

Multilinguale Ein- und Ausgabe am Beispiel der UEDb

István BÁTORI, Krisztián NÉMETH,
Holger PUTTKAMMER, Dorothea SCHÄFER

1. Einleitung

Der Vortrag beschäftigt sich mit den grundlegenden Problemen, die bei der Verarbeitung verschiedensprachigen Textmaterials auftreten. Exemplarisch werden die Schwierigkeiten behandelt, die beim Projekt "Computerlinguistische Erschließung des Uralisch-Etymologischen Wörterbuchs" (CLUE) bei der Ein- und Ausgabe des Sprachmaterials auftreten.

CLUE ist ein Kooperationsvorhaben zwischen der Uralischen Abteilung des Sprachwissenschaftlichen Instituts der Ungarischen Akademie der Wissenschaften und dem Institut für Computerlinguistik des Fachbereichs Informatik der Universität Koblenz-Landau, Abteilung Koblenz. Ziel des Projektes ist es, das von der Budapester Gruppe herausgegebene Uralisch-Etymologische Wörterbuch (UEWb; RÉDEI 1988-1991) als intelligentes Computerwerkzeug zur Verfügung zu stellen. Der Text des UEWbs wird dafür in einer Datenbank (UEDb) gespeichert. Mit Hilfe einer graphischen Oberfläche werden die Informationen des UEWbs besser zugänglich gemacht und aufbereitet. Das gezielte Zusammenstellen der Informationen erleichtert die Arbeit des Linguisten (BÁTORI et al. 1998a, BÁTORI et al. 1998b).

2. Multilinguales Sprachmaterial im UEWb

Bei einem etymologischen Wörterbuch handelt es sich um ein multilinguales Korpus mit den daraus entstehenden Problemen für die maschinelle Verarbeitung, wie sie in diesem Abschnitt ausgeführt werden. Das UEWb enthält eine komplette, strukturierte und übersichtlich geordnete Sammlung der uralischen Etymologien. Diese Etymologien sind in den 25 Hauptsprachen belegt und werden durch Dialektformen ergänzt, wenn dies erforderlich ist. Insgesamt werden im UEWb rund 170 uralische Sprachvarietäten referiert.

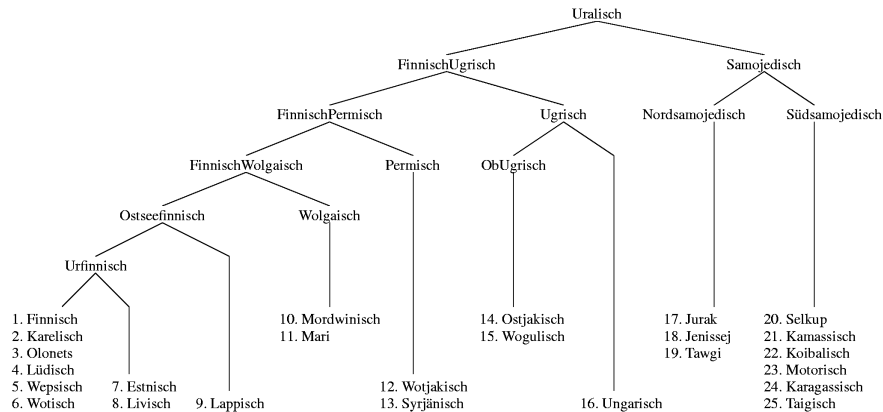


Abbildung 1: Die uralischen Sprachen

In dem gedruckten Werk sind zur Wiedergabe der Belege zwei Tendenzen zu beobachten: Zum einen die quellentreue Übernahme der Orthographie der Wörter, zum anderen eine Vereinheitlichung der Belegwörter. Da die konsequente Benutzung der einen Methode die andere ausschließt, gibt es im UEWb ein gemischtes Verfahren für die schriftliche Fixierung der Belege.

- Die Sprachen mit Schrifttradition (Estnisch, Finnisch, Norwegisch-Lappisch und Ungarisch) werden in der schriftsprachlichen Form zitiert.
- Alle anderen Sprachen werden phonetisch transkribiert. Das Alphabet wurde von den Autoren des UEWb festgelegt und folgt nicht dem Standard der International Phonetic Association (IPA). Die Herausgeber berufen sich hierbei auf POSTI und ITKONEN (1973). Dabei gibt es jedoch folgende Ausnahmen:
 - Die Belege aus sieben Wörterbüchern (Ahlqvist, Genetz, Grundström, Itkonen, Nielsen, Szilasi, und Munkácsi) werden mit drucktechnisch bedingten Vereinfachungen übernommen (s. RÉDEI 1988-1991, XX).
 - Samojedische Belege werden, wenn in der Quelle so geschrieben, in kyrillischer Schrift zitiert.

Das ergibt zwölf verschiedene Schreibweisen für die erfaßten 25 uralischen Sprachen, wenn man die undurchsichtige Lage der samojedischen Sprachen außer acht läßt. Dabei werden 464 Zeichen¹ benötigt. Durch Bereinigung des Belegmaterials kann sich diese Zahl noch leicht ändern. Für eine einheitliche Nutzung des Materials muß die unterschiedliche Verschriftlichung in der Eingabe und Anzeige überwunden werden.

Es stellt sich nun die Frage, wie mit dieser großen Anzahl an Zeichen gearbeitet werden kann, wie sie organisiert und dargestellt werden können, wenn die Plattformunabhängigkeit als Bedingung erfüllt werden soll.

3 Eingabe – virtuelle Tastaturen

In diesem Abschnitt wird die Frage geklärt, wie mit dieser Vielzahl verschiedener Zeichen gearbeitet werden kann. Wie muß demnach die Eingabe in das System gestaltet werden?

Da die große Anzahl an Zeichen nicht auf der Tastatur zu finden ist, benötigt man weiterführende Methoden zur Bewältigung dieses Problems. Mögliche Lösungen wären:

1. eine Spezialtastatur mit allen benötigten Zeichen;
2. eine sequentielle Eingabe über mehrere Tasten, wie beispielsweise “ä” für “ä” in HTML;
3. Tastenkombinationen wie “Alt-Gr” + “q” für “@”;
4. Eingabe per Mausclick auf eine virtuelle Tastatur am Bildschirm.

Die erste Lösung ist unpraktikabel und viel zu teuer. Für die nächsten beiden Methoden müßte der Benutzer die Zuordnungen für alle Zeichen lernen. Bei der erwähnten großen Anzahl an Zeichen ist das unzumutbar (NÉMETH 1999).

Deshalb wurde die vierte Lösung realisiert. Bei den virtuellen Tastaturen entfällt die Lernphase, die Zeichen werden auf dem Bildschirm

¹ Unter Zeichen verstehen wir Schriftzeichen, also Buchstaben mit und ohne diakritische Zeichen und andere Variationen sowie Interpunktionszeichen.

direkt angeboten. Die Arbeit ist ohne Vorleistung des Benutzers möglich. Von den Benutzern, die ständig mit dem System arbeiten, ist eine Mischung von normalen und virtuellen Tastatureingaben erwünscht, indem die oft benutzten Buchstaben per Tastatur und nur die selten benutzten per Mausklick eingegeben werden.

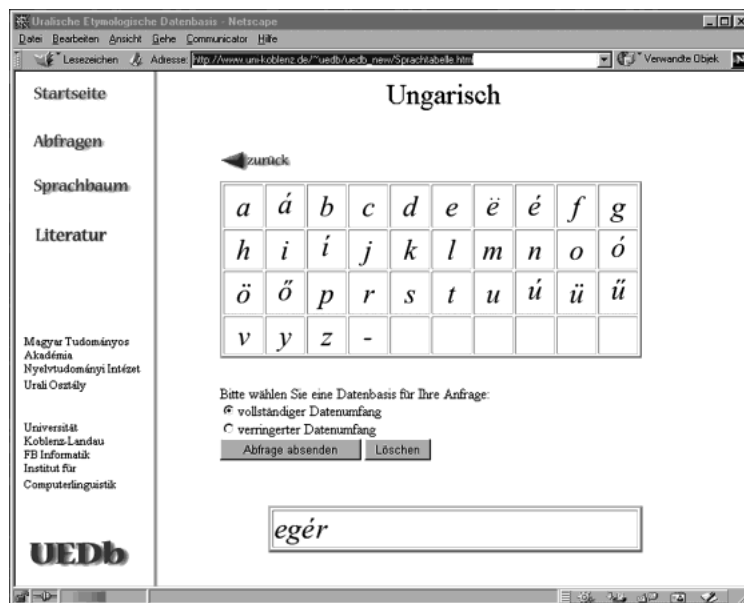


Abbildung 2: Screenshot einer virtuellen Tastatur

Für eine bessere Übersicht wird anstelle einer vollständigen Tastatur mit gleichzeitig allen Zeichen eine Reihe von einzelsprachlichen Tastaturen erstellt. Die Benutzung der sprachspezifischen Tastaturen ist weniger fehleranfällig, weil nicht alle Zeichen in allen Sprachen zulässig sind. Dabei werden auf den virtuellen Tasten analog zu normalen Tastaturen nur die einzelnen Zeichen und nicht die Alphabete dargestellt². Die einzelsprachlichen Tastaturen enthalten lediglich die in der jeweiligen Sprache

² Alphabete enthalten neben den Einzelbuchstaben auch Mehrfachbuchstaben wie ungarisch "cs", "sz", "ny" usw. oder Norwegisch-Lappisch "ggj", "llj" u.ä.

gebräuchlichen Zeichen. Da die Lautvorräte der Dialekte einer Sprache und die dafür benötigten Schriftzeichen teilweise unterschiedlich sein können, wird auch eine dialektabhängige Wahl der Tastatur angeboten. Die dialektabhängige Zugehörigkeit eines Zeichens wird zusätzlich farblich hervorgehoben.

Die virtuellen Tastaturen wurden mit Internet-fähigen Werkzeugen erstellt. Die Tastaturen wurden mit JavaScript implementiert. Damit die Mausclicks abgefragt werden können, sind alle Zeichen HTML-Links.

Abschließend läßt sich sagen, daß virtuelle Tastaturen den Umgang mit einem großen Zeichenvorrat ohne Einarbeitungsphase ermöglichen und sich leicht mit gängigen Werkzeugen erstellen lassen.

4. Ausgabe – Zeichencodierung und -sortierung

In dem folgenden Abschnitt wird ausgeführt, wie die verschiedenen Zeichen in Fontdateien verwaltet und auf dem Bildschirm dargestellt werden. Der letzte Unterabschnitt befaßt sich zusätzlich mit der Sortierung des multilingualen Materials.

Die Bereitstellung eines adäquaten Zeichenvorrats für die UEDb kann erst durch die Einführung und umfassende Nutzung des geplanten internationalen Standards Unicode gelöst werden. Bis zur Einführung dieses Systems wurden provisorische Lösungen gefunden, wobei die Kompatibilität zu Unicode berücksichtigt wurde.

4.1 Organisation der Zeichen

Für die Darstellung der Ergebnisse und der virtuellen Tastaturen auf dem Bildschirm werden spezielle Fonts benötigt. Die Zeichen werden dafür auf acht Fontdateien verteilt. Dabei werden nur die Positionen benutzt, die in keinem gängigen Zeichensatz (ISO8859, Windows, Macintosh, Macintosh Expert) für Steuerzeichen vorbehalten sind. Jeder Font enthält alle von einem Vokal oder mehreren Konsonanten abgeleitete Zeichen in alphabetischer Ordnung (z.B.: TimesNewRomanUral_a, TimesNewRomanUral_bh usw. ...).

Selbst im Unicode, der 65 536 Zeichen verwaltet und der auch schon IPA enthält, sind nicht alle Zeichen vorgesehen³, die in der UEDb benötigt werden. Die noch fehlenden, nicht zum UNICODE-Standard gehörenden Zeichen werden daher in einem nach UNICODE 2.1 kodierten Font in der sogenannten "private use area" abgelegt (Unicode 1996).

4.2 Darstellung der Zeichen

Der umfangreiche Zeichensatz des UEWbs muß auf verschiedenen Ausgabemedien dargestellt werden können. Die UEDb ist für eine Internetnutzung konzipiert und wird deshalb mit Hilfe eines Internet-Browsers bedient. Für einen solchen Rahmen ermöglicht die TrueDoc-Technologie eine plattformunabhängige Schriftanzeige.

Bei der Verwendung dieser Technologie müssen im ersten Schritt aus TrueType- oder Postscript-Fonts die Dynamic-Fonts, die sogenannten Portable Font Resources (kurz PFR) mit geeigneten Werkzeugen (z.B. Typograph TM 2.0 von Hexmac(c)) erstellt werden. Eine solche Fontdatei ist nur 10-15 kB groß. Um die Fonts vor unberechtigter Nutzung zu schützen, wird beim Erstellen der PFRs in die Datei die Adresse des WWW-Servers eingetragen; der Zeichensatz funktioniert nur dann, wenn die Datei von diesem Server geladen wird. Auf der Client-Seite werden dann mit Hilfe des Character Shape Players (standardmäßiger Bestandteil der Browser von Netscape und Microsoft seit den jeweiligen Versionen 4.0) aus den PFRs wieder die Rasterbilder der Schriftzeichen erstellt. Die Information, welcher Font benutzt wird, ist in den HTML-Code eingebettet.

4.3 Interne Codierung der Zeichen

Letztendlich müssen die Zeichen systemintern unabhängig von den verwandten Werkzeugen zur Repräsentation und speichereffizient abgelegt

³ Abgesehen von funktionalen Abweichungen deckt IPA nicht alle in dem UEWb benutzten Zeichen ab; so fehlen die unterbestimmten Vokale ɜ, ɘ, ɚ der Rekonstruktionen und die Halbvokale ɛ, ɜ, ɞ sowie möglicherweise auch noch andere in der Uralistik gebräuchliche Zeichen.

werden. Die folgende Abbildung veranschaulicht die verschiedenartigen Kodierungen der einzelnen Zeichen.

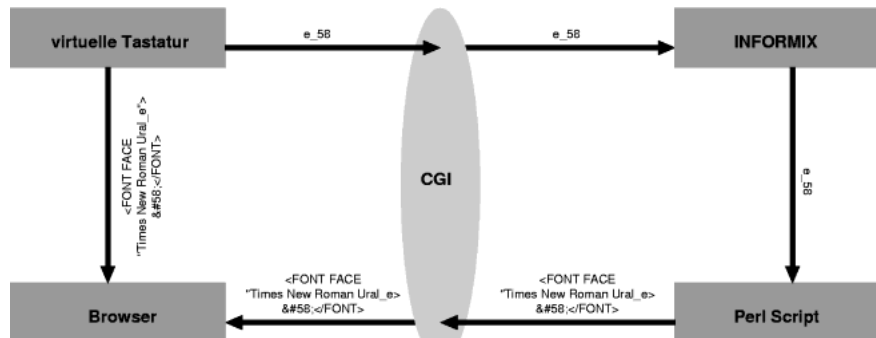


Abbildung 3: Codekonversion in der UEDb

Ein Java-Script erzeugt den HTML-Code für die Darstellung der Zeichen auf den virtuellen Tasten im Internet-Browser. Die Eingaben über die virtuellen Tastaturen im Browser werden der Datenbank in Form eines numerischen Codes übergeben. Dieser numerische Code ist angelehnt an die Organisation der Fontdateien, ist jedoch trotz der Verteilung auf mehrere Fonts durchgängig. Die Zeichencodes bestehen aus zwei Feldern. Das erste steht immer für den erwünschten Font, das zweite Feld für den ASCII-Code des benötigten Zeichens. Z.B. "e 45" für "ë". Diese zusätzliche Codierung dient der Unabhängigkeit der Datenbank von zukünftigen Anpassungen der Zeichencodierung wie zum Beispiel an Unicode. In der anderen Richtung übernimmt ein PERL-Script die Übersetzung der Ausgabe der Datenbank in den HTML-Code für die Anzeige im Browser.

4.4 Zeichensortierung

Eine einheitliche alphabetische Sortierung der Abfrageergebnisse ist aufgrund des vielsprachigen Materials unmöglich. Da die Abfrageergebnisse nach einheitlichen Kriterien sortiert werden müssen, wurde dafür ein mehrstufiges Sortierverfahren ausgearbeitet: Ein ähnliches mehrstufiges Verfahren ist auch für UNICODE Version 2.1 vorgesehen (Unicode 1996).

- Die Sprachen und Dialekte erhalten eine kanonische Ordnung. Die Belege der verschiedenen Sprachen und Dialekte werden immer

nach dieser Ordnung aufgelistet. D.h. an der ersten Stelle stehen immer die finnischen, an der zweiten die karelischen Belege usw. analog zur Numerierung im Sprachbaum in Abbildung 1.

- Innerhalb von Sprachen mit Schrifttradition wie Estnisch, Finnisch, Norwegisch-Lappisch und Ungarisch wird nach dem jeweiligen Alphabet sortiert. Grundzeichen mit Diakritika sind entweder Bestandteil des Alphabets oder kommen in der jeweiligen Rechtschreibung nicht vor. Im letzteren Fall werden die Nebenzeichen bei der Sortierung ignoriert. Beispielsweise wird “á, í, ő” zur Kennzeichnung der Verlängerung eines Vokals benutzt. Diese Zeichen sind jedoch nicht Bestandteil des ungarischen Alphabets.
- Für alle anderen Sprachen wird eine Ordnung aller vorkommenden Zeichen aufgestellt, das heißt alle benutzten Alphabete werden vereinigt. Um die große Zeichenmenge sortieren zu können, wird ein zweistufiges Verfahren verwendet. Es werden zwei Ordnungen erstellt: eine Ordnung für die Grundzeichen und eine für die Diakritika sowie andere Variationsmöglichkeiten. In der Ordnung aller Zeichen erscheinen dann zunächst Einzelzeichen, gefolgt von den Zeichen, die sich durch Kombination mit einem (weiteren) Diakritikum aus dem ersteren Zeichen ergeben, und entsprechend weiter in der Reihenfolge der Diakritika, vorausgesetzt, das jeweilige kombinierte Zeichen ist nicht schon vorher in der Sortierreihenfolge erschienen. So ergibt sich beispielsweise aus “á, í, ő” und “a b c ...” die Reihenfolge “a á á̂ á̃ ... á á̄ ... à ... b ...”.

Die sprachspezifische Sortierung gewährleistet, daß Muttersprachler die gewohnte Sortierung vorfinden. Bei den Sprachen mit Schrifttradition führt dies aber zu einer abweichenden Sortierung, da sie nach ihrem Schriftbild und die transkribierten Sprachen nach ihrem Lautbild sortiert werden.

5. Fazit

In dem Vortrag werden die Probleme bei der Arbeit mit mehrsprachigen Daten erläutert, wie sie auch in anderen multilingualen Informationssystemen vorkommen. Die Lösungen für das CLUE-Projekt werden aufge-

zeigt und vorgeführt. Die UEDb spielt dabei wegen ihres überschaubaren Sprachmaterials im Gegensatz zu anderen Sprachfamilien wie den indogermanischen Sprachen eine Vorreiterrolle.

Literatur

- BÁTORI et al. (1998a): I. B., K. NÉMETH und H. PUTTKAMMER, Lautrepräsentation in etymologischen Wörterbüchern anhand der Uralischen Etymologischen Datenbasis. In: B. SCHRÖDER, W. LENDERS, W. HESS und Th. PORTELE (Hrsg.), Computer, Linguistik und Phonetik zwischen Sprache und Sprechen: Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache, Frankfurt/M u.a.: Peter Lang Verlag. <http://www.uni-koblenz.de/öbatori/archiv/lare5.html>.
- BÁTORI et al. (1998b): I. B., K. NÉMETH und T. VÁRADI, Computerwerkzeuge für die Sprachforschung: Aufbau und Funktionen einer etymologischen Datenbasis für den uralischen etymologischen Wortbestand. In: *Studia Iranica, Mesopotamica et Anatolica* 3, 3-14 (Vortrag auf der 6. Internationalen Konferenz "Use of Computers in Historical and Comparative Linguistics", Frankfurt a.M., 21.-24. Oktober 1997).
- NÉMETH, K. (1999): Die virtuellen Tastaturen der Uralischen Etymologischen Datenbasis. Studienarbeit, Universität Koblenz-Landau, Fachbereich Informatik, Koblenz.
- POSTI, L. und ITKONEN, T. (Hrsg., 1973): FU-Transkription yksinkertaistaminen, Volume 7. Castrenianumin toimitteita, Helsinki.
- RÉDEI, K. (Hrsg., 1988-1991): Uralisches Etymologisches Wörterbuch I-III. Budapest und Wiesbaden: Akadémiai Kiadó / Otto Harrassowitz.
- Unicode (1996): The Unicode Consortium. The Unicode Standard, Version 2.0. Reading, Mass.: Addison-Wesley Developers Press.