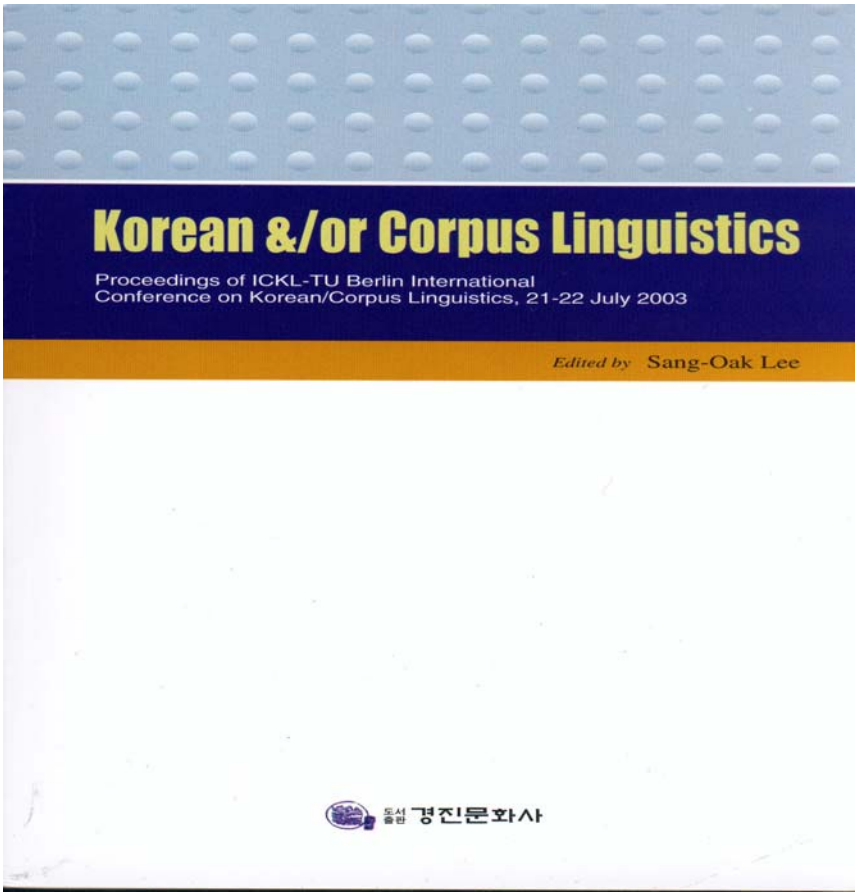
Das Zaza-Deutsche Korpus und die Berechnung sprachlicher Häufigkeiten

(The Zaza-German Corpus and the calculation of linguistic frequencies), in: Koean &/or Corpus Linguistics, Ed.: Sang-Oak Lee, Proceedings of ICKL-TU Berlin International Conference on Korean/Corpus Linguistics, 21-22 July 2003, Seoul 2004, S. 87-149.

ACHTUNG! Dieser Artikel ist eine Internetausgabe. Beim Zitieren ist neben dem Autorennamen die obige bibliografische Quelle anzugeben.



Das Zaza-Deutsche Korpus und die Berechnung sprachlicher Häufigkeiten

The Zaza-German Corpus and the calculation of linguistic frequencies

von Zülfü Selcan

Technische Universität Berlin
Institut für Sprache und Kommunikation
Berlin, 21.07.2003

1. Zusammenfassung

Das Zaza-Deutsche Textkorpus[1] wurde als Forschungsprojekt von der Deutschen Forschungsgeimeinschaft (DFG) zwischen Jan. 2001 – Dez. 2002 gefördert[2].

Für die Dokumentation mündlicher Literatur und gesprochener Texte in der Zaza-Sprache lag im Jahr 2002 folgendes Sprachmaterial vor:

- 203 Stunden Tonbandaufnahmen
- ca. 2 000 Sprüche und Sprichwörter in handschriftlichen Notizen
- eine Sammlung von mehreren tausend selten gebrauchten Wörtern in handschriftlichen Notizen

---

1 Dieser Vortrag wurde auf der *International Conference on Korean/Corpus Linguistics* an der Technischen Universität Berlin, 21.-22 Juli 2003, gehalten.

2 Das Projekt wurde von Herrn Prof. Dr. Peter Erdmann geleitet und von Prof. Dr. Ronald E. Emmerick - Universität Hamburg – mitunterstützt.

Die Aufgabenstellung des Projekts war, dieses Sprachmaterial *fonetisch* zu transkribieren, den Text computermäßig zu erfassen und ins Deutsche zu übersetzen. Weiterhin sollte der Text nach literarischen Kategorien gegliedert und nach linguistischen Aspekten aufbereitet werden, um daraus ein Zaza-Deutsches Textkorpus zu erstellen. Ein Teil der Tonbandaufnahmen besteht aus Volksliedern, die nach dem Grundsatz der Einheit von Text und Melodie, sowohl durch Transkription als auch durch Notation dokumentiert wurden.

Eine weitere Aufgabe des Projekts bestand darin, die Tonbandaufnahmen, welche durch Tonverlust gefährdet sind, zu digitalisieren.

Bisher konnten folgende Ergebnisse erzielt werden:

- 70 fonetische Transkripte fertig gestellt
- 16 Übersetzungen gefertigt
- 27 Stunden Tonbandaufnahmen digitalisiert
- 100 Lieder durch Notation erfasst

2. Digitalisierung des Tonbandmaterials

Die Tonbänder unterliegen dem Einfluss von Temperatur, Feuchtigkeit, Magnetismus und Verschleiß beim Abspielen. Bei der Benutzung reißen sie sogar, wenn das Bandmaterial schwach ist. Bei mühsamen Flickarbeiten geht manchmal ein Teil des Tons sogar verloren.

Ein anderer Nachteil besteht darin, die Bänder etwa alle 6 Monate umzuspulen, damit die Magnetschicht keinen Schaden nimmt und kein Tonverlust eintritt.

Aus diesen Nachteilen ergibt sich die Notwendigkeit, die Tonbänder zu digitalisieren.

Bei ersten Digitalisierungsversuchen stellte sich heraus, dass die Tonbänder einen erheblichen Rauschton haben. Aufnahmen mit sehr guter Tonqualität enthalten kaum Rauschen, der Anteil solcher ist jedoch sehr gering.
Auch unter dem Gesichtspunkt Töne ohne Rauschen bzw. mit minimalem Anteil zu gewinnen, um eine bessere Abhörqualität, folglich auch präzisere Transkription zu erreichen, musste die Tonaufnahme entrauscht werden.
Für die Säuberung des Rauschanteils wurde die Software *Clean* benutzt, die es jedoch nicht restlos, sondern nur teilweise entfernen konnte. Es erfordert oft mehrere Wiederholungen, bis man ein befriedigendes Ergebnis erzielt. Dieser Mangel wurde später durch den Einsatz der leistungsfähigeren Software *Wawes Restoration* beseitigt, mit dem eine effektivere Rauschunterdrückung erzielt wurde. Die beim Digitalisieren erforderlichen Arbeitsgänge sind:

1. Abspielen der Tonbandkassette und Aufnahme in PC
2. Schnitt der Gespräche/Lieder mit Toneditor
3. Bearbeitung der Aufnahme mit Toneditor (Schnitt, Pegel, u. ä.)
4. Normalisieren (Pegelausgleich)
5. Zusammensetzen der Titel
6. Brennen der Titel auf CD

Die meisten Tonbandkassetten der Sammlung haben eine Spieldauer von 60 Minuten. 90-minütige Kassetten sind gering. Die Audio-CDs haben eine Kapazität bis 80 Min. In der letzten Zeit sind auch 90-Minuten-CDs auf dem Markt. Ihr Nachteil besteht aber darin, dass sie nur mit einem kleinen Teil von CD-Spielern gehört werden können.

Es wurden vorrangig älteste Aufnahmen und die Bänder mit schwachem Ton digitalisiert. Insgesamt wurden rund 27 Stunden Tonbandaufnahmen mit Gesprächen und Liedern digitalisiert und auf 20 CDs übertragen.
Die Praxis ergab, dass die Bearbeitungszeit für die Digitalisierung einer Tonbandkassette, je nach Aufnahmequalität, 8-10 Stunden dauert: d. h. für eine Kassette bzw. eine CD atwa einen Arbeitstag.

Vor Projektbeginn lagen uns keine Digitalisierungserfahrungen vor. Daher stellte uns dieser enorme Zeitaufwand vor eine entscheidende Schlussfolgerung: Es musste von dem ursprünglichen Ziel, sämtliches Sprachmaterial von rund 200 Stunden zu digitalisieren, abgerückt werden, weil es nach der ermittelten Bearbeitungszeit etwa 200 Tage benötigen würde.
Da unser Hauptziel jedoch darin bestand, Texte für das Korpus zu gewinnen und aufzubereiten, entschieden wir uns dafür, vorrangig Gesprächstexte zu transkribieren.

3. Textgewinnung und -aufbereitung

3.1 Die fonetische Transkription

Die fonetische Transkription erfordert ein genaues und mehrmaliges Abhören der Gesprächsaufnahmen, um die Aussprache, d. h. die einzelnen Laute sowie den Wortakzent möglichst präzis zu erfassen. Diesem sind jedoch Grenzen gesetzt: Bei schlechter Tonqualität, undeutlicher Artikulation der Sprecher wird die genaue Erfassung der Aussprache bzw. Laute erschwert. Im Zaza ist dies besonders bei Plosiven der Fall. Die Laute *p, t, k* weisen jeweils vier Varianten auf: *nicht palatal, palatal und nicht ejektiv, ejektiv.*

	nicht ejektiv	*ejektiv*
nicht palatal	[k]	[k’]
palatal	[kʲ]	[k’ʲ]

Die gesprochenen Texte, welche digitalisiert und auf CD übertragen wurden, sind auf einem CD-Spieler abgehört und handschriftlich transkribiert worden. Die Transkription erfolgte nach dem internationalem Standard IPA. Dabei wurde die Wortbetonung, Nasalität, Ejektivität und Palatalisierung, welche im Zaza existieren, berücksichtigt. Bei der fonetischen Verschriftlichung sind mehrere Hilfszeichen benutzt worden, um zusätzliche Informationen über den Gesprächsablauf sowie dessen Elemente zu geben.

3.1.1 Verwendete Textmarkierungen

Besondere Fälle in gesprochenen Texten sind durch Markierungen ausgezeichnet worden. Dazu gehören solche Stellen:

unverständlich, undeutbar, lückenhaft, versprochen, Sprechpause, Fremdwörter

1. unverständliche Stellen

(Text) Nur teilweise verständliche, hörbare Wörter, meistens Endsilben, werden in runde Klammern gesetzt. Beispiel aus dem Text *Ots 001:*

<16> ... Wɛr'te ʦor: 'ãsmi (də) hok'mat	<16> ... Innerhalb von vier Monaten wurde die

'darija wɛ.	Regierung(sherrschaft) beseitigt.

2. undeutbare Stelle und Lücke

_ _ _ Akustisch unverständliche bzw. undeutbare Stellen und Tonlücken werden mit drei untergesetzten, getrennten Bindestrichen gekennzeichnet. Tonlücken von mehreren Silben oder Wörtern müssen als solche markiert werden. Undeutbare Stellen an den Endsilben können aufgrund des Kontextes und der Sprachkompetenz meistens identifiziert werden, aber die Aussprache kann nicht genau erfasst werden. Daher werden solche Textstellen im Zaza-Korpus durch Einsetzung in runde Klammern gekennzeichnet. Beispiel aus dem Liedtext *L 009 Ismail 2:*

soga'jiga 'sɛwti'malə',	Es ist Sogayige, die Verbrannte,
ja, 'ʤigɛre mɨ, pul u dʒyno.	Ja, mein Herz, es ist Hügel und Dreschplatz.
_ _ _ 'sɛro 'lɛmɨ,	_ _ _ auf, o weh,
ʒe mal u ga'u 'nino.	Es kommt nicht wie Kleinvieh und Rinder.

3. versprochene Textstelle

~ Das Tilde-Zeichen gibt an, dass Sprecher/in sich verspricht. In den gesprochenen Texten kommt gelegentlich auch vor, dass Sprecher sich manchmal versprechen, aber wieder korrigieren. Solche Textdeffekte zu markieren ist zweckmäßig, damit

derartige Passagen hinsichtlich der linguistischen Analyse oder des Spracherwerbs nicht als Regelfall übernommen werden. Beispiel aus dem Märchen *M 008 Lazek ve hokımdare ra:*

'vano 'niạdə, ɛz pa'sao; 'hɛte mi də vɨ'lɛ na ro, ɛz hu'rɛndi də zõno. tɨ ki ~, sɨ'ma ki [qa'nat]e 'mɨne.	(Er) sagt ‚siehe, ich bin König; auf meiner Seite legte (er) den Hals hin (und) schätzte mich. (Und) du ~, ihr seid meine [Flügel].

4. Sprechpause

… Drei Punkte markieren die Sprechpause. Beispiel aus dem Text *Ots 001:*

'vakʾə "vazə!' 'vakʾə 'ha ha'lene q... <ʤi'nɨkə ku'xɛna> Ha'lene ... õn'der a 'ʦɨka? <ʤi'nɨkə:> <5> 'qɛzə, 'qɛzə!	sagte „sag es!" (Er) sagte „dort im Nest von E ... <Frau hustet> (Im) Nest ... wie heißt denn jene verdammte? <Frau> <5> Ente, Ente!

5. Fremdwörter und fremdsprachigen Zitate

[Text] Fremdwörter – meistens türkisch -, deren Zazaform zwar besteht, aber von Sprechern aufgrund ihrer Mehrsprachigkeit nicht bevorzugt wird, werden in eckige Klammern gesetzt. Bei mehrsprachigen Sprechern – vor allem bei männlichen – kommt vor,

dass Fremdwörter aus dem Türkischen benutzt werden, obwohl deren Zaza-Äquivalenz existiert und verwendet werden könnte. Bei weiblichen Sprechern ist der Gebrauch von türkischen Fremdwörtern relativ gering.
Manchmal treten auch fremdsprachige Zitate – meistens türkisch – auf, welche durch Einschließung in eckige Klammern markiert werden. Beispiel aus dem Text *Ots 001:*

<31> ɛj ɛkʾə ˈr:aə ma sɛr ˈnearda, ˈvakʾə ‘[sɛn gʲɛl buriˈja, gʲɛl buriˈja!]’ ˈʦao kʾə o ʤeˈra wɛ, ma võz da.	<31> Er hat also den Weg zu uns nicht gefunden und sagte „[komm du hierher, komm her!]“ In dem Moment, als er aufwärts ging, liefen wir weg.
<52> … ˈAla ˈvɨndə, [ˈʣanɨm]!	<52> ... Sei doch ruhig [Mensch]!

6. Zusätzliche Informationen

<Text> Anmerkungen zu Sprechernamen bei Dialogen werden in spitze Klammern gesetzt.

<Zahl> Laufende Textnummern werden bei Märchen, Erzählungen u. ä. Gesprächstexten etwa nach allen drei Sätzen in spitze Klammern gesetzt. Dies ist vor allem bei gedrucktem Text, aber auch in elektronischem hilfreich, um die eine gesuchte Belegstelle leicht zu finden. Vgl. dazu obige Beispiele.

3.2 Gefertigte Transkripte

Aufgrund der Arbeitsteilung, die Transkription und die Notation parallel durchzuführen (s. u. Abschnitt 5), ergab sich die Notwendigkeit, vorrangig die Lieder zu bearbeiten. Daher ist die Anzahl der Transkripte bei den Liedern höher (54) als bei anderen Textarten. Andererseits ist die Textlänge von manchen Märchen größer als bei den Liedern. Die Texte der handschriftlichen Transkripte wurden mit dem internationalen Standard, dem Unicode-Zeichen in PC eingegeben.
Die bisher gefertigten Transkripte haben eine Gesamtlänge von rund 10000 Textwörtern.
Bisher sind 70 Texte fonetisch transkribiert und 16 davon ins Deutsche übersetzt worden:

Abkürzung Textart	Textart	Anzahl Texte	Anzahl Übersetzung
L	Lieder	54	8
M	Märchen	9	2
E	Erzählungen	2	2
Gtel	Telefongespräch	1	-
Ots	Genozid-Bericht	1	1
Spw	Sprichwörter	1	1
A	Anekdoten	2	2
	Summe	70	16

3.3 Übersetzung ins Deutsche

Bei der Übertragung der Zazatexte ins Deutsche ist als Orientierung folgender Aspekt beachtet worden:

1. Lexikalisch: in der Zielsprache möglichst dieselbe Entsprechung oder ein dem Original näheres Synonym zu benutzen.
2. Syntaktisch: gleichen oder ähnlichen Satzaufbau anzustreben.
3. Zusätzliche stilistische Verbesserung oder Ergänzung vermeiden.

Eine Übertragung nach diesen Kriterien hat den Vorteil, das Sprachverständnis des Originals in den Vordergrund zu stellen, was für die linguistische Forschung - vor allem Lexikographie, u. a. – von Nutzen ist. Die stilistische Wiedergabe im Deutschen wurde nur sekundär berücksichtigt.

3.4 Fonemisierung der fonetischen Transkripte

Die fonetische Verschriftlichung des gesprochen Zaza hat den Vorteil, die Aussprache exakt zu dokumentieren. Dies befriedigt vor allem die Bedürfnisse der Linguisten im Hinblick auf die ausführliche grammatische, dialektologische, und lautgeschichtliche Untersuchungen.
Für die Forscher anderer Fachgebiete wie Literaturwissenschaft, Ethnologie, Geschichte u.a. sowie weiterer Zielgruppen wäre die Benutzung fonemischer Korpustexte äußert schwierig. Denn diese sind nicht an der wissenschaftlichen Erforschung des Zaza, sondern an den gesellschaftlichen Informationen der Korpustexte interessiert. Für diese Benutzergruppe wäre die fonemische bzw. orthografische Schreibung geeignet.

Mit der fonetischen Schreibung wäre das *Zaza-Deutsche Textkorpus* nur auf einen kleinen Kreis von Linguisten beschränkt und würde die Benutzung durch weitere Zielgruppen sehr erschweren. Damit auch diese das Korpus leichter nutzen können, sollten die Texte zusätzlich mit fonemischer Schreibung aufbereitet werden.

Zu diesem Zweck wurden fonetische Verschriftlichungen mittels Programmierung in fonemische Schreibung umgewandelt. Insofern stellt die automatisierte Fonemisierung von fonetischen Transkripten eine Bereicherung des Korpus dar.

Nach der Erfassung der fonetischen Transkripte am PC (s. o. Abschn. 4), sollen diese zusätzlich fonemisiert werden. Auf der nächsten Seite wird ein Text in beiden Schreibformen mit Übersetzung gezeigt.

Cani cani	**Cani Cani**	**'ʤani 'ʤani**
Namiê ma Mercano,	Diesseits von uns ist Mercan,	'Namie ma Mɛr'ʤano,
Domiê ma Tercano.	Jenseits von uns ist Tercan.	'Domie ma Tɛr'ʤano
Ez ke yare dawete keri,	Wenn ich die Liebste einlade,	Ɛzkʾə 'jarə da'wɛtə kɛri,
Kam mı ra se vano.	Wer kann mir was sagen.	Kam mɨ ra sɛ 'vano.
Gozagunê şêney rake,	Knöpfe deine Brust auf,	Goza'gune ʃe'nɛj 'rakʾə,
Heto zu mêẓidano, zu çêregano.	Eine Seite ist (bestückt mit) Mez,idia, eine Seite (mit) Çêrega	'Hɛto zu meʒi'dano, zu ʧerɛ'gano.
Erê cani cani, Melema mı cani.	Du, die Seele, Mein Balsam, die Seele.	Ɛ're 'ʤani 'ʤani, Mɛ'lɛma mɨ 'ʤani.

Fekê to qutiye, Dıdonê to mırcani.	Dein Mund (ist wie) die Dose, Deine Zähne (wie) Perlen.	ˈFɛkʲe to quˈtijə, Dɨˈdõne to mɨrˈʤani
Çımê şiaê gılori, Buri seɨa qeytani.	Die schwarzen runden Augen, Die Brauen darüber ein Bogen.	ˈʦɨme ˈʃi̯ae gɨˈlori, Buˈri ˈsɛrːa qɛjˈtani.
Suɨeta to vêsena, Soa Gumısxani.	Dein Gesicht strahlt, (Wie) der Apfel von Gumisxan	Suˈrːɛta to veˈsɛna, ˈSõa Gumɨsˈxani.
Pırnıka to ṗona, Qundaxê yunani.	Deine Nase ist platt, (Wie) Yunans Kolben.	Pɨrːˈnɨka to pʾõna, Qunˈdaxe juˈnani.
Poɨê tuyo nermo, İpegê suḳa Vani.	Dein Haar ist weich, (Wie) die Seide der Stadt Van.	ˈPorːe tujo ˈnɛrmo, İˈpɛgʲe ˈsukʾa ˈVani
Saɨê to sero wanino, Yetımxanê Elemani.	Auf deinem Kopf ist zu lesen: Deutsches Waisenhaus.	Saˈrːe to ˈsɛro waˈnino, Jɛtɨmxaˈne Ɛlɛˈmani.
Namê to gırano, Nêşkinu wedari.	Dein Name ist edel, Ich kann es nicht aussprechen.	Naˈme to gɨˈrano, ˈNeʃkinu ˈwɛdari.
Şênê to şis keno, Voɨa Koê Mırcani.	Deine Brust strahlt weiß, (Wie) der Schnee von Mırcan-Berg.	ʃeˈne to ʃis ˈkɛno, ˈVorːa ˈKoe Mɨrˈʤani.

3.5 Morfemische Texte

Hinsichtlich der vielseitigen Nutzungsmöglichkeiten der Zaza-Texte wurde auch der Versuch unternommen, zur Durchführung ausführlicher grammatischer Untersuchungen auf der Korpusbasis, Texte auch mit morfemischer Schreibung aufzubereiten. Zu diesem Zweck ist ein Transkript manuell in morfeme zerlegt worden. Die Praxis ergab, dass eine

derartige Textaufbereitung sehr zeitaufwendig ist und soll im Rahmen dieses Projekts lediglich als ein Versuch betrachtet werden. Bisher wurde ein Text morfemisch aufbereitet: Dɛ're 'Laʧi.

Dɛ'r-e 'Laʧ-i	**Das Latsch-Tal**
Də 'wɛla 'wɛla, 'Hal-e ma 'jaman-o.	O weh, o weh, Unsere Lage ist furchtbar.
Or:'di gur'laɣ a'm-o, Dor'm-e ma qa'pan-o.	Das Militär mächtig angerückt, Wir sind umzingelt.
'Bɛxt-e Hɛj'dɛr u Dɛ'mɛn-i re, Kʲes xɨ'rav-ə 'ne-van-o.	Über die Heyderiden und Demeniden Sagt man nichts schlechtes.
Or:'di-j-e 'Tɨrk-i 'zãf-o, 'ʤa-e wɛ'laɣ-ə ma 'ne-da-n-o.	Die türkische Armee ist groß, (Und) gibt uns keinen freien Weg.
Ma zu'vini qɨr: 'kɛ-mə, 'ʦɛm-e Mu'zɨr-i ʤɛn'dɛg u lɛs-'u 'a-n-o.	Wir schlachten uns gegenseitig; Der Muzır-Fluß trägt Leichen und Tote.
Õn'der-i də 'da-mə 'pero, Tɛ'dɛ ʃin u ʃi'wan-o.	Wir kämpfen in dem Verdammten, Darin ist Klage und Trauer.
Dɛ'r-e 'Laʧ-i 'bɨ-ves-o, İ'vɨs-e mɨ ga'van-o.	Verbrennen soll Latsch-Tal, Mein İvıs, es ist (ein) Engpass.
'Bɨra 'pero-d-e, na qɛw'ɣ-a a'ʃir-ə 'ni-j-a,	Kämpft Brüder, dies ist kein Stammeskrieg,
Mɛrɛ'v-e Kɨrman'ʤ-un u zaɬɨ'm-an-e Tɨr'kʾ-an-o.	Es ist Kampf der Kirmanc und der türkischen Tyrannen.
'Dɛst-e xo ra xo 'mɛ-ʤer-e, Sar ma re qo'la-ə 'va-n-o.	Steht nicht da mit leeren Händen; Man wird schlechtes von uns sagen.

P'ɛ'p'ug 'ber-o 'bɨ-nis-o,	Der (Vogel) Wiedehopf käme (und) niste ein,
ʤen'ʣ-un-e ma re 'bɨ-wa-n-o.	Und singe (trauernd) über unsere Jugendlichen.
Qɛ'mɛr-e Hɛ'sɛn-i 'vɛr-e mɨɣa'ra də gɨ'n-o 'war:o,	Qemer (Sohn) von Hesen ist vor der Höhle gefallen,
'Mal-o 'ʃer-e 'mɨn-o bɛ'ran-o.	O, mein rasender Löwe.
Hɛs-e Kal-i 'ku-n-o qɛwɣa,	Hesê Kali geht in den Kampf,
Bɛʃ'li-u̯ və 'doʃi sɛr 'a-n-o.	(Und) trägt die Gewehre auf der Schulter herbei.
'Hɛm-e 'ʣɨv-e 'Kʲeʒ-i pɛr's-ɛn-e,	Fragt ihr nach Hem dem Sohn von Dzive Keji,
'Xɨsm-e or:'di u tawɨ'r-an-o.	Er ist Gegner von Armeen und Bataillonen.

4. Notation der Zaza-Lieder

4.1 Literarischer Aspekt

Nach der Tradition werden Empfindungen über die Ereignisse gesellschaftlicher, politischer und individueller Art in Lieder und Dichtungen zum Ausdruck gebracht. Die (Volks-)Sänger und Poeten erzählen in ihren Liedern und Dichtungen das Ereignis von Anfang bis Ende, wobei meistens auch Einzelheiten über die daran beteiligten Personen sowie deren Beweggründe, die gesellschaftliche Bewertung und den Zeitgeist in wohlgeformter Artikulation.

Insofern erscheinen die Zaza-Lieder nicht nur als literarisches Zeugnis, sie konservieren auch Informationen über die Sozialgeschichte. Damit bilden die Lieder eine informative Quelle für die Oralhistory der Zaza. Hier ein typisches Beispiel, in dem Informationen zum Korea-Krieg (1950) geliefert werden:

Mame'k'ija, Mame'k'ija, 'Dae, lɛ'mɨn, Mamɛ'k'ija.	Es ist Mamekiye, es ist Mamekiye, O Mutter, o weh, es ist Mamekiye*.
Za'ɬɨmi ma 'dajmə a're, 'bɛrdimə, Dɨsɛup'õn'ʤas mor'dɛm 'pija.	Der Tyrann trieb uns zusammen (und) brachte fort, Zweihundertfünfzig Leute zusammen.
	Mamekiye ist der einheimische Name der Stadt *Tunceli.*

Hier erfährt man, dass aus Dersim 250 Personen als Soldat durch die türkische Regierung zum Korea-Krieg geschickt wurden. Diese klare Information ist aus schriftlichen Quellen selten, vielleicht auch gar nicht zu erfahren. Weil in der Zaza-Gesellschaft keine Schriftkultur existierte, erhalten die Lieder neben ihrem literarischen auch eine besondere Bedeutung als historisches Zeugnis. Die nächsten Verse bezeugen, dass am Anfang des 20. Jahrhunderts ein *Deutsches Waisenhaus* im Zazaland existiert hat.

Sa'r:e to 'sɛro wa'nino: Jɛtɨmxa'ne ɛlɛ'mani.	Auf deinem Kopf ist zu lesen: Deutsches Waisenhaus.

'Por:e tujo 'nɛrmo,	Dein Haar ist weich,
İ'pɛgʲe 'sukʾa 'Vani.	(Wie) die Seide der Stadt Van.

Die traditionelle Zaza-Dichtung besteht in der Regel aus Zweizeilern. Die zweiten Verse sind gereimt. Die als Endreim benutzte Silben sind verschiedener Art, welche verschiedene Strophen bilden. Beispiel aus dem Liebeslied *L 046 Cani cani:*In seltenen Fällen kommen auch solche Lieder vor, die aus Dreizeilern bestehen. Dies beruht vor allem auf der besonderen Form der musikalischen bzw. melodischen Ausdruckweise. Als Beispiel zum Dreizeiler-Lied sei das Lied von *Hemed (Xosım),* gesungen von Sıleman aus Yeresk, zu nennen.
Ein anderes Merkmal der Zaza-Lieder ist, dass sie keine feste Verslänge haben.
Bei erzählenden Liedern liegt der Schwerpunkt vielmehr auf dem Text als auf Melodie. Hier gilt nämlich, dem Zuhörer den Ablauf der Handlung, die Zusammenhänge und das Mitgefühl sprachlich, eingebettet in eine Melodie, zu artikulieren.

Beim Gesang wird die Länge der Verse durch Variierung der Tonlänge einander ausgeglichen. Kürzere Zeilen werden durch Dehnung einer oder mehrerer Silben und längere Verse in schnellerem Tempo auf derselben Tonhöhe gesungen.

4.2 Musikalischer Aspekt

Bisher konnten 100 Lieder durch Notation musikalisch dokumentiert werden. Die Zaza-Lieder sind durch eigenartige Struktur von Melodie und Rhythmus gekennzeichnet. Die meisten Lieder haben eine ungleichmäßige

bzw. freie Taktart. Die ungleichmäßigen Rhythmen bestehen aus den Takten 6/8, 7/8 u. ä. So wird z. B. der Takt 9/8 2 + 2 + 2 + 3 benutzt. Diese dreier Gruppe kann natürlich am Ende, manchmal am Anfang oder sogar in der Mitte erscheinen. Bei der europäischen Musik wird der 9/8-Takt in der Regel in Form von 3 + 3 + 3 verwendet. Bei der türkischen kommt dies als 2 + 2 + 2 + 3 vor. Weil bei der Zaza-Musik der Text eine entscheidende Rolle spielt, kann die Dreier Gruppe ihre Stellung, je nach der Taktart wechseln. Ein anderes interessantes Merkmal ist, dass in einem Lied bei einem Takt neuere Taktzahlen eingefügt werden. Die Notation der Zaza-Musik wurde nach dem Grundsatz, möglichst originalgetreue Takte zu benutzen, durchgeführt. Das in der Musikbearbeitung übliche Verfahren, symmetrische, d.h. ständig gleichbleibende Taktzahlen zu benutzen, wurde vermieden.
Die Zaza-Lieder lassen sich nach ihrer rhythmischen Struktur in zwei Hauptgruppen einteilen: Lieder mit bestimmter und unbestimmter / freier Taktart.

4.2.1 Lieder mit bestimmter Taktart

Diese Lieder werden in einem bestimmten Rhythmus gesungen und gespielt. Bei der Notation sind sie durch Taktstriche und Abschnitte geteilt. Ein Teil dieser Lieder kommt mit gleichbleibendem Takt vor, während ein anderer Teil aus unterschiedlichen Takten zusammengesetzt ist.
Die Erscheinung von verschiedenen Taktarten in einem Lied ist meistens durch unterschiedliche Silbenzahl der Zeilen bedingt. Die Verschiedenheit der Verslänge führt zu unterschiedlichen Taktzahlen. Die Gleichheit der Silbenzahl in den Zeilen hätte darin auch gleiche Taktzahl zur Folge. Unter diesem Aspekt ließe sich die Liedergruppe mit bestimmter Taktart in gleich- und ungleichmäßige Gruppe zwar unterteilen, aber dies ist jedoch

primär sprachlich, d. h. durch Verschiedenheit der Silbenzahl der Zeilen bedingt, und nicht musikalisch: ungleichmäßiger Takt, gleichmäßiger Takt. Rund 77 % der bearbeiteten Stücke hat bestimmte Taktart, von denen 40 % einen gleichmäßigen und 37 % einen ungleichmäßigen Takt aufweisen.

4.2.2 Lieder mit unbestimmter / freier Taktart

Diese Lieder besitzen zwar einen eigenen Rhythmus, aber es ist nicht einfach, diesen genau zu bestimmen. Die Grundmelodie ist im Allgemeinen deutlich und die Spieler bzw. Sänger können dennoch ihrer Emotion und Fähigkeit entsprechend den Takt variieren. Bei den Melodien können Dehnung wie Verkürzung frei angewandt und mit dem Instrument Zwischenmusik unterschiedlicher Länge gespielt werden. Bei der Notation solcher Lieder werden Taktzahl und Taktstrich weggelassen, weil es als eine angenäherte Notation zu betrachten ist.
Etwa 20 % der dokumentierten Lieder gehören zu der Gruppe mit unbestimmter bzw. freier Taktart. Die Zahlenangaben hier dienen nur zur Erläuterung der rhythmischen Struktur der dokumentierten Stücke und können nicht für die Gesamtheit des Liederbestandes verallgemeinert werden. *Cuŕe* und *Ṭomır*

4.2.3 Probleme und Methoden der Notation

Die herkömmliche Notationsform von jeder Musik Anatoliens ist aus heutiger Sicht und Erfahrung unbefriedigend. So wurde bei der Ermittlung des Bestandes türkischer Volksmusik an rund 10 000 Liedern Notation durchgeführt, aber ernsthafte notationstechnische Probleme dabei nicht beseitigt. Das unter Zazas, Türken und Kurden am meisten benutzte Musikinstrument ist *Ṭomır* und seine Variante *Cuŕe (*tür. *saz, bağlama).*

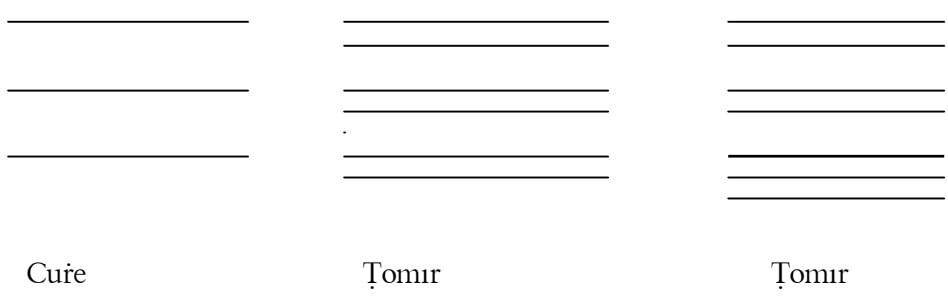


Bild 1: Saitenanordnung der Zupfinstrumente

Bei *Cuŕe* und *Ṭomır* werden die Saiten auf dem Instrument in dreier Form aufgespannt: *unten, mitte, oben*.
Die Saiten können nach der Reihenfolge *unten – mitte - oben* aus der Kombination 1–1–1, 2–2–2 und 3–2–2 bestehen.
Die bei traditionellen Zazaliedern benutzten Instrumente sind Geige und das Zupfinstrument *Cuŕe*, welches drei Saiten hat: 1–1–1 (Abbildung 1).
Jede Saite wird mit unterschiedlichem, aber bestimmtem Ton so gestimmt, das beim Zupfen aller drei Saiten ein harmonischer Klang entsteht. Beim Spielen werden in der Regel die drei Saiten gleichzeitig gespielt.

4.2.4 Die traditionelle und die angewandte Notationsform

Bei der traditionellen Notationsart wird aber *nur einer* von den drei Saitentönen geschrieben, während die beiden anderen Töne weggelassen werden. Es wird so geschrieben, als wenn nur eine einzelne Saite gespielt wird. Daher ist die bisherige Notationsform unvollständig.
Um die musikalische Dokumentation von Zaza-Liedern möglichst genau und vollständig durchzuführen, wurden Töne aller drei Saiten berücksichtigt. Denn nur so lassen sich gehörte Töne auch originalgetreu dokumentieren. Ein wichtiger Vorteil dieser Schreibweise wäre z.B., den Musikern, vor allem solchen, die *Cuṙe* oder *Ṭomır* spielen, zu ermöglichen, die Stücke originalnah nachzuspielen. Auf dem folgenden Notenbild ist die angewandte Notenschreibung der traditionellen Form gegenübergestellt.

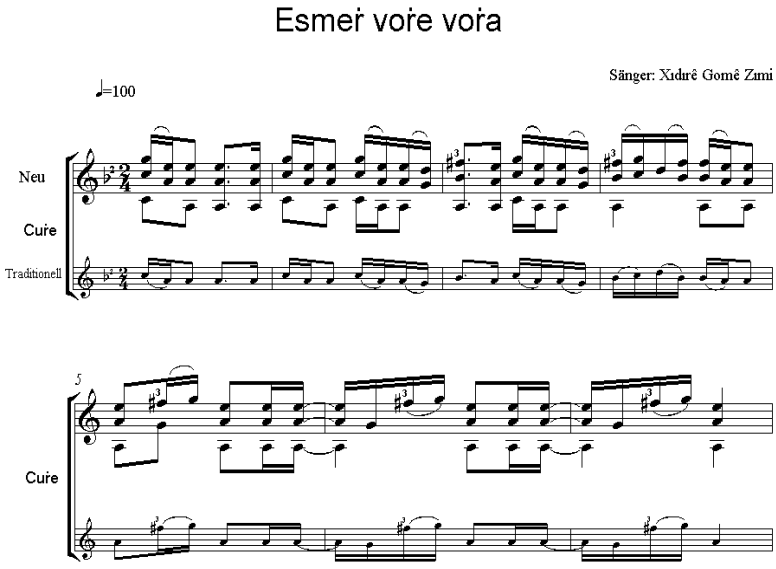
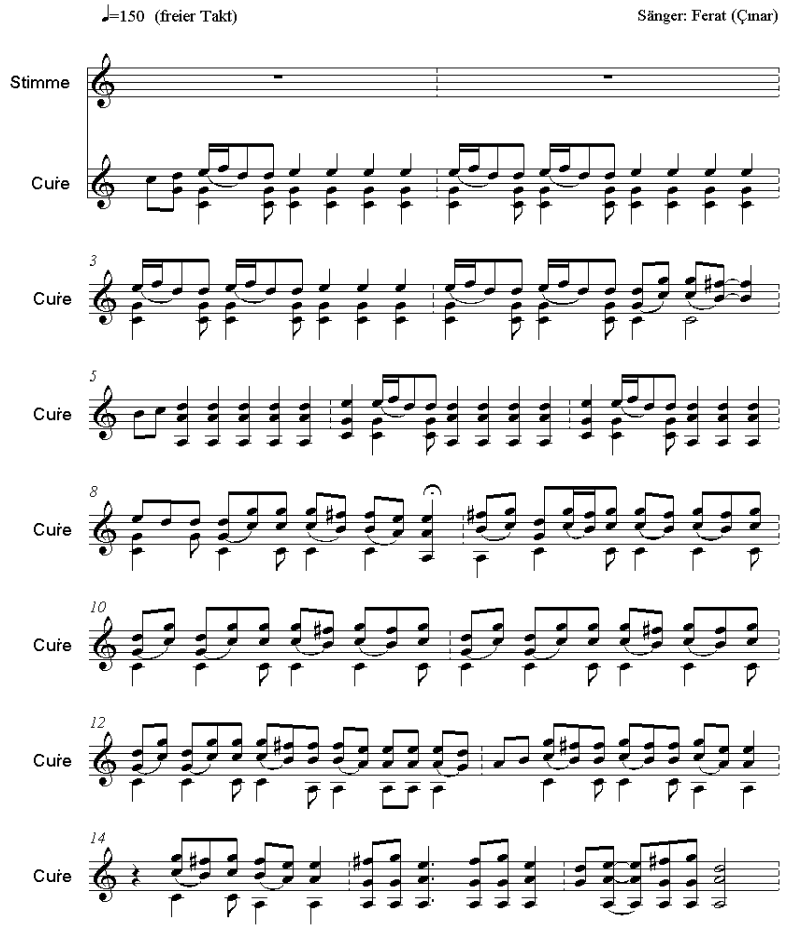
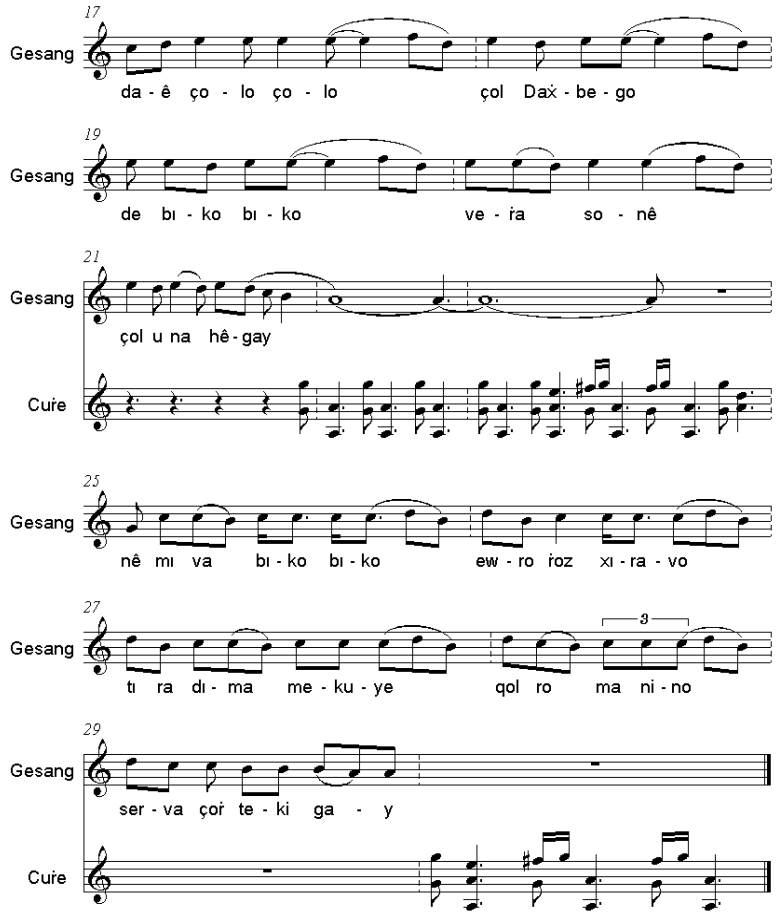


Bild 2: Gegenüberstellung der traditionellen und der neuen Notation von *Cuṙe*

Daẋbe


♩=150 (freier Takt)
Sänger: Ferat (Çınar)
Stimme
Cuŕe
3
Cuŕe
5
Cuŕe
8
Cuŕe
10
Cuŕe
12
Cuŕe
14
Cuŕe

2


17
Gesang
da - ê ço - lo ço - lo çol Daẋ - be - go
19
Gesang
de bı - ko bı - ko ve - ɨa so - nê
21
Gesang
çol u na hê - gay
Cuɨe
25
Gesang
nê mı va bı - ko bı - ko ew - ro ɨoz xı - ra - vo
27
Gesang
3
tı ra dı - ma me - ku - ye qol ro ma ni - no
29
Gesang
ser - va çoɨ te - ki ga - y
Cuɨe

5. Die sprachlichen Häufigkeiten

Die Fragestellung, nach welcher empirischen Gesetzmäßigkeit die Häufigkeit von Wortformen, Fonemen und Wortlängen auftreten und wie diese mathematisch begründet werden können, ist anhand von Korpusdaten untersucht worden. Außer den Zaza- und deutschen Texten sind zusätzlich Daten von Englisch und Spanisch einbezogen worden, um eine allgemeinere Schlussfolgerung aus den gewonnenen Erkenntnissen zu erzielen.

5.1 Die Wortformenhäufigkeit

Aus dem Korpusanteil des Zaza mit 10059 Textwörtern (tokens) ergaben sich 2654 Wortformen (types) und aus dem deutschen Übersetzungstext mit 9249 Textwörtern resultierten 2048 Wortformen. Die Ermittlung der Wortformenhäufigkeit des Zaza im Rahmen dieses Projekts stellen den Beginn einer solchen Untersuchung dar, während dies im Deutschen sowohl in einzelnen Arbeiten als auch in zahlreichen Korpora bedeutend fortgeschritten sind.

Die Erforschung der Wortformenhäufigkeit im Deutschen geht schon auf das Jahr 1897/98 zurück, in dem F. W. Kaeding das *Häufigkeitswörterbuch der deutschen Sprache* veröffentlichte. Kaedings Wortzählung ist von H. Meier 1964 als eine überarbeitete und fonemisch transkribierte *Rangbuch der geläufigsten deutschen Wortformen* publiziert. W. D. Ortmann veröffentlichte 1975 7995 hochfrequente (grafemische) Wortformen aus der Kaeding-Liste in verschiedenen Anordnungen. Das Gesamtkorpus Kaedings umfasste insgesamt 10910777 Wortformen (Ortmann, S. 3).

Tabelle 4 und Tabelle 5 zeigen die auf Korpusdaten beruhende Wortformen im Zaza und im Deutschen in absteigender Reihenfolge der Häufigkeit. Danach sind die ersten drei häufigsten Wortformen sind im Zaza *mı* '*1. Person sg. obliquus:* ich, mich, mir, ...', *ra* '*Postposition:* von' und *ma '1. Person pl. nom./obl.:* wir, uns'. Im Deutschen sind diese (nach diesem Korpus): *(und), der, die.* In der Wortliste von Kaeding und Ortmann ist das erste häufigste Wort jedoch *die* und die ersten drei Ränge bestehen aus *die, der, und,* welche mit dem des Zaza-Korpus übereinstimmen. Allerdings ist die Rangordnung im Zaza-Korpus verschoben, was auf die Besonderheit der Zaza-Texte und deren Übersetzungsform zurückzuführen ist. Weil im gesprochenen Zaza das Satzverbindungselement *und* fehlt, wurde es in der Übersetzung meistens mit in Klammern gesetzter Form wiedergeben. Dieser Zusammenhang ist auch in der Wortliste dadurch gekennzeichnet worden, indem es in Klammern gesetzt wurde.

Eine interessante Fragestellung ist, wie viel bzw. welche Wortformen *die Hälfte* des (laufenden) Korpustextes bilden. Die Antwort dazu liefert die Summenhäufigkeit in Tab. 4 und Tab. 5: bei der Summenhäufigkeit von 0,5, d.h. der Texthälfte, wird im Zaza von den ersten 160 Wortformen und im Deutschen von 80 Wortformen erfasst. 80 % (0,8) des Textes wird im Zaza von bis zu 900 Wortformen, und im Deutschen von 600 gebildet.

Die von der Häufigkeitsliste gelieferten Informationen auf verschiedene Fragestellungen sind hier zwar hilfreich, es sollte jedoch erwähnt werden, dass daraus noch keine allgemeine Schlussfolgerungen gezogen werden sollten, weil das Zaza-Deutsche Korpus sich noch in der Aufbaufase befindet und erst nach Realisierung eines Korpusumfangs von etwa 1 Million Textwörtern einen Anspruch auf Repräsentativität erheben kann. Dennoch ist die Aussagekraft trotz genannter Einschränkung tendenziell richtig.

Zur Verdeutlichung der Zusammenhänge von sprachlichen Häufigkeiten sind auch englische und spanische Texte einbezogen und untersucht worden. Dazu diente ein englischer Text aus dem ACE-Korpus mit 41869 laufenden Textwörtern (5998 Wortformen) und ein spanischer Berichtstext aus dem Internet mit 3348 Textwörtern (1261 Wortformen) benutzt. Die aus diesen Texten resultierenden Wortformen und deren Häufigkeiten sind in Tab. 6 und Tab. 7 aufgelistet. Das häufigste Wort ist im Englischen *the* und im Spanischen *de* '*Präposition für Herkunft, Herbewegung ...*: von, aus, ...'. Bemerkenswert ist auch, dass die Texthälfte, d. h. bei der Summenhäufigkeit von 0,5, im Englischen von den ersten 800 Wortformen gebildet wird, während dies beim Spanischen schon mit 60 Wortformen erfolgt.

Die sich aus fallender Häufigkeit in Tab. 4. und Tab. 7 ergebende Reihenfolgennummer wird den entsprechenden Wortformen als Rangzahl zugewiesen. Ordnet man den Wortformenrang *x* und die zugehörige relative Häufigkeit *h(x)* zueinander, so erhält man die grafischen Darstellungen in Bild 3, 4, 5.

Aus den empirischen relativen Wortformenhäufigkeiten sind durch Addieren auch die rel. Summenhäufigkeiten errechnet und in der Spalte 6 der Tab. 4 und Tab. 5 aufgeführt und in Bild 6 und Bild 7 dargestellt, woraus der Verlauf der empirischen Daten sichtbar wird. Die nächste Aufgabe besteht darin, den mathematischen Zusammenhang zwischen *x* und *h(x)* bzw. *H(x)* so zu bestimmen, dass die Ausgleichskurve durch die empirischen Werte verläuft.

Zu Beginn der Berechnungen wurde zuerst von der Annahme $h(z) = h_0 e^{u(z)}$ ausgegangen - $h(x = 1, z = 0) = h_0$, $z = \ln x$ - und versucht, aus der umgestellten Gleichung $\ln \frac{h_0}{h(z)} = u(z)$ den empirischen

Verlauf von *u(z)* zu ermitteln. Als nächstes wurde eine Näherungskurve angesetzt, wobei einige Ansätze zwar zu befriedigenden Resultaten führten und es sich dadurch eine Gleichung für die rel. Häufigkeit *h(x)* aufstellen ließ. Aber von dieser auch die rel. Summenhäufigkeit $H(z) = \int h(z)dz$ zu bilden, war nur durch komplizierte und aufwendige rekursive Integralformeln erreichbar, was nicht effektiv war. Dies führte zu der Einsicht, nicht von *h(z),* sondern von *H(z)* ausgehend eine Regressionskurve aufzustellen. Als Annahme wurde die Exponentialfunktion $H(z) = h_0 e^{u(z)}$ und deren Umformung $\ln \frac{H(z)}{h_0} = u(z)$ benutzt. Aus den empirischen Werten für *h(z), z=lnx, x=1 ... x_{max} ,* ergab sich der Verlauf von *u(z)*, wofür zunächst die Gleichungen $u(z) = az + bz^2$, $u(z) = az + bz^3$ angesetzt wurden, deren Anpassungsgenauigkeit jedoch unbefriedigend war.

Als Maß für die Abweichung zwischen den empirischen und den gerechneten Daten bzw. deren Übereinstimmung gilt bekanntlich der mittlere quadratische Fehler, welcher mit Gl. (1) und (2) ausgedrückt wird: *S* für die Summenhäufigkeit, *s* für die Häufigkeit.

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(H_{i\,emp} - H_{i\,ger})^2} \qquad (1)$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h_{i\,emp} - h_{i\,ger})^2} \qquad (2)$$

Bedeutend präzisere Resultate lieferte der Ansatz $u_3(z) = az + bz^2 + cz^3$ für Zaza, Deutsch und Spanisch. Für Englisch zeichnete sich $u_5(z) = az + bz^3 + cz^5$ hervorragend aus: mit S_5=2,32 10^{-3} gegebüber

S_3=1,11 10^{-2} (Tab. 1). Die Anwendung von $u_5(z)$ auf Zaza, Deutsch und Spanisch führt zwar auch zu ausreichend guten Ergebnissen, aber dessen Abweichungsgrad ist im Vergleich zu $u_3(z)$ etwas schlechter. So ist das Verhältnis S_3 : S_5 bei Zaza 6,04 10^{-3} : 8,58 10^{-3}, bei Deutsch 3,39 10^{-3} : 4,78 10^{-3} und bei Spanisch 5,44 10^{-3} : 8,48 10^{-3} (Tab. 1).

Ähnliches gilt auch für den Häufigkeitsverlauf *h(x)*.

Aus dieser Erfahrung wird deutlich, dass die optimalen Exponenten von *z* in *u(z)* je nach Sprache variieren können. Während *m* zwischen 2 und 3 schwankt, wechselt *n* zwischen 3 und 5, so dass sich daraus vier Kombinationsmöglichkeiten ergeben. Daher wird *u(z)* im folgenden in allgemeiner Form, d. h. mit den Exponenten *m* und *n* geschrieben:

$$u(z) = az + bz^m + cz^n \qquad (3)$$

Man könnte das Polynom *u(z)* durch ein viertes Glied d^q erweitern, dessen Lösung aber zu Determinanten vierter Ordnung führt und damit den Rechenaufwand ziemlich erhöht. Daher ist es zweckmäßig, *u(z)* in Gl. (3), welches mit Determinanten dritter Ordnung leichter lösbar ist, bei drei Gliedern zu belassen. Dabei liefert es eine ausreichend hohe Genauigkeit, so dass zusätzliche Berechnungen nicht dringend erforderlich sind.

Aus dem Ansatz

$$\ln\frac{H(z)}{h_0} = u(z)\,, \quad H(x) = h_0 e^{u(z)} \text{ und } H'(x) = h(x) = h_0 u'(z) e^{u(z)}$$

ergibt sich

$$H_1(x) = h_0 e^{az+bz^m+cz^n} \qquad (4)$$

Darin bedeuten $z = \frac{\ln x}{\ln x_m}$, x=Wortformenrang und x_m der maximale Wortformenrang bzw. Anzahl der Wortformen (s. Tab. 3).
Aus der Ableitung von $H_1(x)$ erhält man die Häufigkeit

$$H'_1(x) = h_1(x) = h_0 \frac{1}{x \ln x_m}(a + mbz^{m-1} + ncz^{n-1})\, e^{az+bz^m+cz^n} \qquad (5)$$

Hier muss Gl. (5) die Forderung der Randbedingung für den Anfang, d. h. bei $x=1$ ($z=0$), $h_1(z=0)=h_0$ erfüllen, was in vorliegender Form noch nicht möglich ist, weil es $h_1(x=1,\, z=0) = h_0 \frac{a}{\ln x_m}$ ergibt.

Dies lässt sich durch ein ergänzendes Glied $H_2(x)$ beheben, welches sowohl in $H(x)$ als auch in $h(x)$ die Anfangsbedingung erfüllt:

$$H_2(x) = h_0(\ln x_m - a)\frac{1}{k}\ln(1+kz) \qquad (6)$$

$$H'_2(x) = h_2(x) = h_0(\ln x_m - a)\frac{1}{x \ln x_m (1+kz)} \qquad (7)$$

Die Konstante k, welche hier $k=100x_m$ gewählt wird, hat die Aufgabe, H_2 und h_2 so zu minimieren, dass sie bei $x>1$ ($z>0$) kaum mehr Einfluss auf die Häufigkeitsverteilung ausüben und vernachlässigbar klein werden.
Mit der Abkürzung

$$r = h_0(\ln x_m - a) \qquad (8)$$

lassen sich diese wie folgt vereinfachen:

$$h_2(x) = \frac{r}{x \ln x_m (1+krz)} \qquad (9)$$

$$H_2(x) = \frac{1}{k}\ln(1+krz) \tag{10}$$

Für den Fall $r < 0$, welcher bei Fonemen vorkommt (s. u.), sollte es bei Gl. (9) im Argument des Logarithmus ein Minus gesetzt werden: *ln(1-krz)*.
Damit gehen die Summenhäufigkeit und die Häufigkeit in folgende allgemeine Form über:

$$H(x) = h_0\, e^{az+bz^m+cz^n} + \frac{1}{k}\ln(1+krz) \tag{11}$$

$$h(x) = \frac{h_0}{x\ln x_m}(a+mbz^{m-1}+ncz^{n-1})\, e^{az+bz^m+cz^n} + \frac{r}{x\ln x_m(1+krz)} \tag{12}$$

Die zweiten Glieder dienen nur dazu, die Anfangsbedingung zu erfüllen, wobei sie mit steigendem x bzw. z rasch auf eine unbedeutend kleine Größe sinken. Schon bei $x=2$ nehmen sie einen minimalen Wert an: $H_2(2)$=(1,5 … 5,6)10^{-5}, $h_2(2)$=(1,2 … 5,7)10^{-6}.

Durch ihr schnelles Sinken haben H_2 und h_2 bei zunehmendem x keinen relevanten Einfluss auf die Häufigkeitsverteilung. Bei $x=x_m$ weisen H_2 und h_2 gegenüber H_1 und h_1 einen verhältnismäßig sehr geringen Wert auf: $H_2(x_m)$=(1,9 … 7,4)10^{-5}; $h_2(x_m)$=(1,8 … 8,8)10^{-6} . Bei x_m ist das Verhältnis h_1/h_2 = 2,3 10^5 … 3,2 10^6.

Demnach wird der Häufigkeitsverlauf in Gl. (11) und Gl. (12) praktisch nur durch das erste Glied bestimmt. Bei maximalem x-Wert, d. h. für $x=x_m$ $(z=1)$ nehmen H und h unter Vernachlässigung der zweiten Glieder folgende Werte an:

$$H(x_m) = h_0\, e^{a+b+c} = 1 \qquad (13)$$

$$h(x_m) = \frac{a + mb + nc}{x_m \ln x_m} \qquad (14)$$

Bestimmung der Konstanten:

Mit Hilfe der Fehlerrechnung lassen sich die Konstanten *a, b, c* in Gl. (11), (12) nach der Methode der kleinsten Quadrate ermitteln.

Bei $\ln \frac{H_i}{h_0} = u_i$ und unter der Annahme $u = u_i + v_i$, ergibt sich v_i, die Abweichung zwischen dem empirischen und dem theoretischen Wert:

$$v_i = u - u_i \qquad (15)$$

Das Quadrat der Abweichung ist dann

$$v_i^{\,2} = u^2 - 2u_i u + u_i^{\,2} \qquad (16)$$

Nach dem Einsetzen von *u* aus Gl. (3) in Gl. (16) wird die Forderung gestellt, dass die Summe der Abweichungsquadrate ein Minimum erreicht. Dieser Fall tritt dann ein, wenn die partielle Ableitung von v_i^2 nach *a, b, c* Null wird:

$$\frac{\partial(v_i^2)}{\partial a} = 0 \;,\; \frac{\partial(v_i^2)}{\partial b} = 0 \;,\; \frac{\partial(v_i^2)}{\partial c} = 0 \qquad (17\ \text{a, b, c})$$

Aus der Ableitung resultiert ein Gleichungssystem mit den drei Unbekannten *a, b, c.* Die in eckige Klammern gesetzten Ausdrücke sind mit der Summe $\sum$ identisch, welche jeweils über $i=1 \ldots x_m$ gebildet sind.

$$[z^2]a + [z^{1+m}]b + [z^{1+n}]c = [u_i z] \qquad (18)$$

$$[z^{1+m}]a + [z^{2m}]b + [z^{m+n}]c = [u_i z^m] \qquad (19)$$

$$[z^{1+n}]a + [z^{m+n}]b + [z^{2n}]c = [u_i z^n] \qquad (20)$$

Mit der Lösung der sich daraus ergebenden Determinanten *D, D_a , D_b ,D_c* können *a, b, c* nach den Gl.n (25, a, b, c) berechnet werden. Dabei müssen die Exponenten *m* und *n* in verschiedenen Kombinationen (s. o.) vorgegeben werden, bis ein Anpassungsoptimum, welches sich bei kleinstem Abweichungsfehler *S* ergibt, erreicht wird.

$$D = \begin{vmatrix} [z^2] & [z^{1+m}] & [z^{1+n}] \\ [z^{1+m}] & [z^{2m}] & [z^{m+n}] \\ [z^{1+n}] & [z^{m+n}] & [z^{2n}] \end{vmatrix} \qquad (21)$$

$$D_a = \begin{vmatrix} [u_i z] & [z^{1+m}] & [z^{1+n}] \\ [u_i z^m] & [z^{2m}] & [z^{m+n}] \\ [u_i z^n] & [z^{m+n}] & [z^{2n}] \end{vmatrix} \qquad (22)$$

$$D_b = \begin{vmatrix} [z^2] & [u_i z] & [z^{1+n}] \\ [z^{1+m}] & [u_i z^m] & [z^{m+n}] \\ [z^{1+n}] & [u_i z^n] & [z^{2n}] \end{vmatrix} \qquad (23)$$

$$D_c = \begin{vmatrix} [z^2] & [z^{1+m}] & [u_i z] \\ [z^{1+m}] & [z^{2m}] & [u_i z^m] \\ [z^{1+n}] & [z^{m+n}] & [u_i z^n] \end{vmatrix} \qquad (24)$$

$$a = \frac{D_a}{D} \ , \ b = \frac{D_b}{D} \ , \ c = \frac{D_c}{D} \qquad (25, \text{a, b, c})$$

Die Konstanten *a, b, c,* die optimalen Exponenten *m* und *n* sowie die mittleren quadratischen Fehler *S* und *s* für die jeweiligen Sprachen sind in Tab. 2 angegeben.

Die empirischen und die gerechneten Werte der Summenhäufigkeit und der Häufigkeit sind für Zaza und Deutsch in Bild 3, 4, 5 sowie in Bild 6, 7 veranschaulicht. Aus den Bildern 6 und 7 wird deutlich erkennbar, welchen Textanteil *H(x)* die Wortformen bis zu einem bestimmten Wortformenrang *x* umfassen. In Bild 10 fällt auf, dass die Kurve der Summenhäufigkeit nach der ca. 1000. Wortform beim Englischen einen geraden Verlauf annimmt. Dies beruht darauf, dass dieser Bereich durch Wortformen mit geringer Häufigkeit belegt ist. Hier zeigt sich ganz besonders, welche hervorragend gute Übereinstimmung mit den empirischen Werten durch Gl. (11) und Gl. (12) erzielt wird.

Bisher wurden die Häufigkeiten in Abhängigkeit des absoluten Wortformenranges *x* dargestellt. Nimmt man aber statt Absolutwert den normierten Wortformenrang $z = \frac{\ln x}{\ln x_m}$, so ergibt sich eine andere Veranschaulichung, in der die gesammte Spannweite der Abzisse mit *z=0…1* erfasst ist. Ein solches Diagramm ist in Bild 11a zu sehen, worin die Häufigkeitsverteilung der behandelten vier Sprachen eingezeichnet ist. Hier wird die sprachspezifische Bedeutung der Maximalhäufigkeit h_o , welche den Verlauf primär bestimmt, besonders deutlich. Sie ist beim Spanischen mit rund 0,077 am höchsten und beim Zaza mit 0,03 am neidrigsten. Bei Deutsch und Englisch sind sie sich sehr nah (s. a. Tab. 2).

Tabelle 1: Vergleich der Abweichung von Ansätzen bei Wortformen

	u(z)	$az+bz^2$	$az+bz^2+cz^3$	$az+bz^3+cz^5$
Zaza	S	$1{,}08\ 10^{-2}$	$6{,}04\ 10^{-3}$	$8{,}58\ 10^{-3}$
	s		$2{,}44\ 10^{-4}$	$3{,}51\ 10^{-4}$
Deutsch	S	$1{,}36\ 10^{-2}$	$3{,}39\ 10^{-3}$	$4{,}78\ 10^{-3}$
	s		$1{,}16\ 10^{-4}$	$1{,}99\ 10^{-4}$
Englisch	S	$4{,}56\ 10^{-2}$	$1{,}11\ 10^{-2}$	$2{,}32\ 10^{-3}$
	s		$1{,}43\ 10^{-4}$	$8{,}35\ 10^{-5}$
Spanisch	S	$2{,}82\ 10^{-2}$	$5{,}44\ 10^{-3}$	$8{,}48\ 10^{-3}$
	s		$3{,}82\ 10^{-4}$	$6{,}81\ 10^{-4}$

Tabelle 2: Konstanten und Abweichungen der Wortformenhäufigkeit

	h_o	a	b	c	m	n	S	s
Zaza	0,02982	7,03245	-5,11763	1,59422	2	3	$6{,}040\ 10^{-3}$	$2{,}443\ 10^{-4}$
Deutsch	0,04271	6,18429	-3,42716	0,39059	2	5	$2{,}599\ 10^{-3}$	$1{,}264\ 10^{-4}$
Englisch	0,04130	5,06278	-4,82223	2,95100	3	5	$2{,}318\ 10^{-3}$	$8{,}357\ 10^{-4}$
Spanisch	0,07706	5,95404	-6,46764	3,06657	2	3	$5{,}444\ 10^{-3}$	$3{,}825\ 10^{-4}$

Tabelle 3: Wortzahlen und die mittlere Wortform

	Anzahl Textwörter	Anzahl Wortformen	mittlere Wortform
	n tw	n wf	m wf
Zaza	10059	2654	486,93
Deutsch	9249	2048	327,04
Englisch	41869	5998	1755,65
Spanisch	3348	1261	268,40

Bild 3: Häufigkeit der Wortformen im Zaza

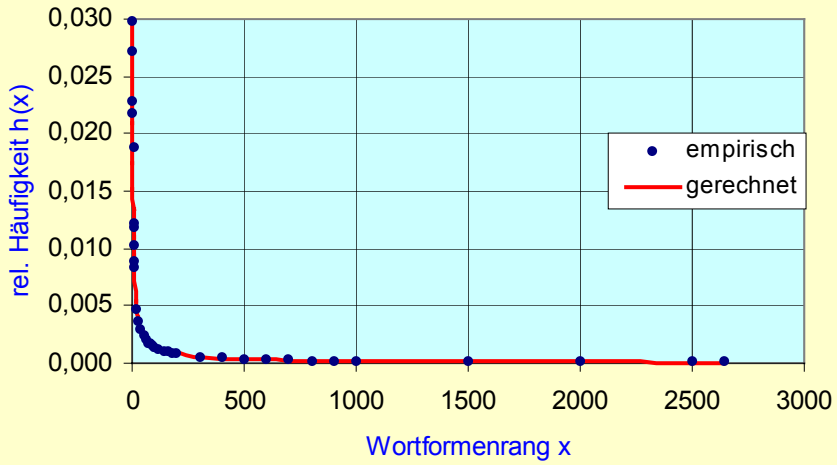


Bild 4: Häufigkeit der Wortformen im Zaza

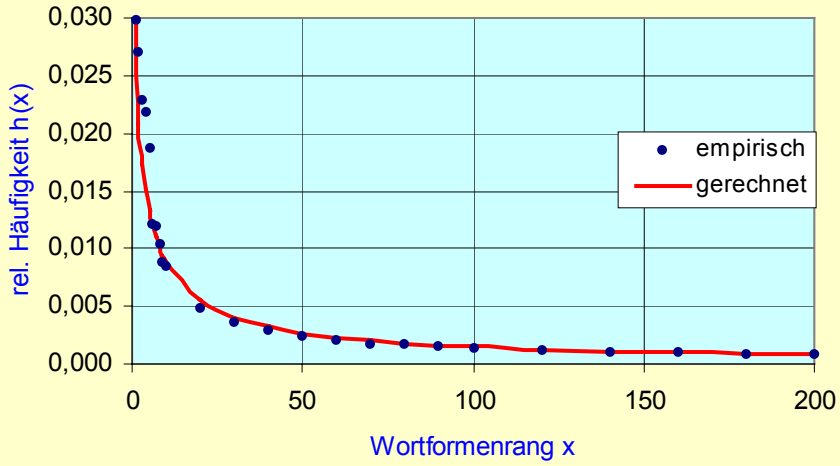


Bild 5: Häufigkeit der Wortformen im Deutschen

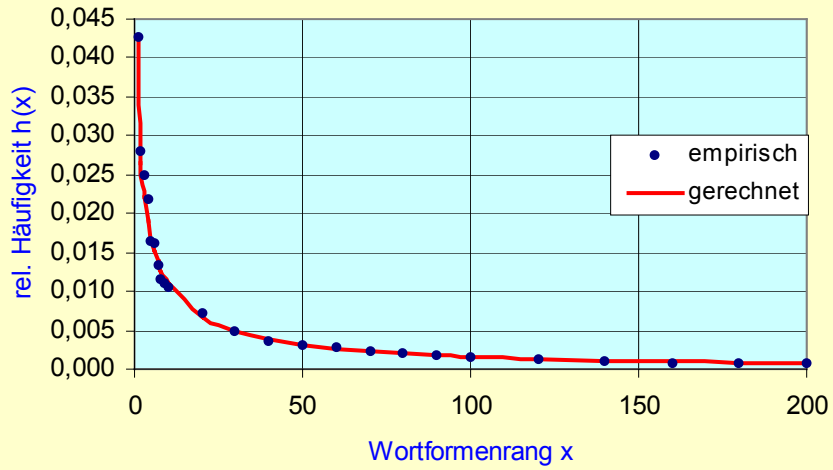


Bild 6: Summenäufigkeit der Wortformen im Zaza

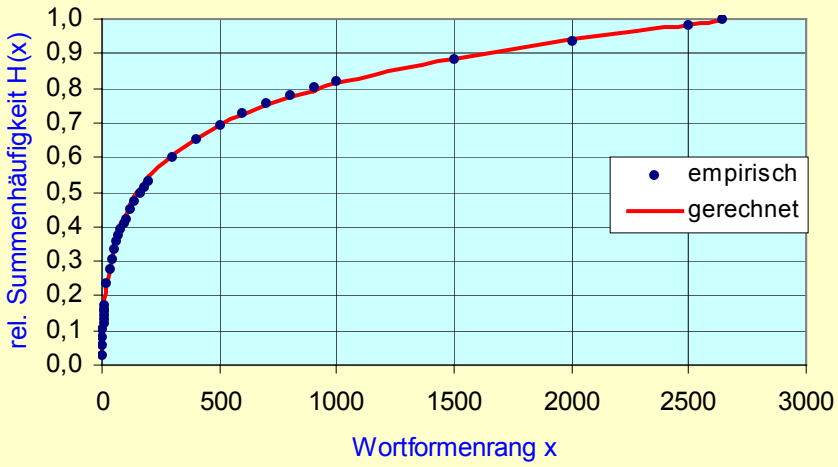


Bild 7: Summenäufigkeit der Wortformen im Deutschen

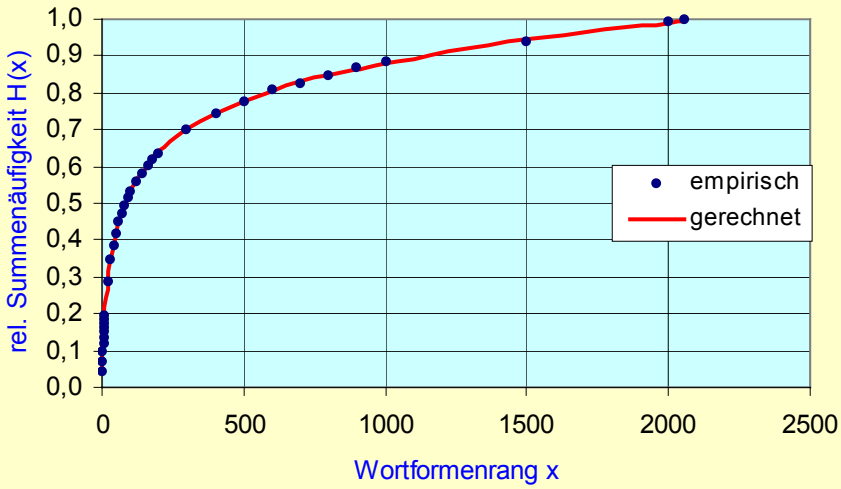


Tabelle 4: Häufigkeit der Wortformen im Zaza

		Häufigkeit			rel. Summenhäufigkeit	
		empirisch		gerechnet	emp.	ger.
Wortformenrang		absolut	relativ	rel.	rel.	rel
mı	1	300	0,02982	0,02982	0,02982	0,02982
ra	2	273	0,02713	0,02085	0,05695	0,05328
ma	3	230	0,02286	0,01742	0,07981	0,07228
de	4	220	0,02186	0,01510	0,10167	0,08847
ke	5	189	0,01878	0,01339	0,12045	0,10268
xo	6	122	0,01212	0,01208	0,13258	0,11539
vake	7	120	0,01193	0,01103	0,14450	0,12693
u	8	104	0,01034	0,01017	0,15484	0,13752
to	9	89	0,00885	0,00945	0,16369	0,14733
a	10	85	0,00845	0,00883	0,17213	0,15646
çê	20	48	0,00477	0,00547	0,23783	0,22508

bıraê	30	36	0,00358	0,00403	0,27818	0,27182
keno	40	29	0,00288	0,00321	0,30908	0,30769
kerd	50	24	0,00239	0,00268	0,33482	0,33696
biye	60	21	0,00209	0,00230	0,35679	0,36174
vez	70	18	0,00179	0,00202	0,37547	0,38326
çino	80	17	0,00169	0,00180	0,39276	0,40231
ala	90	16	0,00159	0,00162	0,40896	0,41939
her	100	14	0,00139	0,00148	0,42357	0,43490
amo	120	12	0,00119	0,00126	0,44981	0,46220
meso	140	11	0,00109	0,00110	0,47337	0,48571
mıno	160	10	0,00099	0,00097	0,49493	0,50638
vınde	180	9	0,00089	0,00087	0,51451	0,52483
bervê	200	9	0,00089	0,00079	0,53240	0,54150
erzeno	300	6	0,00060	0,00055	0,60177	0,60699
gêreke	400	5	0,00050	0,00042	0,65395	0,65467
inu	500	4	0,00040	0,00034	0,69459	0,69232
çiyê	600	3	0,00030	0,00029	0,72739	0,72351
dustê	700	3	0,00030	0,00025	0,75721	0,75020
aşire	800	2	0,00020	0,00022	0,78086	0,77356
yamu	900	2	0,00020	0,00020	0,80074	0,79437
vêsaiye	1000	2	0,00020	0,00018	0,82061	0,81315
intamê	1500	1	0,00010	0,00012	0,88581	0,88711
qızvane	2000	1	0,00010	0,00010	0,93550	0,94151
seři	2500	1	0,00010	0,00008	0,98519	0,98505
qeydker	2646	1	0,0001	0,0001	0,9997	0,9963

Tabelle 5: Häufigkeit der Wortformen im Deutschen

	Wortformenrang	Häufigkeit empirisch absolut	Häufigkeit empirisch relativ	Häufigkeit gerechnet rel.	rel. Summenhäufigkeit emp. rel.	rel. Summenhäufigkeit ger. rel
(und)	1	395	0,04271	0,04271	0,04271	0,04271
der	2	259	0,02800	0,02657	0,07071	0,07287
die	3	231	0,02498	0,02203	0,09569	0,09700
ist	4	202	0,02184	0,01902	0,11753	0,11744
ich	5	152	0,01643	0,01682	0,13396	0,13530
es	6	151	0,01633	0,01514	0,15029	0,15125
er	7	124	0,01341	0,01379	0,16369	0,16569
auf	8	108	0,01168	0,01269	0,17537	0,17891
sie	9	103	0,01114	0,01176	0,18651	0,19112
von	10	98	0,01060	0,01097	0,19710	0,20247
in	20	66	0,00714	0,00665	0,28781	0,28684
man	30	44	0,00476	0,00481	0,34523	0,34313
wenn	40	34	0,00368	0,00376	0,38674	0,38554
soll	50	28	0,00303	0,00308	0,41983	0,41955
Bruder	60	26	0,00281	0,00261	0,44870	0,44789
Seite	70	21	0,00227	0,00226	0,47346	0,47216
bin	80	19	0,00205	0,00199	0,49540	0,49333
unsere	90	17	0,00184	0,00177	**0,51508**	0,51210
Erde	100	15	0,00162	0,00160	0,53227	0,52893
sprach	120	11	0,00119	0,00133	0,56071	0,55808
schon	140	10	0,00108	0,00114	0,58374	0,58269
komm	160	8	0,00086	0,00099	0,60363	0,60394
drei	180	8	0,00086	0,00088	0,62093	0,62261
einem	200	7	0,00076	0,00079	0,63769	0,63922

schaut	300	5	0,00054	0,00051	0,70083	0,70230
Geistliche	400	4	0,00043	0,00038	0,74516	0,74611
Bleistift	500	3	0,00032	0,00030	0,77846	0,77959
begegnen	600	2	0,00022	0,00025	**0,80711**	0,80670
Mensch	700	2	0,00022	0,00021	0,82874	0,82952
jagte	800	2	0,00022	0,00018	0,85036	0,84927
etwa	900	2	0,00022	0,00016	0,87199	0,86672
fluche	1000	1	0,00011	0,00015	**0,88669**	0,88241
lässt	1500	1	0,00011	0,00010	0,94075	0,94414
Mama	2000	1	0,00011	0,00008	0,99481	0,99044
wertvolle	2056	1	0,00011	0,00008	1,00000	0,99439

Bild 8: Häufigkeit der Wortformen im Englischen

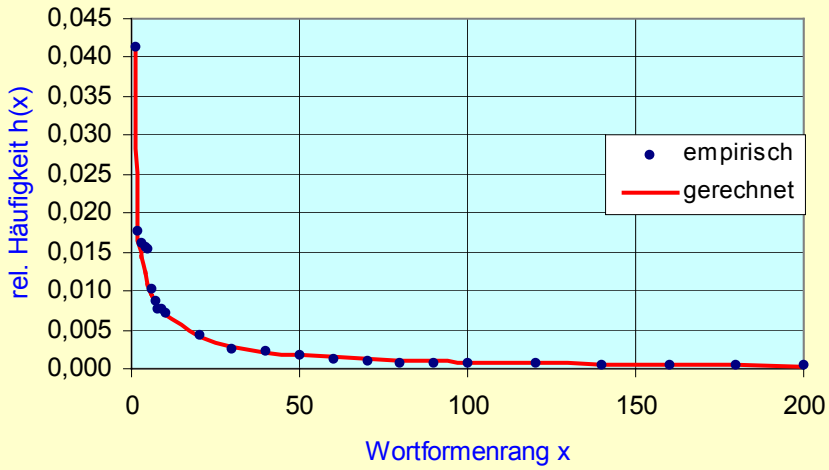


Bild 9: Häufigkeit der Wortformen im Spanischen

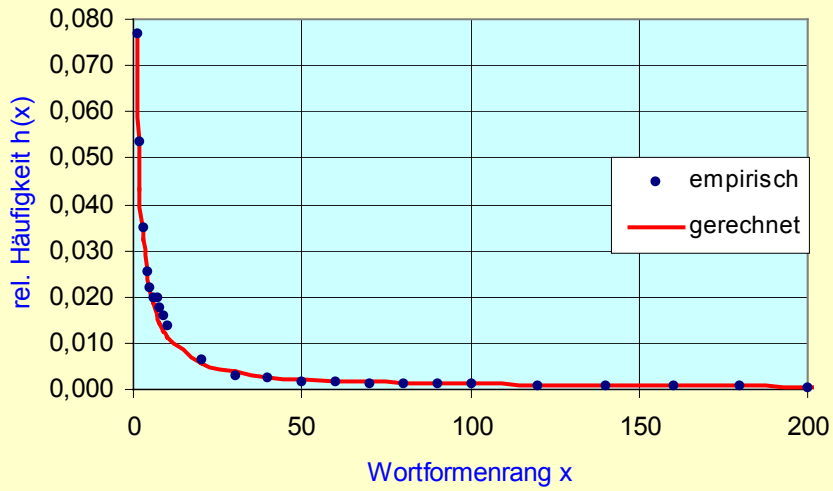


Bild 10: Summenäufigkeit der Wortformen im Englischen

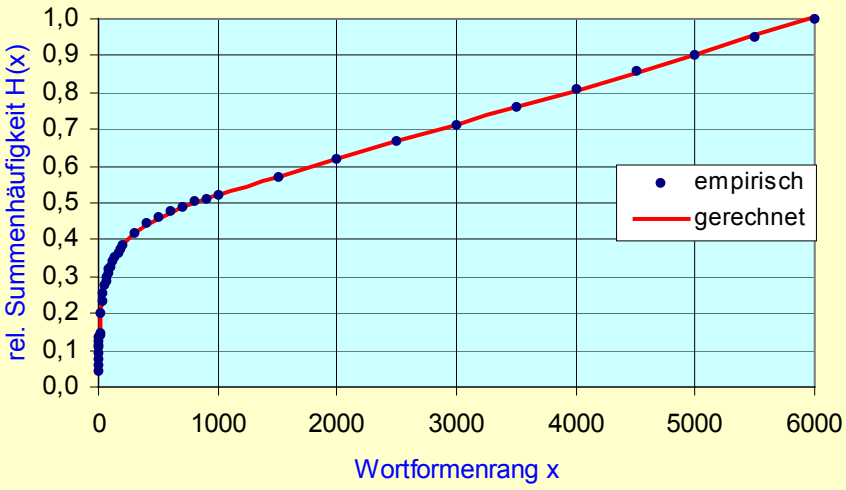


Bild 11: Summenäufigkeit der Wortformen im Spanischen

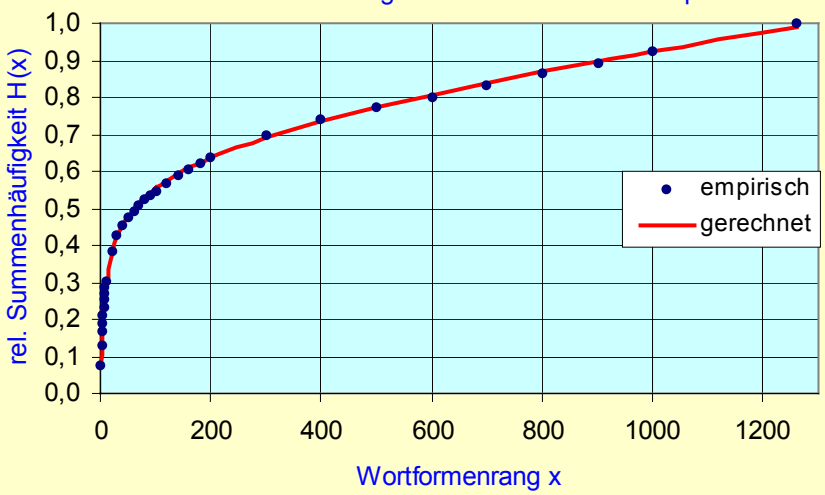


Bild 11a: Wortformenhäufigkeit in normierter Darstellung

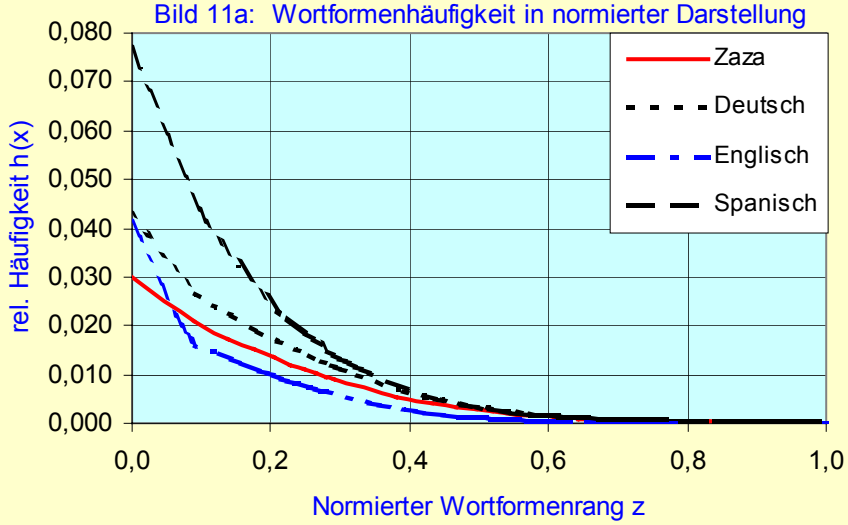


Tabelle 6: Häufigkeit der Wortformen im Englischen

		Häufigkeit			rel. Summenhäufigkeit	
		empirisch		gerechnet	emp.	ger.
	Wortformenrang	absolut	relativ	rel.	rel.	rel
the	1	1729	0,04130	0,04130	0,04130	0,04130
and	2	740	0,01767	0,01762	0,05897	0,06168
to	3	676	0,01615	0,01436	0,07512	0,07753
a	4	659	0,01574	0,01227	0,09085	0,09079
of	5	648	0,01548	0,01078	0,10633	0,10228
in	6	436	0,01041	0,00964	0,11675	0,11246
her	7	363	0,00867	0,00874	0,12541	0,12164
he	8	326	0,00779	0,00800	0,13320	0,13000
I	9	326	0,00779	0,00739	0,14099	0,13768
she	10	300	0,00717	0,00686	0,14815	0,14480
they	20	182	0,00435	0,00402	0,20187	0,19670
by	30	108	0,00258	0,00282	0,23244	0,23022
would	40	95	0,00227	0,00215	0,25694	0,25479
there	50	72	0,00172	0,00173	0,27584	0,27404
said	60	52	0,00124	0,00143	0,29002	0,28976
who	70	44	0,00105	0,00122	0,30120	0,30297
It's	80	37	0,00088	0,00106	0,31085	0,31432
your	90	34	0,00081	0,00093	0,31928	0,32424
good	100	32	0,00076	0,00083	0,32709	0,33303
sand	120	28	0,00067	0,00068	0,34121	0,34801
looked	140	25	0,00060	0,00057	0,35360	0,36045
didn't	160	22	0,00053	0,00049	0,36478	0,37106
white	180	21	0,00050	0,00043	0,37493	0,38028
though	200	19	0,00045	0,00039	0,38422	0,38845
Far	300	12	0,00029	0,00025	0,41988	0,41932

box	400	9	0,00021	0,00019	0,44434	0,44124
gun	500	7	0,00017	0,00016	0,46302	0,45877
lift	600	6	0,00014	0,00014	0,47849	0,47378
takes	700	5	0,00012	0,00013	0,49165	0,48721
service	800	4	0,00010	0,00012	**0,50340**	0,49957
hour	900	4	0,00010	0,00011	0,51296	0,51119
led	1000	4	0,00010	0,00011	0,52251	0,52226
leaned	1500	4	0,00010	0,00010	0,57028	0,57307
speak	2000	4	0,00010	0,00009	0,61805	0,62049
type	2500	4	0,00010	0,00009	0,66581	0,66687
parodically	3000	4	0,0001	0,00009	0,7136	0,71315
deaths	3500	4	0,0001	0,00009	0,7614	0,75976
district	4000	4	0,0001	0,00010	**0,8091**	0,80699
writing	4500	4	0,0001	0,00010	0,8569	0,85500
Delia's	5000	4	0,0001	0,00010	**0,9047**	0,90362
cascading	5500	4	0,0001	0,00010	0,9524	0,95381
Spluttering	5998	4	0,00010	0,00010	1,00000	1,00457

Tabelle 7: Häufigkeit der Wortformen im Spanischen

Wortformenrang		Häufigkeit empirisch absolut	Häufigkeit empirisch relativ	Häufigkeit gerechnet rel.	rel. Summenhäufigkeit emp. rel.	rel. Summenhäufigkeit ger. rel
de	1	258	0,07706	0,07706	0,07706	0,07706
la	2	180	0,05376	0,04344	0,13082	0,12966
y	3	117	0,03495	0,03263	0,16577	0,16720
que	4	85	0,02539	0,02604	0,19116	0,19629
a	5	74	0,02210	0,02160	0,21326	0,21998
el	6	67	0,02001	0,01840	0,23327	0,23990
en	7	66	0,01971	0,01600	0,25299	0,25705

los	8	60	0,01792	0,01413	0,27091	0,27208
por	9	54	0,01613	0,01263	0,28704	0,28543
un	10	46	0,01374	0,01141	0,30078	0,29743
para	20	21	0,00627	0,00566	0,38620	0,37649
han	30	10	0,00299	0,00370	0,42712	0,42196
Este	40	8	0,00239	0,00274	0,45400	0,45371
internacional	50	6	0,00179	0,00218	0,47431	0,47811
quien	60	6	0,00179	0,00181	0,49223	0,49797
aún	70	5	0,00149	0,00156	**0,50777**	0,51477
bosniacos	80	5	0,00149	0,00137	0,52270	0,52938
ello	90	4	0,00119	0,00123	0,53584	0,54236
puede	100	4	0,00119	0,00112	0,54779	0,55406
aquel	120	3	0,00090	0,00095	0,56900	0,57461
pues	140	3	0,00090	0,00083	0,58692	0,59239
sólo	160	3	0,00090	0,00075	0,60484	0,60819
Hollywood	180	3	0,00090	0,00068	0,62276	0,62250
memoria	200	2	0,00060	0,00063	0,63650	0,63565
Karadjic	300	2	0,00060	0,00048	0,69624	0,69026
alejarse	400	1	0,00030	0,00041	0,74283	0,73421
ciudadanía	500	1	0,00030	0,00036	0,77270	0,77234
relajación	600	1	0,00030	0,00033	**0,80257**	0,80673
pena	700	1	0,00030	0,00031	0,83244	0,83850
alusión	800	1	0,00030	0,00029	0,86231	0,86832
testigos	900	1	0,00030	0,00028	**0,89217**	0,89662
negro	1000	1	0,00030	0,00027	0,92204	0,92368
agregó	1261	1	0,00030	0,00024	1,00000	0,98994

5.2 Die Fonemhäufigkeit

Die ersten ausführlichen Angaben über die Fonemhäufigkeit im Zaza sind in Selcans *Grammatik der Zaza-Sprache, Nord-Dialekt* (1998), S. 203-205, enthalten. Diese Häufigkeitswerte basieren auf einer Datenmenge von 103619 Fonemen (S. 203). Das Foneminventar des Zaza besteht aus 25 Konsonanten und 8 Vokalen:

/p, b, t, d, c, ç, k, g, q, f, v, s, z, ṣ, ẓ, x, ẋ, h, w, y, m, n, l, r, ṙ /
/i, ê, e, ü, ı, u, o, a/.

Die Diftonge */ia, iu, üa/* wurden auf */i, a, u, ü/* verteilt. Dort ist die Fonemhäufigkeit auch grafisch dargestellt, aber deren mathematische Beschreibung nicht behandelt. Diese Lücke soll nun hier unter Berücksichtigung der deutschen Fonemhäufigkeit geschlossen werden.
Mit der Fonemhäufigkeit des Deutschen befassten sich Lindner (1963) und Hug (1979), wobei auch die Diftonge mit der Transkriptionsform a^{o}, a^{e}, $o^{ü}$ berücksichtigt wurden. Meinhold/Stock jedoch interpretieren diese in ihrer Arbeit (S. 88) als bifonematisch und betrachten den zweiten Vokal als Realisation der Kurzvokale, der Reihe nach */ʊ, ɪ, ʏ/*. Ferner behandelten sie die von Lindner angegebenen beiden r-Allophone, nämlich das Zungenspitzen-r [r] und das Zäpfchen-r [ʁ] als ein Fonem (vgl. S. 131 f.). Der Fonembestand des Deutschen umfasst 20 Konsonanten, 8 kurze und 8 lange Vokale:

/p, b, t, d, k, g, f, v, s, z, ʃ, ç, x, h, j, m, n, ŋ, l, r/
/i, i:, e:, ɛ, ɛ:, a, a:, ü, ü:, ö, ö:, ə, u, u:, o, o:/.

Hug versucht in seiner Arbeit, die Fonemhäufigkeit mit Hilfe der Normalverteilung so zu bestimmen, indem er die Anzahl der Vokale

innerhalb jeder 100-Fonemfolge ermittelt (S.121). Es wird also weder der mathematische Zusammenhang zwischen dem Fonemrang und der Häufigkeit bestimmt noch die Vorkommenshäufigkeit einzelner Foneme. Hugs Korpusdaten bestehen aus nur 5000 Fonemen (S. 121), was für eine statistische Auswertung zu gering ist und folglich für eine befriedigende oder repräsentative Aussage nicht ausreicht. Dies schlägt sich in abweichenden Häufigkeitswerten, wenn die Resultate von Hug und Lindner gegenübergestellt werden (Hug, S. 126; Lindner, S. 118 f.):

	% aller Foneme	
	Lindner	Hug
d	5,2	7,175
t	9,2	11,525
s	5,4	5,915
n	9,5	16,545

Hier fällt auf, dass Hugs Werte höher sind als die Lindners. Aufgrung der 4-fach größeren Fonemzahl Lindners, dürften dessen Ergebnisse relativ zuverlässiger sein.
Die von Meinhold/Stock korrigierten Häufigkeitsangaben, welche auf 20000 Fonemen beruhen, werden hier als Grundlage für die mathematische Beschreibung verwendet.

Für die Berechnung der Fonemhäufigkeit gilt dieselbe Methode wie bei Wortformenhäufigkeit. Hier wird der sich aus sinkender Reihenfolge der Häufigkeit ergebender Fonemrang (Tab. 9, 10) als Variable *x* benutzt.
Die relative Fonemhäufigkeit für Zaza und Deutsch ist in Bild 12, 13 und die Summenhäufigkeit in Bild 14, 15 abgebildet. Aus Tab. 8 sind die

Konstanten *a, b, c,* die Exponenten *m, n* sowie die mittleren quadratischen Abweichungen *S* und *s* zu entnehmen.

Die Anwendung der Gleichungen (11) und (12) auf die Häufigkeitsverteilung von Wortformen und Fonemen hat sich nach den hier gewonnenen Ergebnissen sehr gut bewährt. Sie sind nicht nur auf die sprachlichen Häufigkeiten beschränkt, sondern können auch auf andere, ähnlich verlaufende Verteilungsform, welche von einem bestimmten Anfangswert $h_o=h$ *(x=1, z=0)* ausgehend bei steigendem *x* ständig abklingen, allgemein angewandt werden. Außer den Wortformen- und der Fonemhäufigkeit bilden sie auch für weitere Verteilungen in der Linguistik, z. B. Silbenhäufigkeit u.a., eine gute mathematische Grundlage.

Die nach den Gl.n (11) und (12) berechneten Näherungskurven sind in Bild 12, 13 und in Bild 14, 15 dargestellt, welche mit dem empirischen Verlauf gut übereinstimmen.

Tabelle 8: Die Konstanten und Abweichungen der Fonemhäufigkeit

	h_o	a	b	c	m	n	S	s
Zaza	0,12282	3,00428	-0,56351	-0,34393	2	5	$5{,}128\ 10^{-3}$	$4{,}95\ 10^{-3}$
Deutsch	0,09141	3,62672	-1,62295	0,40181	3	5	$6{,}292\ 10^{-3}$	$3{,}305\ 10^{-3}$

Bei näherer Betrachtung der Fonemhäufigkeit in Tab. 9 und Tab. 10 ist festzustellen, dass das am häufigsten vorkommende Fonem im Zaza */e/*[ɛ] und im Deutschen */n/* ist. Aus der Summenhäufigkeit ist zu entnehmen, dass bei 0,5, d. h. bei der Hälfte, der ausgewerteten Fonemfolge bzw. des Textes, von sechs Fonemen im Zaza und von sieben Fonemen im

Deutschen umfasst werden (s. auch Bild 14**,** 15). Diese sind im Zaza der Reihe nach */a, n, r, i, ê/* und im Deutschen */n, t, ə, r, a, i, s/*.

Bild 12: Häufigkeit der Foneme im Zaza

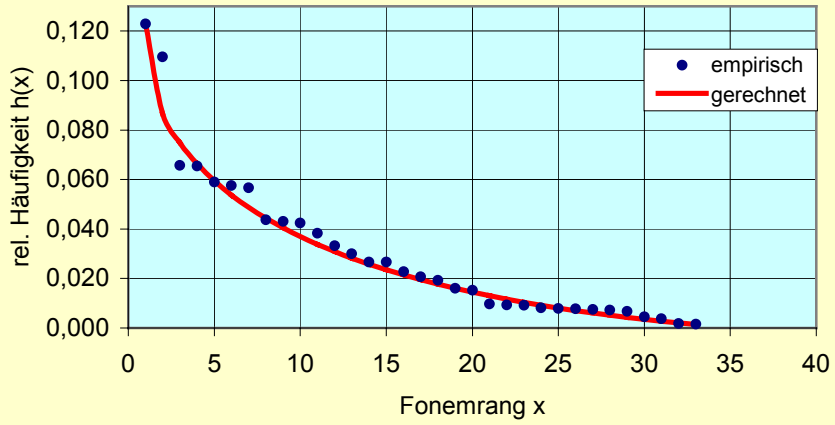


Bild 13: Häufigkeit der Foneme im Deutschen

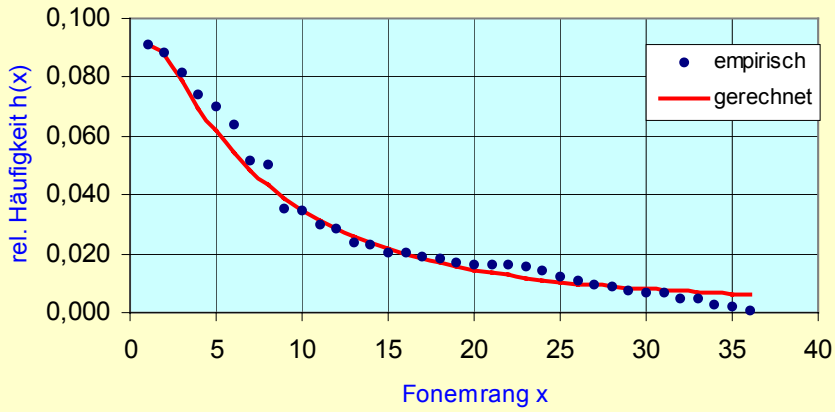


Bild 14: Summenäufigkeit der Foneme im Zaza

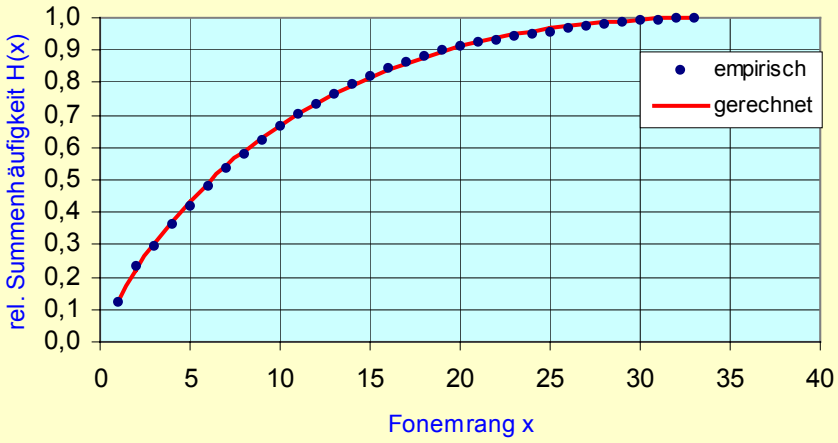


Bild 15: Summenäufigkeit der Foneme im Deutschen

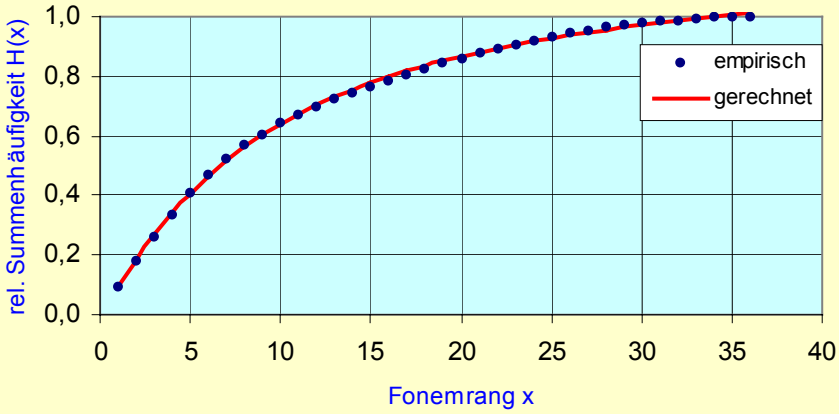


Tabelle 9: Fonemhäufigkeit im Zaza

Fonem	Fonemrang	Häufigkeit % aller Foneme	rel. Häufigkeit		rel. Summenhäufigkeit	
			empirisch	gerechnet	empirisch	gerechnet
e	1	12,823	0,12263	0,12282	0,12282	0,12282
a	2	10,618	0,10935	0,08664	0,23235	0,21938
n	3	6,177	0,06561	0,07492	0,29806	0,29985
r	4	6,044	0,06531	0,06632	0,36347	0,37032
i	5	5,897	0,05892	0,05945	0,42248	0,43311
ê	6	5,877	0,05752	0,05369	0,48010	0,48962
o	7	5,494	0,05652	0,04873	0,53671	0,54079
m	8	4,719	0,04364	0,04436	0,58042	0,58730
k	9	4,352	0,04304	0,04047	0,62352	0,62969
ı	10	3,992	0,04234	0,03697	0,66593	0,66838
d	11	3,609	0,03825	0,03379	0,70424	0,70374
t	12	3,212	0,03315	0,03089	0,73745	0,73606
s	13	2,672	0,02996	0,02823	0,76745	0,76560
y	14	2,572	0,02666	0,02578	0,79416	0,79259
u	15	2,562	0,02656	0,02352	0,82076	0,81722
v	16	2,53	0,02267	0,02143	0,84347	0,83969
z	17	2,045	0,02067	0,01949	0,86417	0,86014
l	18	1,907	0,01927	0,01769	0,88348	0,87872
h	19	1,59	0,01608	0,01602	0,89958	0,89556
b	20	1,375	0,01528	0,01446	0,91488	0,91079
w	21	1,195	0,00969	0,01300	0,92458	0,92451
c	22	1,042	0,00929	0,01164	0,93389	0,93683
p	23	0,997	0,00919	0,01037	0,94309	0,94782

ç	24	0,962	0,00819	0,00918	0,95129	0,95760
g	25	0,945	0,00789	0,00807	0,95919	0,96622
q	26	0,892	0,00779	0,00703	0,96699	0,97377
ŗ	27	0,872	0,00749	0,00606	0,97449	0,98030
ş	28	0,75	0,00719	0,00515	0,98170	0,98590
x	29	0,75	0,00669	0,00429	0,98840	0,99061
f	30	0,38	0,00449	0,00349	0,99290	0,99450
ẓ	31	0,305	0,00369	0,00273	0,99660	0,99760
ẋ	32	0,297	0,00180	0,00203	0,99840	0,99998
ü	33	0,147	0,00160	0,00137	1,00000	1,00167

Tabelle 10: Fonemhäufigkeit im Deutschen

Fonem	Fonemrang	Häufigkeit % aller Foneme	rel. Häufigkeit		rel. Summenhäufigkeit	
			empirisch	gerechnet	empirisch	gerechnet
n	1	9,14	0,09141	0,09141	0,09141	0,09141
t	2	8,85	0,08851	0,08750	0,17992	0,18258
ə	3	8,18	0,08181	0,07863	0,26172	0,26594
r	4	7,41	0,07411	0,06969	0,33583	0,34015
a	5	7,03	0,07031	0,06163	0,40614	0,40580
i	6	6,36	0,06361	0,05460	0,46974	0,46388
s	7	5,2	0,05200	0,04853	0,52175	0,51542
d	8	5	0,05000	0,04331	0,57175	0,56131
l	9	3,56	0,03560	0,03880	0,60735	0,60233
ɛ	10	3,46	0,03460	0,03489	0,64196	0,63915
u	11	2,98	0,02980	0,03150	0,67176	0,67233
f	12	2,89	0,02890	0,02855	0,70066	0,70235

i:	13	2,41	0,02410	0,02596	0,72477	0,72959
m	14	2,31	0,02310	0,02369	0,74787	0,75440
a:	15	2,02	0,02020	0,02168	0,76807	0,77708
e:	16	2,02	0,02020	0,01991	0,78827	0,79786
b	17	1,92	0,01920	0,01833	0,80747	0,81698
k	18	1,83	0,01830	0,01692	0,82577	0,83460
g	19	1,73	0,01730	0,01567	0,84308	0,85090
z	20	1,64	0,01640	0,01455	0,85948	0,86600
ʃ	21	1,64	0,01640	0,01354	0,87588	0,88005
v	22	1,64	0,01640	0,01263	0,89228	0,89313
o	23	1,54	0,01540	0,01181	0,90768	0,90536
ç	24	1,44	0,01440	0,01108	0,92208	0,91680
h	25	1,25	0,01250	0,01041	0,93458	0,92755
p	26	1,06	0,01060	0,00980	0,94519	0,93765
o:	27	0,96	0,00960	0,00925	0,95479	0,94718
u:	28	0,87	0,00870	0,00875	0,96349	0,95617
ŋ	29	0,77	0,00770	0,00829	0,97119	0,96469
ü	30	0,675	0,00675	0,00787	0,97794	0,97278
x	31	0,67	0,00670	0,00749	0,98464	0,98046
ü:	32	0,48	0,00480	0,00714	0,98944	0,98778
j	33	0,48	0,00480	0,00683	0,99424	0,99477
ɛ:	34	0,29	0,00290	0,00653	0,99714	1,00146
ö:	35	0,19	0,00190	0,00627	0,99904	1,00786
ö:	36	0,096	0,00096	0,00602	1,00000	1,01401

5.3 Die Wortlängenhäufigkeit

Ein anderer Aspekt der sprachlichen Häufigkeitsuntersuchung ist, die Vorkommensintensität der verschiedenen Wortlängen zu ermitteln. Diese Frage wurde anhand der Korpusdaten für Zaza und Deutsch behandelt und durch Berücksichtigung weiterer Sprachen – Englisch und Spanisch – erweitert, um eine überblickende Beobachtung der Wortlängenhäufigkeit zu ermöglichen.

Die Zählungsergebnisse von vier Sprachen sind in Tabelle 13 bis 16 aufgelistet und die Wortlängenhäufigkeiten in Bild 16 bis 20 grafisch dargestellt. In Bild 16 fällt ein besonderes Merkmal des Zaza auf; bei der Wortlänge 3 ist eine Mulde im Häufigkeitsverlauf zu beobachten, während die Wortlängen 2 und 4 ein herausragendes Maximum bilden (Tab.13): 0,266, 0,253. Die Häufigkeit der 3er Wortlänge sinkt etwa bis zur Hälfte des Maximums: 0,123. Beim Deutschen dagegen kommt nur ein Maximum vor (Bild 17), welcher bei 3er Wortlänge auftritt und den Wert 0,32 erreicht (Tab. 14).

Eine sprachvergleichende Untersuchung der Wortlängenhäufigkeit erfordert, dass dies unter gleichen Bedingungen erfolgt: streng genommen sollte die fonemische Schreibung als einheitliche Basis dazu dienen. Die Schreibung des Zaza ist fonemisch und des Deutschen grafemisch. Nach dieser theoretischen Überlegung müssten deutsche Texte in fonemische Transkription umgewandelt werden. Denn für die Foneme wie /ʃ, ʧ, p, t, k, .../ die aus mehreren Zeichen bestehenden Grafeme <*sch, tsch, pp, tt, ck,* ...> zu benutzen, würde bei der Zählung die tatsächlich gesprochene Wortlänge verzerren. Zur Klärung dieses Problems wurde der deutsche Text fonemisch aufbereitet und mittels Programmierung ausgewertet. Das Resultat dieses Tests ist in Bild 18 eingezeichnet. Bei näherer

Betrachtung stellt man fest, dass die Maximumwerte der Wortlängen beider Schreibungsformen übereinstimmen. Die Kurve der fonemischen Schreibung ist nur geringfügig nach links verschoben, wobei die Verlaufsform der Häufigkeit tendenziell erhalten bleibt.

Aus Bild 19 und 20 geht hervor, dass die maximale Häufigkeit im Englischen bei der 3er Wortlänge, beim Spanischen dagegen bei den 2ern vorkommt. Ein eigenartiges Kennzeichen des Spanischen ist bei der 4er Wortlänge sichtbar, wo es zu einem muldenartigen Häufigkeitsverlauf kommt und bei 5er Wortlänge wiederum ein Maximum erreicht.

In vergleichender Bewertung lässt sich sagen, dass die Verlaufsform des Zaza und des Spanischen verschiedene Verteilungstypen für sich bilden, die des Deutschen und Englischen dagegen als ein gemeinsames Typ klassifiziert werden können.

Als nächstes soll die mathematische Erfassung der Häufigkeitsverteilung von Wortlängen behandelt werden. Ein solcher Versuch ist von Piotrowski u. a. (1985, S. 281 f.) unternommen worden, worin ein wissenschaftlicher Text in Deutsch mit tausend Textwörtern ausgezählt und dabei versucht wurde, dies auch mathematisch zu begründen. Hier wird der Häufigkeitsverlauf in logarithmische Normalverteilung transformiert, was jedoch mit einer großen Diskrepanz zwischen den empirischen und gerechneten Werten der Verteilung behaftet ist. Unter dem Gesichtspunkt der Allgemeingültigkeit ist die Anwendung der logarithmischen Normalverteilung zur Bewertung der Wortlängenhäufigkeit ungeeignet, wenn man die besondere Verteilungsform des Zaza und des Spanischen berücksichtigt.

Aus Bild 16 ist deutlich erkennbar, dass die Kurve des Häufigkeitsverlaufs sich aus mehreren Exponentialkurven der Art e^{-x^2} zusammensetzt, welche waagerecht um n_1 , n_2 , n_3 , n_4 nach rechts

versetzt sind und deren Maximum durch entsprechende Konstante a_1, a_2, a_3, a_4 bestimmt wird. Die Dehnung bzw. Stauchung dieser symetrischen Elementarkurven in der Breite, d. h. nach rechts und links wird durch den Faktor b_1, b_2, b_3, b_4 festgelegt. Aus den Maximumpunkten lässt sich entnehmen, wieviel Elementarkurven sich zusammen bilden. Demnach besteht die Elementarkurve aus der Formel

$$h_i(x) = a_i e^{-b_i (x-n_i)^2} \qquad (26)$$

i=1, 2, 3, 4, (5), x=Wortlänge

Hierbei lässt sich a_i leicht gewinnen, wenn $x=n_i$ gesetzt wird:

$$h_i(n_i) = a_i e^0 = a_i.1 = a_i$$

Im zweiten Schritt kann man b_i ermitteln:

$$\ln \frac{a_i}{h_i} = b_i (x-n_i)^2, \quad b_i = \frac{\ln \frac{a_i}{h_i}}{(x-n_i)^2}$$

Da die einzelnen Exponentialkurven sich überlagern, ist eine explizite und genaue Bestimmung der Konstanten a_i, b_i in einem Zug zwar nicht möglich, aber sie lassen sich nach mehreren iterativen Schritten optimal bestimmen. Die nach diesem Verfahren ermittelten Konstanten sind in der Tabelle 11 zusammen geführt.

$$h(x) = h_1(x) + h_2(x) + h_3(x) + h_4(x) \qquad (27)$$

$$h(x) = a_1 e^{-b_1 (x-n_1)^2} + a_2 e^{-b_2 (x-n_2)^2} + ... = \sum_{i=1}^{k} a_i e^{-b_i (x-n_i)^2} \qquad (28)$$

Die Anzahl *i* der Elementarkurven bewegt sich zwischen 4 und 5; bei der *Zaza-Sprache* und beim *Englischen* ist k=4, bei *Deutsch* und *Spanisch* 5.

Tabelle 11: Die Konstanten der Wortlängenhäufigkeit

		i				
		1	2	3	4	5
Zaza	a	0,26	0,25	0,115	0,017	0
	b	1,6	1,14	1	0,5	
	n	2	4	6	8	
Deutsch	a	0,315	0,115	0,07	0,041	0,0187
	b	1	0,9	1	0,8	0,5
	n	3	5	6	8	10
Englisch	a	0,23	0,07	0,07	0,013	0
	b	0,42	0,75	0,35	0,5	
	n	3	5	7	10	
Spanisch	a	0,238	0,085	0,025	0,05	0,005
	b	1	0,2	0,6	0,2	0,3
	n	2	5	7	9	13

Tabelle 12: Die mittlere Wortlänge

	mittlere Wortlänge
Zaza	3,886
Deutsch	4,619
Englisch	4,411
Spanisch	4,953

Bild 16: Wortlängenhäufigkeit im Zaza

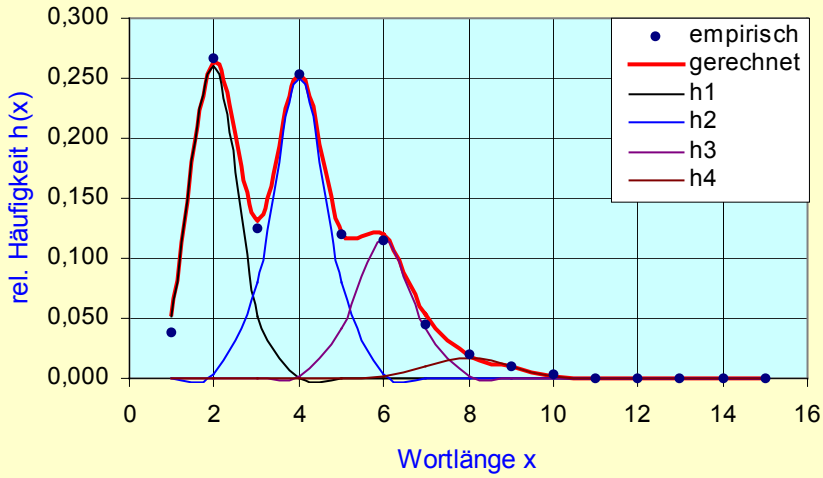


Bild 17: Wortlängenhäufigkeit im Deutschen

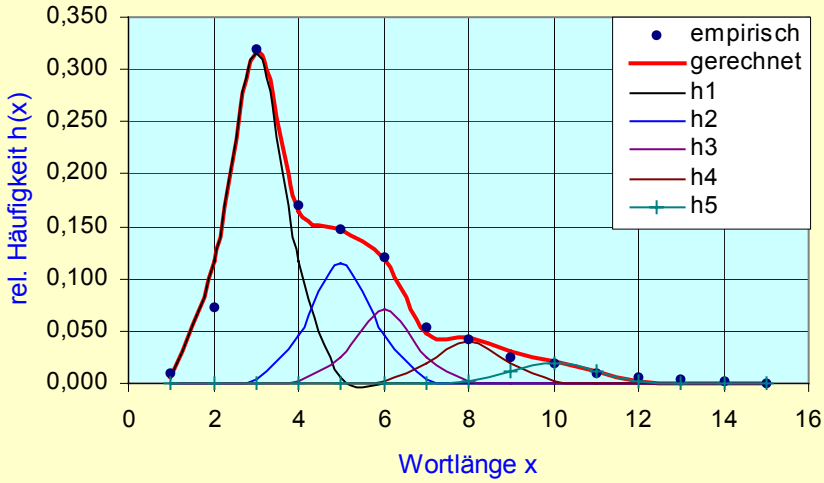


Bild 18: Wortlängenhäufigkeit im Deutschen grafem./fonem.

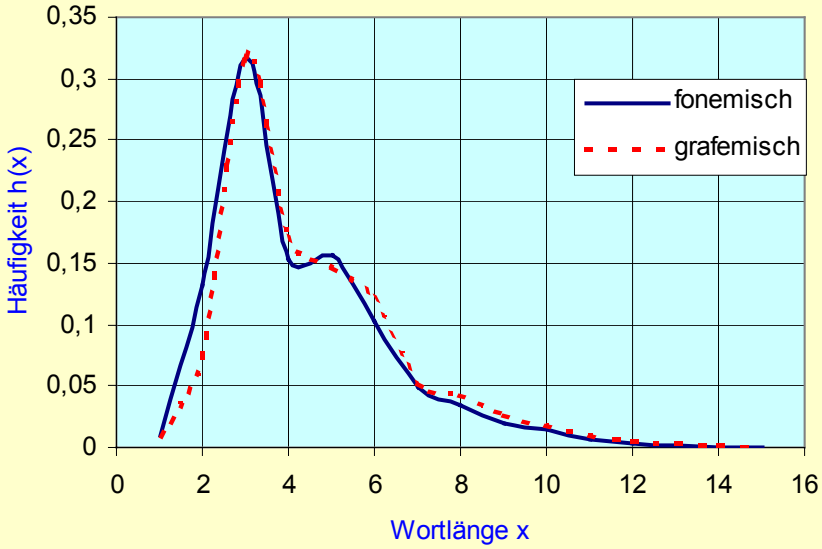


Bild 19: Wortlängenhäufigkeit im Englischen

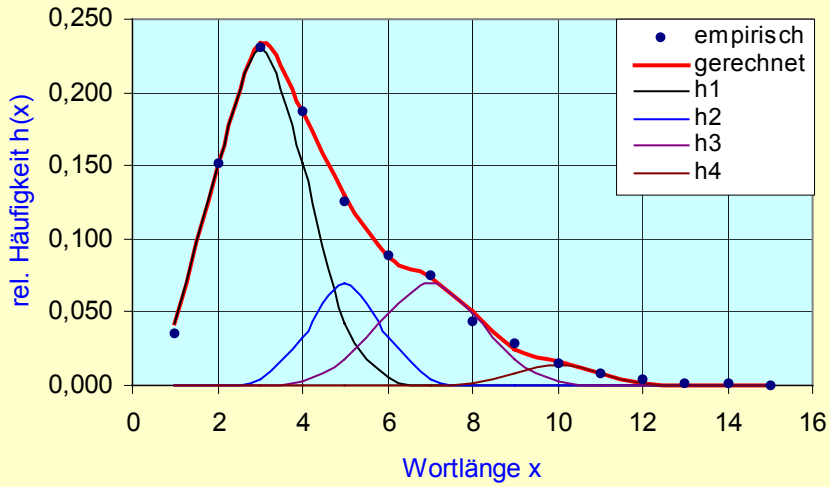


Bild 20: Wortlängenhäufigkeit im Spanischen

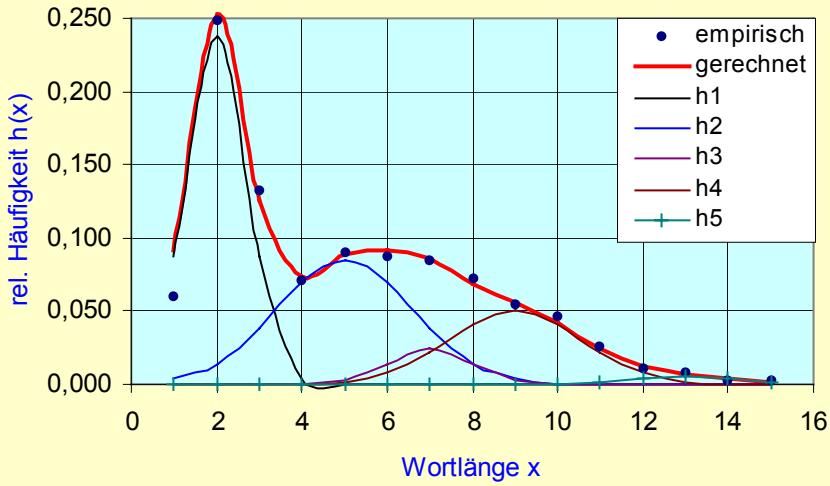


Tabelle 13: Wortlängenhäufigkeit im Zaza

Wortlänge	Häufigkeit			Summen-häufigkeit
	empirisch		gerechnet	emp.
x	absolut	relativ	rel.	rel.
1	382	0,038	0,0525	0,038
2	2677	0,2661	0,2626	0,3041
3	1265	0,1257	0,1325	0,4298
4	2550	0,2535	0,2525	0,6833
5	1209	0,1202	0,1224	0,8035
6	1160	0,1153	0,1199	0,9188
7	455	0,0452	0,0526	0,964
8	206	0,0205	0,0191	0,9845
9	99	0,0098	0,0103	0,9943
10	40	0,004	0,0023	0,9983

11	6	0,0006	0,0002	0,9989
12	7	0,0007	0,0000	0,9996
13	2	0,0002	0,0000	0,9998
14	1	0,0001	0,0000	0,9999
15	1	0,0001	0,0000	1

Tabelle 14: Wortlängenhäufigkeit im Deutschen

Wortlänge	Häufigkeit			Summen-häufigkeit
	empirisch		gerechnet	emp.
x	absolut	relativ	rel.	rel.
1	91	0,0099	0,0058	0,0099
2	677	0,0733	0,1159	0,0832
3	2951	0,3195	0,3182	0,4027
4	1573	0,1703	0,1639	0,573
5	1358	0,147	0,1466	0,72
6	1108	0,12	0,1185	0,84
7	488	0,0528	0,0475	0,8928
8	391	0,0423	0,0448	0,9351
9	237	0,0257	0,0298	0,9608
10	173	0,0187	0,0204	0,9795
11	95	0,0103	0,0114	0,9898
12	49	0,0053	0,0025	0,9951
13	28	0,003	0,0002	0,9981
14	11	0,0012	0,0000	0,9993
15	6	0,0006	0,0000	0,9999

Tabelle 15: Wortlängenhäufigkeit im Englischen

Wortlänge	Häufigkeit			Summen-häufigkeit
	empirisch		gerechnet	emp.
	absolut	relativ	rel.	rel.
1	1020	0,0356	0,0429	0,036
2	4361	0,1521	0,1512	0,188
3	6636	0,2314	0,2337	0,419
4	5381	0,1876	0,1872	0,607
5	3618	0,1261	0,1301	0,733
6	2556	0,0891	0,0876	0,822
7	2172	0,0757	0,0739	0,898
8	1254	0,0437	0,0512	0,941
9	825	0,0288	0,0251	0,97
10	449	0,0157	0,0160	0,986
11	230	0,008	0,0081	0,994
12	106	0,0037	0,0018	0,997
13	44	0,0015	0,0001	0,999
14	22	0,0008	0,0000	1
15	7	0,0002	0,0000	1

Tabelle 16: Wortlängenhäufigkeit im Spanischen

Wortlänge	Häufigkeit			Summen-häufigkeit
	empirisch		gerechnet	emp.
	absolut	relativ	rel.	rel.
1	202	0,0605	0,0910	0,06
2	831	0,2487	0,2521	0,309
3	441	0,132	0,1258	0,441

4	239	0,0715	0,0744	0,513
5	302	0,0904	0,0893	0,603
6	292	0,0874	0,0916	0,691
7	285	0,0853	0,0857	0,776
8	242	0,0724	0,0687	0,848
9	184	0,0551	0,0558	0,903
10	153	0,0458	0,0420	0,949
11	88	0,0263	0,0240	0,975
12	37	0,0111	0,0120	0,987
13	26	0,0078	0,0070	0,994
14	9	0,0027	0,0040	0,997
15	10	0,003	0,0015	1

Bibliographie

Altmann, G., K.-H. Best; Länge der Wörter in deutschen Texten, in: Glottometrika, 15, 1996, 166-180.

Hug, Marc; Die Phonemverteilung im Deutschen; La Distribution des Phonemics en Français, Genève 1979.

Kaeding, F. W.; Häufigkeitswörterbuch der deutschen Sprache, Berlin 1997/98.

Lindner, Gerhard; Distinktive Merkmale in Kontrastellung in zusammenhängendem deutschen Text, in: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung (Berlin), 16.1963, S. 117-126.

Meier, H.; Rangbuch der geläufigsten deutschen Wortformen, 1964.

Meinhold, G./E. Stock, Phonologie der deutschen Gegenwartssprache, Leipzig 1982.

Ortmann, W.-D.; Hochfrequente deutsche Wortformen, I-II, München 1975.

Piotrowski u.a.; Mathematische Linguistik, aus dem Russischen übers. von A. Falk, Bochum 1985.

Selcan, Zülfü; Grammatik der Zaza-Spracher, Berlin 1998, S. 203-205.

Sparmann, H.; Die Wahrscheinlichkeit des Auftretens häufiger Wörter in einem bestimmten Textumfang, in: Zeitschr. für Phonol, Sprachwiss. und Kommunik., 26.1973, 698-699.