0,33	39
6	**DNE**	0,31	37
7	**NIK**	0,31	37
8	**ODN**	0,30	35
9	**PRE**	0,28	33
10	**ACI**	0,27	32
11	**KLI**	0,27	32
12	**KTO**	0,26	31
13	**ENI**	0,26	31
14	**TOR**	0,26	31
15	**INA**	0,25	30
16	**INI**	0,25	30
17	**STA**	0,25	29
18	**AME**	0,24	28
19	**EPR**	0,24	28
20	**ENT**	0,23	27
21	**ANI**	0,23	27
22	**EST**	0,23	27
23	**RAV**	0,23	27
24	**APR**	0,23	27
25	**NAM**	0,23	27
26	**ICH**	0,22	26
27	**OST**	0,22	26
28	**ALI**	0,22	26
29	**AJU**	0,21	25
30	**LIK**	0,21	25

Table 27.5 Trigramme der slowakischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**KTOR**	0,25	29
2	**LINI**	0,23	27
3	**INIK**	0,22	26
4	**KLEIN**	0,21	25
5	**KYSL**	0,18	21
6	**YSLI**	0,18	21
7	**LOON**	0,17	20
8	**SLIK**	0,17	20
9	**PLOD**	0,16	19
10	**KYCH**	0,15	18

Table 27.6 4-Gramme der slowakischen Sprache

4.4.5 Vergleich der slawischen Sprachen

Die Zusammenstellung beider Sprachen kann das häufige Auftreten der Vokale überhaupt in slawischen Sprachen bestätigen.

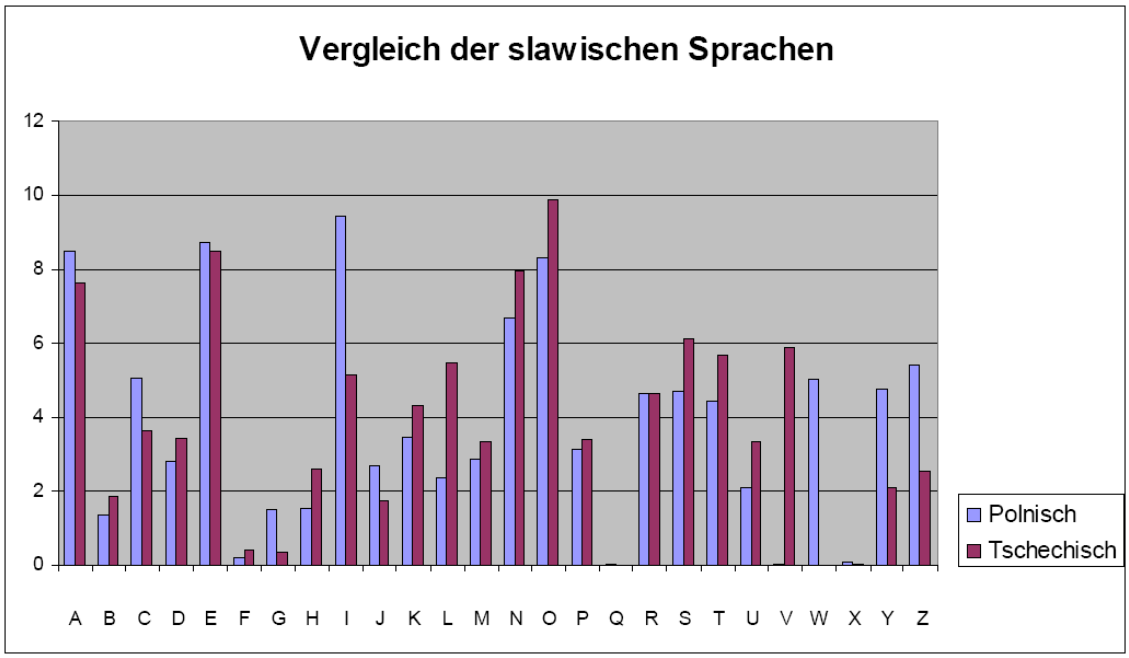


Diagramm 15 Vergleich der slawischen Sprachen

4.5 Irisch:

Die irische Sprache zählt zu den Indogermanischen Sprachen, mit der weiteren Untergliederung Keltisch, Inselkeltisch, Goidelisch und Irisch.
Der Ursprung der irischen Sprache kann nicht genau datiert werden, aber die frühesten Inschriften (Ogam-Schrift) stammen aus dem 4. Jahrhundert und weisen eindeutig nach, dass zu dieser Zeit auf Irland schon Irisch gesprochen worden ist.
Diese Entwicklung des Irischen setzte sich in den Jahrhunderten fort und verschiedene Sprachen vermischten sich mit dem Irischen. Geschichtlich gesehen wurden erst durch die Romanisierung von Britannien und der Rückkehr irischer Mönche von der Missionierung neue Wörter der altirischen Sprache zugeführt.
Ende des 8. Jahrhundert war der nordische Einfluss der Wikinger sehr stark, und aus dieser Belagerungszeit resultierte eine Vermischung des Altirische mit skandinavischen Begriffen und umgekehrt. Wichtig für die heutige irische Sprache war der Einfall der Normannen ab 1169, wobei sich das klassische Irisch entwickelte. Zwischen dem 15. und dem 17. Jahrhundert wurde trotz Ansiedlung von fremdsprachigen Siedlern das Irisch in der Unterschicht sowie teilweise in der Oberschicht (Adel) gesprochen. Aufgrund politischen Unruhen (1607) musste der Adel von der Insel fliehen und es wurde nach dieser Zeit nur noch in der Unterschicht Irisch gesprochen und zwar das „moderne Irisch“. Der entscheidendste Faktor für den Rückgang der irischen Sprache war die Hungersnot auf dem Lande, die 1845-49 eine Vielzahl von irischsprachigen Menschen umbrachte. Als Folge dieser Katastrophe immigrierten viele Iren ins Ausland. (u.a. USA) Versuche nach der Unabhängigkeitserklärung Irlands (1922) die irische Sprache wieder Zu beleben scheiterten, so dass heutzutage nur noch 1,66 Millionen Menschen Irisch sprechen.

Das irische Alphabet umfasst 31 Zeichen. Davon sind 5 kurze Vokale und die dazugehörigen 5 langen Vokale. Zudem kommen 13 Konsonanten und 8 weitere Konsonanten, die ausschließlich in Fremd- und Lehnwörtern auftauchen.

Zusammensetzung:

A, E, I, O, U (kurze Vokale) Á, É, Í, Ó, Ú	(lange Vokale)
B, C, D, E, F, G, H, L, M, N, P, R, S, T	(verwendeten Konsonanten)
J, K, Q, V, W, X, Y, Z	(Fremd oder Lehnwörter)

Bei der folgenden Erhebung würde eine Text von 17461 Zeichen analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.

Ersetzungsvorschrift: Alle langen Vokale werden durch kurze Vokale ersetzt.

Eine Besonderheit des Irischen hat der Buchstabe H, welcher selbstständig nur in Fremd- oder Lehnwörtern vorkommt. Zudem dient er als Vorschlag vor Vokalen bei einer Art syntaktischen Umgebung (z. B. wir aus álainn → **h**álainn).

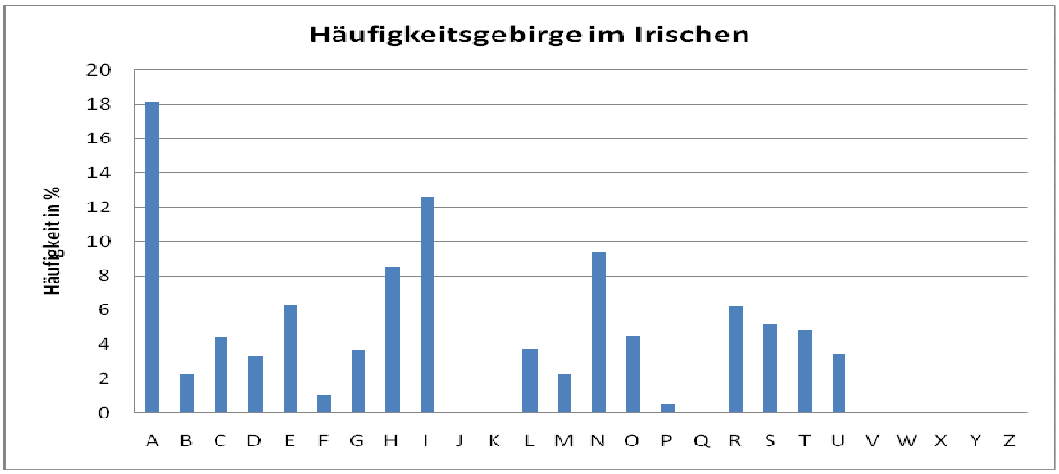


Diagramm 16 Die Buchstaben-Häufigkeitsverteilung eines irischen Textes

Aus der Häufigkeitsgebirge lässt sich eindeutig erkennen, dass es sich bei den Buchstaben A und I um die dominierenden Buchstaben in der irischen Sprache handeln.
Beide zusammen machen anteilig einen Prozentsatz von 30,68 % aus.
Da in dem bearbeiteten Text kaum oder gar keine Lehn- bzw. Fremdwörter aufgetaucht sind, kann man dementsprechend die Buchstaben J, K, Q, V, W, Y und Z für die statistische Betrachtung vernachlässigen. Damit grenzt sich aber die Irische Sprache auf eine Basis von nur 18 Buchstaben ein, was sich auch im Informationsgehalt wiederspiegelt (3.18).
Zu erwähnen ist das in dem Trigramm der Teilstring ACH gegenüber den anderen sehr oft aufgetaucht ist. (1,49 %)

Die statistischen Größen der irischen Sprache lauten:

Entropie: 3.81 (*maxmögliche :* 4.70)
KI = 0.087302368508014 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**AI**	3,52	614
2	**AN**	3,25	568
3	**EA**	2,73	477
4	**NA**	2,68	468
5	**HA**	2,66	465
6	**CH**	2,59	453
7	**IN**	2,33	407
8	**AR**	1,99	348
9	**IR**	1,85	323
10	**AC**	1,76	308
11	**IS**	1,59	278
12	**EI**	1,59	277
13	**TH**	1,51	263
14	**RA**	1,48	258
15	**SA**	1,37	240
16	**BH**	1,36	238
17	**AG**	1,35	235
18	**NN**	1,32	230
19	**LA**	1,17	205
20	**AD**	1,09	191
21	**IA**	1,08	189
22	**TA**	1,02	178
23	**HI**	1,01	176
24	**RI**	0,95	166
25	**AS**	0,93	162
26	**NI**	0,93	162
27	**DH**	0,92	161
28	**IL**	0,90	157
29	**OI**	0,89	155
30	**IG**	0,87	152

Table 28 Bigramme der irischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**ACH**	1,49	260
2	**AIN**	0,89	155
3	**EAN**	0,74	130
4	**AGU**	0,74	130
5	**GUS**	0,73	128
6	**HAI**	0,73	127
7	**CHA**	0,68	118
8	**INN**	0,65	113
9	**THA**	0,62	108
10	**ABH**	0,54	95
11	**ADH**	0,54	94
12	**ANN**	0,53	92
13	**ARA**	0,52	90
14	**SAN**	0,52	90
15	**ATH**	0,50	87
16	**ANA**	0,50	87
17	**ITH**	0,50	87
18	**CHT**	0,49	86
19	**AIR**	0,49	85
20	**LAI**	0,48	84
21	**NNA**	0,47	82
22	**BHI**	0,46	80
23	**REA**	0,46	80
24	**IRE**	0,45	78
25	**NAI**	0,44	77
26	**INA**	0,44	77
27	**EIR**	0,42	74
28	**IRI**	0,42	74
29	**RAN**	0,41	72
30	**IGH**	0,41	71

Table 29 Trigramme der irischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**AGUS**	0,73	128
2	**ACHA**	0,41	72
3	**ACHT**	0,37	64
4	**EACH**	0,36	63
5	**IREA**	0,36	63
6	**AINN**	0,35	61
7	**INNI**	0,34	60
8	**NACH**	0,34	59
9	**ANNA**	0,33	58
10	**CHTA**	0,32	56

Table 30 4-Gramme der irischen Sprache

4.6 Baltische Sprachen

Hier wurden die Sprachen Lettisch und Litauisch ausgewertet und verglichen.

4.6.1 Lettisch:

Die lettische Sprache zählt zu den Indogermanischen Sprachen, mit der weiteren Untergliederung Baltisch, Ostbaltisch und Lettisch.
Der Ursprung der lettischen Sprache liegt wahrscheinlich in den Sprachen der Letgallen und Semgallen.
Die Lettgallen bewohnt ganz Latgale und Vidzemes sowohl im Zentrum als auch landläufig. Hingegen bewohnten die Semgallen nur das Zentrum des Landes. Ein weiteres Detail auf den Sprachverlauf gibt die Sprache der Liven, welche sich mit der lettischen Sprache mischte und eine eigene Mundart erzeugte, das Livisch. Aus dieser Zeit würde aus dem Livischen viele Wörter entlehnt und in die lettischen Sprache eingegliedert (Beispiele hierfür: *joma* oder *kaija).*
Aufgrund der Hauptdialekte und den vielen Mundarten läst sich anhand der Aussprache die regionale Herkunft bestimmen. Charakteristisch trotz gleicher Schriftart ist die Betonung der kurzen Vokale.
Nach verschiedenen Einflüssen vom Deutschen bis zum Russischen erlangt im 20. Jahrhundert das Englische immer mehr Bedeutung in der lettischen Sprache.

Das lettische Alphabet besteht aus 33 Zeichen.

a ā b c č d e ē f g ģ h i ī j k ķ l ļ m n ņ o p r s š t u ū v z ž

Bei der folgenden Erhebung würde eine Text von 16.188 Zeichen analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.

Ersetzungsvorschrift:

ā ersetzt durch a	i ersetzt durch i	Â ersetzt durch a
ē ersetzt durch e	ņ ersetzt durch n	ļ ersetzt durch l
ž ersetzt durch z	š ersetzt durch s	č ersetzt durch c
ķ ersetzt durch k	ū ersetzt durch u	j ersetzt durch j
ģ ersetzt durch g	ī ersetzt durch i	ç ersetzt durch c
à ersetzt durch a	æ ersetzt durch ae	ž ersetzt durch z

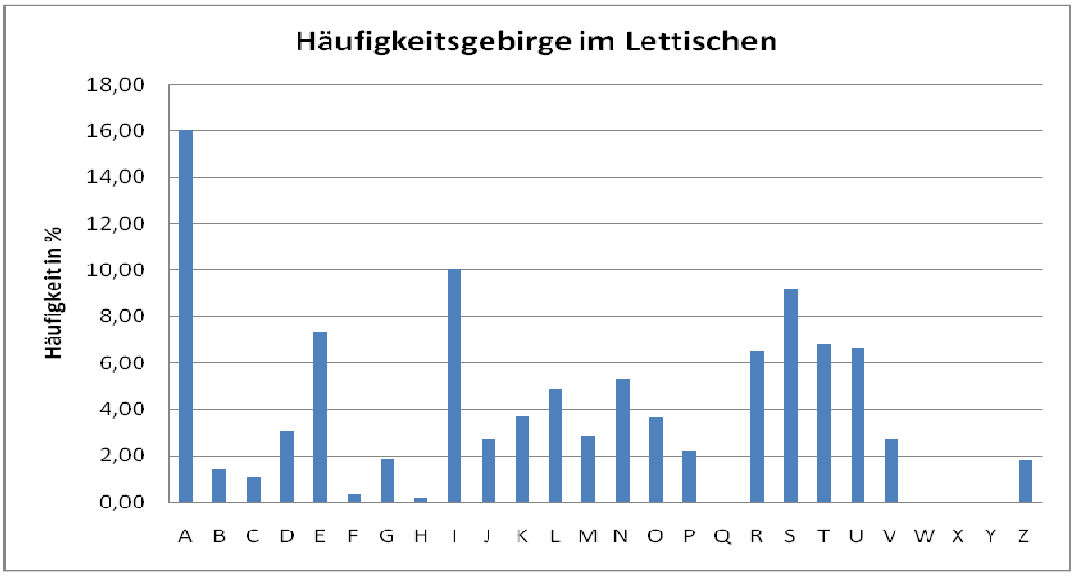


Diagramm 17 Die Buchstaben-Häufigkeitsverteilung eines lettischen Textes

Aus der Graphik lässt sich erkennen, dass in der lettischen Sprache vielfach Wörter mit den Buchstaben A, E, I, S gebildet werden. Sie machen einen Gesamtanteil von 42,53 % aus. Die Wörterbildung erfolgt mit Ausnahmen von Fremdwörtern ausschließlich ohne die Buchstaben Q, W, X und Y. Zudem treten die Buchstaben F und H eher selten im Sprachgebrauch der Letten auf.

Die statistischen Größen der lettischen Sprache lauten:

Entropie: 4.03 (*maxmögliche :* 4.70);*KI* = 0.075113557666289 (0.0384615384615385)

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**AS**	3,08	498
2	**RA**	2,29	370
3	**IE**	2,15	348
4	**AT**	1,63	263
5	**ST**	1,61	261
6	**AR**	1,50	242
7	**SA**	1,50	242
8	**AL**	1,47	238
9	**TU**	1,45	235
10	**TA**	1,39	225
11	**JA**	1,38	223
12	**AN**	1,36	220
13	**LA**	1,29	208
14	**ES**	1,16	188
15	**LI**	1,13	183
16	**NA**	1,08	175
17	**NI**	1,04	169
18	**IJ**	1,03	167
19	**UR**	1,02	165
20	**TI**	1,02	165
21	**MA**	1,00	162
22	**DA**	0,98	159
23	**VI**	0,98	158
24	**ER**	0,96	156
25	**IS**	0,96	155
26	**IT**	0,92	149
27	**IN**	0,90	146
28	**TE**	0,88	143
29	**KA**	0,88	143
30	**RI**	0,85	137

Table 31 Bigramme der lettischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**IJA**	0,84	136
2	**JAS**	0,74	119
3	**TUR**	0,60	97
4	**ATU**	0,56	91
5	**URA**	0,53	85
6	**ATV**	0,47	76
7	**ASA**	0,47	76
8	**TVI**	0,46	75
9	**LAT**	0,46	74
10	**NIE**	0,45	73
11	**LIT**	0,45	73
12	**IES**	0,43	70
13	**STA**	0,43	70
14	**VIE**	0,43	69
15	**TER**	0,41	67
16	**RAS**	0,41	67
17	**DZE**	0,41	67
18	**ITE**	0,41	66
19	**ERA**	0,40	65
20	**VAL**	0,39	63
21	**ZEJ**	0,38	62
22	**SAN**	0,37	60
23	**AMA**	0,36	59
24	**IEK**	0,35	56
25	**RAT**	0,34	55
26	**STU**	0,33	53
27	**ALA**	0,33	53
28	**INA**	0,33	53
29	**AST**	0,32	52
30	**ESU**	0,32	51

Table 32Trigramme der lettischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**IJAS**	0,57	93
2	**ATVI**	0,47	76
3	**LATV**	0,46	75
4	**LITE**	0,41	66
5	**TURA**	0,40	65
6	**DZEJ**	0,39	63
7	**ITER**	0,38	62
8	**TERA**	0,38	62
9	**ATUR**	0,35	56
10	**ERAT**	0,35	56

Table 33 4-Gramme der lettischen Sprache

4.6.2 Litauisch:

Die litauische Sprache zählt zu den Indogermanischen Sprachen, mit der weiteren Untergliederung Baltisch, Ostbaltisch und Litauisch.
Der Ursprung der litauischen Bevölkerung geht bis auf das Jahr 2000 v. Chr. zurück. In dieser Zeit drängten baltische Völker in finn-urgisches Gebiet ein und verdrängten die Völker der Region. Unklar ist aber die Herkunft der Sprache, einige vermuten, dass die Sprache, welche dem Sanskrit sehr nahe steht, aus dem Punjab-Gebiet (Nordindien) stammt. (vgl. Ludwig...)
Die weitere Entwicklung der Sprache ist durch die friedliche Expansion durch Erweiterung der Handelswege rund um das Schwarze Meer und entlang der Flüsse geprägt. Der zunehmende wirtschaftliche Einfluss lies die Bevölkerung schnell wachsen und sie besiedelten das Gebiet zwischen Ostsee und Wolgaraum.
Diese Expansion wurde im 6. Jh. durch skandinavische Völker die in diese Gebiete eindrangen gebremst. In der weiteren Entwicklung hätte die baltische Bevölkerung gegen die ins Land vorstoßenden Wikinger und später (um 854) gegen die Dänen und Schweden zu kämpfen, welche in Kurland einfielen.
Trotz dem Versuch der Christianisierung zum Ende des 14. Jahrhundert, bewahrten sich die Litauer als einziges baltisches Land seine Selbständigkeit und eine intakte Machtpolitik.
Spätere Kriege und Umwerfungen im Lande führten dann dazu, dass Litauen sich vollkommen zersplitterte und mit dem engen Bündnis mit Polen die litauische Staatlichkeit und den Niedergang der Macht Litauens zur Folge hatte. (vgl. Quelle
Nachdem es im 18. Jh. eine russische Einflussphase gab, brachte die Oktoberrevolution die endgültige Trennung und führte zur Gründung eines eigenen Staates, der auch durch die Sowjetunion anerkannt wurde. Die nationale Identität prägte sich im Wesentlichen in einer kurzen Phase der Eigenstaatlichkeit zwischen 1918 und 1940 aus.
Nach dem 2. Weltkrieg wurde Litauen wieder durch die Sowjetunion besetzt, infolge dessen kam es durch litauische Partisanen immer wieder zu Widerstand.
1990 erklärte Litauen seine Unabhängigkeit und wurde 1991 durch die EG- Staaten anerkannt.
In der weiteren unabhängigen Entwicklung wurde Litauen 2004 Mitglied der EU.
Das litauische Alphabet besteht aus 32 Zeichen.

Kleine Buchstaben: a ą b c č d e ę ė f g h i į y j k l m n o p r s š t u ų ū v z ž
Große Buchstaben: A Ą B C Č D E Ę Ė F G H I Į Y J K L M N O P R S Š T U Ų Ū V Z Ž

Bei der folgenden Erhebung würde eine Text von 13.568 Zeichen analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder Sonderzeichen.

Ersetzungsvorschrift:

ų	ersetzt durch U oder u	š	ersetzt durch S oder s
č	ersetzt durch C oder c	ž	ersetzt durch Z oder z
ė	ersetzt durch E oder e	ą	ersetzt durch A oder a
ū	ersetzt durch U oder u	į	ersetzt durch I oder i
ę	ersetzt durch E oder e		

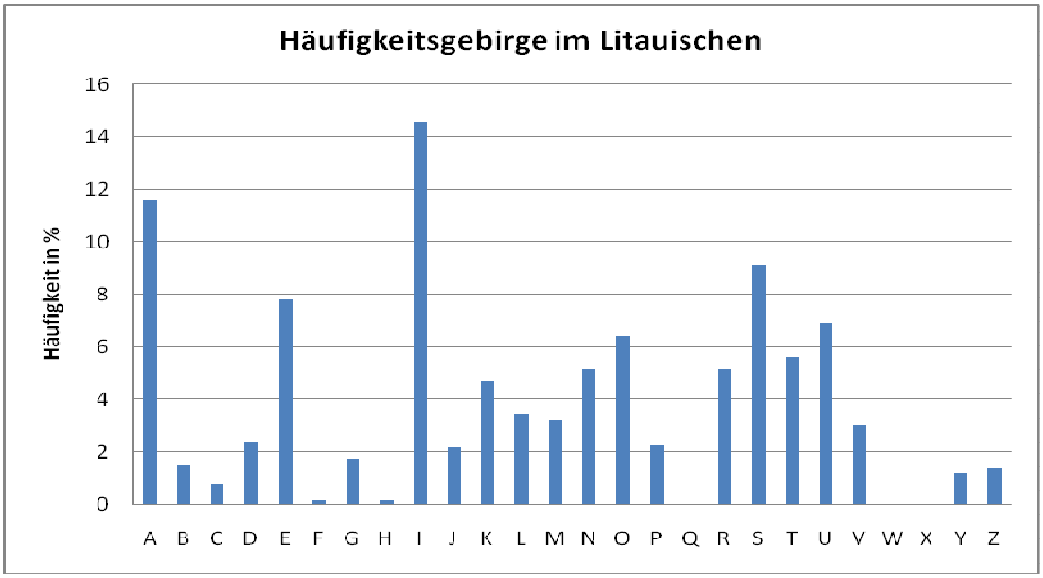


Diagramm 18 Die Buchstaben-Häufigkeitsverteilung eines litauischen Textes

Aus der Häufigkeitserhebung kann man deutlich erkennen, dass die Vokale A und I vermehrt in der litauischen Sprache erscheinen, hier liegt der Prozentsatz bei 14,56 % für I und 11,56% bei A. Hingegen Buchstaben wie W, X und Q verschwindend gering genutzt werden oder gar nicht in der Sprache vorkommen. Anwendung finden diese Buchstaben nur in Fachtexten, wo sie zur originalgetreuen Schreibung fremdsprachlicher Eigennamen gebraucht werden.

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**IN**	2,03	276
2	**IS**	1,98	269
3	**AI**	1,88	255
4	**OS**	1,68	228
5	**SI**	1,62	220
6	**IA**	1,44	195
7	**LI**	1,39	188
8	**AS**	1,33	181
9	**IE**	1,33	180
10	**US**	1,32	179
11	**ST**	1,30	177
12	**AU**	1,27	172
13	**KA**	1,25	170
14	**TI**	1,25	170
15	**NI**	1,19	162
16	**ES**	1,12	152
17	**RA**	1,11	151
18	**IR**	1,10	149
19	**IU**	1,07	145
20	**TA**	1,05	142
21	**AR**	1,05	142
22	**ET**	1,04	141
23	**RI**	1,01	137
24	**EN**	0,99	134
25	**NE**	0,96	130
26	**IK**	0,91	123
27	**TU**	0,91	123
28	**ME**	0,83	112
29	**SK**	0,83	112
30	**AL**	0,82	111

Table 34 Bigramme der litauischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**INI**	0,58	79
2	**USI**	0,52	70
3	**IAU**	0,44	60
4	**INE**	0,41	55
5	**IET**	0,4	54
6	**ELI**	0,4	54
7	**AUS**	0,39	53
8	**AIS**	0,39	53
9	**STA**	0,38	52
10	**ETU**	0,38	51
11	**IST**	0,37	50
12	**TIN**	0,35	48
13	**ISK**	0,35	47
14	**INK**	0,32	43
15	**STO**	0,31	42
16	**VOS**	0,31	42
17	**JOS**	0,29	40
18	**KAL**	0,29	40
19	**UVO**	0,29	40
20	**RAS**	0,29	40
21	**LIE**	0,29	39
22	**SKA**	0,29	39
23	**IJO**	0,29	39
24	**KAI**	0,29	39
25	**CIA**	0,29	39
26	**NIN**	0,27	37
27	**IEN**	0,27	37
28	**VIE**	0,27	37
29	**MEN**	0,25	34
30	**NES**	0,25	34

Table 35 Trigramme der litauischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**IAUS**	0,25	34
2	**ININ**	0,24	33
3	**NINK**	0,23	31
4	**LIET**	0,23	31
5	**IETU**	0,21	29
6	**ETUV**	0,21	29
7	**UVOS**	0,20	27
8	**TORI**	0,19	26
9	**AUSI**	0,19	26
10	**IJOS**	0,18	24

Table 36 4-Gramme der litauischen Sprache

Die statistischen Größen der litauischen Sprache lauten:

Entropie: 4.05 (*maxmögliche :* 4.70);

KI = 0.073704144855511 (0.0384615384615385)

4.6.3 Vergleich der baltischen Sprachen

Auffällig ist das in den N-Gramme deutliche Unterschiede in den Buchstaben bestehen.
Gleich sind die Vokalverteilung A-E-I bei beiden Sprachen.

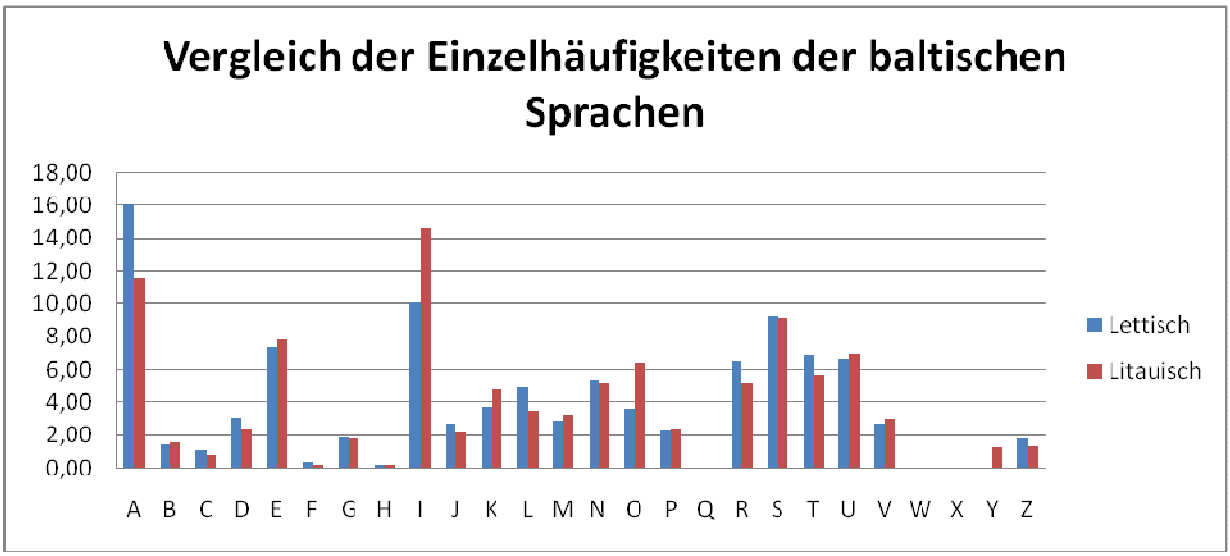


Diagramm 18.1Vergleich der baltischen Sprachen

Aus dem Vergleich anhand der Einzelverteilungen ist nur der Unterschied der häufigeren Verwendung des Buchstaben A im Lettischen bzw. umgekehrt das I im Litauischen.
Zudem wird im Litauischen das O vermehrt verwendet als im Lettischen. Die anderen kleinen Abweichungen sind zu vernachlässigen.

4.7 Maltesisch als Vertreter der Afroasiatischen Sprache

Eines der ersten Schriftstücke geht bis in das Jahr des Frühägyptischen zurück. (ca. 4. Jt v. Chr.). Weitere Überlieferungen sind aus dem 2. Jt. V. Chr. zu belegen (westsemitische Idiome[5]).

4.7.1 Maltesisch:

Die maltesische Sprache zählt zu den Afroasiatischen Sprachen, mit der weiteren Untergliederung Semitisch und Maltesisch.
Der Ursprung der maltesischen Sprache liegt wahrscheinlich in dem maghrebinischen Arabisch, welches zur Zeit der Belagerung der Insel durch die Phönizier als Sprache etabliert wurde. Während den Belagerungen und Eroberungen wechselte sich die arabische Sprache mit der italienischen Sprache ab, bis 1914 die Engländer das lateinische Alphabet für Maltesisch einführten und es 1934 zur Koamtssprache wurde (neben Englisch).Heutzutage sprechen 330.000 Menschen die einheimische Sprache Malti.

Das maltesische Alphabet setzt sich aus 30 Buchstaben zusammen, welche insgesamt 28 separaten Schriftzeichen umfasst.
Dabei zu beachten ist, dass die beiden Digraphen *għ/G*Ħ (ein Konsonant) und ie/IE (ein Vokal) als eigenständige Buchstaben zählen.
Demzufolge setzt sich das maltesische Alphabet aus folgenden Buchstaben zusammen:

A, B, Ċ, D, E, F, G, Ġ, Għ, H, Ħ, I, IE, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, Ż
Wobei Ċ/ċ, Ġ/ġ, Ħ/ħ, Ż/ż *(*Għ/għ*)* als maltesische Sonderzeichen gelten.

Bei der folgenden Erhebung würde eine Text von 11439 Zeichen analysiert.

Folgende Symbole wurden durch ein Leerzeichen ersetzt . , : ! ? „ ” ; - ( ) /`‘und sonstige Zeichensetzungen oder SonderzeichenErsetzungsvorschrift:

Ċ	wird ersetzt durch C/c	Ġ	wird ersetzt durch G/g
Għ	wird ersetzt durch Gh	Ħ	wird ersetzt durch H
Ż	wird ersetzt durch Z		

Eine Besonderheit ist, dass Umlaute wie Ä, Ö, Ü, sowie die Konsonanten ß, C und Y existieren im Maltesischen nicht.

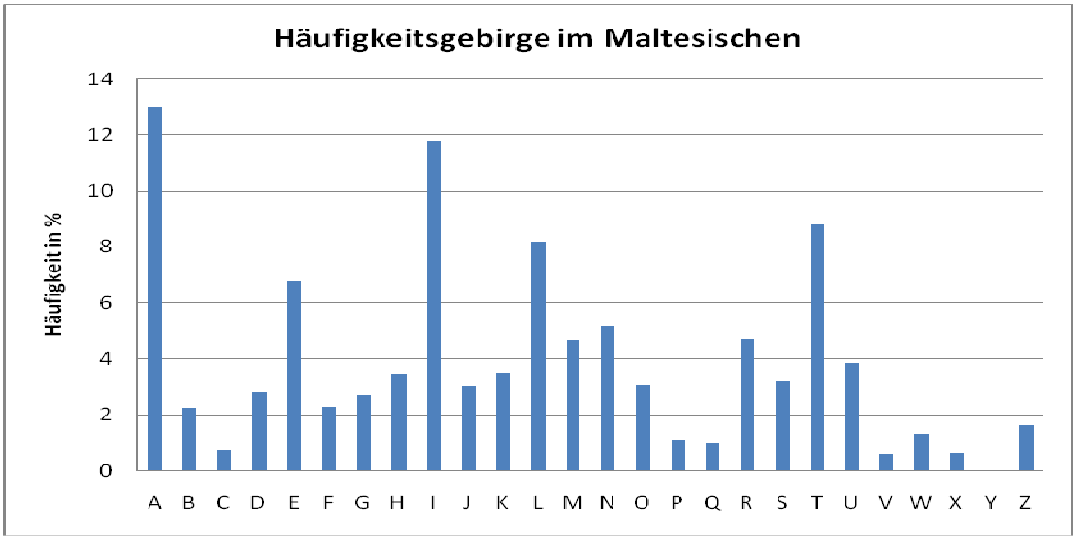


Diagramm 19 Die Buchstaben-Häufigkeitsverteilung eines maltesischen Textes

Aus der Häufigkeitsanalyse kann man sehen, dass die beiden Vokale A und I im Maltesischen oft genutzt werden. Zudem ist eine starke Ausprägung des Buchstaben T und vor allem bei dem Buchstaben L zu erkennen. Die durch das Auftreten von Fremdwörtern entstandenen Erhebungen der Buchstaben C und Y sind in dieser Betrachtung zu vernachlässigen, da sie in der maltesischen Sprache nicht vorkommen. Auffällig ist das im Vergleich zu den meisten anderen Sprachzweigen der Buchstabe Q verwendet wird. Bei dem 4-Gramm fällt auf das mehrere Buchstaben in den Teilstrings häufiger Verwendung finden als vorher als separater Buchstabe (z.B. AKKA oder ZZJON). Zudem fällt auf das der häufigste String ein Bestandteil des Namens der Insel ist. (unabhängig davon das Malta in dem analysierten Text eher relativ wenig vorkam)

Die statistischen Größen der maltesischen Sprache lauten:

Entropie: 4.22 (*maxmögliche :* 4.70)
KI = 0.066681310781703 (0.0384615384615385)

[5] Mundart

Bigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**TA**	2,64	301
2	**AL**	2,51	287
3	**LI**	1,97	225
4	**IL**	1,75	200
5	**AT**	1,66	190
6	**HA**	1,65	188
7	**TI**	1,62	185
8	**IE**	1,61	184
9	**TT**	1,40	160
10	**AR**	1,35	154
11	**GH**	1,30	148
12	**IT**	1,28	146
13	**MA**	1,24	142
14	**ET**	1,17	134
15	**MI**	1,16	133
16	**RA**	1,15	131
17	**NI**	1,13	129
18	**IS**	1,04	119
19	**EN**	1,03	118
20	**IN**	1,02	116
21	**LL**	0,98	112
22	**JA**	0,95	108
23	**ON**	0,91	104
24	**EM**	0,83	95
25	**LA**	0,77	88
26	**AK**	0,75	86
27	**KA**	0,75	86
28	**AM**	0,74	84
29	**ER**	0,74	84
30	**AN**	0,73	83

Table 37 Bigramme der maltesischen Sprache

Trigramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**GHA**	0,76	87
2	**TAL**	0,72	82
3	**ATA**	0,63	72
4	**MAL**	0,58	66
5	**ALI**	0,53	61
6	**JON**	0,48	55
7	**ALT**	0,46	53
8	**ITT**	0,46	52
9	**ETT**	0,45	51
10	**ONI**	0,45	51
11	**IEN**	0,44	50
12	**ITA**	0,43	49
13	**LIS**	0,42	48
14	**AKK**	0,40	46
15	**HAL**	0,39	45
16	**TTA**	0,39	45
17	**ZJO**	0,39	44
18	**IST**	0,39	44
19	**ENT**	0,38	43
20	**ILL**	0,38	43
21	**AGH**	0,36	41
22	**ZZJ**	0,35	40
23	**TTI**	0,35	40
24	**MIN**	0,34	39
25	**LTI**	0,33	38
26	**IET**	0,33	38
27	**KKA**	0,32	37
28	**ATT**	0,32	36
29	**KIE**	0,32	36
30	**TAT**	0,32	36

Table 38 Trigramme der maltesischen Sprache

4-Gramme			
Nr.	Teilstring	Häufigkeit (in %)	Häufigkeit
1	**MALT**	0,43	49
2	**JONI**	0,43	49
3	**ZJON**	0,38	43
4	**ZZJO**	0,34	39
5	**ALTI**	0,32	37
6	**KEIN**	0,31	36
7	**AKKA**	0,31	35
8	**ATAL**	0,28	32
9	**KKAD**	0,25	29
10	**KADE**	0,25	29

Table 39 4- Gramme der maltesischen Sprache

5 Zusammenfassung

„ No matter how resistant the cryptogram, all that is really needed is an entry, the identification of one word, or of three or four letters.“

Helen Fouche Gaines 1939

Die Unregelmäßigkeiten, die in allen Sprachen vorkommen, können genutzt werden, um die Chiffren zu knacken. Für einen Kryptoanalytiker haben lebende, gesprochene Sprachen zwei ungeheure Vorteile: Zum einen ergeben die meisten Kombinationen von Buchstaben lediglich ein sinnloses Wirrwarr, zum anderen tauchen in längeren Texten Buchstaben stets mit einer bestimmten Wahrscheinlichkeit auf. Noch mehr als die Einzelzeichenhäufigkeit prägt die Häufigkeit von Multigrammen einer Sprache, die wesentlich weniger ausgeglichen sind. Jede Sprache hat charakteristische statistische Eigenschaften. Auβer der hier dargestellten können auch mittlere Wortlänge, Vokalhäufigkeit, Häufigkeit der Konsonanten, Häufigkeit eines Buchstabens in Abhängigkeit von seiner Länge im Wort als Kenngrössen der Sprache dienen. Anhand deren lässt sich erkennen in welcher Sprache die Mitteilung, die zum Entschlüsseln steht, geschrieben wurde. Man kann jetzt auf die Frage, die am Anfang dieser Arbeit steht, antworten: zusammenfassend für das Entschlüsseln sind Kenntnisse der jeweiligen Sprache notwendig!

6 Anhang 1:

Häufigkeit der Einzelbuchstaben (in %)

	Deutsch	Englisch	Niederländisch	Spanisch	Französisch	Italienisch	Polnisch	Tschechisch	Finnisch
A	5,4459	7,1948	7,1749	6,6900	6,8163	10,7441	8,5080	7,6304	12,9235
B	1,7497	1,5793	1,4143	0,7129	0,7002	0,8864	1,3348	1,9000	0,0114
C	3,3694	4,0486	1,7846	3,5163	3,3018	5,0546	5,0773	3,6433	0,0798
D	5,1121	3,1148	6,8532	4,0304	3,7136	3,5701	2,8194	3,4273	0,9800
E	16,8857	13,0484	18,8418	15,9230	15,6301	13,2070	8,7325	8,4945	7,5853
F	1,2789	2,4192	0,7770	1,0967	1,1326	1,3127	0,2121	0,4026	0,0228
G	3,7629	2,3439	2,9440	1,5697	0,8443	1,0520	1,4970	0,3732	0,1483
H	5,2632	4,7067	2,7498	1,2201	0,5903	1,4967	1,5469	2,6122	2,0646
I	8,5131	7,7087	6,8654	7,3206	7,1115	9,8025	9,4311	5,1360	11,0756
J	0,1757	0,0877	1,4993	0,1577	0,1922	0,0000	2,7071	1,7480	2,1558
K	1,5073	0,5829	1,9242	0,0480	0,0000	0,0123	3,4681	4,3209	5,0987
L	3,7664	3,7227	4,1459	5,3122	4,8531	5,7570	2,3827	5,4699	5,4523
M	2,2205	2,5445	1,8757	2,5636	3,2194	2,9782	2,8568	3,3890	3,5700
N	10,4244	7,8090	9,9065	7,1424	9,4248	7,5696	6,7116	7,9544	9,2392
O	3,1059	7,5207	5,8456	6,0114	6,0818	9,6553	8,3458	9,8792	7,1860
P	0,6324	2,3001	1,3597	3,5301	3,2057	2,6285	3,1562	3,3978	1,5855
Q	0,0105	0,1003	0,0243	1,3572	1,7367	0,6932	0,0250	0,0000	0,0114
R	7,1429	6,4114	6,4951	7,0258	5,8141	6,0943	4,6657	4,6352	2,6691
S	6,2434	6,4929	4,4494	9,4386	9,5277	5,9410	4,6781	6,1377	8,3609
T	6,0818	9,2191	6,0155	7,3069	7,3174	5,8950	4,4286	5,6565	9,3989
U	3,4045	2,8328	1,7725	5,7235	6,9193	2,9475	2,0709	3,3389	5,0644
V	0,8924	0,8649	2,6648	1,1241	1,0571	1,6409	0,0250	5,8627	2,8630
W	1,6408	1,0654	1,3961	0,0548	0,0000	0,0000	5,0399	0,0000	0,3194
X	0,0246	0,4512	0,0182	0,7060	0,4187	0,0184	0,0749	0,0491	0,0000
Y	0,0738	1,7298	0,0850	0,3564	0,3569	0,0000	4,7779	2,0819	2,0760
Z	1,2719	0,1003	1,1169	0,0617	0,0343	1,0428	5,4300	2,5100	0,0570

Literaturverzeichnis

1. Friedrich L. Bauer: Kryptologie, Methoden und Maximem; Berlin: Springer-

Verlag, 1993; ISBN 3-540-57771-8

2. Claus Schönleber: Verschlüsselungs-Verfahren; München: Franzis-Verlag GmbG,

1995; ISBN 3-7723-5043-7

3. www.wikipedia.org.de

4. http://www-ivs.cs.unimagdeburg.de/bs/lehre/wise0102/progb/vortraege/hbeier/hbeier05.html

5. http://de.wikibooks.org/wiki/Kryptographie

6. *CrypTool 3.05*
7. http://www.dankultur.de/daenemark-info/sprache.htm#Verbreitung%20der%20Sprache
8. v. Pistohlkors, 1991, 12f. / v. Rauch, 1970, 25f.
9. http://www.creation-evolution.eu/html/spracheevolution.html
10. http://www.dankultur.de/daenemark-info/sprache.htm#Verbreitung%20der%20Sprache
11. v. Pistohlkors, 1991, 12f. / v. Rauch, 1970, 25f.
12. http://www.zeno.org/Meyers-1905/A/Rum%C3%A4nische+Sprache+und+Literatur
13.http://siebenbuergen.heim.at
14.http://www.sprachendienst.de/de/schwedisch/index.shtml
15. http://www.azm-lu.si/index.php?option=com_content&task=view&id=292&Itemid=374&limit=1&limitstart=1
16.Susanna Vykoupil, Slowakei, München 1999 (Beck)
17. http://www.fachuebersetzungen.de/Ungarisch/ungarisch.html
18. http://www.zappmedia-gmbh.de/sprachen-information.html
19. http://lexikon.meyers.de/meyers/Ungarische_Sprache

Tabellenverzeichnis

Diagrammverzeichnis

Abbildungsverzeichnis