

Multilingual text retrieval: Requirements and solutions

Jost GIPPERT (Tbilisi, 19.9.1997 / Frankfurt, 23.10.1997)

1. Aims of computational text retrieval in general:

- 1.1. Linguistic: Texts as coherent sources to be analyzed as to informations about phonetic, morphological, syntactical and lexico-semantic features of the language(s) involved
- 1.2. Literary: Texts as coherent structures representing ideas, attitudes, etc. to be analyzed with respect to the relationship between their contents and the shape they are presented in; closely connected (or intersecting) with linguistic approaches (discourse structure, metrics);
- 1.3. Historical (in the widest sense, including political, sociological, juridical approaches etc.): Texts as coherent sources to be searched for data about certain periods of time, represented, e.g. by names of persons and places or other key words;
- 1.4. Typical results to be produced:

- 1.4.1. Statistical analyses (phonetic, lexical, metrical); example: Phonetic statistics of the Old Georgian Šatberdi codex (Xth century), cf. Table 1
- 1.4.2. Grammatical analyses (phonetic, morphological, syntactical)
- 1.4.3. Word indices / concordances ("Key-Word-In-Context"), partial or total; example: KWIC-index of names beginning with *O-* from the Šatberdi codex (cf. Table 2)

('ა')	132098	('ნ')	34800	('ლ')	5556
('ბ')	14505	('ო')	8436	('ყ')	9316
('გ')	15896	('ო')	27635	('შ')	8105
('დ')	36053	('პ')	2131	('ჩ')	2517
('ე')	62739	('ც')	700	('ც')	9082
('ვ')	18720	('რ')	39813	('ძ')	3617
('ზ')	3239	('ს')	58289	('წ')	7916
('ც')	1443	('ტ')	5861	('ჭ')	853
('თ')	35149	('უ')	26067	('ხ')	7843
('ი')	74355	('კ')	5603	('ჯ')	2168
('კ')	6615	('ყ')	4145	('ჯ')	1122
('ლ')	29901	('ქ')	6113	('ჭ')	1542
('მ')	47023	('ჭ')	74		

Table 1: Phonetic statistics of the Šatberdi codex (Xth cent.)

2. Conditions and requirements of computational text retrieval:

- 2.1. Preparation of electronic texts in a structured way:
 - 2.1.1. Unique encoding of linguistic elements (letters, but also words) corresponding to the script system to be represented
 - 2.1.2. Unique encoding of structural elements of texts (e.g., chapters, paragraphs, pages, lines, strophes, verses, sentences, phrases, speakers)
- 2.2. Problems that must be solvable for retrieval software:
 - 2.2.1. Displaying, printing and keyboard handling of script(s) involved
 - 2.2.2. Sorting of data according to the alphabetic order involved (or another meaningful order)
 - 2.2.3. Morphological tagging as a basis for morphological and syntactical analysis (starting with the distinction of proper names and other nouns; this is a special problem of Georgian when written in Mxedruli script, cf. Table 2, but also a problem of languages that use capital letters both for proper names and for sentence-initial words)

- 2.3. The "lemma dilemma":
 - 2.3.1. Word forms not matching in form have to be treated as one word: the problem of prefixes, suffixes, infixes, ablaut, suppletivism etc.; cp. German *schreiben* vs. *schrieb* vs. *geschrieben*; English (*to*) *be* vs. (*I*) *am* vs. (*we*) *are* etc.
 - 2.3.2. Word forms identical in form but different in meaning (total or partial homonyms) have to be treated separately; cp. Russian *мыка* "pain" (*múka!*) vs. *мыка* "flour" (*muká!*) or Georgian *šen* Pers.Pron. "you" vs. *šeni* Poss.Pron. "your" vs. *ašenebs* "(he) constructs"

268:22	– #იორამ, #იორამს – #ოზია, #ოზიას – #იოათამ, #იოათამს
198:2	წელ: #აზარია, რომელსა #ოზია-ცა ერქუა, რომელი განკეთორა – ნბ- #ოზიას (1)
268:22	#იორამ, #იორამს – #ოზია, #ოზიას – #იოათამ, #იოათამს – #აკჰს, აქს #ოზიასა (1)
338:28	მამის-მამისა ჩემისა #ოზიასა, ვითარმედ: "იგივე #ოზიელ (1)
197:40	#ელისე, #აბდია, #იოველ, #ოზიელ, #ელაზარ, #აზარია; #იორამ – ც- #ოლდად (1)
198:14	#იერემია, #სოფონიას, #ოლდად, #ზარუქ; #იოაქს – გ- თთუე. #ომეროს (1)
196:18	ასოდ. ამისთვის-ცა მამისევისა #ომეროს რიცხვ იგი კბ-თა მათ #ონორი (1)
201:38	– იე~ წელ, #არკადი – კვ~ წელ; #ონორი – იე~ წელ; #თევდოსი მცირც #ოსე (2)
198:3	წინაღწარმეტყულებდეს #ოსე, #ამოს, #ესაია, #იონას; #იოათამ
198:6	მოუვდა #ისრაელსა და წარტყუნა #ოსე მეფეც #ისრაელისადა ათსა #ოსცსითგან (1)
198:25	ყოველნი წელნი ტყუნვითგან #ოსცსითგან მეფისა #ისრაელისადა და #ოსია (1)
198:13	– ნე~ წელ; #ამოს – ბ~ წელ; #ოსია – ლა~ წელ, #ოსოტერ (1)
200:19	– კვ~ წელ; #პტოლემეოს #ოსოტერ – კზ~ წელ; #პტოლემეოს #ოქოზია (1)
197:40	#აზარია; #იორამ – ც~ წელ; #ოქოზია – ა~ წელ; #გოლოლია, #ომრავე (1)
320:5	ქლაქი, #კასპი, #ურბნისი და #ომრავე, და ციხენი მათნი: ციხც #ომრავისადა (1)
320:6	ციხც #კასპისა, #ურბნისისა და #ომრავისადა. დაუკვრდა #ალექსანდრეს

Table 2:

KWIC-index of names beginning with *O-* from Šatberdi-codex

- 2.4. Sequential searching vs. preindexation (among others, a question of size limits)

3. Evaluation of an integrated software solution based on preindexation: Wordcruncher (Brigham Young University / Johnston and Co.): DOS versions since 1985 (last release: 4.6), Windows version since 1996 (latest release: 5.2β; cf. <http://www.wordcruncher.com>); following examples taken from a text database that has been established within the TITUS project (cf. <http://titus.uni-frankfurt.de/texte/texte.htm>) since 1986

3.1. Preparation of texts for DOS version ("BYB"-text, Table 3) and Windows version ("ETA"-text, Table 4): marking of text structure elements (according to 2.1.2) by |x where x is any letter (upto 3 levels in WCDOS, upto 10 levels in WCWin); additional marking of text formatting elements in WCWIN by ▶X(name)◀ where ▶◀ is a pair of letters to be defined and X is a certain letter; definitions collected and described in "SIF" file

```

|bGrig.Nis.Buneh.
|p67
|11 tkumuli emidisa da netarisa mamisa [|] čuenisa #grigoli
|12 #nosel ebiskoposisy kacisa šesakmisatws, romeli
|13 miučera zmasa twssa #petres ebiskopossa sabas̄tielsa*
|14 (1) ukuetu-mca žer-iqo didebay satnovebis-momgebeltay pařivita
|15 monagebtayta, ipovnes-mca qovelni sikadulni sapasetani undo, ražams-mca
|16 ševařqut satnovebata šenta tkumulisa misebr #solomon
|17 brznisa, rametu uzeštaes pařivisa mis sapasetaysa ars madli igi ģirsi,
|18 šenda. da ač macuves čuen dęsascauli. emidisa da didebulisa aġvsebisay
|19 čueulebisaebr p̄rvelisa močqalebasa megobrebay šeni da čuen
|110 čina-uřopt didebulebasa šensa. o kaco ģmrtisao, zġuensa kninsa da
|111 vitar-igi šenda ģirs-ars, xolo ara tu ukninessa zalisa čuenisa[sa]. da
|112 ese zġueni tkumul ars vitarca samoseli šeuracxi moksovili gonebisagan
|113 glaxakisa šromita. da mizeza amis tkumulisa vhgoneb tu
|114 mravalta aġonon, vitarmed silařqit ikma čueni ese dačqebay, romlisa-igi
|115 ara ģirs-viřvenit da tana-mdeb. xolo ara-ve šors ars tana-mdebebisagan
|116 samartlisa emiday řbasili - mamay da mozġuari qoveltay - igi
|117 xolo marřoy gamoikulevs dabadebulta šina ģmrtisata martlad
|118 dabadebuli mšgavsad ģmrtisa čęšmarřad, romlisay-igi sulī ars vitarca
|119 xaři dambadebelisa twsisay. gamoacxada mcnabay dabadebulta
|120 ģmrtisatay daparultay da qvna igini gamočinebul da ġulixmis-sařopel
|121 mattws, romelni eřiebden mas. xolo čuen davaklebt sařwrvelebata
|122 (2) misa mimart, romeli zeda guedebis didebulebatagan da
|123 sařinelebata mista.
  
```

Table 3: Preparation of text file for Wordcruncher (DOS)

```

SIF=|etc\tituscx.sif

▶Pcenter◀▶Title◀Textus homiletici et exegetici▶Tn16◀

▶Tsubtitle◀e Codice Satberdiensi▶Tn16◀
|ACod.Satb.

▶Tw122◀Gregorius Nyssenus, De hominis opificio▶Tn16◀

▶Pnormal◀
|bGrig.Nyss.Buneh.▶Tcegi16◀
|P67
|11 tkumuli emidisa da netarisa mamisa [|] čuenisa Grigoli
|12 Nosel ebiskoposisy kacisa šesakmisatws, romeli
|13 miučera zmasa twssa Petres ebiskopossa sabas̄tielsa*
|14 (1) ukuetu-mca žer-iqo didebay satnovebis-momgebeltay pařivita
|15 monagebtayta, ipovnes-mca qovelni sikadulni sapasetani undo, ražams-mca
|16 ševařqut satnovebata šenta tkumulisa misebr Solomon
|17 brznisa, rametu uzeštaes pařivisa mis sapasetaysa ars madli igi ģirsi,
|18 šenda. da ač macuves čuen dęsascauli. emidisa da didebulisa aġvsebisay
|19 čueulebisaebr p̄rvelisa močqalebasa megobrebay šeni da čuen
|110 čina-uřopt didebulebasa šensa. o kaco ģmrtisao, zġuensa kninsa da
|111 vitar-igi šenda ģirs-ars, xolo ara tu ukninessa zalisa čuenisa[sa]. da
|112 ese zġueni tkumul ars vitarca samoseli šeuracxi moksovili gonebisagan
|113 glaxakisa šromita. da mizeza amis tkumulisa vhgoneb tu
  
```

Table 4: Preparation of text file for Wordcruncher (Windows)

3.2. Display of text in DOS version (Fig. 1; note that correct display of special characters requires special VGA adaptation) and Windows version (Fig. 2 and, with text level names displayed and with different font requiring different markings in preparational text, Fig. 3):

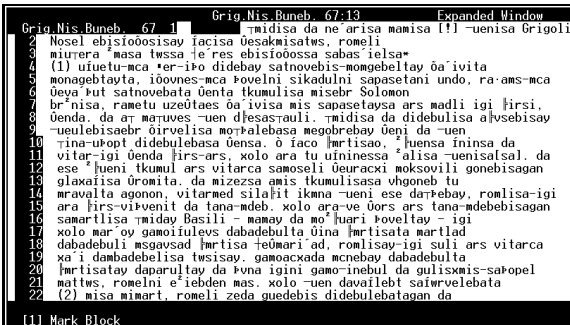


Fig. 1: Main text window in WC-DOS

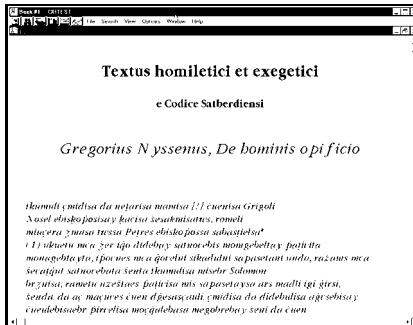


Fig. 2: Main text window in WC-Win

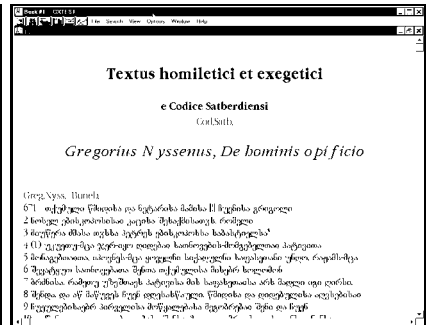


Fig. 3: Same, with mxedruli font

3.3. Preindexation presupposing definition of alphabetic character sequence, text levels (WCDOS: "BYC"-file; WCWIN: "LST"-file, cp. Fig. 4, and "ETX"-file, cp. Fig. 5), and formatting prescriptions (WCWIN only: "SIF"-file, cp. Table 5).

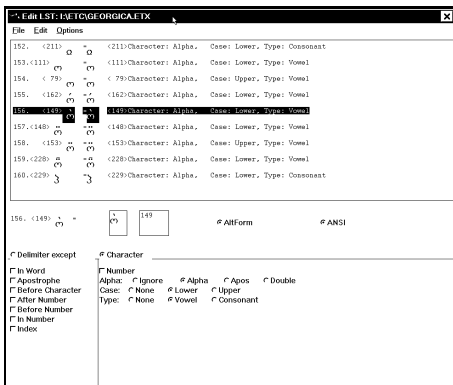


Fig. 4: Definition of character sequence

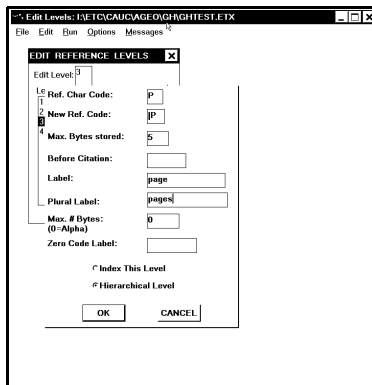


Fig. 5: Definition of text levels

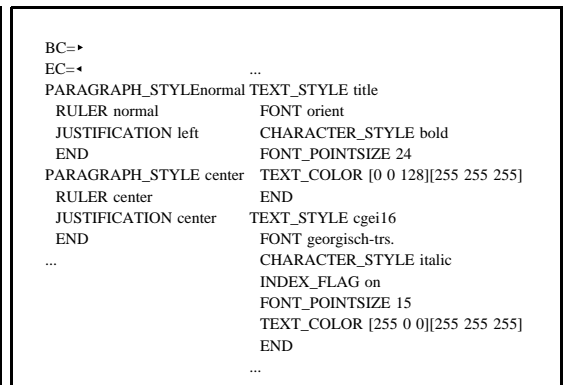


Table 5: Definition of formatting elements

3.4. Features of the search engine (searching by double-clicking upon a certain word or using word wheel):

- 3.4.1. Search for single word forms (cp. Fig. 6), context search (cp. Fig. 7), search with wild card letters (cp. Fig. 8 and Fig. 9) and for substrings (cp. Fig. 10 and Fig. 11), combined search of word forms belonging to one lemma (cp. Fig. 12), save list function (cp. Fig. 13)

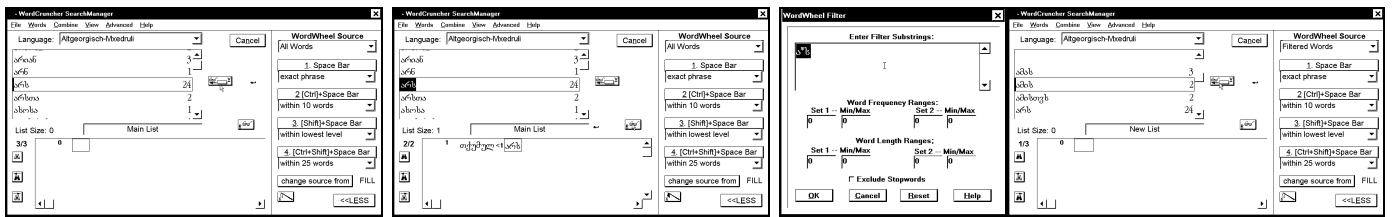


Fig. 6: Search of single word form

Fig. 7: Combined search

Fig. 8: "Wild cards"

Fig. 9: Filtered "word wheel"

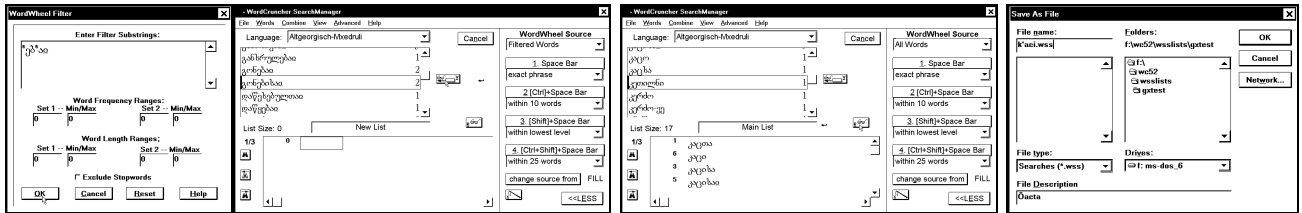


Fig. 10: Substring search

Fig. 11: Filtered word wheel

Fig. 12: Related word forms collected

Fig. 13: Save list function

3.4.2. Data entry presupposing keyboard definition (cp. Fig. 14) to be included as a "character map table" in the "SIF"-file (cp. Table 6):



Fig. 14: Keyboard definition as usable in search engine

CHARACTER_MAP Georgisch	
MAPS 61=61 63=63 96=96 65=192 83=234 68=68 70=70 71=204 72=208 74=74 75=214 76=217	
MAPS 80=229 220=246 42=42 89=89 88=88 67=194 86=86 66=66 78=78 77=77 59=59 58=58	
MAPS 95=95 94=196 35=35 124=197 39=39 97=97 115=115 100=100 102=102 103=103 104=104	
MAPS 99=99 118=118 98=98 110=110 109=109 44=44 46=46 45=45 123=92 91=123 93=125	
MAPS 106=106 107=107 108=108 246=148 228=132 113=113 119=119 101=101 114=114	
MAPS 116=116 122=122 117=117 105=105 111=111 112=112 252=129 43=43 121=121 120=120	
MAPS 125=182 92=184 64=231 126=93 181=219 124=91 131=159 225=160 237=161 243=162	
MAPS 228=132 224=133 229=134 231=135 234=136 235=137 232=138 239=139 238=140	
END	

Table 6: Character map table definition within "SIF"-file (extract)

3.4.3. Results given in reference list with sorting (cp. Fig. 15) and text-out function (cp. Fig. 16)

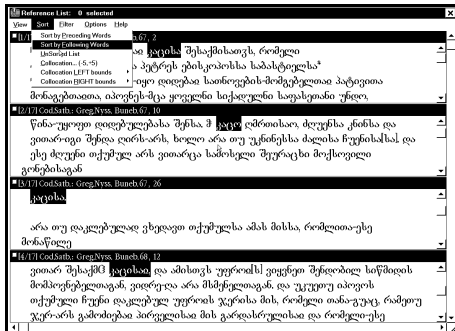


Fig. 15: Sorting of reference list

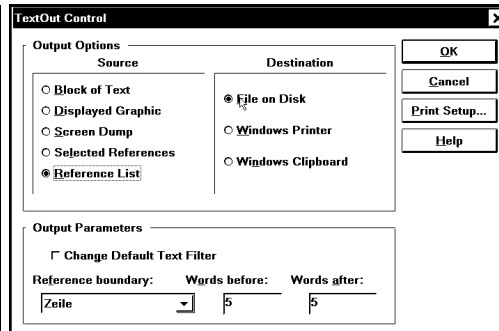


Fig. 16: Text-out function

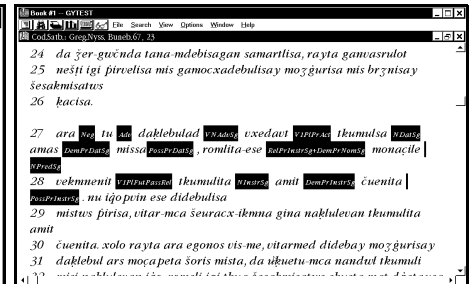


Fig. 17: Morphological tagging (provisional)

3.4.4. Other features, limitations and shortcomings:

- 3.4.4.1. Embedding of graphics (cp. Fig. 18: Old Georgian palimpsest from Mt. Sinai) and other text hyperlinks (cp. Fig. 20)
- 3.4.4.2. Virtually no size limits of texts to be preindexed and handled in Windows version (practically: 2 GB)
- 3.4.4.3. Number of unique word forms in a text limited to 16,777,216
- 3.4.4.4. No practicable solution existing yet for morphological tagging (cp. Fig. 17 for a provisional method)
- 3.4.4.5. No KWIC generation possible yet with Windows version (outstanding feature of DOS version; largest index produced in recent years: Index Galenicus (published: Verlag J.H.Röll, Dettelbach 1997) based on ca. 17 MB of Greek text (cf. <http://titus.uni-frankfurt.de/lexica/galeninx.htm>))

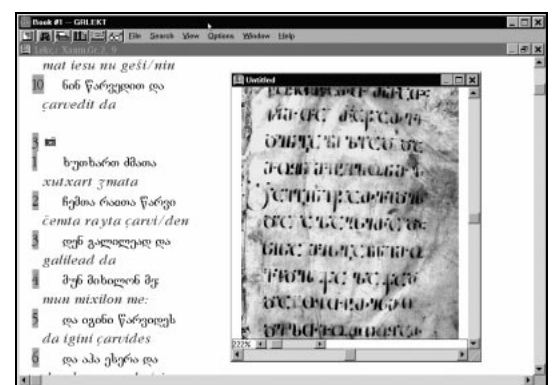


Fig. 18: Text with graphics hyperlink

4. Aims of multilingual text retrieval:

4.1. Case 1: Texts constituted by multilingual elements, containing, e.g. quotations from foreign languages within them (cp., e.g., Fig. 19 showing an Old Maldivian document containing Arabic words);

- 4.1.1. Special task: separation of elements belonging to each language
- 4.1.2. Special requirements:

- 4.1.2.1. Encoding of language boundaries within texts (delimiters)
- 4.1.2.2. Handling of different scripts, script directions etc.

4.2. Case 2: Texts in different languages that are interrelated with each other in a certain way, e.g. in that

- 4.2.0.1. one text is translated from the other (or both represent translations from a third one etc.; typical case: Bible translations), or
- 4.2.0.2. texts referring to the same contents (e.g., historical data), but independently from another (typical case of chronicles vs. eyewitness reports)

- 4.2.1. Special task: Establishing interdependencies between linguistic elements (e.g., names, translational word pairs, syntactical units)
- 4.2.2. Special requirements: Establishing interdependencies between text elements (e.g., chapter structure, sentence structure)
- 4.2.3. Unified text structure (e.g., text formatted in columns) vs. synchronized texts

5. Evaluation of solutions as offered by Wordcruncher for Windows; examples taken from Bible translation and other multilingual text environments

5.1. Marking up languages by text styles using transcriptions (cp. Fig. 20: the Maldivian example, comments added as hyperlinked notes) or original scripts including right-to-left directed ones (cp. Fig. 21: passage from the gospel of Matthew in Georgian, Greek, Armenian, and Syriac):



Fig. 19: Maldivian document containing Arabic words

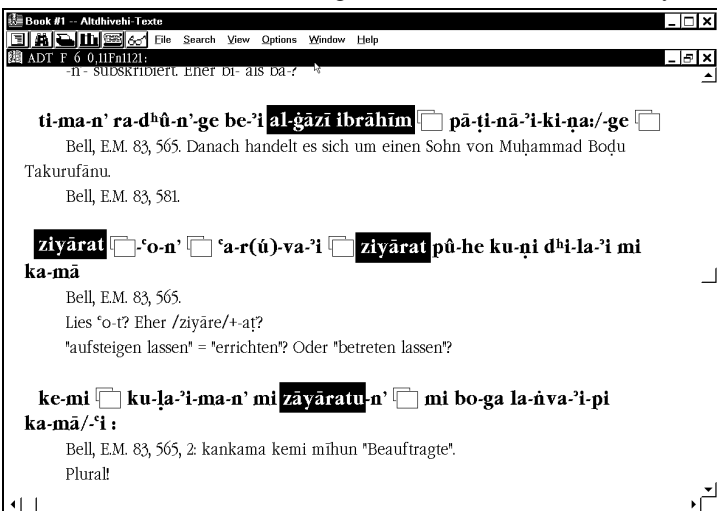


Fig. 20: Maldivian/Arabic mixed text (transcription)

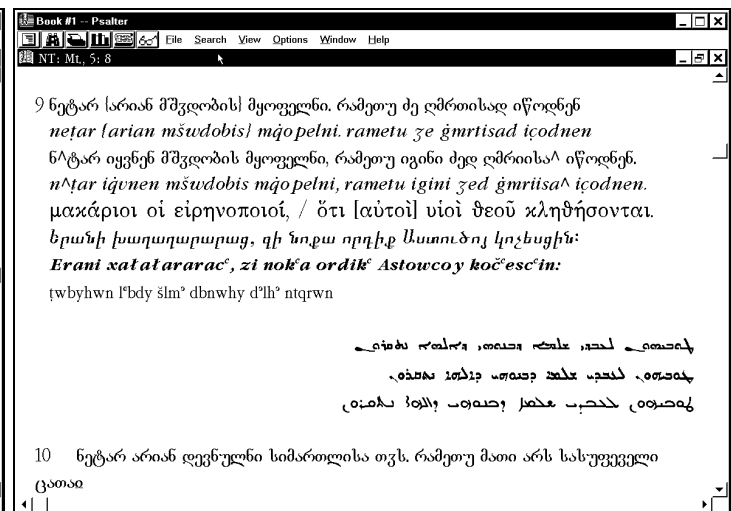


Fig. 21: Multilingual representation of Mt. 5,9

5.2. Search engine adaptable to every language contained, when searching by double-click (cp. Fig. 22) and when searching using the word-wheel, the latter including combined searches (cp. Fig. 24):

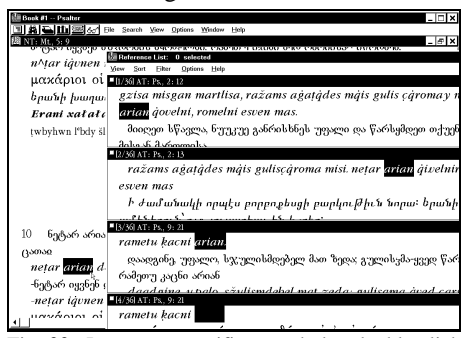


Fig. 22: Language-specific search by double-click upon key word

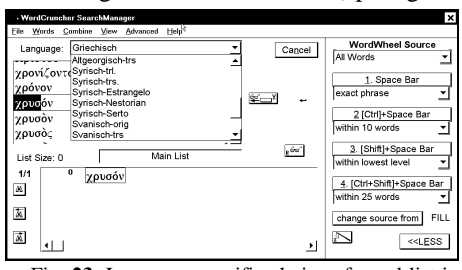


Fig. 23: Language specific choice of word list in word wheel

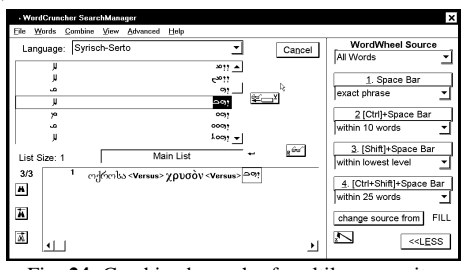


Fig. 24: Combined search of multilanguage items

- 5.3. Alternate solution for parallel texts: synchronized arrangement (cp. Fig. 25: several Bible versions)
- 5.4. Related problem: Synoptical arrangement of various sources in one language, treated as different "languages" (cp. Fig. 26: St. Nino's legend in five Old Georgian versions with an Armenian and a Latin parallel)

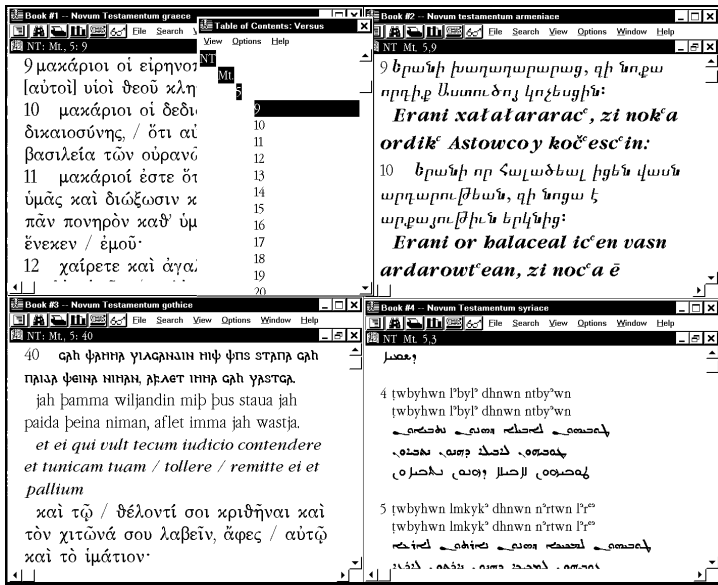


Fig. 25: Bible versions arranged synchronized

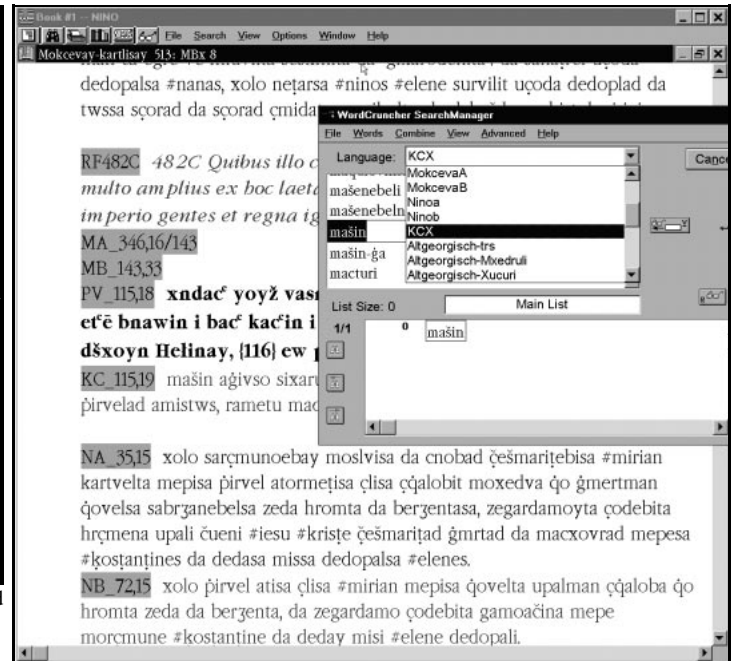


Fig. 26: St. Nino's legend in various sources

6. The encoding problem: Font mapping vs. unique encoding

6.1. The Wordcruncher solution:

- 6.1.1. Language definitions depending on text style definitions in the "SIF" file::
 - TEXT_STYLE definition (Table 7) referring to FONT declaration (Table 8),
 - FONT declaration (Table 8) referring to LANGUAGE definition (Table 9),
 - (LANGUAGE definition using TEXT_STYLE definition for items to be displayed in word wheel)

<pre> TEXT_STYLE mge12 FONT georgisch-mroveli INDEX_FLAG on FONT_POINTSIZE 12 TEXT_COLOR [0 128 0][255 255 255] END TEXT_STYLE mxge16 FONT georgisch-mxedruli INDEX_FLAG on FONT_POINTSIZE 16 TEXT_COLOR [128 0 0][255 255 255] END TEXT_STYLE mxge22 FONT georgisch-mxedruli INDEX_FLAG on FONT_POINTSIZE 24 TEXT_COLOR [128 0 0][255 255 255] END ... </pre>
--

Table 7: TEXT STYLE definitions >>>

<pre> FONT georgisch-mxedruli FONT_NAME TITUS-Mxedruli FONT_FAMILY roman CHAR_SET ansi PITCH proportional DIRECTION left-to-right FONT_TYPE TrueType LANGUAGE Altgeorgisch-Mxedruli END FONT georgisch-trs. FONT_NAME TITUS-Christlicher Orient FONT_FAMILY roman CHAR_SET ansi PITCH proportional DIRECTION left-to-right FONT_TYPE TrueType LANGUAGE Altgeorgisch-trs END ... </pre>
--

Table 8: FONT declarations >>>

<pre> LANGUAGE Altgeorgisch-Mxedruli LANGUAGE_ID GEORGIAN CHARACTER_MAP Georgisch TEXT_STYLE mxge16 LST_FILENAME \ETC\GEORGICA.ETX END LANGUAGE Altgeorgisch-Xucuri LANGUAGE_ID GEORGIAN CHARACTER_MAP Georgisch TEXT_STYLE axge16 LST_FILENAME \ETC\GEORGICA.ETX END LANGUAGE Altgeorgisch-trs LANGUAGE_ID GEORGIAN CHARACTER_MAP Georgisch TEXT_STYLE cgei16 LST_FILENAME \ETC\GEORGICA.ETX END ... </pre>
--

Table 9: LANGUAGE definitions

- 6.1.2. Disadvantage: No unique encoding for separate languages / scripts possible because of font mapping (plain 8-bit-encoding on Windows-ANSI-basis); cp. Table 10 and Table 11 showing text passages rendered in plain (DOS/) ASCII format (text markups represented in bold characters)

112	►Tiodd16◄ti-ma-n' ra-d#û-n'-ge be-&i ►Tvoar16◄al-Σ#zΩ ibr#hΩm►Tiodd16◄ ►HC{char}{ADT F 6 0 12Fn1122}◄ p#-i-n#-&i-ki-≡a:/-ge ►HC{char}{ADT F 6 0 12Fn1123}◄►Tn16◄
112Fn1122	►Tfn12◄Bell, E.M. 83, 565. Danach handelt es sich um einen Sohn von Murammad Bo u Takuruf#nu.►Tn16◄
112Fn1123	►Tfn12◄Bell, E.M. 83, 581.►Tn16◄
113	►Tiodd16◄►Tvoar16◄ziy#rat►Tiodd16◄ ►HC{char}{ADT F 6 0 13Fn1124}◄-%o-n' ►HC{char}{ADT F 6 0 13Fn1125}◄ %a-r(û)-va-&i ►HC{char}{ADT F 6 0 13Fn1126}◄ ►Tvoar16◄ziy#rat►Tiodd16◄ pû-he ku-≡i d#i-la-&i mi ka-m#►Tn16◄
113Fn1124	►Tfn12◄Bell, E.M. 83, 565.►Tn16◄
113Fn1125	►Tfn12◄Lies %o-t'? Eher /ziy#re/+-a≈'?►Tn16◄
113Fn1126	►Tfn12◄"aufsteigen lassen" = "errichten"? Oder "betreten lassen"?►Tn16◄
114	►Tiodd16◄ke-mi ►HC{char}{ADT F 6 0 14Fn1127}◄ ku-φa-&i-ma-n' mi ►Tvoar16◄z#y#ratu►Tiodd16◄-n' ►HC{char}{ADT F 6 0 14Fn1128}◄ mi bo-ga la-μva-&i-pi ka-m#/-%i :►Tn16◄
114Fn1127	►Tfn12◄Bell, E.M. 83, 565, 2: kankama kemi

Table 10: Maldivian text in DOS/ASCII representation

►Pnormal◄p9
►Tdge16◄ozia Ωva ioatam. ioatam Ωva akaz. akaz Ωva eze #ria.►Tn16◄
►Tcgei16◄ozia Ωva ioatam. ioatam Ωva akaz. akaz Ωva eze #ria.►Tn16◄
►Tmge16◄ozia Ωva ioatam; -ioatam Ωva akaz; aka^ Ωva eze #ria.►Tn16◄
►Tcgei16◄ozia Ωva ioatam; -ioatam Ωva akaz; aka^ Ωva eze #ria.►Tn16◄
►Tgr16◄äO#f#@ rè / #ΣΣΣ# Θ#Σ ΩδΣ äI#π, äI#π äπ rè #ΣΣΣ# Θ#Σ ΩδΣ äAφ# / äAφ# rè #ΣΣΣ# Θ#Σ ΩδΣ ÄE#αf#Σ.►Tn16◄
►Thy16◄Ozia cnaw zyova±am: Yova±am cnaw za#az: A#az cnaw zezekiay.►Tn16◄
►Thyti16◄Ozia cnaw zyova±am: Yova±am cnaw za#az: A#az cnaw zezekiay.►Tn16◄
►Tsk16◄%wzy& &wld lywtm ywtm &wld l&#z &#z &wld l#zqy&►Tn16◄
►Pright◄
►Tse16◄'8zy" ^wjd j08TM y8TM ^wjd j"x7 ^x7 ^wjd j+7q0"►Tn16◄
►Tsn16◄'8zy\$ ^wjd j08TM y8TM ^wjd j\$x7 ^x7 ^wjd j+7q0\$►Tn16◄
►Tss16◄'8zy" ^wjd j08TM y8TM ^wjd 1'7 ^7 ^wjd j+7q0"►Tn16◄

Table 11: Passage from Matthew in DOS/ASCII representation

6.2. Future prospect: The Unicode encoding solution

- 6.2.1. Present stage: many scripts encodable in UTF-8 transformation, all letters contained in **one** font; text displayable as web-page in HTML-format using Netscape Communicator 4.0 (cf. <http://titus.uni-frankfurt.de/unicode/unitest.htm>)

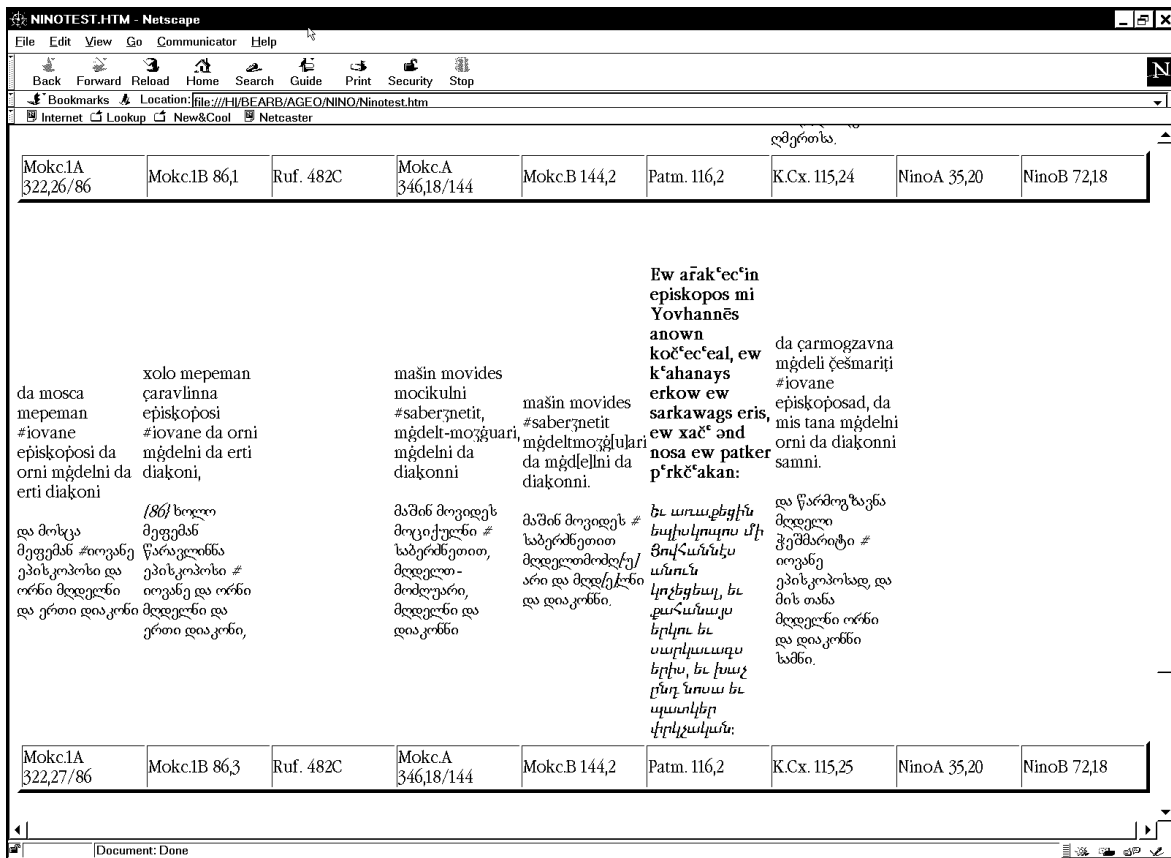


Fig. 27: Several sources of St. Nino's legend Unicode (UTF-8) encoding, displayed synoptically

- 6.2.2. Shortcomings: No retrieval (input/output) possible except for pure reading (for the time being)
- 6.2.3. Disadvantage: Encoding based on scripts, not languages; languages using the same writing system are not distinguishable formally, homographs remaining ambiguous; cp., e.g., German *hier* "here" and French *hier* "yesterday" or French *haut* "high" vs. German *haut* "strikes" (vs. German *Haut* "skin").