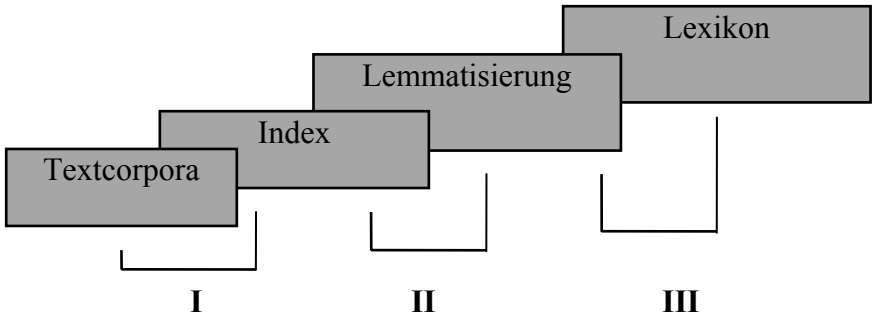
Umwandlung von Texten in Lexika:

Textkorpus, Index, Lemmatisierung, Lexikon

1. Elektronische Datenverarbeitung als zeitgemäßes Forschungsverfahren.

Mit der Entwicklung der Computerlinguistik und den dadurch gegebenen Möglichkeiten der Erforschung und Verarbeitung von Sprachmaterialen hat sich eine eigenständige und in methodologischer Hinsicht neue Etappe der Sprachwissenschaft entwickelt. Dies gilt für die Forschungsgeschichte der kaukasischen Sprachen, für die sich durch die Computerlinguistik interessante neue Perspektiven ergeben.

2. Vom Text zum Lexikon über Indizierung.



Die Schaffung einer Textdatenbank für die kaukasischen Sprachen ist ein primäres Ziel des Projekts „ARMAZI“ (Kaukasische Sprachen und Kulturen: Grundlagen ihrer elektronischen Dokumentation), das von verschiedenen wissenschaftlichen Institutionen Georgiens in Verbindung mit dem Institut für Vergleichende Sprachwissenschaft der Universität Frankfurt durchgeführt wird; mit der zu schaffenden Textdatenbank soll die Grundlage für die computergestützte linguistische Erforschung der Sprachen gelegt werden. Ein sekundäres, aber in der Tat wesentliches wissenschaftliches Resultat davon wird die automatische Verarbeitung der textuellen Daten sein, die in der Transformation der Texte in ein Lexikon besteht.

2.1. Etappen bei der Transformation textualer Materialien in ein Lexikon. Die Umwandlung des Inhalts von Texten in ein Lexikon erfordert mehrere Schritte:

I. Vom Text zum Index;

II. Vom Index zur Lemmatisierung;

III. Von der Lemmatisierung zum Lexikon.

I. Der erste Schritt bei der Transformation schließt die Bearbeitung des Textes im Hinblick auf seine Strukturierung ein. Jeder Text besteht aus rein linguistischen Daten und zusätzlichen Informationen, dem Metatext. Innerhalb des ersten Schrittes muß der Text strukturiert werden, wobei es den rein sprachliche Gehalt vom Metatext zu scheiden gilt. Der so gewonnene strukturierte Text wird dann indiziert – dabei entsteht eine andere, nicht-textuale Art von Datenbank, nämlich eine Liste aller Wortformen, die im Text vorkommen.

II. Die nächste Etappe der elektronischen Verarbeitung besteht im Transfer vom **Index zur Lemmatisierung.** Diese setzt die Analyse des Indexes mit einem die Morphologie abbildenden Modell voraus.
In der Praxis der elektronischen Datenverarbeitung natürlicher Sprachen werden zwei verschiedene Prinzipien angewendet, die als **Thesaurusprinzip** und **Generierungsprinzip** bezeichnet werden können. Die erste Methode geht von der Analyse einer textuellen Datenbank aus, die eine möglichst große Anzahl von Wörtern enthält, während die zweite Art des Herangehens die Entwicklung von Modellen der Morphologie der jeweiligen Sprache zugrundelegt, mit denen es möglich ist, alle denkbaren Wortformen zu generieren. Beide Prinzipien haben ihre Vor- und Nachteile, und beide konnten für bestimmte Sprachen, je nach deren morphosyntaktischen Eigenheiten, bereits in verschiedenen informationstechnologischen Verfahren effektiv eingesetzt werden.
Die Methode einer elektronischen Verarbeitung der linguistischen Daten, d.h. der Transformation der Daten vom Text zum Lexikon, die wir in Zusammenarbeit mit den deutschen Kollegen W. Schulze und J. Gippert entwickelt haben und von denen wir auf dem Server des **TITUS**- und **ARMAZI**-Projekts nunmehr eine Testversion im Internet bereitstellen können, beruht auf einem Konzept, das ein Gemisch beider Prinzipien darstellt. Einerseits wird eine Datenbank im Sinne eines Wortindexes erzeugt, die als ein auf die Textcorpora bezogener Thesaurus zu betrachten ist, andererseits werden die im Index enthaltenen Wortformen durch Modellierung analysiert, um den Index zu lemmatisieren.

III. Von einer lemmatisierten Datenbank aus ein Lexikon zu erzeugen, ist der letzte Transformationsschritt. Die Frage der konkreten Verarbeitungswege hängt davon ab, was für ein Lexikon erstellt werden soll: ein morphologisches, ein Synonym-, ein zwei- oder mehrsprachiges Lexikon oder alle zusammen; das elektronische Medium gestattet es, verschiedene Aspekte der Lexikographie miteinander zu verbinden.

3. Die udische Sprache.

Die Schritte auf dem Weg zur Transformation eines Textes in ein Lexikon werde ich anhand des Beispiels einer ostkaukasischen Sprache, nämlich des Udischen demonstrieren, das zusammen mit neun weiteren Sprachen (Lezgisch, Tabasaranisch, Agulisch, Rutulisch, Caxurisch, Arcinisch, Kryzisch, Buduxisch, Xinalugisch) die lezgische Untergruppe der OKS bildet. Die Wahl des Udischen ist dabei durch die folgenden Aspekte motiviert:

1. Das Udische ist eine nichtverschriftete Sprache; das bedeutet, dass es sich nicht um eine normierte Standardsprache, sondern um eine lediglich in zwei mündlichen Varietäten, dem Wartaschen- und dem Nij-Dialekt, vorliegende Sprache handelt. Abgesehen davon, dass die Dialekte Varietäten im phonetischen, morphologischen und sogar im syntaktischen Bereich aufzuweisen haben, beruht unser Interesse an dieser Sprache darauf, inwieweit es das Udische gestattet, von elektronischen Texten ausgehend ein Lexikon automatisch zu erzeugen, also eine Frage der Methodologie.

2. Das Udische ist durch besondere grammatische Merkmale in typologischer Hinsicht charakterisiert: es besitzt äußerst komplexe Deklinations- und Konjugationssysteme.

3. Das Udische kann möglicherweise auf eine längere Überlieferungsgeschichte zurückblicken: Die Sprache kann als „Relikt“ des sog. kaukasischen „Albanischen“ aus dem Mittelalter angesehen werden, von dem mit den 1975 gefundenen Sinai-Palimpsesten noch unentzifferte Zeugnisse existieren.

3.1. Die Komplexität der Deklination und Konjugation des Udischen

Bevor ich ganz konkret über die oben genannten Schritte auf dem Weg vom Text zum Lexikon spreche, möchte ich kurz auf die Deklinations- und Konjugationssysteme des Udischen eingehen.

Ein morphologisches Charakteristikum des Udischen (wie auch anderer OKS) bildet das Kasussystem, dessen Ausprägung ein interessantes Objekt der allgemeinen Kasustheorie darstellt. Man kann im Udischen gemeinhin zwischen primären oder grammatischen Kasus unterscheiden, durch die die Beziehungen zwischen den primären Aktantenpositionen (v.a.

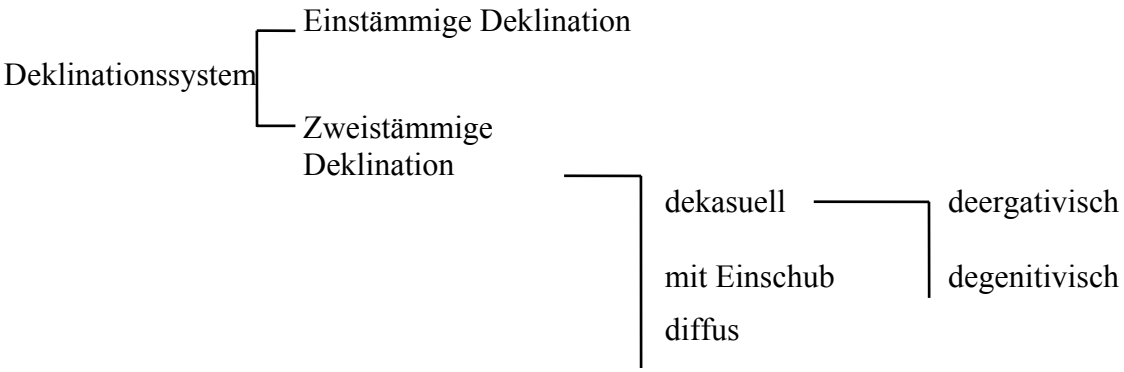
Agens und Patiens[1]) wiedergegeben werden, und Lokalkasus, durch die verschiedene adverbiale Beziehungen zum Ausdruck kommen.

Um das Kasussystem des Udischen zu beschreiben, müssen wir zuerst auf das Zwei-Stamm-Prinzip eingehen, das ein grundlegendes Charakteristikum der Deklinationssysteme in den OKS darstellt. Dieses Prinzip reflektiert die Tatsache, daß innerhalb des Flexionsparadigmas ein formaler Unterschied zwischen einem „Absolutiv-" und einem „obliquen" Stamm besteht.

Innerhalb der zweistämmigen Deklination kann man in den daghestanischen Sprachen zwischen einem „diffusen", einem „dekasuellen" und einem Deklinationstyp „mit Einschub" unterscheiden (Topuria 1985). Im Udischen lassen sich alle drei Typen der zweistämmigen Deklination beobachten: die diffuse Deklination, die dekasuelle Deklination und die Deklination mit Einschub.

Graphisch lässt sich die Deklination des Udischen wie folgt darstellen:

Schema 1



Vom **dekasuellen Deklinationstyp** sind im Udischen der deergativische[2] und der degenitivische Untertyp anzutreffen. Dabei stellt beim deergativische Untertyp die Ergativform die Bildungsgrundlage für alle markierten Grundkasus dar, beim degenitivischen Untertyp die Genitivform.

Die **Deklination mit Einschub** (so der Terminus von Žirkov) stellt den zweiten Untertyp des zweistämmigen Deklinationssystems dar. Die Einschübe, die bei anderen Autoren auch als Determinanten oder Stammerweiterungen bezeichnet werden (Žirkov, Burčuladze, Schulze), sind als ein Element zu betrachten, das in allen obliquen und Lokalkasus zwischen Stamm

[1] Oder, in der Terminologie von Schulze 2000, den „relationalen Primitiven S (subjectiv), A (agentiv), O (objectiv)".

[2] Dieser Deklinationstyp ist bei Schulze (1989) unter dem Stichwort „oblique inflection" (OI) erfasst. Schulze unterscheidet zwei Abarten dieses Typs: eine, bei der der Einschub in allen „obliquen Kasus" auftritt, und eine, wo sich der Einschub nicht in allen „obliquen Kasus" zeigt; der Ergativ bleibt unerweitert. Solche Fälle

und Kasusmorphem tritt und offensichtlich zur Markierung der „obliquen Kasus“ dient (s. Schulze, 1988).

Als „**diffuse Deklination**“ wird der Deklinationstyp bezeichnet, bei dem Ergativ und Genitiv nicht differenziert sind; dieser Deklinationstyp ist im Udischen allerdings nur selten anzutreffen und hat offensichtlich Reliktcharakter, da offenbar nur vier Nomina so dekliniert werden (*bul-* „**Kopf**“, *tur-* „**Bein**“, *pul-* „**Auge**“ und *kul-* „**Hand**“); zudem gibt es, wie Parallelformen zeigen, eine Tendenz, das Flexionssystem zu vereinheitlichen.

Ungeachtet dessen, daß bereits umfangreiche grammatische Darstellungen zum Udischen existieren, gibt es kaum einen Bereich der udischen Grammatik, über dessen Beurteilung zwischen allen Forschern Einigkeit herrschen würde. Dies betrifft gerade auch die Frage der grammatischen Kasus des Udischen. So gehören nach Schiefner zu den Grundkasus ein Nominativ, Instruktiv, Genitiv, Dativ, Affektiv. Nach Dirr umfassen die Grundkasus einen Nominativ, Instrumentalis [später Ergativ-Instr.], Genitiv, Dativ, Akkusativ. Pančvidze unterscheidet einen Nom., Erg., Gen., Dat.1 und Dat.2. Jeiranišvili reduziert die Zahl der Grundkasus weiter auf vier, nämlich Nom., Erg., Gen., Dat.; Sixarulidze bleibt bei dieser Einteilung. Bei Schulze (1982) wird demgegenüber die Zweiteilung des Dativs morphosyntaktisch und funktional begründet, womit der Autor erneut zu einer fünffachen Gliederung des Kasussystems kommt.

Diese Einteilung in fünf Grundkasus (Absolutiv, Ergativ, Genitiv, Dativ1 und Dativ 2) werde ich im folgenden für das Udische übernehmen. Alle andere Kasus (Ablativ, Komitativ, Adessiv, Allativ, Superessiv, Kausalis) bilden das System der Lokalkasus und bleiben hier unberücksichtigt.

Schema 2

Kasus	Einstämmige Deklination	Zweistämmige Deklination			
	A	B1		B2	B3
		deergativisch	degenitivisch	mit Einschub	Diffus
Abs.	ST + 0	ST + 0	ST + 0	ST + 0	ST + 0
Erg.	ST + **Erg**	ST + **Erg**	ST + G+**Erg**	ST + E+**Erg**	ST + **E/G**
Gen.	ST + **G**	ST+Erg+**G**	ST + **G**	ST + E+**G**	ST + **E/G**
Dat1.	ST + **D1**	ST+Erg+**D1**	ST + G+**D1**	ST + E+**D1**	ST + **D1**
Dat2.	ST + **D2**	ST+Erg+**D2**	ST + G+**D2**	ST + E+**D2**	ST + **D2**

ST - *Stamm**, Erg., G., D1., D2. -** Kasuszeichen**, E -** Einschub;* ***E/G -*** *Gemeinsames Zeichen für Ergativ und Genitiv.*

betrachten wir als Beispiele der deergativischen Deklination: das -**n**, das in Gen., Dat.1 und Dat.2 erscheint, erklären wir als reduzierte Form des Ergativzeichens -**en**.

3.2. Wichtige Parameter (Eigenschaften) der Deklination:

Für die automatische Analyse und Synthese der Deklination sind vier Parameter (Eigenschaften) von Bedeutung:

(A1) Die Modelle der Grundkasus im Singular: Im Vergleich zu Schema 2 sind im Schema 3 die Deklinationstypen im Sinne der Deklinationsmuster erweitert. So haben wir z.B. bei der Einstämmigen Deklination vier Deklinationsmuster zu unterscheiden: einen **i**-Genitiv, einen **un**-Genitiv, einen **ay/ey**-Genitiv und einen **in**-Genitiv; bei der Deklination mit Einschub sind der **un**-Genitiv und der **ay/ey**-Genitiv anzutreffen.

Schema 3

	A				**B1**		**B2**		**B3**
	i-Gen.	**un-Gen.**	**ay-Gen.**	**in-Gen.**	**Deerg.**	**Degen.**	**un-Gen**	**ay-Gen.**	**Diffus**
Abs.	-	-	-	-	-	-	-	-	-
Erg.	-en	-en	-en / -n	-en	**-en**	**-in**-en	**-n**-en	**-n**-en	-in
Gen.	-i	-un	-a(y) -e(y)	-in	**-n**-a(y) **-n**-e(y)	**-in**	**-n**-un	**-n**-a(y) **-n**-e(y)	-in
Dat1.	-V	-V	-V	-V	**-n**-V	**-in**-V	**-n**-V	**-n**-V	-V
Dat2.	-Vx	-Vx	-Vx	-Vx	**-n**-Vx	**-in**-Vx	**-n**-Vx	**-n**-Vx	-Vx

(A2) Das Modell der Grundkasus im Pl. Im Plural genügt ein einziges Modell für alle Nomina:

Abs. -
Erg. **-on**
Gen. **-oy / -o**
Dat1. **-o**
Dat2. **-ox**

Dabei ist jedoch die Allomorphie in der Pluralbildung selbst zu beobachten. Es gibt insgesamt 8 Allomorphe, die bei Substantiven die Pluralität markieren:

(B) Pluralität:

Schema 4

	Pl.1	Pl.2	Pl.3	Pl.4	Pl.5	Pl.6	Pl.7	Pl.8
Abs.Pl.	**-ux**	**-mux**	**-imux**	**-rux**	**-urux**	**-xox**	**-mxox**	**-rxox**
Erg.Pl.	-ɣ	-muɣ	-imɣ	-ruɣ	-urɣ	-ɣ	-mx	-rx
Gen.Pl.	-ɣ	-muɣ	-imɣ	-ruɣ	-urɣ	-ɣ	-mx	-rx
Dat1.Pl.	-ɣ	-muɣ	-imɣ	-ruɣ	-urɣ	-ɣ	-mx	-rx
Dat2.Pl.	-ɣ	-muɣ	-imɣ	-ruɣ	-urɣ	-ɣ	-mx	-rx

(C) Die Modelle der Lokalkasus im Sing. und Pl.: Bei den Lokalkasus unterscheiden wir 3 Basisstämme:

Dat1. - Dativ(1)form (Stamm + Dativ(1)zeichen) - wird für Ad. Al. und Super. verwendet;

Dat2. - Dativ(2)form (Stamm + Dativ(2)zeichen) - wird für Abl. und Komit. verwendet;

Erg. -Ergativform (Stamm+Ergativzeichen) - wird für Kaus. verwendet.

Schema 5

Kasus	Singular	Plural
Abl.	Dat2. + **-o**	Dat2. + **-o**
Kom.	Dat2. + **-olan/-ol**	Dat2. + **-olan/-ol**
Ad.	Dat1. + **-ṣta**	Dat1. + **-ṣta**
Al.	Dat1. + **-č̣**	Dat1. + **-č̣**
Sup.	Dat1. + **-l**	Dat1. + **-l**
Kaus.	Erg. + **-ķena/-ķ**	Erg. + **-ķena/-ķ**

Um die Struktur der substantivischen Wortformen zu bestimmen, haben wir die folgenden, absteigend nach der Länge von rechts nach links sortierten Morphemslots (bei Substantiven im Grundkasus) definiert:

Sing. **St + E + Kz + Fok. + Pz.**
St + E + Kz + Pz.
St + E + Kz + Fok.
St + Kz + Fok. + Pz.
St + Kz + Pz.
St + Kz + Fok.
St + E + Kz
St + Kz

Pl. **St + Pl. + Kz + Fok. + Pz.**
St + Pl. + Kz + Fok.
St + Pl. + Kz + Pz.
St + Pl. + Kz

4. Die udische Konjugation. Das Konjugationssystem im Udischen lässt sich relativ leicht modellieren, soweit es um die Tempus-Modus Kategorien geht. Nur die Personalzeichen stellen eine Schwierigkeiten dar.
In der überwiegenden Mehrzahl der OKS wird das Verbum nicht nach Personen, sondern nach Klassen konjugiert. Eine Ausnahme bilden nur das Udische, Lezgische, Agulische und teilweise das Tabasaranische, nämlich dessen nordtabasaranischer Dialekt.

Das Verb ist im Udischen monopersonal: nur das Subjektzeichen ist in der Verbalform zu finden. Will man die verschiedene Konjugationssysteme des Udischen modellieren, so sind

zwei Merkmale zu berücksichtigen: **die Serien der Personalzeichen** und die **Verbalkonstruktion.**

Zur Bezeichnung des Subjekts bedient sich das udische Verb verschiedener **Serien von Personalzeichen**, die mit den entsprechenden Personalpronomina bzw. Demonstrativpronomina identisch sind:

a) Die Personalzeichen der I. Serie entsprechen den Personalpronomina im Nominativ, diejenige der II. Serie solchen im Dativ und diejenige der III. Serie solchen im Genitiv:

Schema 6

		Serien	
	I	II	III
1 pers. (sing.:pl.)	*zu / z : jan*	*zax / za : jax / ja*	*bezi /bez : beši/beš*
2 pers. (sing.:pl.)	*nu / n : nan*	*vax / va : va'x / va'*	*vi : efi / ef*
3 pers. (sing.:pl.)	*ne : qun*	*ṭux / ṭu : qox / qo*	*ṭay / ṭa : qoy / qo*

b) Syntaktische Entsprechungen:

Schema 7

Verbalkonstruktionen	Serien der Personalzeichen	Kasus des Subjekts
Nominativ-Konstruktion	I (N/E-Pz)	Nominativ
Ergativ-Konstruktion	I (N/E-Pz)	Ergativ
Dativ-Konstruktion	II / I (D-Pz) / (N/E-Pz)	Dativ / Ergativ
Genetiv-Konstruktion	III / II (G-Pz) / (D-Pz)	Genetiv

c) Verbalkonstruktionen: Im Udischen lassen sich vier Arten von Verbalkonstruktionen beobachten: **Nominativ-Konstruktion** (N-K), **Ergativ-Konstruktion** (E-K), **Dativ-Konstruktion** (D-K) und **Genitiv-Konstruktion** (G-K). Dabei bedingen die transitiven (**agentiven**) Verben eine ergativische Konstruktion (E-K), während intransitive (**faktitivische**) Verben mit der absolutivischen bzw. nominativischen Konstruktion verbunden sind. Die in allen OKS anzutreffende Gruppe „**affektiver**" Verben (zumeist ***verba sentiendi***) bedingt eine „**inversive**" oder **Dativ-Konstruktion**. Die weniger zahlreichen „**possessivischen**" Verben (***verba habendi***) verlangen eine „**possessivische**" Konstruktion.

Schema 8

Verbalkonstruktion:	Verbtyp:
Nominativ-Konstruktion (N-K)	intransitiv („**faktitiv**")
Ergativ-Konstruktion (E-K)	transitiv („**agentiv**")
Dativ-Konstruktion (D-K) („inversiv")	affektiv (***verba sentiendi***)
Genetiv-Konstruktion (G-K)	possessiv (***verba habendi***)

d) Die Slots der Personalzeichen.

Das Udische unterscheidet zwischen „einfachen" und zusammengesetzten Verbalstämmen, die auch als einkomponentige und zweikomponentige Verbalstämme bezeichnet werden

können. Die größte Zahl der udischen Verben ist zusammengesetzt. Sie bestehen aus zwei Bestandteilen (Komponenten), einer meist nominalen Basis und einem verbalen Stamm.
Die Stellung der Personalzeichen am Verb variiert stark. Sie sind regelmäßig infigiert, können jedoch in Abhängigkeit von dem Verbalstamm oder der Verbalkonstruktion, auch als Suffixe oder Präfixe erscheinen.

Je nach der Position der Personalzeichen im Sinne der Slots sind insgesamt sieben Varianten möglich:

Schema 9

	Pz-Kv	*zax- ḍu < zax-bu*	*besun*	haben	Präs.
1 a	Kn-**Pz**-Hv	*umud-bez-bu*	*umud-besun*	hoffen	Präs
b					
2 a	Kv-M1-T-**Pz**	*bak-s-a-ne*	*baksun*	sein	Präs.
b	Kn- Hv-M1-T-**Pz**	***aš**-be-s-a-ne*	***aš**-besun*	arbeiten	Präs.
3a	Kv1-**Pz**-Kv2-M1-T	*ba-ne-k-s-a*	*baksun*	sein	Präs.
b	Kn-**Pz**-Hv-M1-T	***aš**-ne-b-s-a*	***aš**-besun*	arbeiten	Präs.
4 a	Kv-T-**Pz**	*bak-i-ne*	*baksun*	sein	Aorist
b	Kn-Hv-T-**Pz**	***aš**-b-i-ne*	***aš**-besun*	arbeiten	Aorist
5 a	Kv1-**Pz**- Kv2-T	*ba-ne-k-i*	*baksun*	sein	Aorist
b	Kn-**Pz**-Hv-T	***aš**-ne-b-i*	***aš**-besun*	arbeiten	Aorist
6 a	Kv1-P-**Pz**-Kv2-T	*kar-qa-z-x-i*	*karxesun*	leben	Part.Konju. I
b	Kn-P-**Pz**-Hv-T	***umud**-qa-z-b-i*	***umud**-besun*	hoffen	Part.Konju. I
c	P-**Pz**-Kv1-Kv2-T	*gi-z-kar-x-ey*	*karxesun*	leben	Part.Konju. II
d	Kn-P-**Pz**-Hv-T	*ǯahil-gi-z-bak-ey*	***ǯahil**-baksun*	Jung sein	Part.Konju. II
7 a	Pc-**Pz**	*karxal-zu*	*karxesun*	leben	Fut.II
b	Pc-**Pz**-T	*karxal-zu-y*	*karxesun*	leben	pir.natv.

Darüber hinaus können die Personalzeichen aber auch außerhalb der Verbalformen auftreten, nämlich:

***I**. bei Nomina:*
a) als Kopula:

(1) ṭevaxṭa šägirden, maṭuxte buṭuḳsay Isusa, pine Ṗeṭrax: mono **bixaǯuġ-ne (N)**. (Io. 21. 7)
Da spricht der Jünger, den Jesus liebhatte, zu Petrus: es ist der **Herr**!

b) bei Fokussierung:
(2) vaᶜ ägänä zu **č̣ev-zu-ḳsa** (**V**) ǯinurɣox veevzevulun zoren, ef ġarmuġon gena ši **zoren-qun (N) č̣evḳesa (V)** ? (Mt. 12. 27)
Und wenn ich die bösen Geister durch Beelzebuls Kraft austreibe, durch welche **Kraft** treiben eure Söhne sie aus?

(3) manorte ṗinaxo teḳun, laśag buḳsunaxo tene, išu buḳsunaxo tene, amma **bixoġoxo-ġun** (**N**) **bake** (**V**). (Io. 1. 13)
Die nicht aus dem Blut, noch aus dem Willen des Fleisches, noch aus dem Willen eines Mannes, sondern von **Gott** geboren sind.

***II**. bei den Pronomina:*

(4) ägänä elen aćesbain ič zorrux, ṭevaxṭa **eṭin-nu (Prn.) elenbo (V)** šoṭux? (Mt. 5.13.)
Wenn das Salz seine Kraft verliert, **womit** soll man **es** salzig machen?

***III**. bei Partikeln:*

(5) va^{c} pine šoṭu isusen: „śulurġoy buḳoy kur, va^{c} gögnä ḳušurġoy mecurux; amma adamari ġarey **te-ṭa (Prt.) bu (V)** ma ḳoḍbane ič bex". (Mt. 8. 20)
Und Jesus sagt zu ihm: „Die Füchse haben Gruben und die Vögel des Himmels Nester, aber der Menschensohn hat **nichts,** (wohin) er seinen Kopf legen könnte".

5. Textcorpora des Udischen. Als elektronischen Text haben wir die Anfang des 20. Jhs. von den Brüder Bezhanov übersetzten vier Evangelien digitalisiert. Der Text wurde von uns vor zwei Jahren gescannt und korrigiert; dabei wurden offensichtliche Druckfehler beseitigt. Von der Korrektur anderer Arten von Fehlern halten wir uns aus einfachem Grund jedoch noch zurück; denn wir dürfen auf keinen Fall scheinbare Fehler korrigieren, die in Wirklichkeit eine sprachwirkliche phonetische oder graphematische Variation reflektieren. Es scheint uns sinnvoller sein, zunächst zwei Varianten des Evangelientexts zu erzeugen: einen, der dem Originaltext der Brüder Bezhanov mit seinen unsystematischen Fehlen entspricht, und einen von uns vereinheitlichten Text.
Damit erst kommen wir zu der eigentlichen Frage der elektronischen Datenverarbeitung, nämlich zur automatischen Erzeugung eines Lexikons auf der Basis eines elektronischen Texts bei nichtnormierten Sprachen bzw.Dialekten.
Die Transformation vom Text zum Lexikon soll in den oben genannten drei Schritten so verwirklicht werden, dass die im Text gegebene linguistische Information nicht verloren geht. Dies ist durch vier für Datenbanken charakteristische Komponenten erreichbar:
1. Integrität;
2. Logische und morphologische Widerspruchslosigkeit;
3. Grammatische Vollständigkeit;
4. Minimale Redundanz.

5.1.Vom Text zu Index. Probleme der Indizierung bei der automatischen Erzeugung resultieren meist aus **Inkonsistenzen des Ausgangstexts.** Um einen Index korrekt zu erzeugen und dann durch die Modellierung erst ein lemmatisiertes Verzeichnis und dann ein Lexikon zu ermöglichen, ist es erforderlich, den Text zu **vereinheitlichen**. Dieses Problem ist bei unserem Text in folgenden Fällen aufgetaucht:

a) Verschiedenheit bei der Schreibung von Lexemen, wie z.B.:

färišṭa / farišṭä / färišṭä / färišta - Engel
čaṗluɣ / čapluɣ / čablug - Weingarten
böhär / boᶜhär / buhär / bühar / bühär - Frucht
bütün / bütun / büṭun - alle, alles
ait / äit / aiṭ / äiṭ - Wort

b) Uneinheitlichkeit bei der Übernahme / Wiedergabe fremder Eigennamen bzw. Toponymik, wie z.B.:

Ioan vs. **Ioann** (abs.) - Iohannes
Daneil (abs.) vs. **Danil-**un (gen.)- Prophet Daniel
Arimafei-axo vs. **Arimofei-**axo (abl.) - Arimathäa
Zavedeev-i - Zebedäus
Aminadav-*en* (Erg.) vs. **Aminadavov-***i* (Gen.)
Aram-*en* (Erg.) vs. **Aramov-***i* (Gen.)
wo der in der russischen Vorlage im Genitiv stehende Eigennamen einen udischen Genitiv mit dem Kasuszeichen **–i** bedingt hat. Weitere Beispiele: Adamov-i, Alfeev-i, Amosov-i, Aviev-i, Avraamov-i, Davidov-i.

David und alle mögliche Kasusformen davon: **David**-i, **David**-en, **David**-axo, **David**-al-cirik vs. **Davidov**-i machen es unmöglich, den Stamm und entsprechend die Lemmaform automatisch zu identifizieren.

c) *Uneinheitlichkeit bei der Schreibung identischer Morpheme:*
(bei Verben) Präs. ***-ay / -ai****: buṭay / buṭai* „er ist“
(bei Substantiven) Gen. ***-ay / -ai, -ey / -ei:*** *ġarei / ġarey* „des Sohnes“

d) Uneinheitlichkeit bei der Behandlung von Komposita:
ṗaᶜ ćola (wtl. „Zwei Gesicht(e)“**) - Heuchler**

(6) vaᶜ evaxte afrenexa, ma baka, etärte **ṗaᶜ ćola or** (Pl.) (Mt. 6,5)
Und wenn ihr betet, sollt (ihr) nicht sein, wie die **Heuchler**.

(7) bixoġon pine šoṭu ǯuġab: **ṗaᶜ ćola**! (Sing.) (Lk. 13.15)
Gott sagt ihm die Antwort: (Ihr) **Heuchler**!

e) Uneinheitlichkeit bei der Schreibung zusammengesetzter Verben:

(8) meṭin **murdal-le besa** adamarax (Mt.15.20) zu **murdalbesun** - besudeln, entweihen;
Dies **entweiht** den Menschen.

5.2. Probleme bei der Modellierung: Nachdem wir den Text möglichst vereinheitlicht haben, versuchen wir, einen **morphologischen Index** zu erzeugen, wobei die oben genannten

Modellen zur Anwendung kommen. Unter dem Begriff **morphologischer Index** verstehen wir eine Liste von im Text anzutreffenden Wortformen mit automatisch erzeugter morphologischer Analyse.
Hierfür gibt es zwei Möglichkeiten:
a) Es wird einerseits eine Liste von Morphemen und andererseits eine Liste von Lemmata des Udischen erstellt.
b) Im Index verzeichnete Wortformen werden automatisch analysiert, um die Lemmata und den Deklinationstyp zu bestimmen.
Wir haben die erste Möglichkeit der Verarbeitung gewählt, weil wir, um das Lemma und den Deklinationstyp desselben bestimmen zu können, das Lexem in mindestens 5 Flexionsformen benötigen: Im Absolutiv (Sing)., einer Pluralform, einer Erg.- bzw. Kausativform, einem Genitiv und entweder einen Dativ oder einer beliebigen Lokalkasusform. Nur in diesem Fall ist es möglich, das Lemma automatisch zu identifizieren und den Deklinationstyp zu bestimmen.
Andererseits besteht der Grund, warum wir auf zweite Möglichkeit der Analyse zur Zeit verzichten, darin, dass wir im Rahmen der Morphologie noch keine Regeln haben aufstellen können, um die folgenden homophonen Kasusformen zu unterscheiden:

a) **nana -** Mutter
nana als Abs.
nana (< **nanay**)als Gen.
nana als Dat1.
b) **kul** - Hand **(diffuse Deklination)**
kin als Erg.
kin als Gen.

Für das Erkennen der Gen.-Form können zwei Regeln hilfreich sein:

a) Es ist eine Genitiv-Form zu erwarten, falls der Wortform eine der Postpositionen ***toġol, baxṭin, laxo*** oder ***boš*** folgt;

oder

b) falls das nächste Wort wieder ein Substantiv ist, so dass das davor stehende **nana** oder **kin** ein Attributiv im Genitiv sein kann.
Um eine Abs.-Form von einer Dat1.-Form zu unterscheiden, brauchen wir schon ein syntaktisches Parsing, da wir nur im Verbindung mit dem finiten Verb feststellen können, ob dieses, als intransitives Verb, ein Subjekt im Abs. oder, als *verbum sentiendi*, ein Subjekt im Dat1 oder ein freie Ergänzung (Lokativ) erfordert. Alle diese Regeln verlangen von uns eine automatische Kontextanalyse, die noch nicht entwickelt worden ist.

Bei der Lemmatisierung treten die folgenden Probleme im phonetischen und morphologischen Bereich auf:

4.1. Im phonetischen Bereich:
***a) Allophonie**:*
Gen.: **-ay / -ey (im Sing.);**

4.2. In morphologischen Bereich:
a) Allomorphie:

Schema 10

Kasus	Singular	Plural
Abs.	------	-ux, -mux, -rux, -urux, -xox, -mxox, -rxox
Erg.	-en, -n, -in	-ɣon, -muɣon, -ruɣon, -urɣon, -ɣon, -mxon, -rxon
Gen.	-i/-y, -ay/-a, -ey/-e, -un , -in	-ɣoy/ɣo, -muɣoy/-muɣo, -ruɣoy /-ruɣo, -urɣoy/urɣo, -ɣoy/-ɣo, -mxoy/-mxo, -rxoy/-rxo
Dat1.	-a, -e, -i, -u	-ɣo, -muɣo, -ruɣo, -urɣo, -ɣo, -mxo, -rxo
Dat2.	-ax, -ex, -ix, -ux	-ɣox, -muɣox, -ruɣox, -urɣoy, -ɣox, -mxox, -rxox
Abl.	(-ax, -ex, -ix, -ux) + -o	(-ɣox, -muɣox, -ruɣox, -urɣox, -ɣox, -mxox, -rxox) + o
Kom.	(-ax, -ex, -ix, -ux) + -olan/-ol	(-ɣox, -muɣox, -ruɣox, -urɣox, -ɣox, -mxox, -rxox) + -olan/-ol
Ad.	(-a, -e, -i, -u) + -ṣta	(-ɣo, -muɣo, -ruɣo, -urɣo, -ɣo, -mxo, -rxo) + -ṣta
Al.	(-a, -e, -i, -u) + -č̣	(-ɣo, -muɣo, -ruɣo, -urɣo, -ɣo, -mxo, -rxo) + -č̣
Sup.	(-a, -e, -i, -u) + -l	(-ɣo, -muɣo, -ruɣo, -urɣo, -ɣo, -mxo, -rxo) + -l
Kaus.	(-en, -n, -in) + -ķena/-ķ	(-ɣon, -muɣon, -ruɣon, -urɣon, -ɣon, -mxon, -rxon) + -ķena/-ķ

b) Schwankungen in der Allomorphie:

Ioani vs. Ioanun (i-Gen. vs. un-Gen)

(9) **ioani** iči buṭay partal buše popnuxo vaᶜ toxġa ṭollay ič bačanel, xorag gena šeṭay buney ṭeķal vaᶜ čolla uč̣. (Mt. 3,4)
Er aber, **Johannes**, hatte ein Gewand aus Kamelhaaren an und einen ledernen Gürtel um seine Lenden; seine Speise aber waren Heuschrecken und wilder Honig.

(10) ṭevaxṭa eġunsa šeṭa ṭoᶜġoᶜl **ioanun šägirdux** vaᶜ exġuni: eṭabaxṭin ian vaᶜ fariseyġon gölö ġiruxian efsa, amma vi šägirdġon gena ġirux teġun efsa? (Mt. 9,14)
Da kamen **die Jünger des Johannes** zu ihm und sprachen: Warum fasten wir und die Pharisäer so viel, und deine Jünger fasten nicht?

c) Schwankung zwischen den Deklinationsmustern (**meistens bei Lehnwörtern**):

aġa - *Herr, Besitzer*

Einstämmige Deklination:

(11) šeṭa **aġan** (*erg*) pine šoṭu: šelle irähmlu vaᶜ doġri nökär! (Mt. 25,21)
Da spricht sein **Herr** zu ihm: Recht so, du tüchtiger und treuer Knecht ...

(12) šägird abuz tene učiṭelaxo, vaᶜ nökär abuz tene ič **aġaxo**: (*abl*) (Mt. 10,24)
Der Jünger steht nicht über dem Meister und der Knecht nicht über seinem **Herrn**.

Zweistämmige (degenitivische) Deklination:

(13) hametär tavaxqabanan exna **aġinax** (*dat2*), te iaqabane balṭġox ič exnu. (Mt. 9,38)
Darum bittet **den Herrn** der Ernte, daß er Arbeiter in seine Ernte sende.

(14) evaxte baneki bias, čaṗluġun **aġinen** (*erg*) pine ič iräšṗärä: ḳalpa fählig̈o vaᶜ tada šoṭġo ahq, axrunṭxo burqi beśunṭulciriḳ*. (Mt. 20,8)
Als es nun Abend wurde, sprach **der Herr** des Weinbergs zu seinem Verwalter: Ruf die Arbeiter und gib ihnen den Lohn und fang an bei den letzten bis zu den ersten.

axçima - *Ostern, Osterfest, Osterlamm:*

Einstämmige Deklination:

(15) süfṭäumʒi ačamun gena iśaqun baki isusi ṭoġol* šägirdux vaᶜ piqun šoṭu: man buyruġbesa ia hazirbaian venḳ **axçima** (*dat1.*)? (Mt. 26,17)
Aber am ersten Tage der Ungesäuerten Brote traten die Jünger zu Jesus und fragten: Wo willst du, daß wir dir das **Passalamm** zum Essen bereiten?

(16) šeṭin pine: takenan šähärä fulanṭa ṭoᶜġoᶜl vaᶜ upanan šoṭux: učiṭelen exne: bez vädä iśane, vi ḳua balzu **axçima** (*dat1*) bez šägirdġoxol. (Mt. 26,18)
Er sprach: Geht hin in die Stadt zu einem und sprecht zu ihm: Der Meister läßt dir sagen: Meine Zeit ist nahe; ich will bei dir das **Passa** feiern mit meinen Jüngern.

Zweistämmige Deklination (degenitivisch)

(17) šägirdġon biqun ṭetär, etärte buyruġnebe šeṭġo isusen, vaᶜ häzirqunbi **axçiminax**. (*dat1*) (Mt. 26,19)
Und die Jünger taten, wie ihnen Jesus befohlen hatte, und bereiteten das **Passalamm**.

(18) **axçimin äziz** ġimxox zafbalun buṭay ixṭiar barḳaney xalxna baxṭin sa ṭussaġa, maṭuxte buqoqsay; (Mt. 27,15)
Zum (Oster-)Fest aber hatte der Statthalter die Gewohnheit, dem Volk einen Gefangenen loszugeben, welchen sie wollten.

färišṭä - *Engel*

Einstämmige Deklination:

(19) **färišṭän** (*erg.*) gena, äytax čupqoč* taradi, pine: ma qavaᶜqéᶜbi, šeṭabaxṭinte vaᶜn furunanexa isusax, čärčäräz biṭux (Mt. 28,5)
Aber **der Engel** sprach zu den Frauen: Fürchtet euch nicht! Ich weiß, daß ihr Jesus, den Gekreuzigten, sucht.

(20) evaxte monor taq̇unci, - me **färišṭa** (*dat1.*) bixoġoy aķneci neṗeboš iosifa vaᶜ pine: ayza, aq̇a aylax vaᶜ ič nanax, ṭiṭa egiṗṭa, ṭia baka zu vax uķa ma (Mt. 2,13)
Als sie aber hinweggezogen waren, siehe, da erschien **der Engel** des Herrn dem Josef im Traum und sprach: Steh auf, nimm das Kindlein und seine Mutter mit dir und flieh nach Ägypten und bleib dort.

Zweistämmige Deklination (degenitivisch)

(21) neṗaxo ayzeri, iosifen bine etärte bixoġo **färišṭinen** (*erg.*) buyruġnebi šeṭu, vaᶜ aneq̇i čubġox iči. (Mt. 1,24)
Als nun Josef vom Schlaf erwachte, tat er, wie ihm **der Engel** des Herrn befohlen hatte, und nahm seine Frau zu sich.

(22) šeṭabaxṭinte šono šonone, maṭalaxote camne: zu iaq̇azbesa bez **farišṭinax** (*dat2*) vi beś, maṭinte häzirballe vi iaq̇ax vi beś (Mt. 11,10)
Dieser ist's, von dem geschrieben steht: "Siehe, ich sende meinen **Boten** vor dir her, der deinen Weg vor dir bereiten soll."

iaq̇ - *Weg*

Zweistämmige Deklination (deergativisch):

(23) vaᶜ neṗeboš aq̇unq̇i buyruġ, irodun ṭoᶜġoᶜl nu qaybakaq̇un, q̇eyri **iaq̇en** (*erg.*) taq̇unci ičġo ölkina. (Mt.2,12)
Und Gott befahl ihnen im Traum, nicht wieder zu Herodes zurückzukehren; und sie zogen auf einem andern **Weg** wieder in ihr Land.

(24) vaᶜ evaxte afrenexa ma baka, etärte ṗaᶜ ćola or, maṭġote buq̇oq̇sa meč̣itġᶜo, vaᶜ **iaq̇na** (*gen.*) künʒimuġo čurpi afrepesax (Mt. 6,5)
Und wenn ihr betet, sollt ihr nicht sein wie die Heuchler, die gern in den Synagogen und an den **Straßen**ecken stehen und beten ...

Einstämmige Deklination:

(25) šeṭabaxṭinte šono šonone, maṭalaxote camne: zu iaq̇azbesa bez färišṭinax vi beś, maṭinte häzirballe vi **iaq̇ax** (*dat2*) vi beś (Mt. 11,10)
Dieser ist's, von dem geschrieben steht: "Siehe, ich sende meinen Boten vor dir her, der deinen **Weg** vor dir bereiten soll."

xod - *Baum*

Einstämmige Deklination:

(26) har šel bühär nuṭ ečal **xodax** (*dat2.*), boṭi boq̇unsesa arġo boš. (Mt. 7,19)
Jeder **Baum**, der nicht gute Früchte bringt, wird abgehauen und ins Feuer geworfen.

Zweistämmige Deklination (degenitivisch):

(27) vaᶜ tavaral **xodin** (*gen.*) tumexne biti: har xod šel bühär nuṭ ečalṭux boṭi boq̇unsesa arġo boš. (Mt. 3,10)
Es ist schon die Axt **den Bäumen** an die Wurzel gelegt. Darum: jeder Baum, der nicht gute Frucht bringt, wird abgehauen und ins Feuer geworfen.

Zweistämmige Deklination (deergativisch):

(28) teṭu bako šel **xoddu** (*dat1.*) ečes pis bühär, vaᶜ pis **xoddual** (*dat1+ Fokus*) teṭu bako ečes šel bühär. (Mt. 7,18)

Ein guter **Baum** kann nicht schlechte Früchte bringen, und ein fauler **Baum** kann nicht gute Früchte bringen.

5.4. Die genannten Probleme können teilweise durch Dialektunterschiede, nämlich als Nijismen im Evangelium erklärt werden:

a) Phonetische Dialektunterschiede sind grundsätzlich bei Lautwandelerscheinungen zu beobachten; z.B.:

1) Palatalisierung der Vokale.

Wart.-Dialekt	**Nij-Dialekt**	
ḳaša	ḳäšä	Finger
toš	töš	Draußen

2) *Spirantisierung der Konsonanten*:

Wart.-Dialekt	**Nij-Dialekt**	
ḳoǯ	ḳož	Haus
bač	baš	Hundert

3) Desaffrizierung: z.B. **ç̣** > **č** : **pasç̣ȧ** / **pasčaȧ** - König

b) Morphologische Dialektunterschiede:
Dies betrifft Unterschiede im morphologischen Inventar, z.B. bei der Deklination der Substantive:

Wart.-Dialekt	**Nij-Dialekt**
Dat2. - ax, -ux, -ix	- a(x)
Abl. - aç̣	- ač

2) Unterschiede bei den Deklinationstypen: für den **Wart**.-Dialekt sind der Deklinationstyp mit Einschub und der deergativische Untertyp typisch, die der **Nij**-Dialekt nicht kennt:

z.B. **zor - „***Kraft“*

Wart.-Dialekt	**Nij-Dialekt**
Abs zor	zor
Erg. zor -en	zor - en
Gen. zor - **n** -ay	zor - ay
Dat1. zor - **n** - u	zor - a
Dat2. zor - **n** - ux	zor - a

c) Syntaktische Dialektunterschiede:

Verbalkonstruktionen und syntaktische Entsprechungen: die Nominativ-Konstruktion (**N-K**), Ergativ-Konstruktion (**E-K**), Dativ-Konstruktion (**D-K**) und Genetiv-Konstruktion (**G-K**).

Schema 11

Verbalkonstruktionen	Serien der Personalzeichen		Kasus des Subjekts
	Wart.-Dialekt	**Nij-Dialekt**	
Nominativ-Konstruktion	I (N/E-Pz)	I (N/E-Pz)	Nominativ
Ergativ-Konstruktion	I (N/E-Pz)	I (N/E-Pz)	Ergativ
Dativ-Konstruktion	II (D-Pz)	II (D-Pz)	Dativ
Genetiv-Konstruktion	III (G-Pz)		Genetiv

6. Vom Index zur Lemmatisierung: Probleme und Lösungen. Von Index ausgehend muss die linguistische Information so in ein Lemma-Verzeichnis transferiert werden, dass keine morphologische Eigenschaft der Lexeme verloren geht. Dazu dienen spezielle Programme mit Feldstrukturen, die dafür sorgen, dass die Information in der Datenbank getrennt gespeichert bleibt, wobei die Verknüpfung zwischen der im Text anzutreffenden Wortform und dem Lemma fest bleibt.

6.1. Das Prinzip der Feldstrukturen.

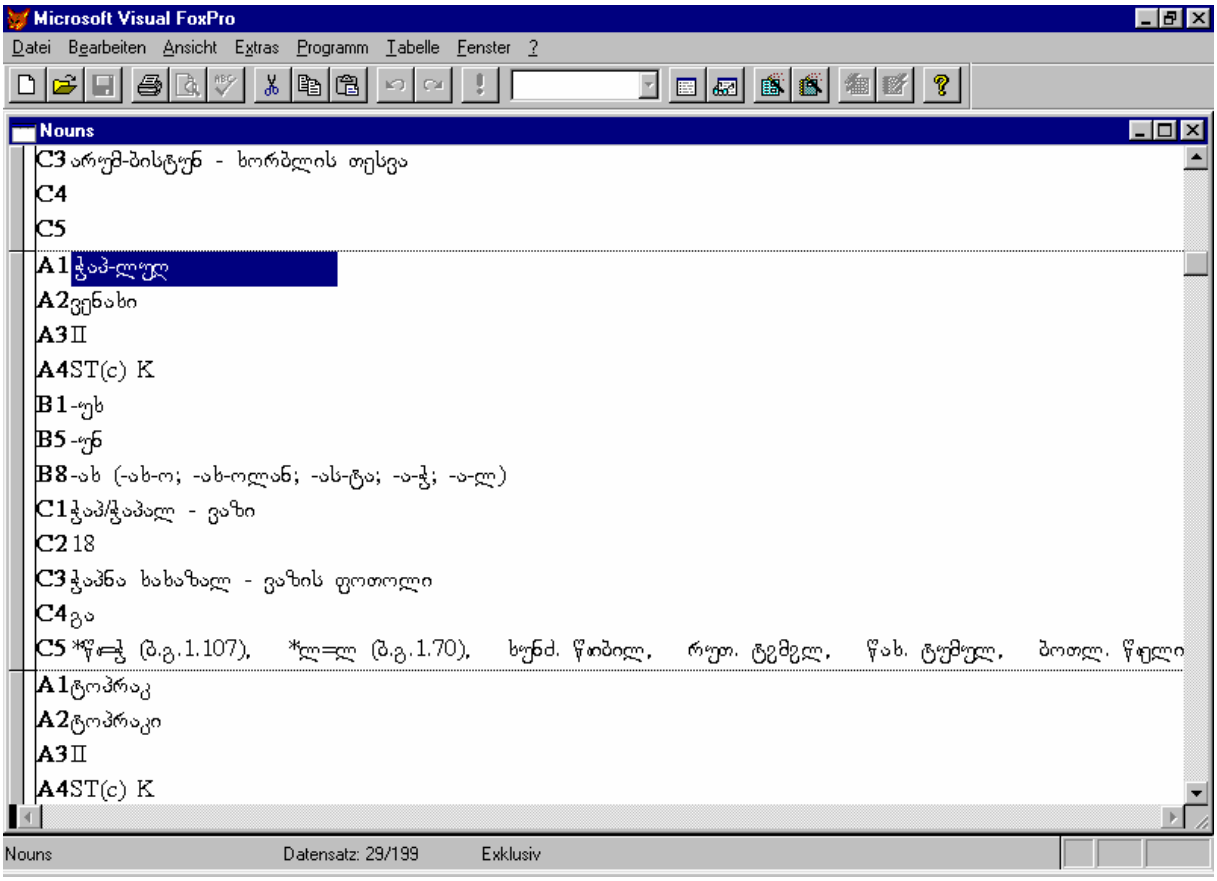
I. Erste Variante: Datenbank für das Udische auf der Grundlage von FoxPro, 1997 (nicht auf Textcorpora bezogene Datenbank)

Feldstruktur: A-Grundfeld, B-morphologisches Feld, C-lexikalisches Feld.

Schema 12

Ground form	Pl.	Class	declension type	Gen.	Dat.2	translation
baba	-ux	I	One-stem dec.	-i	-ax	father
viči	-mux	I	Two-stem dec.	-ey	-ex	brother
čur	-xox	II	with insert	-r-ay	-r-ex	cow
aṣ̌	-rux	II	geergative	-n-ex	-n-ax	work

I. Klasse - Eigennamen bzw. Verwandtschaftstermini, **II. Klasse** - Appellativa;



II. Neue Variante: Datenbank der udischen Evangelien auf der Grundlage von Access, 2001, in Vorbereitung (auf ein Textcorpus bezogene Datenbank):

scheme 13

Basic form	Phon. Variant	Altern. Stamm	2. stem	Pl.	Declension type	gen.	dat2.	translation
färišṭä	färišṭa farišṭä färišta		färišṭ	-ux	A B1(degen)	-un -in	-ax -ax	angel
ç̣ȧpluγ	ç̣apluγ ç̣ablug			-ux	A	-un	-ax	vineyard
böhär	boᶜhär buhär bühar bühär			-ux	A	-un	-ax	fruit
Ait	äit aiṭ äiṭ			-urux	A	-un	-ax	word
xod				-urux	A B1(deerg) B1(degen)	-un -ux -in	-ax -ux -ax	tree
Ioan	Ioann				A	-in -un	-ax	Iohannes
David		Davidov			A	-i	-ax	David
Aram		Aramov			A	-i	-ax	Aram
pasç̣aġ	pasčaġ			-ux	A	-un	-ax	king
kul			k	-mux	B3	-in	-ex	hand
vädä			väd	-imux	B1(degen)	-in	-ax	time