

# DATENBANK FÜR URALISCHE SPRACHEN

Pirkko Suihkonen

## Einleitung

In meinem Artikel werden die Völker und Volksgruppen, die uralische Sprachen sprechen, sowie die durch den Server für elektronische Korpora der Universität Helsinki Mitte 1998 erhältlichen elektronischen Korpora der sowohl eng als entfernt verwandten Sprachen des Finnischen vorgestellt. Weiter werden die Anwendungsmöglichkeiten dieser elektronischen Korpora in Forschung und Unterricht durch Beispiele präsentiert.

## 1. Datenbank für uralische Sprachen<sup>0</sup>

Die **Datenbank für uralische Sprachen** ist ein elektronisches Archiv für das in dieser Form gespeicherte Material in uralischen Sprachen. Das Archiv selbst befindet sich im Server für das multilinguale Korpus des Instituts für Linguistik der Universität Helsinki (University of Helsinki Language Corpus Server, UHLCS) gespeichert und ist mit dem internationalen Computernetz verbunden. Die elektronisch verwendbaren Materialsammlungen der kleinen uralischen Sprachen werden wiederum **Datenbank für bedrohte uralische Sprachen** genannt. An mehreren Universitäten sind gegenwärtig Projekte für die Sammlung und Dokumentation des Materials in den vom Aussterben bedrohten Sprachen im Gange. In diesem Fall kann die **Datenbank für bedrohte uralische Sprachen** als ein Teil der **Datenbank für uralische Sprachen** betrachtet werden. Auch der Server für elektronische Korpora der Universität Helsinki enthält gesammeltes Material in zahlreichen weiteren bedrohten Sprachen. Die **Datenbank für uralische Sprachen** und die **Datenbank für bedrohte uralische Sprachen** sind ein Teil der im Server für elektronische Korpora der Universität Helsinki gespeicherten **multilingualen Datenbank**. In ihrer gegenwärtigen Form besteht die Datenbank der uralischen Sprachen hauptsächlich aus mehrsprachigem Material in unterschiedlichen Texttypen. Weiter sollten die Datenbanken verschiedener Sprachen auch Grammatiken und Wörterbücher der betreffenden Sprachen enthalten. Eine Datenbank dieser Art kann aber auch demographische und kulturbezogene Information über eine ethnisch bzw. sprachlich definierte Gruppe beinhalten. Eine gelungene mehrsprachige Datenbank ist stets ein international anwendbares Archiv für diese Sprachen und die Kultur ihrer Sprecher.

---

0. Pirkko Suihkonen, Institut für Linguistik, Postfach 4 (Vuorikatu 6 A 10), 00014 Universität Helsinki, Finnland (Homepage: <http://www.ling.helsinki.fi/~suihkone/>). Der Artikel besteht aus Vorträgen an den Tagen der Wissenschaft, organisiert von der Universität Helsinki am 9. und 10. Januar 1997, und dem Kongreß der IFUSCO 1997 am 24. April 1997. Die deutsche Übersetzung wurde von Frau Magister Irmeli Helin und Herrn Doktor Helmut Diekman besorgt. Ihnen beiden sage ich meinen herzlichsten Dank. Meinen Dank ich möchte auch an Frau Prof. Dr. Inse Cornelssen für ihre Kommentare über diesen Artikel aussprechen.

Die Benennung **Datenbank für uralische Sprachen** wird auch als Oberbegriff verwendet, und zwar für diverse Projekte zur Sammlung bzw. Verarbeitung des elektronischen Materials in uralischen Sprachen, welches dann auf den Server für elektronische Korpora der Universität Helsinki geliefert und für Forschungs- und Unterrichtszwecke zur Verfügung gehalten wird. Ein solches Unternehmen war das von Suomen Akatemia (der Akademie von Finnland) und dem Finnisch-ugrischen Institut der Universität Helsinki gemeinsam finanzierte Projekt unter dem Namen **Struktur und Typologie der uralischen Sprachen**. Während des Projekts wurde Material in den Sprachen Komi (Syrjänisch), Udmurtisch (Wotjakisch), Hantisch (Wogulisch) und Nenzisch (Jurakisch) gesammelt. Ein Teil des Materials wurde im Lauftext Wort für Wort morphologisch und der Wortart gemäß kodiert.

Seit 1996 besteht auch ein weiteres, entsprechendes Projekt, die **Datenbank der bedrohten finnougri-schen Sprachen**, an dem auch schwedische und norwegische Forscher teilnehmen. Dieses Projekt wird von *Suomen Akatemia* (der Akademie von Finnland), den Finnisch-ugrischen und Linguistischen Instituten der Universität Helsinki sowie vom Ausschuß der humanistischen Forschung in den nordischen Ländern, dem *Nordisk Samarbeidsnemd for humanistik forskning* (Joint Committee of the Nordic Research Councils for the Humanities, NOS-H) und dem Nordischen Rat finanziert. Mit Hilfe der Finanzierung vom NOS-H arbeiten neben Schweden und Norwegern auch finnische Forscher. Die Finanzierung vom NOS-H läuft bis Ende 1997 weiter, die von Suomen Akatemia ein Jahr länger. Während des Projekts wird elektronisch anwendbares Material z.B. in den Sprachen Nordsaame (Nordlappisch), Südsaame (Südlappisch), Umeåsaame (Umeålappisch), Lydisch, Livisch, Erzänisch (Mordwinisch), Komi (Syrjänisch), Udmurtisch (Wotjakisch), Selkupisch (Sölkupisch), Nenzisch (Jurakisch) und Enzisch (Jenissei-Samojedisch) verarbeitet. Das kodierte Material wird auf den Server für elektronische Korpora der Universität Helsinki geliefert und den Forschern und Studenten zur Verfügung gestellt. In diesem Zusammenhang wird neben den uralischen Sprachen auch Material in Sprachen aus zahlreichen weiteren Sprachfamilien in den UHLCS geliefert. Ein großer Teil dieses Materials stammt aus den Instituten für Bibelübersetzung (Helsinki und Stockholm). Ein weiteres Ziel der Projekte ist eine grundlegende Erforschung der uralischen Sprachen und der Sprachtypologie.

## 2. Völker und Volksgruppen, die uralische Sprachen sprechen

Die meisten uralischen Sprachen sind Minoritätssprachen, die zu den sog. **bedrohten Sprachen** gehören. Vom Aussterben bedroht sind besonders Sprachen, die nur von kleinen Gruppen gesprochen werden und deren Gebrauchsfunktionen nicht dem gesamten Anwendungsbereich einer Sprache entsprechen. Die Gesamtzahl der in der Welt gesprochenen Sprachen wird unterschiedlich geschätzt und liegt zwischen 4000 und 7000. Sogar noch höhere Zahlen sind erwähnt worden (vgl. Karlsson 1994: 241). Die Anzahl der Sprachen, die auf der nationalen Ebene von mehr als einer Million Menschen als Muttersprache gesprochen werden, bleibt unter 200 (*The World Almanac and Book of Facts 1997*: 642-643). Offizielle Sprachen, d.h. gesetzlich im jeweiligen Land anerkannte Sprachen, gibt es beträchtlich weniger (Joki 1984), und nur einige zehn haben eine zentrale Stellung in der internationalen Zusammenarbeit. Zugespitzt könnte behauptet werden, daß die Mehrzahl der Sprachen der Welt vom Aussterben bedroht ist.

Die folgende Tabelle enthält eine Statistik über die Völker und Volksgruppen, die uralische Sprachen sprechen. Die Zahlen in der ersten Spalte zeigen die Anzahl der Personen, die sich selbst als Mitglieder der jeweiligen ethnischen Gruppe bezeichnen, die in der zweiten stehen für die Sprecher der Sprache der entsprechenden ethnischen Gruppe. Leider ist besonders bei den kleinsten Gruppen keine zuverlässige Statistik verfügbar.

**Tabelle I. Anzahl der Sprecher uralischer Sprachen und der Mitglieder der betreffenden Völker und Volksgruppen am Ende der 80er Jahre des 20. Jh.s.** (*Vestnik statistiki* 1990, 10: 69-75; Laakso (Hrsg.) 1992; Karlsson 1994; Rantala 1994).

	Nationalität bzw. ethnische Gruppe	Anzahl der Sprecher
<b>Lappische Sprachen (Saame)</b>		
	Südsaame	500 ?
	Luleåsaame	2 000 ?
	Piteåsaame	10
	Umeåsaame	10
	Nordsaame	ca. 30 000
	Inarisaame	ca. 400
	Skoltsaame	ca. 300 (Rußland: 20 bis 30)
	Akkalasaame	7 (1993)
	Kildinsaame	650 (1994)
	Turjasaame	6 (1994)
	Saamen im Gebiet der ehemaligen Sowjetunion	1890 797 (1989)
<b>Ostseefinnische Sprachen</b>		
	Livisch	226 99 (10) <sup>2</sup>
	Estnisch	1 026 649 980 033
	Wotisch	ca. 20
	Finnisch	ca. 5 Mill. ca. 5 Mill.
	Ingrisch	820 (2 000) 302
	Karelisch	130 929 62 542
	Weissmeer-Karelisch (Nord-, Dvina-Karelisch)	53 %
	(einschl. des Tver-Karelischen)	
	Olonetzisch	40 %
	Lüdisch (Süd-Karelisch)	7 %
	Wepsisch	12 501 6 355
<b>Wolgafinnische Sprachen</b>		
	Mordwinische Sprachen	1 153 987 773 827
	Erzänisch	ca. 67 %

2. Die Zahl der Sprecher des Livischen in Klammern sowie die des Wotischen stammen von Seppo Suhonen 1997. Die Zahl der Ingrier in Klammern stammt von Manja Lehto.

Mokschanisch		ca. 33 %
Tscheremissische Sprachen (Mari)	670 868	542 160
Ostmari		ca. 90 %
Westmari		ca. 10 %
Permische Sprachen		
Udmurtisch (Wotjakisch)	746 793	520 101
Komi (-Syrjänisch)	344 519	242 515
Permjakisch	152 060	106 531
Ugrische Sprachen		
Ungarisch	ca. 14 Mill.	ca. 14 Mill.
Mansisch (Wogulisch)	8 474	3 140
Hantisch (Ostjakisch)	22 521	13 615
Samojedische Sprachen		
Enzisch		
(Jenissei-Samojedisch)	209	95
Nenzisch (Jurakisch)	34 665	26 730
(Tundranenzisch und Waldnenzisch)		
Nganassanisch (Tawgi)	1 278	1 063
Sölkupisch		
(Ostjak-Samojedisch)	3 612	1 721

Die Tabelle zeigt, daß nur die Sprachen Estnisch, Finnisch und Ungarisch mehr als eine Million Sprecher haben, über 10 Millionen nur das Ungarische. Die Gesamtzahl aller Sprecher der ostseefinnischen Sprachen beträgt nur etwa sieben Millionen. Wie kritisch die Gefahr des Aussterbens der verschiedenen Sprachen schon ist, läßt sich auch anhand der obigen Tabelle feststellen. In unmittelbarer Aussterbegefahr schweben Sprachen, die nur von einzelnen Individuen oder höchstens von einigen Tausenden gesprochen werden. In diesen Fällen konzentrieren sich die Forscher auf die Sammlung und Dokumentation des betreffenden sprachlichen Materials. Besonders bedroht sind von den uralischen Sprachen alle lappischen Sprachen, ausschließlich des Nordsaame, sowie Livisch, Lydisch, Wotisch, Ingrisch, Mansisch, Enzisch, Nganassanisch und Sölkupisch. Auch Wepsisch gehört zu dieser Gruppe. Die nächste Gruppe wird von den Sprachen gebildet, die zwar von mehr als 10 000 Menschen gesprochen werden, bei denen die Sprecherzahl jedoch deutlich die Anzahl der Angehörigen derselben ethnischen Gruppe unterschreitet. Solche Sprachen sind Nordsaame, Tver-Karelisch, Olonetzisch, Hantisch und möglicherweise auch Nenzisch. Die dritte Gruppe besteht aus den wolgafinnischen und permischen Sprachen, die von ziemlich vielen Menschen verwendet werden, deren Funktion als Muttersprache bzw. Hauptsprache trotzdem unter ihnen stark abgenommen hat.



Alle uralischen Sprachen werden in den zentralen Sprachgebieten der indoeuropäischen, im Westen der germanischen, im Osten der slawischen Sprachen gesprochen. Das ungarische Sprachgebiet befindet sich in der Mitte zahlreicher germanischer, slawischer und romanischer Sprachen. Von den Sprachen anderer Sprachfamilien werden besonders viele türkische Sprachen in denselben Gegenden gesprochen, in denen auch uralische Sprachen verbreitet sind (Karte 1. Die Sprachgebiete der uralischen Sprache; Suihkonen 1998a: Map 13. Areal Distribution of the Uralic Languages; Nickel 1994: 7; Laakso (Hrsg.) 1992).

### 3. Elektronische Korpora der uralischen Sprachen im UHLCS

Das aktuelle Datenträgermodell des von den sprachwissenschaftlichen Instituten der Universität Helsinki gemeinsam gepflegten Server für elektronische Korpora ist die Digital AlphaStation 500 mit einem UNIX-Betriebssystem (Digital UNIX 4.0). Der Server für elektronische Korpora der Universität Helsinki wurde vorrangig den Lehrkräften, Forschern und Studenten der Universität zur Verfügung gestellt. Das Recht, den UHLCS zu benutzen, wird mit einem speziellen Formular schriftlich beantragt, wobei der Antragsteller sich verpflichtet, die Korpora nur für Unterrichts- bzw. Forschungszwecke zu verwenden. Die Korpora im UHLCS stehen unter Urheberrecht, und das Benutzungsrecht der Korpora schließt ihre Übertragung an weitere Computer aus, wenn kein von dem Urheberrecht verlangter Sondervertrag vorhanden ist. In allen Forschungsarbeiten, in denen Korpora aus dem UHLCS als Material verwendet werden, müssen diese Korpora im Quellenverzeichnis aufgelistet werden. Werden die elektronischen Korpora für die Forschungsarbeit weiterverarbeitet, z.B. mit neuer grammatischer Information versehen, wird eine Kopie des kodierten Korpus für den UHLCS verlangt.

Schon seit fast 20 Jahren wird multilinguales Material im UHLCS gesammelt. Das erste große finnische Korpus war das von Auli Hakulinen, Fred Karlsson und Maria Vilku gesammelte HKV-Korpus (Hakulinen, Karlsson & Vilku 1980), das eine umfangreiche Sammlung morphologisch und syntaktisch kodierter finnischsprachiger Texte ist. Insgesamt betragen die finnischen Korpora mehrere Millionen Wörter. Eingeschlossen sind ein Textkorpus der gesprochenen Sprache, ein rückläufiges Wörterbuch der finnischen Sprache, Zeitungstexte, die finnische Kulturgeschichte und weitere von den Verlagen Otava und Werner Söderström herausgegebene Literatur. Die Korpora des Finnischen gehören zum Kern des Korpusmaterials im UHLCS, und zwar zusammen mit den großen englischen, deutschen, schwedischen und russischen Korpora sowie einem Swahili-Korpus. Die elektronischen Korpora sind in einem speziellen Korpusverzeichnis aufgelistet, in dem alle Sprachen eigene Unterverzeichnisse haben. Unten wird die Auflistung der wichtigsten Korpusverzeichnisse im UHLCS präsentiert.

Finnisch (Name des Verzeichnisses in Klammern: **fin**), Englisch (**eng**), Schwedisch (**swe**), Swahili (**swa**), Somalisch (**somali**), Russisch (**rus**), Deutsch (**ger**), Spanisch (**espanol**), Estnisch (**viro**), Jiddisch (**yidish**) und Latein (**lat**) haben ihre gesonderten Verzeichnisse. Andere Korpusverzeichnisse sind bis zu einigen Ausnahmen so gruppiert worden, daß die genetisch zusammengehörenden Sprachen jeweils in demselben Verzeichnis enthalten sind. Ihren eigenen Verzeichnispfad haben kaukasische (**caucasian-lgs**), iranische (**iranian-lgs**), mongolische (**mongolic-lgs**), paläosibirische (**palaeo-**

**siberian-lgs**), türkische (**turkic-lgs**), uralische (**uralic-lgs**) und tungusische (**tungusic-lgs**) Sprachen. Die Verzeichnisse verschiedener Sprachen werden je nachdem in Unterverzeichnisse unterteilt welcher Art das Material ist, das in diesen Sprachen zur Verfügung steht. Jede Textsorte wurde in ein eigenes Unterverzeichnis eingeordnet, und neue Unterverzeichnisse wurden je nach Bedarf angelegt. Als Beispiel ist unten das Korpusverzeichnis der uralischen Sprachen aufgeführt worden (i. J. 1998).

## Verzeichnisse der uralischen Sprachen

### Korpusverzeichnis

/corp/

#### Verzeichnis der uralischen Sprachen

uralic-lgs/

#### Verzeichnisse verschiedener

#### Sprachen

enets/

karelian/

New-Testament/

Books-of-Children/

New-Testament/

khanti/

khanti-clauses-references/

khanti-texts-snt/

morphologically-tagged-corpora/

komi/

komi/

permyak/

livonian/

mansi/

mari/

eastern-mari/

western-mari/

mordvin/

erzya/

moksha/

nenets/

saami/

kildin-saami/

northern-saami/

#### Materialverzeichnisse

New-Testament/

Books-of-Children/

Bible-of-Children/

Books-of-Children/

Books-of-Children/

New-Testament/

komi-texts/

komi-texts-snt/

morphologically-tagged-corpora/

Books-of-Children/

New-Testament/

ample/

ife-of-Jesus/

New-Testament/

Books-of-Children/

New-Testament

morphologically-tagged-corpora/

Bible-of-Children/

New-Testament/

morphologically-tagged-corpora/

Books-of-Children/

New-Testament/

New-Testament/

Books-of-Children/

report/

	ume-saami/ tym-dialect/	morphologically-tagged-corpora/
selkup/ udmurt		Books-of-Children/ New-Testament/ udmurt-snt/ udmurt-statistical-data/
vepsian/		Bible-of-Children/ Books-of-Children/ New-Testament/

Das Korpusverzeichnis beinhaltet auch das Korpus des Estnischen (/viro/; bearbeitet von Kazuto Matzumura und Maria Vilkuina) mit Auszügen aus der estnischen Literatur: aus Kurzgeschichten sowie aus Auszügen aus Romanen und Zeitschriften in der estnischen Sprache. Dem Verzeichnis jeder Sprache wurde eine Sonderdatei, **README**, hinzugefügt, mit Auskunft über den Ursprung des Korpus, Quellenverzeichnissen der Veröffentlichungen sowie weiteren wichtigen Informationen über das Korpus. Auch die Kontaktadressen der Verwalter der jeweiligen Korpora sind in der Datei README erhältlich.

Das numerisch kodierte, statistische Korpus des Udmurtischen ist eines der ersten Korpora der entfernt verwandten Sprachen des Finnischen, die in den UHLCS übertragen wurden. Unten folgt ein kurzes Beispiel dieses Korpus, das auch dokumentiert worden ist (Suihkonen 1990). Nach dem verwendeten Kodierungsmuster enthält jeder Abschnitt von zwei Zeilen im Beispiel die kodierte Information für jeweils einen Satz. Die Stelle des Satzes im ursprünglichen Text wurde mit einer fünfstelligen Ziffernfolge am linken Rand der ersten Zeile kodiert: Die zwei ersten Ziffern zeigen die Seitennummer im Text, die zwei nächsten die Nummer des Beispielsatzgefüges im Text, denen noch eine Angabe über die Nummer des Satzes im Satzgefüge folgt. Insgesamt wurden im Satz fast hundert statistische Variablen analysiert, die aus den morphologischen, syntaktischen, semantischen und pragmatischen Eigenschaften der Sprache bestehen. Die über jede Variable erteilte Information wurde mit einer Ziffernfolge aus einer bzw. zwei Ziffern kodiert.

```

07021 1 09 - - - - 1 - - - - - 04 04 - - - - - 4 1 1 - - - - - 3
2 15 - - - - - 1 - - - - - 3 01 04 01 01 02 01 - - 01 03 01 - 1 - - -
07022 1 09 - - - - 1 - - - - - 04 -- 04 -- - - - 4 1 1 - - - - - 3
2 15 - - - - - 1 - - - - - 01 04 01 01 02 01 - - 01 03 01 - 1 - - -
07023 1 09 - - - - 2 - - - - - 04 - - - 04 - - - 4 1 2 - - - - - 3
3 14 - - - - - 1 - - - - - 03 05 01 03 01 01 2 - 01 02 03 - 2 - - -
07031 1 09 - - - - 1 - - - - 2 - 2 - - - 31 31 - - - - - 4 1 1 - - - - -
3 4 13 - - - 17 - - - - - 1 9 - - - - - 2 2 2 03 07 03 03 02 02 1 - 02 04 03 - 2 - - 2
-
07032 1 09 - - - - 2 - - - - 2 - 2 - - - 31 -- 31 - - - - - 4 1 2 - - - - -
3 3 14 - - - 19 - - - - - 1 9 3 - - - - - 2 1 - 04 08 02 04 02 02 2 - 02 04 04 - 2 - - 2
-

```

Die Variablen und Variablenklassen im Korpus werden in den Variablenlisten auf finnisch und englisch erklärt. Diese Listen befinden sich als selbständige Dateien im Korpusverzeichnis. Für die Analyse des in Matrixform kodierten Materials wurden die

Statistikprogramme HYLPS und SAS des EDV-Zentrums der Universität Helsinki verwendet.

Vom verarbeiteten Material des Korpusprojekts **Struktur und Typologie der uralischen Sprachen** wurden die Textkorpora des Komi, des Udmurtischen und des Hantischen in den UHLCS übertragen, und ebenso das Material des Sölkupischen, das als "Nebenprodukt" der von der Finnisch-Ugrischen Gesellschaft finanzierten Wörterbucharbeit (Jarmo Alatalo) behandelt und kodiert wurde. Das Material vom Institut für Bibelübersetzung besteht aus Übersetzungen verschiedener Bücher der Bibel sowie weiterer religiöser Literatur. Das umfangreichste Material stammt aus dem Olonetzischen, Wepsischen und Erzänischen und enthält u.a. Übersetzungen der Kinderbibel, sowie aus dem Udmurtischen, dessen Korpora eine fast vollständige Übersetzung des Neuen Testaments beinhalten. Übersetzungen der Evangelien sind im Tver-Karelischen, Olonetzischen, Wepsischen, Erzänischen, Mokschanischen, Syrjänischen (Komi) und Permjakischen vorhanden. Das Material im Kildinsaame, Enzischen, Nenzischen, und Mansischen enthält jeweils nur einen kurzen Text von einigen zehn Seiten: die einen Texte sind ein Kinderbüchlein über das Leben Jesu, die anderen eine kurze Zusammenfassung des Lukasevangeliums. Auch in dieser Form sind sie eine wichtige Ergänzung der Beispielsammlungen der bedrohten uralischen Sprachen im Server für elektronische Korpora der Universität Helsinki. Von vielen Texten gibt es parallele Übersetzungen in mehreren Sprachen, was außer zum Vergleich der Textübersetzungen auch der kontrastiven Erforschung der Strukturen verschiedener Sprachen dient.

Auch private Forscher und Institutionen haben die elektronischen Korpora der Universität ergänzt. Beispiele dafür sind das Korpus des Nordlappischen, das aus einem Abschnitt der Übersetzung von Irja Seurujärvi-Kari des Berichts des Samischen Ausschusses (*Komiteanmietintö* 1985: 66, Opetusministeriö, Helsinki) besteht und das von Komi, dessen Grundlage das Werk *Jujas da s'ölomjas* der Schriftsteller Ivan Toropov (Syktyvkar 1996; das Werk ist zum elektronischen Korpus von Jack Rueter ausgearbeitet) ist. Einige der elektronischen Korpora der uralischen Sprachen im UHLCS stammen aus den Notizen der Erforscher dieser Sprachen. Korpora dieser Art sind das hauptsächlich aus dem von Kai Donner gesammelten Textmaterial entstandene sölkupische Korpus sowie die meisten hantischen Korpora, die aus den von Károly Rédei, Wolfgang Steinitz und Heikki Paasonen gesammelten Textbeispielen verarbeitet wurden. Zur Zeit wird an einem umeälappischen Korpus gearbeitet, das aus einem auf Band aufgenommenen, transliterierten und in Textform zum elektronischen Korpus verarbeiteten Material besteht.

Die meisten Korpora, die in elektronischer Form erhalten wurden, wie z.B. das Material aus dem Institut für Bibelübersetzung, waren mit kyrillischen Zeichen geschrieben und gespeichert. Weil kyrillisches Material noch nicht als solches in das UNIX-Betriebssystem übertragen werden kann, muß die Übertragung des Materials in zwei Phasen vorgenommen werden. Zuerst werden die Texte in eine solche Form gebracht, daß sie im Computernetz von einem Betriebssystem zum anderen übertragbar sind. Danach wird das Material weiter transformiert. Zuerst muß ein Zeichen im ursprünglichen Material durch nur ein einziges Zeichen in der neuen Form ersetzt werden. Danach erfolgt die Umsetzung der Zeichen in die phonematische Form der betreffenden Sprache. Mit der Umsetzung in die phonematische Form wurde erst angefangen: schon dieser Teil der Korpusarbeit ist ein umfangreicher eigenständiger Forschungsbereich für sich. Das

Ziel der Arbeit ist eine möglichst automatische Umsetzung des Materials in allen Phasen der Verarbeitung.

Weiter wurde kyrillisches Material auch direkt auf einem PC zu Textkorpora transliteriert. Unten ist ein Textbeispiel für die Umsetzung dieser Art aus dem udmurtischen Korpus zu sehen. Der Textabschnitt stammt aus dem udmurtischen Volksmärchen *Amalo dz'its'y* 'Der schlaue Fuchs' (Udmurtisches Textkorpus, UHLCS; *Udmurt kalyk skazkaos*, 7 - 10, Iževsk 1940). Zur Erleichterung der automatischen Verarbeitung des Materials, wie z.B. der Textkodierung oder verschiedener Suchvorgänge, wurden Satzzeichen durch eine Leertaste vom Text getrennt und große Anfangsbuchstaben durch eine Kombination von einem Sternchen und den entsprechenden Kleinbuchstaben ersetzt. Diese Arbeitsphase wird Vorprozessierung des Materials genannt.

\*amalo dz'its'y

\*odig murtlen nokinez vlylmyte . \*utts'any koshkem so aslyz ulon inty . \*baddz'ym n'ulesky injas'kem . \*zökes' ljaljos vylem otyñ . \*ognaz poryny kuzhymez ug tyrmy , pe . \*esh utts'any koshkem .

\*myne , pe , in'i ta malpas'kysa . \*pumitaz gondyr lykte .

- \*gondyr urom , vaj as'meos valts'e ulom , - shuem ta murt .

\*gondyr so shory oskytek uts'kem .

- \*oz'yez , oz'y no ... \*kыз'y ben as'meos valts'e ulom ?

\*ljal porylom . \*tolaltely tynyd no shunyt korka les'tom . \*so ponna tynyd trosges sil' no das'ano luoz na , - valekte , pe , vorgoron .

#### 4. Verarbeitung der Korpora für Forschung und Unterricht

Für Forschungszwecke werden elektronische Korpora unterschiedlich verarbeitet und analysiert. Die erste Phase bei der Kodierung der Korpora ist eine morphologische Analyse des Materials, der eine syntaktische Kodierung folgt. Während der morphologischen Analyse werden die sprachlichen Einheiten mit Kodierungsindexen der grammatischen Kategorien versehen. Die Kodierung der syntaktischen Struktur unterschiedlicher Ebenen erfolgt dann in der syntaktischen Analyse des Materials. Eine gesonderte Phase ist weiter die Kodierung der Textstruktur und die einheitliche Dokumentation der Veröffentlichungsangaben des Materials. Auch eine Übersetzung des Materials ins Englische oder Deutsche wird beabsichtigt. Ein Teil des syrjänischen Korpus (Komi) wurde ins Finnische übersetzt.

Im Verlauf der Korpusprojekte wurde versucht, die beste Kodierungsweise der Korpora zu ermitteln. Die Anstrengungen zur Schaffung und Verarbeitung der elektronischen Korpora in uralischen Sprachen sind keineswegs einzigartig, sondern im gesamten Gebiet der EU sind umfangreiche, von der Union finanzierte Projekte für die Sammlung und Verarbeitung der Textkorpora und die lexikalische Beschreibung aller nationalen Sprachen der EU-Länder im Gange. Auch bei diesen Projekten wurden unterschiedliche Konventionen für die Korpuskodierung entwickelt. In der Kodierungspraxis der Korpora in uralischen Sprachen kann eine Streuung festgestellt werden, die teilweise auf unterschiedliche Methoden in der Korpuskodierung zurückzuführen ist. In der Zukunft wird eine möglichst einheitliche Kodierung angestrebt. Nach zahlreichen Phasen wurde

schließlich vereinbart, ein Kodierungssystem zu entwickeln, das die folgenden Voraussetzungen erfüllt: a) Es sollte die Eigenschaften der uralischen Sprache optimal beschreiben; b) Es sollte ausreichend ikonisch und deutlich sein, um die Verwendung der Korpora zu erleichtern; c) Es sollte für die weiteren verwendeten Formalismen leicht übertragbar sein. Das Hauptgewicht der Kodierung wird auf die Beschreibung der Sprachen gelegt. Auf den nächsten Seiten sind Beispiele für Textkorpora in uralischen Sprachen angegeben, die zu unterschiedlichen Zeitpunkten kodiert wurden (die Ziffernfolgen in den syrjänischen, udmurtischen und erzänischen Korpora weisen auf die ursprüngliche Stelle des Satzes im Text hin).

**Material des Komi (Syrjänischen) mit morphologischer Kodierung (Paula Kokkonen 1994)**

<N'ina Kuratova (1983). Bobön'an' kör, Povest'jas, vis'tjas.  
Komi kn'izhnöj izdatel'stvo, Syktyvkar.>

<FICT\_SH\_Ni\_Ku\_1983\_BK\_186-188/osa novellista Kuzmits>  
<T\_N-P\_Kuz'mits'\_SG\_NOM=Kuzmits (miehen nimi)>

01860001

PRN-PRS\_Mijan\_PL\_GEN=me

N\_s'iktyn\_SG\_IN=kylä

ADJ\_unakod'SG\_NOM=paljonlainen

V\_kutisny\_PRET\_3PL=alkaa

V\_lony\_INF=tulla olemaan

N\_pövjasön\_PL\_INST=lauta

V\_tupkalöm\_3PCP=tukkia

ADJ\_öshin'jasa\_PL\_NOM\_SG=ikkunoilla varustettu

N\_kerkajas\_PL\_NOM=talo

.#Meidän kyläämme alkoi ilmestyä melko paljon taloja, joiden  
ikkunat oli laudoilla suljettuja.#

01860002

ADJ-CAR\_Shushtöm\_SG\_NOM=alakuloinen, ikävä

N\_mövpijas\_PL\_NOM=ajatus

V\_ts'uzhtöny\_PRES\_3PL=syntyä

PRN3\_najö\_PL\_NOM=he, ne

:

PCL\_so=kas

,

V\_dumyshtan\_PRES\_2SG=ajatella

,

ADV\_n'evazhön=äskän

PCL\_na=vielä

V\_olisny\_PRET\_3PL=elää

ADV\_tani=täällä

N\_jöz\_SG\_NOM=ihmiset

,

12

CONJ\_a=mutta

ADV\_öni=nyt

PRN-NEG\_n'ekod\_SG\_NOM=ei kukaan

V-NEG\_oz\_PRES\_3SG-PL=ei

V\_tydav\_NEG\_SG=näkyä

.#Alakuloisia ajatuksia ne synnyttävät, ajattelethan, (että)

vielä äsken täällä asui ihmisiä, mutta nyt ei näy ketään.#

### Material im Erzänischen mit morphologischer Kodierung (Jack Rueter 1994)

00080025

V\_Uchost'\_IND\_PT1\_SUBJ-3PL\_FAB = WAIT

N-P\_Mat'an'\_GEN-OBJ\_IDF = (GIVEN NAME)

,

PRN-REL\_kona\_SG\_NOM\_IDF = WHO

V\_tus'\_IND\_PT1\_SUBJ-3SG\_FAB = LEAVE

N\_chen'ksen'\_GEN-OBJ\_IDF = LIQUOR

V\_veshn'eme\_INF-OBJ = LOOK FOR

.#THEY WERE WAITING FOR MATJA WHO HAD GONE OUT LOOKING FOR  
MOONSHINE.#

00080026

PRN-PER\_Son\_SG3\_NOM = SHE

V\_sas'\_IND\_PT1\_SUBJ-3SG\_FAB = COME

ADJÈN\_chavon'\_GEN-ATR\_IDF = ONE WHO IS EMPTY

N\_kedt'\_PL|NOM-COM\_IDF = HAND

.#SHE CAME BACK EMPTY-HANDED.#

00080027

N\_kedt'\_PL|NOM-COM\_IDF = HAND

00080027

NUM\_Kavto\_SG\_NOM\_IDF = TWO

N\_kudoso\_INE\_IDF = HOUSE

V\_ul'n'in'\_IND\_PT1\_SUBJ-1SG\_FAB = BE

,

PRN-\_mez'ejak\_SG\_NOM\_IDF\_PCL-INCL = ANYTHING

PCL-NEG1\_a = NOT

V\_maksit'\_IND\_PRS\_SUBJ-3PL\_FAB = GIVE

PRN-PER\_ton'\_SG2\_GEN = YOU

N\_kasol'ent'\_SG\_GEN-OBJ\_DEF = BEANS

POP\_kise\_INE\_IDF = FOR

,

N-P\_Pakhom\_SG\_NOM = (GIVEN NAME)

.#I HAVE BEEN TO TWO HOUSES, PAKHOM, BUT THEY WON'T GIVE  
ANYTHING FOR YOUR BEANS.#

**Sölkupisches Korpus mit morphologischer und syntaktischer Kodierung und deutscher Übersetzung (Jarmo Alatalo 1998)**

Itja und Pünegusse. (A8)

T: A8.

iicjä immøljantösä uørkäsjpøkkaqø.

Itja mit seiner Tante lebten.

iicjä N SG NOM SBJ

immøljantösä N SG KOM 3S LDM SAD

uørkäsjpøkkaqø V AOR 3DS FRE KNT VER

T: A8.

okkør taaqøn iicjä mykka immøljaqøntø:

Einmal sagte Itja zur Tante:

okkør P ATR

taaqøn N SG LOK SAD

iicjä N SG NOM SBJ

mykka V AOR 3SS FRE VER

immøljaqøntø N SG DAT 3S LDM SAD

T: A8.

"man kojalang qarøngtølj aaqqøn poqlontø".

"Ich gehe zum Bucht des Schiefen Flußbetts".

man P SBJ

kojalang V OPT 1SS VER

qarøngtølj P IP TRA ATR

aaqqøn N SG GEN ATR

poqlontø N SG DAT SAD

**Udmurtisch** (Pirkko Suihkonen 1998b: 30-31, 48-49, Testkodierung)

**(1) Morphologische Kodierung mit Übersetzungen der Grundformen von Wörtern**

*Dzhog*\_ADV\_MAN

*fast, soon*

*ortts'*+i+z\_V\_-CONT\_-TRA\_+FIN\_IND\_PAST\_SG3

*to pass (away)*

*zarn'i*\_N\_-COUNT\_SG\_NOM|A\_SCAL\_SG\_NOM *golden*

*kuaro*\_A\_SCAL\_SG\_NOM

*with leaves*

*dyr*\_N\_+|-COUNT\_-ANIM\_SG\_NOM

*time*

*dzhog+en*\_ADV\_MAN\_INSM

*fast*

*vu+i+zy*\_V\_-CONT\_-TRA\_+FIN\_IND\_PAST\_PL3

*to come*

*zhob*\_A\_SCAL\_SG\_NOM

*nasty, unpleasant*

*siz'yl*\_N\_+COUNT\_-ANIM\_SG\_NOM|A\_SCAL\_SG\_NOM|ADV\_TIME *autumn*

*nunal+jos*\_N\_+COUNT\_-ANIM\_PL\_NOM

*day*

*In+ez*\_N\_+COUNT\_-ANIM\_SG\_ACC

*heaven, sky*

<i>s'ekyt</i> _A_SCAL_SG_NOM	<i>heavy</i>
<i>pil'em+jos</i> _N_+COUNT_-ANIM_PL_NOM	<i>cloud</i>
<i>shobyrt+i+zy</i> _V_+CONT_+TRA_CAUS_+FIN_IND_PAST_PL3	<i>to cover</i>
.	
<i>Vekts'i</i> _A_SCAL_SG_NOM	<i>thin</i>
<i>puzh</i> _N_+COUNT_-ANIM_SG_NOM	<i>sieve</i>
<i>pyr+ti</i> _PP_SG_TRNS	<i>through</i>
<i>kad'</i> _ADJ_SCAL ADV_MAN CONJ_CMP	<i>as</i>
,	
<i>vistem</i> _ADV_MAN_NEG	<i>without interruption</i>
<i>kis't+o</i> _V_-CONT_-TRA_+FIN_IND_PRES_PL3	<i>to spill, fall</i>
<i>siz'yl</i> _N_+COUNT_-ANIM_SG_NOM A_SCAL_SG_NOM ADV_TIME	<i>autumn</i>
<i>veks'i</i> _A_SCAL_SG_NOM	<i>thin</i>
<i>zor+jos</i> _N_+COUNT_-ANIM_PL_NOM	<i>rain</i>
.	

## (2) Glossierung der Sätze und Übersetzung aus dem Udmurtischen ins Englische

- 0024000101 *Dzhog ortts'+i+z zarn'i kuario dyr,*  
*Fast*\_ADV\_MAN *pass-away*\_V\_-CONT\_TRA\_+FIN\_IND\_PAST\_SG3  
*gold|golden*\_N\_-COUNT\_SG\_NOM|A\_SCAL\_SG\_NOM  
*with-leaves*\_A\_SCAL\_SG\_NOM *time*\_N\_+|-COUNT\_-ANIM\_SG\_NOM,  
*Fast the time of golden leaves passed away,*
- 0024000102 *dzhog+en vu+i+zy zhob siz'yl nunal+jos.*  
*fast*\_ADV\_MAN\_INSM *come*\_V\_-CONT\_-TRA\_+FIN\_IND\_PAST\_PL3  
*unpleasant*\_A\_SCAL\_SG\_NOM *autumn*\_N\_+COUNT\_-ANIM\_SG\_NOM|  
A\_SCAL\_SG\_NOM|ADV\_TIME *day*\_N\_+COUNT\_-ANIM\_PL\_NOM.  
*fast unpleasant autumn days came.*
- 0024000201 *In+ez s'ekyt pil'em+jos zhobyrt+i+zy.*  
*Sky*\_N\_+COUNT\_-ANIM\_SG\_ACC *heavy*\_A\_SCAL\_SG\_NOM  
*cloud*\_N\_+COUNT\_-ANIM\_PL\_NOM  
*cover*\_V\_+CONT\_+TRA\_CAUS\_+FIN\_IND\_PAST\_PL3 .  
*The sky was covered with heavy clouds.*
- 0024000301 *Vekts'i puzh pyr+ti kad', vistem kis't+o siz'yl vekts'i zor+jos.*  
*Thin*\_A\_SCAL\_SG\_NOM *sieve*\_N\_+COUNT\_-ANIM\_SG\_NOM  
*through*\_PP\_SG\_TRNS *as*\_ADJ\_SCAL\_SG\_NOM|ADV\_MAN|CONJ\_CMP ,  
*withouth-interruption*\_ADV\_MAN\_NEG  
*fall-down*\_V\_-CONT\_-TRA\_+FIN\_IND\_PRES\_PL3  
*autumn*\_N\_+COUNT\_-ANIM\_SG\_NOM|A\_SCAL\_SG\_NOM|ADV\_TIME  
*thin*\_A\_SCAL\_SG\_NOM *rain*\_N\_+COUNT\_-ANIM\_PL\_NOM.  
*As through a thin sieve, thin autumn rains keep on falling.*

Die obigen Beispiele sind Auszüge aus manuell kodierten Texten. Zur Zeit wird z.B. an einem Programm für die Analyse des Komi gearbeitet, das die Wortformen automatisch kodiert und identifiziert. Der folgende Abschnitt ist ein finnischsprachiges Beispiel für eine Analyse dieser Art. Der Satz wurde mit Hilfe der von Kimmo Koskeniemi entwickelten Modelle (Two-Level Morphology, Koskeniemi 1983) analysiert. Die

automatische Analyse ergibt zuerst die Wortform im Korpus, dann wird die Interpretation der betreffenden Wortform, d.h. ihre Bedeutungen, aufgelistet. Die Interpretation enthält die Grundform des Wortes und die morphologische Analyse der Wortform. Alle Bedeutungen der jeweiligen Wortform zusammen bilden eine Kohorte (Karlsson 1992: 3; Karlsson 1995: 46). Im UHLCS wird ein automatisches Analyseprogramm des Finnischen verwendet.

### **Automatische Analyse des Finnischen** (Koskenniemi 1983)

(Die Katze schlief auf der Stalltreppe und träumte vom Sommer)

\* kissa nukkui tallin portailla ja uneksi kesästä .

("<\*>")  
 ("<kissa>"  
   ("kissa" N NOM SG))  
 ("<nukkui>"  
   ("nukkua" V PAST ACT SG3))  
 ("<tallin>"  
   ("talli" N GEN SG))  
 ("<portailla>"  
   ("porras" N ADE PL))  
 ("<ja>"  
   ("ja" COORD C))  
 ("<**uneksi**>"  
   ("uneksia" V IMPV ACT SG2)  
   ("uneksia" V PRES ACT NEG)  
   ("uneksia" V PAST ACT SG3)  
   ("uni" N TRA SG))  
 ("<kesästä>"  
   ("kesä" N ELA SG))  
 ("<.>")

Wird eine automatische Kodierung für die Analyse einer Sprache programmiert, kann dieses Programm für die Kodierung weiteren Materials verwendet werden. Die wichtigste Aufgabe der Kodierung, ob automatisch oder manuell, ist die Erteilung möglichst zuverlässigen Wissens über die jeweilige Sprache. Weiter sollte die Kodierung so sorgfältig durchgeführt werden, daß nichts von der Information im ursprünglichen Text verlorenght. Das kodierte Material sollte auch in andere, mit einem verschiedenen Konzept bzw. für typologisch unterschiedliche Sprachen verarbeitete Kodierungssysteme umsetzbar sein.

### **5. Verwendung des Materials zu Forschungszwecken**

Das Material kann in der Praxis vielfach verwertet werden. Wichtig ist die sprachwissenschaftliche Forschung einschließlich der Untersuchung der Sonderbereiche in diesen

Sprachen und der Bearbeitung des Materials für Wörterbücher und Grammatiken. Zu dieser Kategorie gehört u.a. das udmurtische Wörterbuch, das aus dem Material des udmurtischen Textkorpus erstellt wurde. Später wird das Wörterbuch im Server für elektronische Korpora erhältlich sein. Im folgenden Beispiel steht zuerst das udmurtische Wort, dem dann seine englische und finnische Übersetzung folgt.

**Elektronische Korpora als Material für Wörterbücher** (Suihkonen, Zagulyayeva & Tronina 1995: 17)

UDMURT/UDMURTTI  
ENGLISH/ENGLANTI  
FINNISH/SUOMI

ad'ami, N  
man, human being; person.  
ihminen; mies; henkilö.

addz'+em, 1. V PCPL <PAST>, 2. N  
1. s. addz'yyny. 2. seeing.  
1. ks. addz'yyny. 2. näkeminen.

addz'empton, N  
wish to see.  
halu nähdä.

addz'empot+on+tem, ADJ <NEG>  
ks. addz'empton; hateful; loathsome.  
ks. addz'empton; vihattava; vastenmielinen, inhottava.  
addz'empotostem = addz'empotontem.

addz'em#pot+y+ny, V <+CONT> <+TRA> INF  
to want to see.  
haluta nähdä.

addz'is'k+is', V PCPL <PRES>  
s. addz'is'kyny.  
ks. addz'is'kyny.

addz'is'k+on, N  
1. visibility. 2. meeting, encounter. 3. ghost.  
1. näky(väi)sys. 2. kohta; tapaaminen. 3. aave.

addz'is'k+on+tem, PCPL <ACT> <PRES/FUT> <NEG> <NEG>  
s. addz'is'kyny.  
ks. addz'is'kyny.

addz'+is'k+y+ny, V <-CONT> <-TRA> <REC/PSS> INF

1. to see each another, meet each other. 2. to be visible.
3. to seem, appear. 4. to look (like).
1. tavata (toisiaan); nähdä (toisiaan). 2. näkyä, hämöttää.
3. näyttää, tuntua (jlt). 4. näyttää jlt, olla jnk näköinen.

Die Befehlsprachen des UNIX-Betriebssystems, verschiedene Programmiersprachen und Befehle in den Editoren, die "Werkzeuge", bieten zahlreiche Möglichkeiten zur Verwertung des Textmaterials, und zwar von der Alphabetisierung und Kodierung des Materials und den Suchvorgängen der sprachlichen Beispiele bis zur Programmierung der Zergliederer, Parser, in diversen Programmiersprachen. Einen zentralen Teil der Werkzeuge im UNIX-Betriebssystem bilden die Befehle interpretierenden Programme des Betriebssystems. Mit diesen Befehlen können unterschiedliche Suchvorgänge im Material sowie Umsetzungen und Sortierungen der Befunde nach Bedarf durchgeführt werden. Entsprechende Befehle sind auch im **Emacs**, dem Textverarbeitungsprogramm im UNIX-Betriebssystem, enthalten. Weil die Befehle interpretierenden Programme des UNIX-Betriebssystems im eigentlichen Sinne leichte Programmiersprachen eines höheren Niveaus sind, können bei der Definierung des Materials sog. **reguläre Ausdrücke** verwendet werden. Bei regulären Ausdrücken kann eine zu suchende Zeichensequenz durch verschiedene Gruppierungsoperationen erweitert werden, um für unterschiedliche Kombinationen bzw. Alternativen der Zeichensequenzen zu gelten. Durch zusätzliche Zeichen, die einzelne Zeichen oder Ausdrücke ersetzen, wird wiederum die Variation innerhalb der zu suchenden Zeichensequenz definiert. Unten ist ein Abschnitt aus der Datei des udmurtischen Textkorpus angeführt, die sowohl konventionell als auch rückläufig alphabetisiert worden ist. Bei der Definierung der alphabetisierten Zeichensequenzen wurden reguläre Ausdrücke verwendet. Die Zahl vor der alphabetisierten Wortform gibt die Anzahl der jeweiligen Wortform im Text an. Die Zahl am Anfang der Liste steht für die Anzahl der Worteinheiten im ganzen alphabetisierten Material (der Apostroph nach einem dentalen Mitlaut weist auf die Palatalisierung des Konsonanten hin):

### Alphabet

27698  
 2 Abash  
 2 abdras  
 2 Ad'ami  
 65 ad'ami  
 1 ad'amijed  
 3 ad'amijen  
 2 Ad'amijez  
 6 ad'amijez  
 2 Ad'amilen  
 7 ad'amilen  
 1 ad'amiles'  
 2 Ad'amily  
 3 ad'amily  
 5 Ad'amios

### Rückläufiges Alphabet

27470  
 1 vyllas'an'  
 2 berlas'an'  
 1 s'örlas'an'  
 2 taban'  
 1 gan'-gan'  
 7 az'lan'  
 4 mydlan'-az'lan'  
 1 Solan'-talan'  
 4 solan'-talan'  
 1 pedlan'  
 1 kudlan'  
 2 myd-mydlan'  
 1 vallan'  
 5 ullan'

18

25 ad'amios	7 vyllan'
2 ad'amioslen	5 berlan'
1 ad'amiosles'	2 s'örlan'
1 Ad'amiosly	1 urdeslan'
2 ad'amiosly, usw.	1 dzhytlan', usw.

## Konkordanzen

Der Server für elektronische Korpora hat auch fertige Programme, die lokal oder allgemein verwertet werden und mit deren Hilfe Konkordanzen gebildet oder Suchvorgänge im Material durchgeführt werden können. Als Beispiele dafür werden im folgenden drei Abschnitte gezeigt. Die zwei ersten, **kw-snt** und **kw-alg**, sind Suchprogramme. Das Programm **kw-snt** gibt der zu suchenden Zeichensequenz einen Kontext von einem Satz, **kw-alg** den von einer Zeile. Das dritte Beispiel zeigt eine Konkordanz aus dem Korpusmaterial des Saame<sup>3</sup>. In den ersten Beispielen wurden für die Definierung der zu suchenden Zeichensequenzen einfache reguläre Ausdrücke verwendet. Das erste Beispiel ist ein Abschnitt aus der Datei mit den Belegen für den Terminus **tietokonekorpus 'elektronisches Korpus'** aus diesem Text. Der Befehl besteht aus dem Namen der Befehlsdatei (**kw-snt**), der zu suchenden Zeichensequenz mit allen Variablen in Klammern. Der Winkelklammer mit der Spitze nach links folgt der Name der jeweiligen Datei, in der die Sequenz gesucht wird; der Winkelklammer mit der Spitze nach rechts wiederum der Name der Datei, in der die Ergebnisse gespeichert werden sollen. Im Druck steht die gesuchte Zeichensequenz in doppelten Winkelklammern. Die Zahl am Anfang der Zeile zeigt die Nummer des Satzes, in dem die Sequenz im Text gefunden wurde.

**kw-snt 'tietokonekorpu(s|sten|ks(et|i(sta|ksi)))[a-z]\*\040)' <uk-tietopankki >demo**

8: Esittelen artikkelissani Helsingin yliopiston  
<<tietokonekorpuspalvelimessa >> vuoden 1996  
lopulla olleet , suomen lähi- ja etäsukukielten  
<<tietokonekorpuksset >> .

31: Fyysisesti se sijaitsee Helsingin yliopiston yleisen  
kielitieteen laitoksella olevassa  
<<tietokonekorpuspalvelimessa >> , joka on yhdistetty  
kansainväliseen tietokoneverkkoon .

36: Myös Helsingin yliopiston <<tietokonekorpuspalvelimeen  
>> on koottu useiden muiden uhanalaisten kielten aineistoja .

37: Uralilaisten kielten tietopankki ja uhanalaisten

---

3.

Die Programme **kw-alg** und **kw-snt** sind von Kimmo Koskenniemi ausgearbeitet worden, und das Programm **KWIC** wurde von Timo Järvinen auf der Basis des Programms von Kenneth Church weiterentwickelt (Vorlesungsmanuskript 1990).

uralilaisten kielten tietopankki ovat osa Helsingin yliopiston <<tietokonekorpuspalvelimessa >> tallenteilla olevaa eri kielten tietopankkia .

Mit dem Programm **kw-alg** werden die Possessivsuffixe *-ez* und *-yz* der 3. Person im Sg. und Pl. des Udmurtischen als Zeichensequenz gesucht, die auch Suffixe für Akk. sein können. Die Eingabedatei der Konkordanz ist ein morphologisch analysierter Abschnitt aus dem udmurtischen Textkorpus. Das Beispielmateriale ist ein Lauftext, in dem keine Umsetzungen für große Anfangsbuchstaben und Satzzeichen vorgenommen wurden.

**kw-alg '\+(ez|hez|yz)' <Suchpfad/Eingabedatei >Zieldatei**

104: Anaj+ez kosem+ys' gine kyti-oti tölatis'ky+ny  
 125: a, inzhen'er+ly dyshetskon s'ures+ez .  
 118: i+z t'ehn'its'eskoj l'it'eraturaj+ez .  
 48: 'ko+d+-a, myn+a+m tshukaz'e berpum+yz erkyn nunal+e kyl'+i+z.  
 69: Viktor Ivanovits'+len pits'i dyr+yz s'elo+ja+my ortts'+i+z .  
 91: Esh+jos+yz uram+yn kalg+o ,

Das folgende Beispiel ist ein Konkordanzauszug aus dem ganzen Material des nordlapischen Korpus. Als Parameter für das Konkordanzprogramm werden neben der Definierung der Eingabe- und Zieldateien auch Zahlen als Kennzeichen für Wort, Zeichensequenz zwischen zwei Satzzeichen und Länge des Kontexts gegeben.

**KWIC 4 4 </corp/uralic-lgs/saame/northern-saame/report-smiehtamus >Zieldatei**

sápmelas'vuoda ja gávnnaahii , <ahte> sus lea skandinávalas' va  
 igi buoremus dovdomearka lei , <ahte> sus lei ritmalac'c'at njo  
 l ásaidahttojuvvot eanet nu , <ahte> ávdin ja geavatkeahtes ea  
 s dat ordnejuvvui ođđasit nu , <ahte> vearut c'oggojuvvojedje k  
 te sámec'earddaid gaskkas nu , <ahte> viimmát nuortasámiid stii  
 Immos' sáhtá govviduvvot nu , <ahte> váldá ovdan iez'as árbevi  
 ápmelac'c'at leat jurddas'an , <ahte> vuoinjnjat leat sivdnidan  
 Jurddas'uvvui nu <ahte> vuoinjnjat mávssahit ies'  
 z'an s'attai dahkat ja vuovdit <aiddo> " sámegovaid " , maida  
 ássi vuoinjnjaide vuodđduvve <aiddo> dán dihtui luonddu gierd  
 n goit bázi s'addat dovdusin <aiddo> sámi dáiddac'eahppin ,  
 id álgovuđoleamus dáiddahápmiin <aiddo> sárgundáidaga , niibegez  
 o son jámii 28-jahkásaz'z'an , <aiddo> válmmas'tuvvan oahpahead  
 hpaheaddji , journalista Matti <Aikio> ( 1872 - 1929 ) lei vuos

Das Institut für Linguistik an der Universität Helsinki unterhält die Anlagen und sorgt dafür, daß die verwendeten Programme aktuell und optimal anwendbar sind. Das Institut erteilt auch Unterricht in Grundkenntnissen über das UNIX-Betriebssystem und die Verwertung der elektronischen Korpora als Forschungsmaterial. Die sprachwissenschaftlichen Institute der Universität Helsinki bieten während der Semester auch einen

Informationsdienst an, um den Forschern und Studenten bei der Anwendung des Servers für elektronische Korpora bei Bedarf persönlich behilflich zu sein.

## 6. Zum Schluß

An der Universität Helsinki wird schon seit etwa zehn Jahren an der Erarbeitung der elektronischen Korpora in verwandten Sprachen des Finnischen gearbeitet. Trotzdem ist diese Aufgabe bei den meisten Sprachen erst kürzlich in Angriff genommen worden. Das Ziel der Verarbeitung der elektronischen Korpora uralischer Sprachen ist nach wie vor die Sammlung des Materials aus allen uralischen Sprachen in elektronischer Form, die Analyse dieser Korpora und ihre Beschreibung für Forschungs- und Unterrichtszwecke. Weil die Korpora von Experten verschiedener Sprachen zusammengetragen und erarbeitet wurden, liegt das Hauptgewicht der Materialanalyse auf der Beschreibung der Sprache. Die nächste Phase in der Korpusarbeit ist die Analyse der Textstruktur des vorhandenen Materials.

In der Sprachforschung sind elektronische Korpora und umfangreichere Dateibanken für verschiedene Sprachen außerordentlich wichtig. Die elektronischen Korpora bilden ein elektronisches Archiv für das multilinguale Material, das bei der Untersuchung der einzelnen Sprachen helfen und die Arbeit erleichtern kann. Unersetzbar wichtig sind sie für die Sprachtypologie, für die vergleichbare Materialien für kontrastive Analysen unterschiedlicher Sprachen und Sprachfamilien notwendig sind. Die dritte, aber nicht die geringste Anwendungsfunktion der elektronischen Korpora ist ihre Benutzung als Unterrichtsmaterial.

Die Speicherung der elektronischen Korpora uralischer Sprachen im Server für elektronische Korpora der Universität Helsinki zusammen mit den Korpora anderer Sprachen erleichtert und unterstützt die einheitliche Erforschung verschiedener Sprachen. Der UHLCS bietet einen umfangreichen Bezugsrahmen, innerhalb dessen auch für die kleinen elektronischen Korpora die Möglichkeit besteht, in der immer stärker werdenden internationalen Konkurrenz bekannt und geschätzt zu werden, wenn es um die Richtung der Forschung und den Einsatz der finanziellen Mittel geht, die der Forschung zugeteilt werden.

## Quellenverzeichnis

- Hakulinen, Auli, Karlsson, Fred & Vilkuna, Maria (1980). *Suomen tekstilauseiden piirteitä: Kvantitatiivinen tutkimus. Publications 6*. University of Helsinki, Department of General Linguistics. Helsinki.
- Joki, Aulis J. (1984). *Maaailman kielet. Tietolipas 45*. Suomalaisen Kirjallisuuden Seura. Helsinki.
- Karlsson, Fred (1992). *Lexicography and Corpus Linguistics*. Opening Address at 5th Congress of EURALEX, Tampere, August 4, 1992. Yliopistopaino. Helsinki.
- Karlsson, Fred (1994). *Yleinen kielitiede*. Yliopistopaino. Helsinki.
- Karlsson, Fred (1995). The formalism and environment of Constraint Grammar Parsing.

- In Karlsson, Fred, Voutilainen, Atro, Heikkilä, Juha & Anttila, Arto (Hrsg.): *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. 41-88. Mouton de Gruyter. Berlin.
- Koskenniemi, Kimmo (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Publications 11. University of Helsinki, Department of General Linguistics. Helsinki.
- Laakso, Johanna (Hrsg.) (1992). *Uralilaiset kansat. Tietoa Suomen Sukukielistä ja niiden puhujista*. Werner Söderström Osakeyhtiö. Porvoo.
- Nickel, Klaus Peter (1994). *Samisk grammatik*. Davvi Girji O.S. Berlings.
- Rantala, Leif (1994). Samerna på Kola-halvön. Deras situation idag. *Suomalais-Ugri-laisen Seuran Aikakauskirja* 85: 200-204. Suomalais-Ugrilainen Seura. Helsinki.
- Suihkonen, Pirkko (1990). *Korpustutkimus kielitynologiassa sovellettuna udmurttiin*. *Suomalais-Ugrilaisen Seuran Toimituksia* 207. Suomalais-Ugrilainen Seura. Helsinki.
- Suihkonen, Pirkko, Zagulyayeva, Bibinur & Tronina, Galina (1995). *Udmurt-English-Finnish Dictionary with a Basic Grammar of Udmurt*. *Lexica Societatis Fenno-Ugricae* XXIV. Suomalais-Ugrilainen Seura. Helsinki.
- Suihkonen, Pirkko (1998a). An Areal Typological Viewpoint to Languages and Language Groups of Northern and Central Eurasia. Manuskript.
- Suihkonen, Pirkko (1998b). *Documentation of the Computer Corpora of Uralic Languages at the University of Helsinki*. *Technical Reports*, No. TR-2. 30-31. Department of General Linguistics, University of Helsinki. Helsinki.
- Vestnik statistiki* 1990, 10: 69-75.
- The World Almanac and Book of Facts 1997*. *World Almanac Books*. In Imprint of K-III Reference Corporation. A K-III Communications Company. New York 1997.

**Die für den Artikel verwendeten elektronischen Korpora (Server für elektronische Korpora der Universität Helsinki - University of Helsinki Language Corpus Server)**

- Morphologisch kodiertes Korpus des Erzänischen. Kodiert von Jack Rueter. Helsinki 1994.
- Morphologisch kodiertes Korpus des Komi. Kodiert von Paula Kokkonen. Helsinki 1994.
- Morphologisch kodiertes Korpus des Sölkupischen. Kodiert von Jarmo Alatalo. Helsinki 1998.
- Statistisches Korpus des Udmurtischen = Suihkonen 1990.
- Textkorpus des Nordlappischen. *Komiteanmietintö 1985*: 66. Opetusministeriö. Helsinki.
- Textkorpus des Udmurtischen. Verarbeitet von Pirkko Suihkonen & Bibinur Zaguljaeva. Helsinki 1992.