Kabashi, Besim : „Das Albanische Alphabet aus sprachtechnologischer Sicht“. [In:] Demiraj, Bardhyl (Hrsg.): *Der Kongress von Manastir. Herausforderung zwischen Tradition und Neuerung in der albanischen Schriftkultur.* (Akten des 3. Deutsch-Albanischen Kulturwissenschaftlichen Tagung. Ludwig-Maximilians-Universität München, 7.–8. Oktober 2008). Hamburg : Dr. Kovač, 2009. ISBN: 978-3-8300-4705-6. (Philologia – Sprachwissenschaftliche Forschungsergebnisse, Bd. 139).

Besim Kabashi

Das albanische Alphabet aus sprachtechnologischer Sicht

Beim Schreiben oder Lesen von E-Mails in albanischer Sprache ist es nicht selten der Fall, dass zwei Schriftzeichen des albanischen Alphabets nicht benutzt bzw. durch andere ersetzt werden. Was führt dazu, dass die Schriftzeichen *Ç/ç* und *Ë/ë* in vielen Fällen bei der elektronischen Kommunikation oder sogar beim Druck von Dokumenten nicht verwendet werden?
In diesem Artikel wird ein kurzer Überblick gegeben, der das Benutzen des albanischen Alphabets bei der elektronischen Textverarbeitung und Kommunikation samt den damit verbundenen technischen Schwierigkeiten erleichtern soll.

1. Geschichtlicher Hintergrund

Im Jahr 1908 fand vom 14. bis 22. Oktober in der Stadt Manastir (Bitola) ein Kongress statt, der in der albanischen Geschichte als *Kongress von Manastir* bekannt ist. Er begann mit dem Ziel, ein vereinheitlichtes Alphabet für die albanische Sprache zu verabschieden. Bis dahin wurden verschiedene (gemischte) Alphabete benutzt, je nachdem unter welchem kulturellen Einfluss die Schreiber standen – in Gebrauch waren vor allem das lateinisch-italienische, aber auch das griechische, osmanische oder slawische Alphabet. Bis zum Ende des 19. Jahrhunderts benutzten viele Autoren eigene Alphabete. Viele, insbesondere in der Migration lebende Albaner verlangten von dem Kongress, dass das zukünftige Alphabet der albanischen Sprache auf dem lateinischen Alphabet aufbaue, ohne Übernahme von zusätzlichen Zeichen aus Alphabeten anderer Sprachen oder Entwicklung neuer Zeichen.[1] Der Beschluss des Kongresses, an dem Delegierte aus fast allen albanischen Gebieten[2] Südosteuropas und der albanischen Diaspora teilnahmen, war die Konstruktion zweier Alphabete: Das erste hatte eine lateinische Basis, übernahm aber noch Zeichen aus anderen Alphabeten sowie zusätzliche Zeichen, das zweite Alphabet basierte nur auf lateinischer Schrift, das zusätzlich dazu neun darauf basierende Digraphe sowie das *Ç/ç* und das *Ë/ë* beinhaltete. Das erste Alphabet erfülle zwar die Bedürfnisse der Nation, aber für das Drucken von Büchern außerhalb Albaniens und das Verschicken von Telegrammen würde ein anderes, rein lateinisches

[1] „[...] shumica e atyre që hynë në korrespondencë me klubin e Manastirit kërkoi që të caktohej për gjuhën shqipe një alfabet thjesht latin, i papërzier. Alfabetin latin, siç mund të pritej, e përkrahën më shumë shqiptarët jashtë atdheut, të cilët, për shkak të njohjes së gjuhëve evropiane të mbështetura në këtë alfabet i njihnin më mirë epërsitë e tij.“ [Demiraj / Prifti 2004: 71].

[2] Einer der Delegierten des Kongresses, M. Frashëri, sagte bei der Verkündung des Beschlusses „[...] *die Kommission, gewählt aus den Delegierten Albaniens*, beendete ihren Auftrag bemüht um eine Einigung“ (albanisch: „[...] *komisioni, i zgjedhur nga delegatet e Shqipërisë*, e përfundoi detyrën e tij, duke u përpjekur për bashkim“). Zitiert nach [Demiraj / Prifti 2004: 102], Übersetzung sowie Kursivsetzung – B. K.; Mit “Albanien” sind hier die albanischen Gebiete gemeint, denn Albanien erklärte die Unabhängigkeit erst 1912.

Alphabet gebraucht werden[3] – hieß es im Beschluss des Kongresses. Das zweite Alphabet setzte sich im Laufe der Zeit durch und ist auch nach der Normierung der albanischen Sprache (1972) bis heute als Standard anerkannt. Diese Entscheidung war erwünscht und erwartet. Es gab allerdings auch Gegner und Kritiker dieses Kongresses, wie z. B. der Publizist F. Konica, einer der bedeutendsten Intellektuellen jener Zeit. Andere ebenso wichtige Persönlichkeiten, wie z. B. der Sprachwissenschaftler A. Xhuvani, nahmen an dem Kongress nicht teil. Trotz dieser Fakten wird der Kongress von Manastir von den meisten Albanologen und Albanien–Forschern aus heutiger Sicht als Erfolg gewertet. Außer den praktischen Vorteilen, die das Alphabet mit sich brachte, war diese Entscheidung auch ein politisch–kultureller Beweis für die Zugehörigkeit der albanischen Sprache und Kultur zu Westeuropa.[4] Die beiden im Kongress von Manastir entworfenen Alphabete unterschieden sich nur in neun Schriftzeichen, und zwar genau in *dh*, *ë*, *gj*, *ll*, *nj*, *sh*, *th*, *xh*, und *zh* (aus dem heutigen Alphabet, s. u. § 2).[5] Im Beschluss wurde erwähnt, dass beide Alphabete Pflicht in den Schulen sein müssten.[6]

Die Delegierten waren sich der technischen Schwierigkeiten durchaus bewusst, welche die erste Variante mit sich brachte. Die technischen Entwicklungen und Entdeckungen der Zeit bewegten die Delegierten dazu, eine einfachere Variante zuzulassen. Allein das Verschicken von Telegrammen bzw. das Telegraphieren war mit dem ersten Alphabet ohne eine Transliterierung seiner Zusatzzeichen nicht möglich.[7]

2. Das albanische Alphabet

Die von den Delegierten des Kongresses von Manastir einheitlich akzeptierten 36 Laute der albanischen Sprache sind durch die 25 Grapheme des lateinischen Alphabets *A/a*, *B/b*, *C/c*, *D/d*, *E/e*, *F/f*, *G/g*, *H/h*, *I/i*, *J/j*, *K/k*, *L/l*, *M/m*, *N/n*, *O/o*, *P/p*, *Q/q*, *R/r*, *S/s*, *T/t*, *U/u*, *V/v*, *X/x*, *Y/y* und *Z/z*, durch die neun Digraphen *DH/Dh/dh*, *GJ/Gj/gj*, *LL/Ll/ll*, *NJ/Nj/nj*, *RR/Rh/rr*, *SH/Sh/sh*, *TH/Th/th*, *XH/Xh/xh* und *ZH/Zh/zh* sowie die zwei mit Diakritika versehenen Zeichen *Ç/ç* und *Ë/ë* repräsentiert.[8]

Ein Alphabet wird als geordnete endliche Menge von Schriftzeichen definiert. Die Schriftzeichen (Buchstaben) stehen im Albanischen in folgender Reihe[9]: *a*, *b*, *c*, *ç*, *d*, *dh*, *e*, *ë*, *f*, *g*, *gj*, *h*, *i*, *j*, *k*, *l*, *ll*, *m*, *n*, *nj*, *o*, *p*, *q*, *r*, *rr*, *s*, *sh*, *t*, *th*, *u*, *v*, *x*, *xh*, *y*, *z* und *zh*. Die entsprechenden Lautzeichen nach IPA (International Phonetic Alphabet/Association) sind wie folgt: *a* [a], *b* [b], *c* [ts], *ç* [ʧ], *d* [d], *dh*

---

[3] Vgl., [Demiraj / Prifti 2004: 97f.]: "Kur u shapall ky vendim në mbledhjen e hapur të Kongresit më 20 nëntor para dreke, u tha se *alfabeti i Stambollit ishte i mjaftueshëm për të plotësuar nevojat e kombit, porse, që të mund të shtypeshin libra jashtë Shqipërisë si dhe për telegrame jashtë vendit, nevojitej edhe një alfabet tjetër thjesht latin.*" Vgl. auch Seiten 98f. und 145 (Shtojcë dhe dokumente 2: "Vendimi i Komisionit për çështjen e Abecesë").

[4] D. h. vor allem nicht osmanisch, nicht slawisch und auch nicht griechisch. Für weitere Informationen zu der historisch-politischen Lage um den Kongress von Manastir und viele Details zu dessen Ablauf, vgl. [Demiraj / Prifti 2004].

[5] Vgl. hierzu Fotokopie des Beschlusses des Kongresses von Manastir in [Demiraj/Prifti 2004: 100]. Einen Vergleich der damals meist benutzten Alphabete gibt Dr. Pekmezi in seiner *Grammatik der albanesischen Sprache*, publiziert von „Dija" im Jahre 1908 in Wien, vgl. § 10 (Die Alphabete), Seiten 10f.

[6] Vgl. [Demiraj / Prifti 2004: 102].

[7] Die Zeichen *Ç/ç* und *Ë/ë* waren Bestandteil des französischen Alphabets, das die meisten elektrischen Telegraphen der Zeit beherrschten, denn das Französische war in jener Zeit die Sprache der internationalen Diplomatie.

[8] Im Folgenden wird nur die „zweite" Alphabet-Variante – die sich durchsetzte – berücksichtigt.

[9] Der Einfachheit und Übersicht halber werden im Folgenden die Schriftzeichen, falls der Kontext dies erlaubt, nur klein geschrieben angegeben.

[ð], *e* [ɛ], *ë* [ə], *f* [f], *g* [g], *gj* [ɟ], *h* [h], *i* [i], *j* [ʲ], *k* [k], *l* [l], *ll* [ɫ], *m* [m], *n* [n], *nj* [ɲ], *o* [ɔ], *p* [p], *q* [c], *r* [ɾ], *rr* [r], *s* [s], *sh* [ʃ], *t* [t], *th* [θ], *u* [u], *v* [v], *x* [dz], *xh* [dʒ], *y* [y], *z* [z] und *zh* [ʒ], vgl. [BUCHHOLZ / FIEDLER 1987: 26].

Einen ersten Eindruck über die relative Häufigkeitsverteilung der Schriftzeichen in der albanischen Sprache kann aus [KONCERT], einem kleinen Korpus des Albanischen, gewonnen werden. Es hat 172963 laufende Textwörter (*Token*s) und 1034564 Schriftzeichen.

Das Ergebnis, sortiert nach Häufigkeit des Vorkommens, sieht wie folgt aus:[10]

e 72463	*n* 41506	*p* 21150	*q* 8142	***dh*** 5479	***ll*** 3067
ë 70193	*u* 25772	*d* 19629	*b* 7171	***gj*** 5406	***rr*** 2842
t 60686	*m* 25702	***sh*** 19251	*h* 6922	***nj*** 5241	*c* 1757
i 57577	*s* 25672	*j* 18736	*f* 6376	*z* 4679	***zh*** 708
a 52787	*o* 25524	*v* 9872	*g* 5721	***th*** 4636	***xh*** 255
r 43828	*k* 24427	*l* 9391	*y* 5502	***ç*** 3388	*x* 216

Die nicht zum albanischen Alphabet gehörenden Schriftzeichen *w* und *é* kommen 7 bzw. 6 Mal vor, wie in der folgenden Abbildung ersichtlich ist:

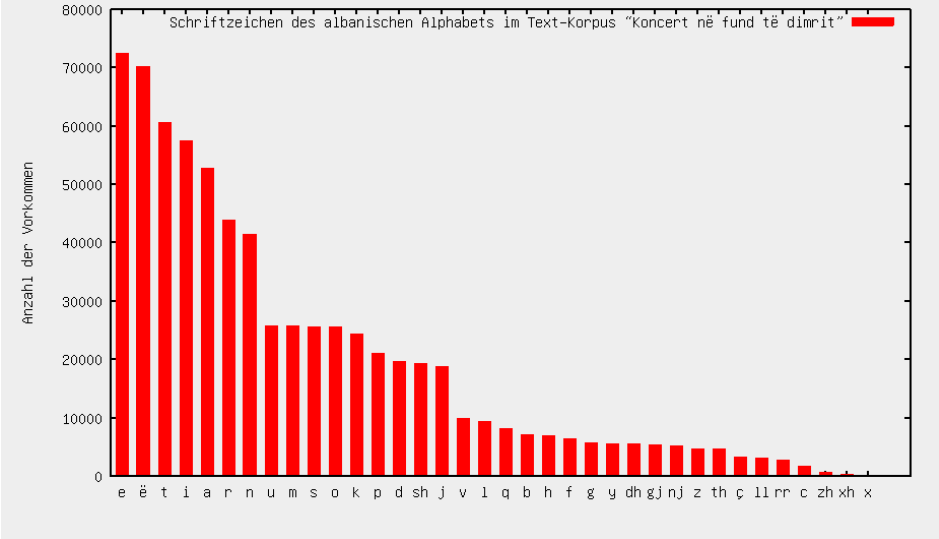

Abbildung 1: Vorkommen der Schriftzeichen des albanischen Alphabets im Text-Korpus „Koncert në fund të dimrit“.

Dabei fällt auf, dass das Zeichen *Ë/ë* sehr häufig vorkommt, besonders in „kurzen Wörtern“ (Funktionswörter: Artikel, Präpositionen, Konjunktionen, ...), die bekanntlich sehr zahlreich[11] vor-

[10] Diese Zahlen wurden mit Hilfe einer automatisch–maschinellen Analyse erstellt und sind dementsprechend als relativ anzusehen. Der untersuchte Text („Koncert në fund të dimrit“), vgl. [KONCERT], stammt aus [ECI / MCI 1994]. Bei dieser Zählung wurde zwischen großen und kleinen Zeichen nicht unterschieden.

[11] Zum Thema Häufigkeitsverteilung der Wörter und Schriftzeichen in Texten vgl. auch die Werke von George Kingsley Zipf, *The Psycho-Biology of Language. An Introduction to Dynamic Philology.* The M.I.T. Press, Cambridge, Mass., 1935 und *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology.*

kommen, z. B. in den Wörtern der Länge 2, wie *të* (8366), *në* (3274), *që* (2589), *më* (1521), *së* (638); Länge 3: *një* (2203), *për* (1882), *atë* (572), *gjë* (433), (*unë* (304), Personalpronomen 1. Person Nominativ Singular), *siç* (201), *çdo* (182), *veç* (144), *nën* (76), *veç* (72), *çka* (16); Länge 4: *ndër* (76). Im Gegensatz zu *Ë/ë* kommt *Ç/ç* deutlich seltener (auch in „kurzen Wörtern“) vor. Die Digraphen, mit Ausnahme von *sh*, treten nur geringfügig auf, was für einen praktischen Vorteil im Umgang mit dem Alphabet spricht.

Das Alphabet dient auch als Kriterium für das *Sortieren* von Zeichenketten in einem Text (alphabetische Sortierung), wobei die Reihenfolge der Schriftzeichen des Alphabets als Ordnungsbegriff (Schlüssel) dafür dient. Auch für die Suche und das schnelle Auffinden von Zeichenketten bzw. Textwörtern, ohne den gesamten Text durchblättern und -lesen zu müssen, dient das Alphabet als Schlüssel.[12]
Die Form und Kombination der Schriftzeichen ist für ihre optische Eindeutigkeit beim Lesen sehr wichtig, in diesem Kontext auch für die *Wortformsegmentierung*, die hier jedoch nicht berücksichtigt wird.[13]

3. Kodierung und Darstellung alphanumerischer Daten

Daten werden am Computer als Folgen von Nullen und Einsen verarbeitet, übertragen und gespeichert. Sie repräsentieren somit eine bestimmte Information. Die Information wird vom Mensch in Form von Ziffern, Schriftzeichen, Interpunktionszeichen oder in einer anderen Form eingegeben. Damit der Computer sie verarbeiten kann, bedarf es ihrer Umwandlung in Folgen von Nullen und Einsen (d. h. „Nein“ und „Ja“), in sogenannte Folgen von Bits. AMELING und KREFT [1996: 1630f.] beschreiben die Information wie folgt:

> „Eine Information I bezeichnet einen abstrakten Sinninhalt. Bei einer Übertragung muß diese zunächst in eine Nachricht N abgebildet (α) werden, die in einem technisch-physikalischen Sinne darstellbar ist; die Nachricht N ist also die Darstellung der Information. Durch eine inverse Abbildung, die Interpretation (α^{-1}), wird die Information wieder zurückgewonnen.“

Wenn zuerst nur Schriftzeichen und Zahlen wie *A/a*, *B/b*, usw. bzw. *1*, *2*, usw. berücksichtigt werden, kann in diesem Sinne das Lautsystem als eine begrenzte Anzahl der Laute einer Sprache als Urbildmenge und die graphischen Zeichen, d. h. das Alphabet einer Sprache als Bildmenge fungieren. Sowohl bei der Abbildung von Urbildmenge in Bildmenge, der *Kodierung*, als auch bei der *Dekodierung*, der Überführung eines Zeichens in Laute, also der Rekonstruktion des ursprünglichen Zeichens, darf die Information nicht verloren gehen.[14]

Addison-Wesley Press, Cambridge, Mass. 1949.

12 Vgl. u. a. lexikalische Suche.

13 Vgl. [NUSHI 1988] für weitere Informationen.

14 Dies ist der Fall, wenn von einem standardisierten Durchschnittswert der Laute und standardisierten Wert der Zeichen einer Sprache ausgegangen wird.
Kodierungen von Zeichen existierten schon vor dem Computerzeitalter: Francis BACON (1561–1626), englischer Philosoph, verwendete für kryptographische Zwecke bereits 1580 einen Binärcode der Länge 5, um das damalige englische Alphabet (24 Schriftzeichen) zu kodieren, wobei die Zeichen lexikographisch angeordnet waren, vgl. Friedrich L. Bauer / Gerhard Goos: *Informatik 1*. 4. Auflage (bearbeitet von Friedrich L. Bauer und Walter Dosch). Berlin / Heidelberg / New York u. a. : Springer-Verlag, 1991. 43f.

* *

Mit einer Bit-Folge der Länge 1 ergeben sich zwei (= 2^1) Kodiermöglichkeiten der Information, d. h. 1 – „ja“ und 0 – „nein“. Mit einer Bit-Folge der Länge 2, ergeben sich dagegen vier, nämlich 00, 01, 10 und 11 (etwa die Dezimalzahlen 1, 2, 3 und 4 oder z. B. die großen Schriftzeichen A, B, C und D). Dies ist aber immer noch weniger als für die Kodierung der Ziffern (0–9), der Schriftzeichen (A–Z und a–z) und der verschiedenen anderen Zeichen, wie denen der Interpunktion, insgesamt benötigt werden. Es werden also längere Bit-Folgen gebraucht. Mit jedem weiteren Bit verdoppeln sich die Möglichkeiten, vgl. $2^3 = 8$, $2^4 = 16$, $2^5 = 32$, $2^6 = 64$ usw. Für die Kodierung der zehn Ziffern sind vier Bit erforderlich (Bit-Folge der Länge 4), wobei sechs Stellen unbelegt bleiben; für die Kodierung der großen Schriftzeichen des lateinischen Alphabets (26) bedarf es einer Bit-Folge der Länge 5, wobei sechs Stellen unbelegt bleiben; mit einer Bit-Folge der Länge 6 wäre es möglich, noch die kleinen Schriftzeichen zusätzlich zu kodieren, wobei selbst hier zwei Stellen unbelegt blieben.

Mit wissenschaftlichen und technologischen Fortschritten kam es zu mehreren Experimenten und zur Entwicklung verschiedener Kodierungen, die mit verschiedener Software und Computern eingesetzt wurden.[15] Die größte Verbreitung erreichte der sog. ASCII–Zeichensatz. ASCII (*American Standard Code for Information Interchange*), der im Jahre 1963 veröffentlicht wurde. Der ASCII–Zeichensatz kodiert die Ziffern, Schriftzeichen und andere spezielle Zeichen sowie Steuerzeichen mit einer Länge von sieben Bit, die zur Fehlererkennung i. d. R. auf acht Bit ergänzt werden. Die effektiv verwendeten 2^7 Stellen ermöglichten 128 Kodierungen in dichtem Code, d. h. bei Belegung aller Stellen. Diese Möglichkeit wurde sehr schnell von vielen Computersystemen mit einer 8-Bit–Architektur für die Darstellung von verschiedenen Zeichen genutzt.

* *

Diese Kodierung wurde von den amerikanischen Behörden unterstützt und von dem *American National Standards Institute* (ANSI) bestätigt. Der Zeichensatz wurde auch für andere Sprachen angepasst, da er diese in seiner ursprünglichen Form nicht abdecken konnte. Die *International Organisation for Standardization* (ISO) normierte den Zeichensatz als ISO/IEC 646.[16]

Die Belegung des ASCII–Zeichensatzes ist in der folgenden Tabelle zu sehen. Die einzelnen Belegungen ergeben sich aus Zeilen und Spalten, wobei die Hexadezimalzeichen in Binärzeichen kodiert sind. Z. B. wird das Schriftzeichen *A* als 41_{16} (hexadezimal) kodiert, als $0100\ 0001_2$ (binär) dargestellt und hat den Wert 65_{10} (dezimal). Entsprechend wird das *Z* als $5A_{16} = 0101\ 1010_2 = 90_{10}$ kodiert.

---

[15] In diesem Abschnitt muss aus Platzgründen auf eine ausführliche Beschreibung und viele Details sowie die Erklärung der fachspezifischen Termini verzichtet werden. Der interessierte Leser vergleiche dazu u. a. [Ameling / Kreft 1996], [Bohn / Flik 2002], [Brockhaus C&IT 2003], [DIN–Normen IT-10], [Duden Informatik], [ECMA-94], [Krückeberg / Spaniol 1990], [Schneider et al. 1997], [Schneider / Werner 2007] und [Zemanek 1967].

[16] IEC steht für *International Electrotechnical Commission*. Vgl. *http://www.iec.ch/* (20.10.2008). Vgl. auch *http://-www.wssn.net/* (20.10.2008).

Code	0 0000	1 0001	2 0010	3 0011	4 0100	5 0101	6 0110	7 0111	8 1000	9 1001	A 1010	B 1011	C 1100	D 1101	E 1110	F 1111
0 0000	NUL 00	SOH 01	STX 02	ETX 03	EOT 04	ENQ 05	ACK 06	BEL 07	BS 08	HT 09	LF 0A	VT 0B	FF 0C	CR 0D	SO 0E	SI 0F
1 0001	DLE 10	DC1 11	DC2 12	DC3 13	DC4 14	NAK 15	SYN 16	ETB 17	CAN 18	EM 19	SUB 1A	ESC 1B	FS 1C	GS 1D	RS 1E	US 1F
2 0010	SP 20	! 21	" 22	# 23	$ 24	% 25	& 26	' 27	(28	) 29	* 2A	+ 2B	, 2C	- 2D	. 2E	/ 2F
3 0011	0 30	1 31	2 32	3 33	4 34	5 35	6 36	7 37	8 38	9 39	: 3A	; 3B	< 3C	= 3D	> 3E	? 3F
4 0100	@ 40	**A** 41	**B** 42	**C** 43	**D** 44	**E** 45	**F** 46	**G** 47	**H** 48	**I** 49	**J** 4A	**K** 4B	**L** 4C	**M** 4D	**N** 4E	**O** 4F
5 0101	**P** 50	**Q** 51	**R** 52	**S** 53	**T** 54	**U** 55	**V** 56	**W** 57	**X** 58	**Y** 59	**Z** 5A	[5B	\ 5C	] 5D	^ 5E	_ 5F
6 0110	` 60	**a** 61	**b** 62	**c** 63	**d** 63	**e** 65	**f** 66	**g** 67	**h** 68	**i** 69	**j** 6A	**k** 6B	**l** 6C	**m** 6D	**n** 6E	**o** 6F
7 0111	**p** 70	**q** 71	**r** 72	**s** 73	**t** 74	**u** 75	**v** 76	**w** 77	**x** 78	**y** 79	**z** 7A	{ 7B	| 7C	} 7D	~ 7E	DEL 7F

Tabelle 1: ASCII / ANSI-Zeichensatz

Diese Kodierung deckt aber viele Alphabete und Zeichen nicht ab, da diese mehr Schriftzeichen besitzen als in der Tabelle Position 41 bis 5A und 61 bis 7A (Schriftzeichenbereiche) zur Verfügung stehen. Es war mehr Platz nötig, um zusätzliche Zeichen anderer Alphabete darzustellen, das £ (Pound–Zeichen) bspw. war nicht vorhanden. Wie aus der Tabelle ersichtlich ist, wurden auch die Zeichen *Ç/ç* und *Ë/ë* des albanischen Alphabets nicht berücksichtigt.

Mit der Verbesserung der Datenverarbeitungstechnologie wurde das Kontroll–Bit, das vor den 7-Bits gesetzt war, überflüssig. Es konnte infolgedessen für die Kodierung weiterer Schriftzeichen benutzt werden. Eine Erweiterung der Kodierung um ein Bit, von 7 zu 8 Bit, bedeutete zusätzliche 128 Positionen (128–255), also insgesamt 256 ($= 2^8$). Somit konnten etliche Zeichen der verschiedenen europäischen Sprachen ihren Platz in einer einzigen Tabelle finden. Dies wurde sehr schnell auch in die Tat umgesetzt. Viele Länder adaptierten die Tabellen, indem sie die Schriftzeichen ihrer nationalen Alphabete in den Tabellen kodierten und so nationale Varianten entwickelten.[17]

Die ISO standardisierte die 8-Bit–Kodierung unter der Norm ISO/IEC 8859 und deckte damit mehrere Sprachen und Regionen ab. Unter dieser Norm wurden die Alphabete der westeuropäischen Sprachen mit der ersten Variante, der ISO-/IEC-8859-1, auch *Latin-1* genannt, kodiert. Vgl. hierzu u. a. [ECMA-94]. ISO-/IEC-8859-2 (Latin-2) deckt die mitteleuropäischen Sprachen ab, ISO-/IEC-8859-3 (Latin-3) die südosteuropäischen Sprachen, ISO-/IEC-8859-3 die baltischen Sprachen. Es folgen die Sprachen, die im kyrillischen Alphabet geschrieben werden (ISO-/IEC-8859-5), Arabisch (ISO-/IEC-8859-6), Griechisch (ISO-/IEC-8859-7), Hebräisch (ISO-/IEC-8859-8) und weitere Sprachen. ISO revidiert und verbessert ständig diese Standards: So sind der ISO-/IEC-8859-15 und ISO-/IEC-8859-16 entstanden, welche die west- bzw. südosteuropäischen Sprachen abdecken.

---

[17] Bei *Drejtoria e Përgjithshme e Standardizimit* (*e Republikës së Shqipërisë*) [DPS] / (*General Directorate of Standardization of the Republic of Albania*), vgl. *http://www.dps.gov.al/* (20.10.2008), konnte keine Information gefunden werden, die mit dem Thema dieser Schrift zu verbinden gewessen wäre.

Code	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0																
1																
2	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	E	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	U	v	w	x	y	z	{	|	}	~	
8																
9																
A	NBSP	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	SHY	®	¯
B	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E	à	á	â	ã	ä	Å	æ	ç	è	é	ê	ë	ì	í	î	ï
F	ð	ñ	ò	ó	ô	Õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Tabelle 2: ISO/IEC 8859-1–Zeichensatz (Latin-1–Zeichensatz)

Während auf der einen Seite das Etablieren eines Standards bzw. einer Kompatibilität angestrebt wurde, erstellten viele Computer- und Softwarehersteller eigene proprietäre Zeichensätze. So entwickelte etwa Microsoft Corp. eigene *Code Pages* (CP) : Für mitteleuropäische Sprachen wurde die CP-1250, die im Wesentlichen ISO-/IEC-8859-2 entspricht, erarbeitet. Für die Sprachen, die in kyrillischen Schriftzeichen/Alphabeten geschrieben werden, die CP-1251, für westeuropäische Sprachen die CP-1252, die im Wesentlichen ISO-/IEC-8859-1 entspricht, für Griechisch die CP-1253, für Türkisch die CP-1254, für Hebräisch die CP-1255, für Arabisch CP-1256, für baltische Sprachen CP-1257, sowie weitere Code Pages für andere Sprachen.

Die Schriftzeichen der albanischen Sprache sind nach den Spezifikationen (*Code Page Global Identifier*, kurz CPGID) von IBM[18] in den folgenden *Code Page*s enthalten: 00912 Latin 2 – ISO; 01250 Windows, Latin 2; 00852 Latin 2 – Personal Computer; 01153 EBCDIC[19] Latin 2 Multilingual with euro; 00870 Latin 2 – EBCDIC Multilingual; 01165 Latin 2 EBCDIC/Open Systems.

Diese Kategorisierung ist regional begründet, denn das albanische Alphabet wird auch z. B. von dem ISO-8859-1–/Latin-1–Zeichensatz, CP 01252, ISO-8859-3–/Latin-3–Zeichensatz (südosteuropäische Sprachen) und weiteren Code Pages abgedeckt, vgl. Tabelle ISO-/IEC-8859-1–Zeichensatz.[20] Auch im Zeichensatz von Mac OS, dem Macintosh Roman, beschrieben in RFC 1345, sind die Schriftzeichen *Ç/ç* und das *Ë/ë* beinhaltet, und zwar auf Positionen 82 das *Ç*, auf 8D das *ç*, auf

[18] Quelle: *http://www-01.ibm.com/software/globalization/cp/cp_language.jsp#Albanian* (20.10.2008).

[19] *Extended Binary Coded Decimals Interchange Code* (EBCDIC), wurde von IBM entwickelt. Dieser arbeitet mit 8-Bit und wird in Großrechnern eingesetzt. Die Zeichensätze dieser Familie werden hier nicht berücksichtigt.

[20] Im Folgenden werden nur die ASCII und die ISO-8859-1 Zeichensätze besprochen – die ISO-8859-2 und andere Zeichensätze dieser Familie sowie CP-1250 (Win Latin 2), CP-1252 (Win Latin 1) usw. funktionieren nach ähnlichen Prinzipien. Die Positionen der *Ç/ç* und *Ë/ë* sind in ISO-8859-1, ISO-8859-2, ISO-8859-3, ISO-8859-4, CP-1250 und CP-1252 gleich, nämlich *Ç* (C7), *ç* (E7), *Ë* (CB) und *ë* (EB). Auf IBM-PC Code Pages bzw. (MS-)DOS Code Pages CP-850 (Latin-1, Westeuropäisch) sowie CP-852 (Latin-2, Mitteleuropäisch), sind die vier Zeichen enthalten und zwar auf anderen Positionen als die in ASCII- bzw. ISO-8859-{1–4}-Tabellen, nämlich *Ç* auf der Position *80*, *ç* auf *87*, *Ë* auf *D3* und *ë* auf *89*.

E8 das *Ë* und auf 91 das *ë*.[21]

Angaben über *Codepage*s zu verschiedenen Sprachen sind auch auf dem Online-Portal der Microsoft Corp.[22] zu finden, vgl. im folgenden diejenigen, welche auf die albanische Sprache zutreffen: (Betriebssysteme *Windows Vista*, *Windows XP* und *Windows Server 2003*): *ANSI codepage 1250*; *OEM codepage 852*.[23]

* *

Mit dem weltweit intensivierten kulturellen und wirtschaftlichen Austausch kam es auch zu einem zunehmenden Datenaustausch zwischen verschiedenen Ländern, Sprachen und wirtschaftlichen Entitäten. Die Datenverarbeitung und -speicherung sowie der Datenaustausch verlangte nach einer Konsistenz dieser Daten. Als Folge entstand in den letzten Jahren ein neuer Standard, der sogenannte Unicode[24], ein zunächst 16-Bit-Code, der 65 536 Zeichen darstellen konnte, ab der Version 2.0 (1996) aber auf 17 Ebenen (engl. *Planes*) erweitert wurde, also auf 1 114 112 (d. h. 17 × 65 536) Zeichen. Version 5.1 definiert 100 713 Zeichen.

Die ISO/IEC 10 646 normierte nach diesem Prinzip das *Universal Character Set* (UCS), eine 4-Byte–Zeichenkodierung, die Platz für 2^{31}, d. h. 2 147 483 648 Zeichen bzw. Zellen (engl. *Code Points*) bietet. UCS besteht aus 128 Gruppen mit je 256 Ebenen, die wiederum in je 256 Reihen mit je 256 Zellen organisiert sind. Der 32. Bit dient als Prüfbit (engl. *Parity-Bit*) für eventuelle Übertragungsfehler, also insgesamt sind es 4 294 967 296 (= 2^{32}) Kodiermöglichkeiten. Die ersten 256 Code Points (der ersten Reihe der ersten Ebene der ersten Gruppe) sind gleich dem ISO/IEC-8859-1–Zeichensatz, vgl. hierzu auch [Bohn / Flik 2002].

Für das Albanische gelten die Zeichenbereiche (engl. *Ranges for the Characters*): A–Z (Code-Positionen 0041–005A), a–z (0061–007A) in *C0, Controls and Basic Latin, Range 000–007F*, wobei das nicht zum albanischen Alphabet gehörende Zeichen W (0057) und w (0077) mit einbezogen sind, sowie die einzelne Code-Positionen *Ç* (00C7), *ç* (00E7), *Ë* (00CB) und *ë* (00EB) in *C1, Controls and Latin-1–Supplement, Range 0080–00FF*.[25]

4. Tastaturbelegung

Eine Tastatur ist bekanntlich ein Gerät, das zur Eingabe von Schrift- und Steuerzeichen dient. Sie hat sich mit der technischen Entwicklung der Hardware und Software zu einer Form und Größe etabliert, auf der ca. 100 Tasten (am häufigsten 101, 102, 104 und 105) Platz finden, die mit beiden Händen bedient werden kann und sich als solche als praktisch erwiesen hat. Da aber eine Tastatur nur eine begrenzte Anzahl an Schriftzeichen fassen und somit nicht beliebig viele Sprachen ab-

---

[21] Vgl. *http://www.unicode.org/Public/mappings/vendors/apple/roman.txt* (20.10.2008) und *http://tools.ietf.org/html/rfc1345* (20.10.2008).

[22] Vgl. *http://www.microsoft.com/* (20.10.2008).

[23] Diese Code Pages entsprechen die *00852 Latin 2 – Personal Computer* und *01250 Windows, Latin 2* bei CPGID–Angaben von IBM, s. o.
Weitere Angaben, bspw. zu Sprach- und Länderidentifikation, sind wie folgt angegeben: Country or Region name abbreviation: *ALB*; Language name abbreviation: *SQI*; LCID Culture Identifier: *0x001C/0x041C*; Culture Name: *sq/sq-AL*; Locale – Language Country/Region: *Albanian*; Language Local: *shqipe*. Vgl. *http://www.microsoft.com/globaldev/nlsweb/default.mspx* (20.10.2008).

[24] Vgl. *http://www.unicode.org/* (20.10.2008).

[25] Vgl. hierzu die ISO/IEC 8859-x to Unicode Mapping Tables unter *ftp://ftp.unicode.org/Public/MAPPINGS/ISO8859/* (20.10.2008).

decken kann, ist es nötig, abweichende Tastaturen zu entwickeln. So entstanden verschiedene Tastaturen für verschiedene Sprachen und Regionen.

Für die Bezeichnung der albanischen Sprache wurde in *Norm ISO 639* der Schlüssel *ISO_639-1_sq* und *ISO_639-2_alb_sqi* vergeben. Für die Tastaturbelegung (engl. *Keyboard Layout*) nach Norm *ISO-3166-1-alpha-2 code*, steht dafür das Identifikationszeichen[26] *AL* zur Verfügung. Es wird aber oft stattdessen oder parallel dazu das Zeichen *SQ* oder *SQI* benutzt, vgl. hierzu z. B. *Microsoft Windows XP*.

Die Tastaturbelegung für die albanische Sprache ist von IBM Corp.[27] unter den Nummern *448* (*KBDID 448*) und *452* (*KBDID 452*) normiert.[28]

Die Tastaturbelegung mit der Nummer 448 (KBDID 448) ist laut IBM eine Revision von Nummer 114. Sie trägt das Registrierdatum 1993-03-10 (letzte Revision am 1999-04-20, ergänzt um das Euro–Zeichen). Die Schriftzeichen *Ç/ç* befinden sich auf Position D 11 L2 bzw. L1, *Ë/ë* auf C 10 L2 bzw. L1, wobei L2 die Shift–Funktion ist, d. h. große Schriftzeichen. Die Position der beiden Zeichen ist auf dieser Tastaturbelegung sehr gut gelungen; sie ist ergonomisch und praktisch. Die meistbenutzte Reihenfolge der Schriftzeichen auf einer Tastaturbelegung ist erhalten worden – auch das Schriftzeichen *W/w*, das nicht zum albanischen Alphabet gehört – und wurde nur erweitert. Das *Ë/ë* ist an der nächstmöglichen Position angebracht, also nicht weit außen, da es sehr häufig vorkommt und so besser und leichter bedient werden kann. Die Position des *Ç/ç* ist auch gut definiert, denn es kommt deutlich seltener vor als *Ë/ë*, vgl. hierzu die Abbildung 1.

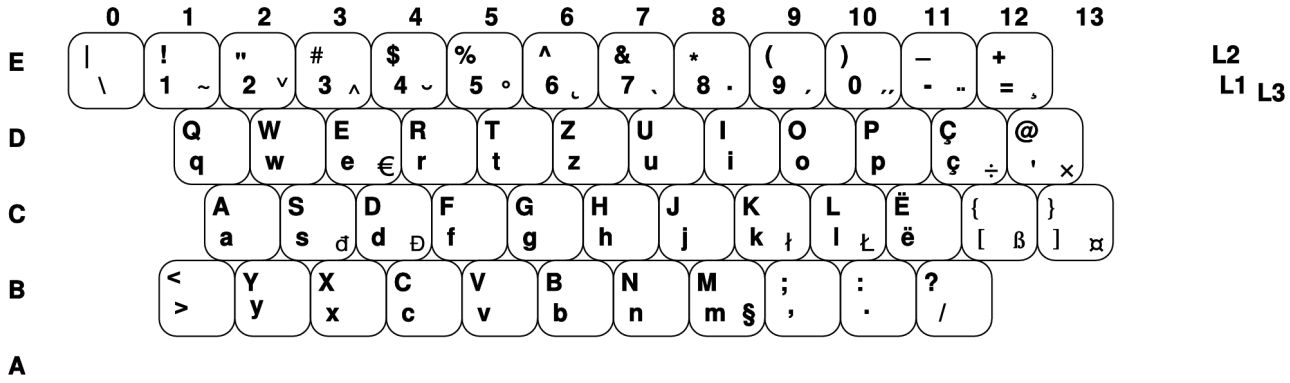

Abbildung 2: Tastaturbelegung für Albanisch (KBD448 – nach IBM)

Auf der Zeile E (vgl. Abbildung Tastaturbelegung für Albanisch – KBD448) sind eine Reihe von Zusatzzeichen (Diakritika) auf dem Level 3 (L3), d. h. Alt Gr-Funktion, angebracht. Sie werden vor allem für die schriftliche Kommunikation in Nachbarsprachen des Albanischen (italienisch, slawische Sprachen mit lateinischem Alphabet oder transliteriert und anderen Sprachen, die dadurch abgedeckt werden) gebraucht. Die Zeichen ¸ (Position E 12 L3) und ¨ (Position E 11 L3), die mit den Zeichen *C/c* bzw. *E/e* kombiniert werden können, um *Ç/ç* bzw. *Ë/ë* einzugeben, werden aber auf einer albanischen Tastatur nicht benötigt, wohl aber für das Schreiben der *Ä/ä*, *Ö/ö* und *Ü/ü*. So

[26] Vgl. *http://www.iso.org/iso/country_codes/iso_3166_code_lists/english_country_names_and_code_elements.htm* (20.10.2008).

[27] Vgl. *http://www.ibm.com/us/* (20.10.2008).

[28] Vgl. *http://www-01.ibm.com/software/globalization/topics/keyboards/registry_index.jsp* (20.10.2008) und *ftp://ftp.software.ibm.com/software/globalization/keyboards/KBD452.pdf* (20.10.2008).

kann mit dieser Tastaturbelegung auch das Deutsche abgedeckt werden – das *ß* (C 11 L3) ist ebenso vorhanden. Auf Position C 2 L3 befindet sich das *đ* und auf C 3 L3 das *Đ*, auf Position C 8 L3 das *ł* und auf C 9 L3 das *Ł*, Zeichen, die alle nicht dem albanischen Alphabet angehören, sondern verschiedenen Alphabeten der slawischen Sprachen. Sie dienen dazu, beabsichtigt oder nicht, die entsprechenden Sprachen auf der albanischen Tastaturbelegung abzudecken, was als Vorteil zu sehen ist.

Die zweite, spätere Tastaturbelegung für Albanisch KBDID 452 zeichnet eine Veränderung der Position der „kritischen" Schriftzeichen des Albanischen, dem *Ç/ç* und *Ë/ë*. Es ist sehr interessant, dass die Position des Zeichens *Ë/ë* hier schlechter angebracht ist. Die Eingabe des Zeichens *Ë/ë* wird deutlich erschwert durch die Positionierung (noch) weiter nach außen und oben. Es bleibt unklar, was die Motivation war, das *Ç/ç* weiter als das Doppelpunkt-Zeichen (C 10 L2) bzw. das Semikolon-Zeichen (C 10 L1) zu positionieren sowie das *Ë/ë* weiter nach außen und oben (D 11 L1 und D 11 L2) zu verlegen. Ebenso bleiben die weiteren Veränderungen unklar.

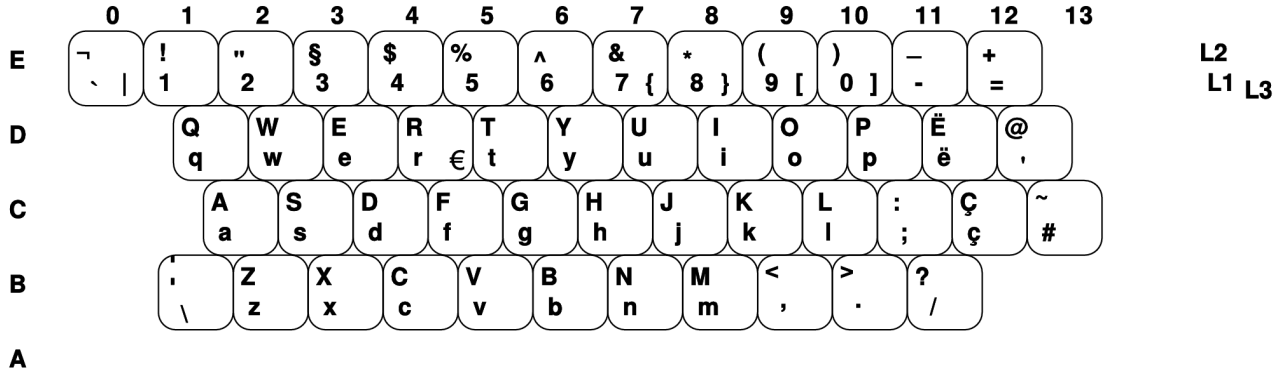

Abbildung 3: Tastaturbelegung für Albanisch (KBD452 – nach IBM)

* *

Die Eingabe der Zeichen ist in verschiedenen Formen z. Z. möglich, sowohl lokal von einer Anwendung oder von dem Programm als auch global, d. h. für alle Programme. Die Eingabe der Sonderzeichen kann z. B. beim Programm *OpenOffice*[29] durch den Menüpunkt Einfügen über den Punkt Sonderzeichen ausgeführt werden. Fast jedes Programm hat eine Möglichkeit, die Kodierung (engl. *encoding*) einzustellen. Unter dem Betriebssystem Linux vgl. hierzu das Programm *Gucharmap* (Gnome Character Map), das zur UCS–/Unicode–Zeicheneingabe dient.[30]

Die Eingabe der beiden Schriftzeichen *Ç/ç* und *Ë/ë* ist bei anderen Tastaturbelegungen auch durch die Kombination von ASCII-Schriftzeichen mit Diakritika möglich, wie die Kombination von ¸ mit *C/c*, die *Ç/ç* ergibt, ebenso wie die Kombination von ¨ mit *e*, die die Eingabe von *Ë/ë* ermöglicht. Auf einer *generischen 105-Tasten-(intl)*–Tastatur, deutsche Tastatur mit *de-default*–Einstellung, ergibt das gleichzeitige Drücken der Tasten Alt Gr und ´ gefolgt von *C/c* das Zeichen *Ç* bzw. *ç* so wie Alt Gr und *Ü/ü* gefolgt von *E* oder *e* das Zeichen *Ë* bzw. *ë*.

In vielen Fällen werden die Stellen des ASCII-Zeichensatzes als ASCII *Position*, z. B. ASCII 65 für das Zeichen A bezeichnet. Sie können in einigen Betriebssystemen und Anwendungen mit dem

[29] Vgl. *http://www.openoffice.org/* (20.10.2008).

[30] Hier können leider nicht alle möglichen Anwendungen und Programme sowie ihre Fähigkeiten und Besonderheiten berücksichtigt werden.

gleichzeitigen Drücken der Alt-Taste und Eingabe der Zeichen-Position als dreistelliger Schlüssel eingegeben werden: Das Schriftzeichen *Ç*, Position 128 (IBM-PC) kann mit der Tastenkombination [Alt + [1 2 8]] eingegeben werden, *ç* mit [Alt + [1 3 5]], *Ë* mit [Alt +[2 1 1]] und entprechend das *ë* mit [Alt + [1 3 7]], wobei die Ziffern im Ziffernblock eingegeben werden müssen.

Mit der zunehmenden elektronischen Kommunikation und dem häufigen Gebrauch von mehreren verschiedenen Tastaturen innerhalb kurzer Zeit kam es zur Idee der Virtualisierung der Tastaturen als Möglichkeit, die gleiche Tastatur für verschiedene Sprachen zu verwenden.

Unter dem Betriebssystem Linux (Desktop KDE) ermöglicht das Hilfsprogramm *kxkb* das Umschalten zwischen verschiedenen Tastaturbelegungen.

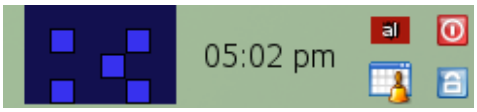

Abbildung 4: kxkb auf der KDE-Kontrollliste unter Linux: Das Identifikationszeichen AL befindet sich auf der albanischen Fahne.

Die Microsoft Corp. stellt auch eine Tastaturbelegung für albanisch zur Verfügung, die mit IBMs Belegung (KBD448) für das Albanische im Wesentlichen übereinstimmt.[31] Die Unterscheidungen liegen im L3. Die Microsoft–Tastaturbelegung hat zusätzlich die Zeichen *[* und *]*, d. h. Doppelbelegung, in den Positionen C 4 L3 bzw. C 5 L3, auf der Position C 10 L3 ein zusätzliches *$*–Zeichen, auf der Zeile B weitere zusätzliche Zeichen: Auf B 4 L3 das *@*, auf 5 B 5 L3 das *{*, auf B 6 L3 das *}*, auf B 7 L3 das *§*, auf B 8 L3 das < und auf B 9 L3 das >.

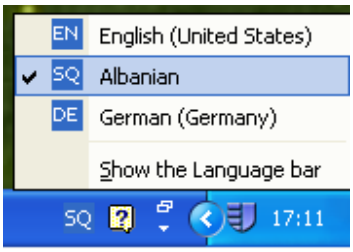

Abbildung 5: Einstellung der Tastaturbelegung unter Microsoft Windows XP: Das SQ ist hier als Identifikationszeichen für Albanisch.

Unter dem Betriebssystem MacOS X von *Apple Inc.* ist bis zur aktuellen Version 10.5 leider noch keine Tastaturbelegung für albanisch vorhanden.

Mobile Telefone und PDAs (*Personal Digital Assistant*), Smartphones, PocketPCs usw. stellen einen Bereich dar, in dem immer noch proprietäre Software eingesetzt wird und kein einheitlicher und reibungsloser Textaustausch stattfindet. Nicht alle Geräte haben die Zeichen der ISO-8859-1–Tabelle, d. h. *Ç/ç* und *Ë/ë* fehlen teilweise.

---

[31] Vgl. *http://www.microsoft.com/globaldev/keyboards/kbdal.htm* (20.10.2008). Mit B 3 L3 kann das ©–Zeichen eingegeben werden, diese Position ist allerdings von Microsoft Corp. nicht dokumentiert. Das €–Zeichen fehlt.

5. Spracheinstellung, Auszeichnungssprachen und Textverarbeitung

Die Spracheinstellung beeinflusst in vielen Hinsichten die Arbeitsumgebung bzw. das Verhalten der Programme, wie z. B. die Sortierung[32], weshalb es wichtig ist, dass die Variablen für die Spracheinstellungen entsprechend richtig gesetzt sind. Bei dem UNIX-artigen Betriebssystemen muss in der *locale* (*Internationalization Variable*) *LANG* auf *LANG=sq_AL* oder *LANG=sq_AL.utf8* für UTF-8 gesetzt werden. Nach dieser Einstellung wird, wie erwartet, *Ç/ç* vor *Ë/ë* richtig sortiert, sonst kommen die vier Zeichen nach den anderen Zeichen, jeweils als das 35. bzw. 36. Schriftzeichen statt des 4. bzw. 8., da sie in der Zeichentabelle höhere Positionsnummern haben als die lateinischen Schriftzeichen (der Bereich 0–127 ASCII). Die Einstellungen für die Sortierung können unter verschiedenen Anwendungen und Umgebungen jeweils lokal oder auch global gewählt werden.

Die Sortierung mit dem albanischen Alphabet ist im Vergleich zu anderen europäischen Sprachen, etwa dem Deutschen, das hier als ähnlicher Fall betrachtet wird, leichter und eindeutig. Im Albanischen ist die Ordnung „*e* vor *ë*“ und „*c* vor *ç*“ eindeutig. Im Deutschen gibt es oft Unklarheiten, insbesondere bei der Sortierung der Namen. Es wird im Normalfall zwischen *ä*, *ö*, *ü* und *ß* und *ae*, *oe*, *ue* und ss nicht unterschieden.[33] Doch in der Praxis kommt es durch verschiedene länderspezifische Regeln (Deutschland, Österreich, Schweiz) sowie verschiedenen Anwendungen (Wörterbücher, Telefonregister usw.) zu unterschiedlichen Sortierungen.
Die Digraphen *dh*, *gj*, *ll*, *nj*, *rr*, *sh*, *th* und *xh* bereiten keine Probleme bei der Erstellung von Sortier- oder Frequenzlisten, da sie einer linearer Ordnung entsprechen, z. B. *gi...*, *gj...*, *gk...*, usw. Bei einer Gesamtsuche in Wörterbüchern gibt es getrennte Kapitel/Schriftzeichen ... *g*, *gj*, *h*, ..., wobei sie nicht ersetzt werden können, wie es z. B. bei *ae* durch *ä* im Deutschen möglich ist. Nur bei einer möglichen Schriftzeichenzählung käme es bei einer falschen Einstellung der *locale* und ähnlichen Variablen zu einer separaten Zählung der Digraphen, z. B. bei *sh* zu getrennter Zählung von *s* und *h*, wodurch die Zählung verfälscht werden würde.[34]
Im Albanischen entspricht ein Zeichen (d. h. auch die Digraphe dh, ...) im Alphabet exakt einem Laut: Die Zeichen variieren nicht nach Kontext, vgl. z. B. im Deutschen *ü/y*, *ei/ai/ay*, *eu/äu*, *dt/t*, *v/f* usw., wodurch im Albanischen das Lesen erleichtert wird. Auch beträgt die maximale Länge der Grapheme zwei, im Vergleich zum Deutschen *sch*, *dsch*, *tsch* usw.[35]

* *

Bei der Auszeichnungssprache HTML (*HyperText Markup Language*)[36], die für die Darstellung von

[32] Viele sprachbedingte Funktionen und Dienste, u. a. Rechtschreibung, zum Teil Silbentrennung usw., werden hier nicht besprochen, da diese nicht direkt mit dem Aphabet in Zusammenhang stehen. Hier wird die Sortierung hervorgehoben.

[33] Vgl. hierzu DUDEN, Die deutsche Rechtschreibung. Mannheim/Leipzig/Wien/Zürich: Dudenverlag, 2000. ISBN : 3-411-04012-2.

[34] Dies kann natürlich durch Erstellung spezieller Programme auch vermieden werden, sofern z. B. s und h auf einer Morphemgrenze aufeinander treffen, s.u.

[35] Vgl. hierzu 69-106. [In:] DUDEN, Das Aussprachewörterbuch. Mannheim/Leipzig/Wien/Zürich: Dudenverlag, 2000. ISBN : 3-411-04064-5.

[36] Für weiterführende Informationen, vgl. *The World Wide Web Consortium* (*W3C*) unter *http://www.w3.org/* . Für HTML-Spezifikationen (Character entity references in HTML 4, *http://www.w3.org/TR/html4/sgml/entities.html*)

Webseiten am meisten verwendet wird, wurde insbesondere für Zeichen, die nicht in der 7-Bit ASCII-Tabelle sind, ein eindeutiger Schlüssel definiert, um die Eingabe zu erleichtern und den verschiedenen Kodierungen von HTML–Dokumenten und Browser–Einstellungen eine Kompatibilität bei der Interpretation zu ermöglichen.

Das Schriftzeichen *Ç* (LATIN CAPITAL LETTER C WITH CEDILLA) ist in HTML als *Ç* oder *Ç* kodiert, *ç* (LATIN SMALL LETTER C WITH CEDILLA) als *ç* oder *ç*, *Ë* (LATIN CAPITAL LETTER E WITH DIAERESIS) als *Ë* oder als *Ë* und *ë* (LATIN SMALL LETTER E WITH DIAERESIS) entsprechend *ë* oder *ë*.

Die Ziffern des Dezimalsystems zeigen die Position der Zeichen in der ASCII–/ISO-8859-1–Tabelle, dort in Hexadezimalsystem $C7_{16}$ = dezimal 199_{10}, $E7_{16} = 231_{10}$, $CB_{16} = 203_{10}$ und $EB_{16} = 235_{10}$.

In LaTeX, einer Auszeichnungsprache zum Textsatz, können u. a. *Ç* als *\{C}*, *ç* als *\{c}*, *Ë* als *\"E* und *ë* als *\"e* eingegeben werden und erscheinen erst nach der Kompilierung als gewöhnliche Schriftzeichen.

* *

Die Speicherung der Texte ist von zentraler Bedeutung, denn oft können die Daten, die in einer bestimmten Kodierung gespeichert wurden, nicht unter anderen Kodierungseinstellungen ohne Probleme und fehlerfrei zur Bearbeitung geöffnet werden. Daher ist es wichtig, die Daten bei der Erstellung und bei der Bearbeitung bzw. Verarbeitung einheitlich in einer entsprechenden kompatiblen Kodierung zu speichern.

Im Folgenden ist in den Abbildungen 6 und 7 die Kombination verschiedener *shell*-Einstellungen, nämlich ISO-8859-1 bzw. UTF-8 (8-Bit *Unicode Transformation Format*)[37] mit der jeweils nicht passenden Kodierung zu sehen, die den identischen Text darstellen:

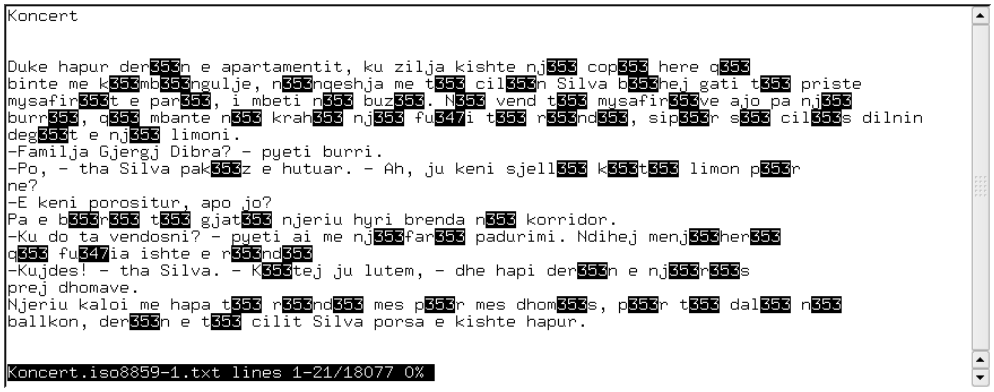

Abbildung 6: Ein ISO-8859-1–kodierter Text unter einer UTF-8-Einstellung in shell.

Die Zahlen 347 und 353 stellen die Schriftzeichen *ç* und *ë* dar. Es handelt sich hierbei um Oktalzahlen: $347_8 = 231_{10} = E7_{16} = ç$ und $353_8 = 235_{10} = EB_{16} = ë$, vgl. hierzu auch die Tabelle 2

(20.10.2008).

[37] UTF-8 ist die z. Z. am häufigsten benutzte Technik für die Kodierung der Unicode-Zeichen in Binärformat, wobei eine Bytekette von variabler Länge verwendet wird. Vgl. hierzu *The Internet Engineering Task Force*, *Request for Comments* Page *http://www.ietf.org/rfc/rfc2279.txt* (20.10.2008) für weitere Informationen zu UTF-8, sowie *http://www.unicode.org/reports/* (20.10.2008) und [UNICODE 2006: §3.9] für zusätzliche Informationen auch zu anderen Kodierungen.

(ISO-8859-1) und den Abschnitt über Auszeichnungssprachen.

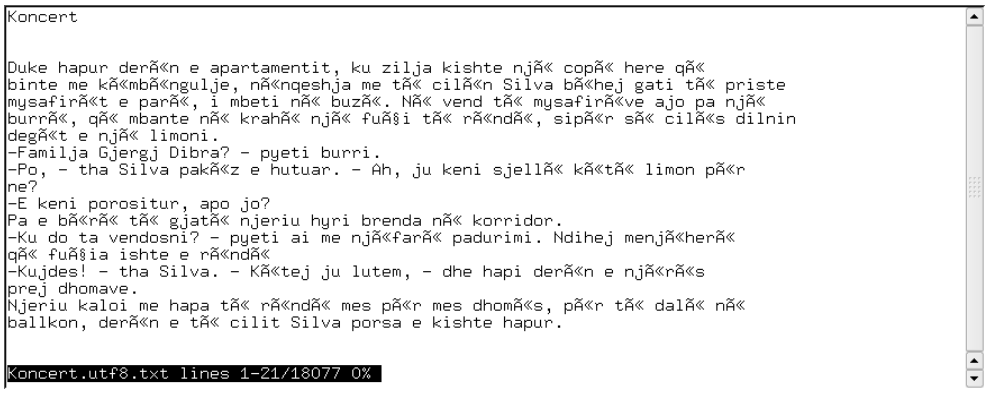

Abbildung 7: Ein UTF-8-kodierter Text unter einer ISO-8859-1–Einstellung in shell.

Das Zeichenpaar *Ã§* stellt die Kodierung von zwei Bytes in UTF8 dar, die im Hexadezimalsystem C3A7 entspricht und im Binärsystem 11000011 10100111. Bei der UTF8–Kodierung der UCS–/Unicode–Zeichen werden, je nach Zahl der Bytes, verschiedene Bit-Muster für die Kodierung gebraucht – bei zwei Byte, wie in diesem Fall, werden die ersten drei Stellen des ersten Byte und die ersten zwei Stellen des zweiten Byte, vgl. 11000011 10100111 gebraucht, d. h. für die Darstellung des Zeichens selbst bleiben die restlichen Bits 00011 100111, d. h. 1110 0111, die im Hexadezimalsystem den Wert E7 ergeben, der schließlich das Zeichen *ç* darstellt. Entsprechend $\tilde{A}\text{«}_{\text{iso8859-1:hex}} = \text{C3AB}_{\text{utf8:hex}} = 11000011\ 10101011_{\text{utf8:bin}} \rightarrow 1110\ 1011_2 = \text{EB}_{16} = \ddot{e}$;

Die Schriftzeichen *Ç* und *Ë* werden in ISO-8859-1–Modus nicht richtig dargestellt. Für beide Zeichen erscheint nur das Zeichen *Ã* (C3). Dahinter stecken die Zahlen *87* (C387) für *Ç* und *8B* (C38B) für *Ë*, welche mit einem Hexadezimaleditor (bspw. Emacs–*hexl-mode*) gesehen werden können. Da aber in der ISO-8859-1–Tabelle die Zeilen mit 8 und mit 9 für Steuerzeichen vorgesehen sind, kann das zweite Byte nicht dargestellt werden, sondern nur das erste, nämlich C3 bzw. das Zeichen *Ã*. So werden beide Schriftzeichen *Ç* und *Ë* mit demselben Zeichen in ISO-8859-1–Mode, mit dem Zeichen *Ã* dargestellt, wodurch der Text bzw. die unsichtbaren Zeichen leichter zerstört werden könnten.

Im Text können auch andere Nicht-ASCII–Zeichen vorhanden sein, die nach diesem Prinzip zu entschlüsseln wären, falls bei der Textverarbeitung bzw. -speicherung Fehler begangen wurden.

Im Folgenden wurde der ISO-8859-1 kodierte Text zur Bearbeitung mit *OpenOffice* geöffnet. Die Schriftzeichen *Ç/ç* und *Ë/ë* werden beide durch das **?**, s. u., ersetzt, also von Zeichen, die nicht dargestellt werden können.

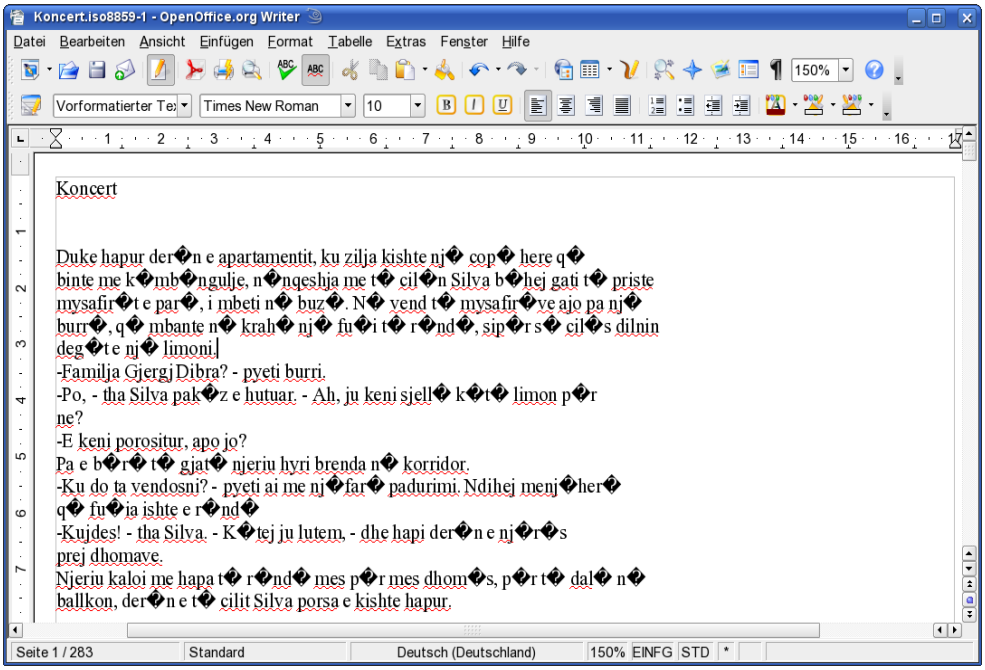

Abbildung 8: Ein ISO-8859-1–kodierter Text geöffnet unter OpenOffice (v.2.4.0).

* *

Die linguistischen Aspekte des Alphabets (und der Rechtschreibung) bilden ein Thema für sich, allerdings können hier kurz einige angesprochen werden, die insbesondere mit den Digraphen zu tun haben:

Bei Digraphen darf bei Zeilenumbrüchen kein Umbruch zwischen den Digraphenbestandteilen stattfinden, wie z. B. zwischen *d* und *h* in *mble**dh**ur*.

Die gelegentlichen Wortbildungen und die zusammengesetzten Neologismen, bei denen die Pseudo–Digraphen an der Stelle der Zusammensetzung entstehen, führen zu Schwierigkeiten und eventuellen Verfälschungen bei der automatischen Sprachverarbeitung.[38]

Bei einer automatischen Wortformsegmentierung und -analyse von Komposita wie *gaz**·h**edhëse*, *gjok**s·h**apur*, *shta**t·h**edhur*, *me**s·h**ollë*, *mo**s·h**yrje*[39] usw. ist darauf zu achten, dass das letzte Schriftzeichen einer Wortform des Kompositums mit dem ersten Schriftzeichen der nächsten Wortform einen Digraphen bildet. An diesen Stellen müsste man beachten, dass die Digraphe richtig erkannt, zugeordnet und analysiert werden.[40] Ein Allomorph-Ansatz der Wortformerkennung wäre von Vorteil in diesem Zusammenhang, vgl. hierzu [Kabashi 2004].

* *

[38] D. h. bei Kontakt der letzten Schriftzeichen der ersten Wortform$_1$ mit dem ersten Schriftzeichen der zweiten Wortform$_2$ der Form $[d|s|t|x|z]_{w1}+[h]_{w2}$, $[g|n]_{w1}+[j]_{w2}$, $[l]_{w1}+[l]_{w2}$ oder $[r]_{w1}+[r]_{w2}$.

[39] Vgl. hierzu [Përnaska / Douchet 2006] für weitere Beispiele.

[40] Es treten weitere Probleme auf, die sowohl mit dem Alphabet als auch mit der Rechtschreibung verbunden sind, wie etwa die Form *lagie*, die eigentlich *lag|je* geschrieben hätte werden müssen – da es aber in diesem Fall um die Vermeidung des Kontaktes *g-j* geht, wird die Wortform, wie ebenso ähnliche Fälle, ortographisch abweichend kodiert. Es könnte hier auch von einer gewissen Sperrfunktion der Digraphen die Rede sein.

Zum Schluss kann noch eine mögliche Aufnahme der Digraphe *dh*, *gj*, *ll*, *nj*, *rr*, *sh*, *th*, *xh* und *zh* in Unicode/UCS als einzelne Zeichen diskutiert werden. Auf der Page *Latin Extended-B* befindet sich der Digraph *NJ* (*Code-Point* 01CA, bzw. als Unicode-*Code-Point* mit *U*+01CA dargestellt; UTF-8, hexadezimal: C78A; Name: LATIN CAPITAL LETTER NJ) bzw. *Nj* (U+01CB; C78B; LATIN CAPITAL LETTER N WITH SMALL LETTER J) und *nj* (U+01CC; C78C; LATIN SMALL LETTER NJ). Ebenso könnten die restlichen Digraphen in UCS/Unicode aufgenommen werden, vgl. hierzu weitere Code-Points dieses Typs. Auch die Aufnahme der Schriftzeichen der alt-albanischen Autoren wäre von Vorteil, vgl. hierzu auch [KABASHI 2007 Z].

6. Schlussbemerkung

Ein rein lateinisches Alphabet hätte den Vorteil gehabt, dass es die bis heute noch andauernden Schwierigkeiten und Probleme, insbesondere mit den Buchstaben *Ç/ç* und *Ë/ë*, nicht gäbe. Beim Umgang mit den vier Zeichen, etwa beim Designen/Prägen von Münzen als auch beim Designen/-Drucken von Banknoten, gibt es 100 Jahre nach dem *Kongress von Manastir* immer noch Schwierigkeiten, trotz des vereinfachten Alphabets.

Der Kongress von Manastir kann aus sprachtechnologischer Sicht dennoch als Erfolg gewertet werden. Die erste Alphabet-Variante hätte deutlich mehr Schwierigkeiten bereitet und eine Transliteration notwendig gemacht. Falls sie sich durchgesetzt hätte, wäre wahrscheinlich erst mit der Durchsetzung des UCS/Unicode eine Normalisierung bzw. Beseitigung der Schwierigkeiten im Umgang mit dem Alphabet gekommen.

Eine Entscheidung, wie etwa die Zeichen *Ç/ç* und *Ë/ë* respektiv durch Digraphe wie etwa *CH/Ch/ch* oder *EH/Eh/eh* darzustellen, wäre eine Möglichkeit. Doch wie sähe das Schreiben des Hilfsverbs *jam* (*sein*) 3. Person Singular Präsens Indikativ Aktiv, *është,* in dieser Schreibweise aus? So: ***Eh****shte****h***! Die optische Seite wäre kein gutes Argument. Und es wären doppelt so viele Zeichen einzugeben – bei zahlreichem Häufigkeitsvorkommen.

7. Referenzen / Literatur

[AMELING / KREFT 1996]

Walter Ameling / Lothar Kreft : „Technische Kodierungen“. 1629–1638 (§ 148) [In:] Hartmut Günther / Otto Ludwig (Hrsg.) : *Schrift und Schriftlichkeit / Writing and Its Use*. Band 10/2. (Handbücher zur Sprach- und Kommunikationswissenschaft. Hrsg. v. Hugo Steger / Herbert Ernst Wiegand.) Berlin / New York : Walter de Gruyter, 1996. ISBN : 3-11-014744-0.

[BOHN / FLIK 2002]

Wilhelm F. Bohn / Thomas Flik : „Zeichen- und Zahlendarstellungen“. 169–190 (§ B1) [In:] Peter Rechenberg / Gustav Pomberger : *Informatik-Handbuch*. 3., aktualisierte Auflage. München / Wien : Carl Hanser Verlag, 2002. ISBN : 3-446-21842-4.

[BROCKHAUS C&IT 2003]

Der Brockhaus *Computer und Informationstechnologie*. Leipzig / Mannheim, F. A. Brockhaus GmbH, 2003. ISBN-10 : 3-7653-0251-1.

[BUCHHOLZ / FIEDLER 1987]

Oda Buchholz / Wilfried Fiedler : *Albanische Grammatik*. Leipzig : VEB, Enzyklopädie, 1987. ISBN-10 : 3-324-00025-4.

[DEMIRAJ / PRIFTI 2004]

Shaban Demiraj / Kristaq Prifti : *Kongresi i Manastirit*. Tiranë : Akademia e Shkencave e Shqipërisë. Instituti i Historisë, 2004. ISBN-10 : 99943-614-5-7.

[DIN–NORMEN IT-10]

DIN Deutsches Institut für Normung e.V. (Hrsg.) : *Zeichenvorräte und Codierung für den Text- und Datenaustausch*. 3. Auflage. Berlin / Wien / Zürich : Beuth, 1998. ISBN-10 : 3-410-12945-6. Vgl. *http://www.din.de/* (10/2008).

[DUDEN INFORMATIK 2006]

Volker Claus / Andreas Schwill : *Duden Informatik A–Z. Fachlexikon für Studium, Ausbildung und Beruf*. Mannheim / Leipzig / Wien / Zürich : Bibliographisches Institut; 4. Auflage, 2006. ISBN-13 : 978-3411052349.

[ECI/MCI 1994]

ELSNET : *European Corpus Initiative, Multilingual Corpus I (ECI/MCI) CD-ROM*. Utrecht : ELSNET, 1994.

[ECMA-94 1986]

European Computer Manufacturers Association (ECMA) : *Standard ECMA-94. 8-Bit Single-Byte Coded Graphic Character Sets*. *Latin Alphabets No. 1 to No 4*. 2nd Edition – June 1986. Geneva : ECMA. Vgl. *http://www.ecma-international.org/publications/standards/* (10/2008).

[KABASHI 2004]

Besim Kabashi : „Analiza automatike e fjalëformave të gjuhës shqipe“. 129–135. [In:] *Seminari Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare, XXIII*. Prishtinë : Universiteti i Prishtinës, 2004. *Libri 23/1*.

[KABASHI 2007 Z]

Besim Kabashi : „Zeichen für Gjon Buzuku. Die Zusammenarbeit zwischen der albanischen Linguistik und der Computerlinguistik.“ 139–146. [In:] Bardhyl Demiraj (Hrsg.) : *Nach 450 Jahren. Buzukus „Missale“ und seine Rezeption in unserer Zeit*. (*Albanische Forschungen 25*). Wiesbaden : Harrasowitz, 2007. ISBN-13 : 978-3-447-05468-3.

[KONCERT]

Der Roman “Koncert në fund të dimrit” von I. Kadaré, publiziert 1988 in Tirana vom Verlag “Naim Frashëri”. [In:] [ECI/MCI 1994]; Korpus-Datei *alb01*.

[KRÜCKEBERG / SPANIOL 1990]

Fritz Krückeberg / Otto Spaniol (Hrsg.) : *Lexikon Informatik und Kommunikationstechnik*. Düsseldorf : VDI-Verlag, 1990. ISBN-10 : 3-18400894-0.

[Nushi 1988]

Pajazit Nushi : *Sistemi i grafisë së tingujve të shqipes dhe vetitë perceptive e përmasat e lexushmërisë së shkronjave të alfabetit të gjuhës shqipe*. Prishtinë : Instituti Albanologjik i Prishtinës, 1988.

[Përnaska / Douchet 2006]

Remzi Përnaska / Zhan-Lui Dyshe (Jean-Louis Douchet) : „Vështrim përfaqësues midis alfabetit fonetik ndërkombëtar (AFN) dhe atij të gjuhës shqipe“. 49–60. [In:] *Studime 12* (*2005*). Prishtinë : Akademia e Shkencave dhe e Arteve e Kosovës, 2006.

[Schneider / Werner 2007]

Uwe Schneider / Dieter Werner (Hrsg.) : *Taschenbuch der Informatik*. 6. Auflage. München : Hanser, 2007. ISBN-13 : 978-3-446-40754-1.

[Schneider et al. 1997]

Hans-Jochen Schneider (Hrsg.) : *Lexikon Informatik und Datenverarbeitung*. 4. Auflage. München / Wien : Oldenburg, 1997. ISBN-10 : 3-486-22875-7.

[Unicode 2006]

Julie D. Allen et al. (Ed.) : *The Unicode 5.0 Standard*. Upper Saddle River [etc.] : Addison-Wesley, 2006. ISBN-13 : 978-0-321-48091-0. Vgl. *http://www.unicode.org/* (10/2008).

[Zemanek 1967]

Heinz Zemanek : *Alphabets and Codes 1967*. München / Wien : R. Oldenburg, 1967.