

Achtung!

Dies ist eine Internet-Sonderausgabe des Aufsatzes
„Segmentation and referencing in electronic text retrieval:
Problems and solutions“
von Jost Gippert (1999).

Sie sollte nicht zitiert werden. Zitate sind der Originalausgabe in
Proceedings [of the] 1999 EBTI, ECAU, SEER & PNC Joint Meeting,
Taipei: Academia Sinica 1999, 339-346
zu entnehmen.

Attention!

This is a special internet edition of the article
“Segmentation and referencing in electronic text retrieval:
Problems and solutions”
by Jost Gippert (1999).

It should not be quoted as such. For quotations, please refer to the original edition printed
in

Proceedings [of the] 1999 EBTI, ECAU, SEER & PNC Joint Meeting,
Taipei: Academia Sinica 1999, 339-346.

Alle Rechte vorbehalten / All rights reserved:

Jost Gippert, Frankfurt 1999-2011

Linking methods as a basis for cross-linguistic text retrieval: Problems and solutions

1. Text segmentation principles:
 - 1.1. Internal vs. external principles of segmentation of texts:
 - 1.1.1. Purely linguistic segmentation (sequence of sentences, words within sentences) vs.
 - 1.1.2. content-based segmentation (e.g., books, chapters, paragraphs, strophes, verses) vs.
 - 1.1.3. segmentation based on "surface" representation (pages, lines of a given edition).
 - 1.2. Superiority of content-based segmentation with respect to larger texts:
 - 1.2.1. More easy referencing by not using too large numbers (e.g., sentence no. 25387);
 - 1.2.2. does not depend on a given printed edition the arrangement of which is always more or less "accidental";
 - 1.2.3. is the only reasonable basis for a cross-textual retrieval (see below).
2. Problems of establishing a solid content-based segmentation for Buddhist texts:
 - 2.1. Example 1: The Pāli Vinayaṭṭaka
 - 2.1.1. Electronic version (public domain) produced by the "Sri Lanka Buddha Jayanti Tripitaka Series" (version 15.1.97) contains referencing to two printed editions, the BJT edition and the PTS edition, but for the latter, only book numbers and pages are indicated. There is no referencing to the content-based segmentation used traditionally (cp. Box 1: this would be Vin. I, 6, 7-10).
 - 2.1.2. Electronic version produced by Mahidol University Computing Center (Bangkok, Thailand, 1.8.1997: "BUDSIR IV") contains referencing to one printed edition, and for this too, only book numbers and pages are indicated. There is no referencing to the content-based segmentation used traditionally (cp. Fig. 1).

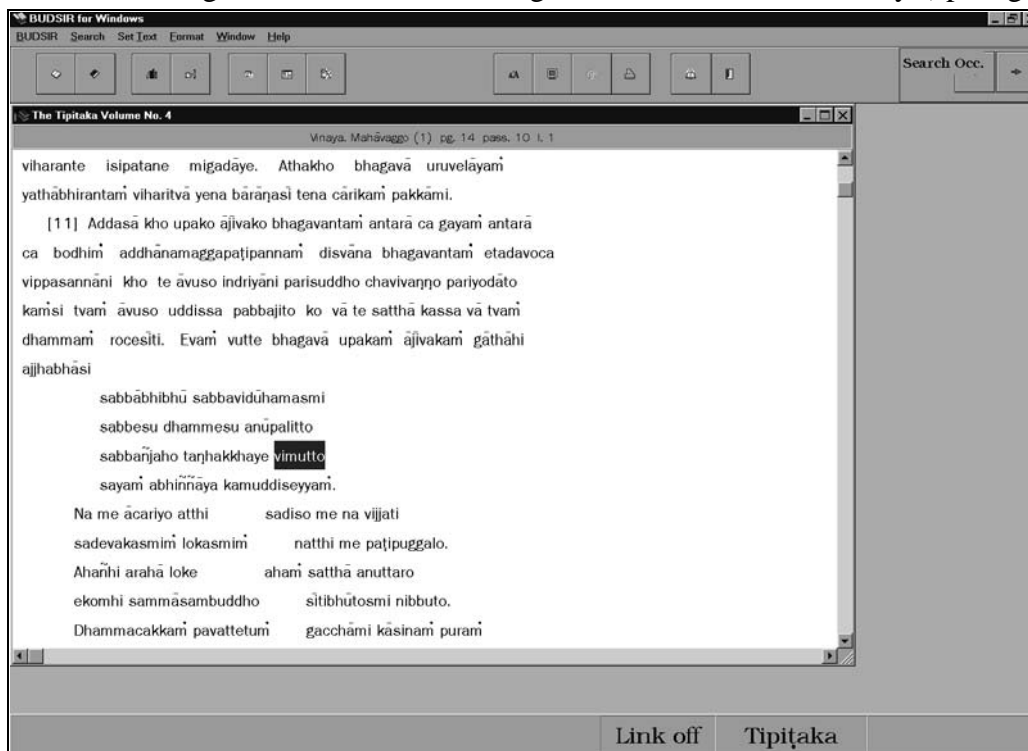


Fig. 1: Mahidol version of Vin. I, 6, 7-8 (screen output)

[BJT Page 016.]

4. Addasā kho upako ājivako1- bhagavantam
antarā ca gayam antarā ca bodhim annamagga
apaṭipannam. Disvāna bhagavantam etadavoca:
"vippasannāni kho te āvuso indriyāni. Parisuddho
chavivaṇṇo pariyodāno. Kam si tvam āvuso
uddissa pabbajito? Ko vā te satthā? Kassa vā
tvam dhammam rovesī?"Ti. Evam vutte bhagavā
apakam ājivakam gāthāhi ajjha bhāsi: -

"Sabbābhibhu sabbavidu'hamasmi
Sabbesu dhammesu anupalitto,
Sabbāñajabho taṇhakkhaye vimutto
Sayam abhiññāya kamuddiseyyam.
Na me ācariyo atthi sadiso me na vijjati,
Sadevakasmiṃ lokasmiṃ natthi me paṭipuggalo.
Aham hi arahā loke aham satthā anuttaro,
Eko'mhi sammāsambuddho sītibhutosmi nibbuto.

Dhammacakkam pavattetum gacchami kāsinaṃ
param,
Andhabhūtasmiṃ lokasmiṃ āhañacham2-
amatadundubhi"ti.

Yathā kho tvam āvuso paṭijānāsi, arahasi
anantajinoti.

"Mā disā ve jinā honti ye pattā āsavakkhaya
m, Jinā me pāpakā dhammā tasmā'ham upakā
jino"ti.

5. Evam vutte upako ājivako ājivako "huveyya-
pāvuso"ti3- vatvā sīsam okampetvā ummagga-
gam gahetvā pakkāmi.

6. Atha kho bhagavā anupubbena cārikam
caramāno yena bārāṇasī isipatanam migadāyo
, yena pañcamaggiyā bhikkhu, tenupasaṃkam
i. Addasaṃsu. Kho pañcavaggiyā bhikkhu
bhagavantam durato'va āgacchantam. Disvāna
a aññamaññam saṇṭhapesum: "ayam āvuso
samaṇo gotamo āgacchatī bāhuliko [PTS Page
009] paṭhānavibbhanto āvatto bāhullāya. So
neva abhivādetabbo. Na paccūṭātabbo. Tassa
pattacīvaram paṭiggahetabbam api ca kho
āsanam ṭhapetabbam, sace ākaṅkhissati, nisīd
issati"ti.

1. "Xjiviko ma. Nu. Pu;a. Ma vi; 2. "Ahañ-
chim āhañachum ityapi

3. "Huveyyāvuso" - katthavi

Box 1:

JT version of Vin. I, 6, 7-10 (raw text)

B

[BJT Page 016.]

|c4. |xVin_I_6,_7

|s1 Addasā k'o upako ājivako1- b'agavanta
m antarā ca gayam antarā ca bodhim
annamaggapaṭipannam. Disvāna b'agava
ntam etadavoca: "vippasannāni k'o te
āvuso indriyāni. Parisuddho c'havivaṇṇo
pariyodāno. Kam si tvam āvuso uddissa
pabbajito? Ko vā te satthā? Kassa vā
tvam d'hammam rovesī?"Ti. |xVin_I_6,_
8 Evam vutte b'agavā apakam ājivakam
gāthāhi ajj'ab'āsi: -

|s2 "Sabbāb'ib'u sabbavidu'hamasmi

|s3 Sabbesu d'hammesu anupalitto,

|s4 Sabbāñajab'o taṇhakk'haye vimutto

|s5 Sayam ab'iññāya kamuddiseyyam.

|s6 Na me ācariyo att'i sadiso me na vijjati,

|s7 Sadevakasmiṃ lokasmiṃ natt'i me paṭi-
puggalo.

|s8 Aham hi arahā loke aham satthā anuttar
o,

|s9 Eko'mhi sammāsambuddho sītib'utosmi
nibbuto.

|s10 Dhammacakkam pavattetum gacc'hami
kāsinaṃ param,

|s11 And'ab'ūtasmiṃ lokasmiṃ āhañac'h'am2-
amatadundub'i"ti.

|xVin_I_6,_9

|s12 Yat'ā k'o tvam āvuso paṭijānāsi, arahasi
anantajinoti.

|s13 "Mā disā ve jinā honti ye pattā āsavak-
k'ayam,

|s14 Jinā me pāpakā d'hammā tasmā'ham upa
kā jino"ti.

|c5. |s1 Evam vutte upako ājivako ājivako
"huveyyapāvuso"ti3- vatvā sīsam okam-
petvā ummaggaṃ gahetvā pakkāmi.

|xVin_I_6,_10 |c6.

|s1 At'a k'o b'agavā anupubbena cārikam
caramāno yena bārāṇasī isipatanam mi-
gadāyo, yena pañcamaggiyā b'ikk'h,
tenupasaṃkamī. Addasaṃsu. Kho pañca-
vaggiyā b'ikk'u b'agavantam durato'va
āgacc'antam. Disvāna aññamaññam saṇṭ
hapesum: "ayam āvuso samaṇo gotamo
āgacc'ati bāhuliko |p9 [PTS Page 009]
paṭhānavibb'anto āvatto bāhullāya. So
neva ab'ivādetabbo. Na paccūṭātabbo.
Tassa pattacīvaram paṭiggahetabbam api
ca k'o āsanam ṭhapetabbam, sace ākaṅk'h

Box 2:

ame, after entering of additional references

S

- 2.1.3. For easy cross-referencing, the greatest amount of information possible should be envisaged, at least with respect to the traditionally used PTS edition (cp. Box 2).
- 2.2. Example 2: The Udānavarga (Collection of Sanskrit strophes of verses, mostly corresponding to strophes of the Pāli Dhammapāda):
- 2.2.1. In spite of relatively simple structure (complete collection consisting of 33 Vargas with a maximum of 87 strophes), great divergencies exist between several editions as to the numbering of strophes (e.g.: Uv. 29,24 in the edition by F. Bernhard, Udānavarga, Göttingen 1965 corresponds to Uv. 29,34 in the edition by R. Pischel, Die Turfan-Recensionen des Dhammapada, Berlin 1908 [and 29,23 in the translation of the Tibetan version by W.W. Rockhill, London 1892]).
- 2.2.2. This may be due to the bad state of preservation of the text, verses and strophes missing in manuscripts, or to secondary additions (cf. Bernhard, o.c., 14).
- 2.2.3. Arrangement and segmentation of the text do not agree at all with the Pāli Dhammapada so that the segmentation of this text cannot be adopted as it is (all the more since for the Pāli Dhammapada itself, two divergent segmentations are used, viz. one counting only strophes (from 1 to 423), and one dividing the text into 26 vargas with a differing amount of strophes; e.g., Uv. 21 [Tathāgatavarga], 1 corresponds to DP 353 ≈ 24 [Taṇhavagga], 20).
- 2.2.4. For the sake of cross-referencing, a common segmentation should be envisaged.

3. Problems of establishing a system for cross referencing between related texts

3.1. Example 1: The Gāndhārī Dharmapada (Collection of Prakrit strophes of verses, mostly corresponding to strophes of the Pāli Dhammapāda and the Sanskrit Udānavarga; edition by J. Brough, London 1962):

3.1.1. Although most of the strophes contained in the text are exact equivalents of strophes as present in the Pāli Dhammapada and/or the Sanskrit Udānavarga, the arrangement is completely different again (e.g., GDP 1 [Brammaṇa], 1ab corresponds to DP 393ab ≈ 26 [Brāhmaṇavagga], 11ab, Uv. 33 [Brāhmaṇavarga], 8ab (Sanskrit version ed. Bernhard) / 33, 11ab (Tibetan version). Cp. Fig. 2 showing the concordance of DP and GDP as present in Brough's edition.

CONCORDANCE II							
PALI DHAMMAPADA							
i. <i>Yamaka</i>		37	137a	74			
1	201	38	137c	75			
2	202	39	137d		vi. <i>Paṇḍita</i>		
3		40	138b		76	231	
4		41	153		77	230	
5		42			78		
6		43			79	224	
7	217				80		
8	218	iv. <i>Puppha</i>			81	239	
9	192	44	301		82	225	
10	193	45	302		83	226	
11	213	46	300		84	324	
12	214	47	294		85		
13	219	48	(294)		86		
14	220	49	292		87		
15	205	50	271		88		
16	206	51	290		89		
17	203	52	291				
18	204	53	293		vii. <i>Arahanta</i>		
19	190	54	295		90		
20	191	55	296		91		
		56			92		
ii. <i>Appamāda</i>		57	297		93		
21	115	58	303		94		
22	116	59	304		95		
23	(128)				96		
24	112	v. <i>Bāla</i>			97		
25	111	60			98		
26	117	61			99		
27	129, 130	62					
28	119	63			viii. <i>Sahassa</i>		
29	118	64	233		100	306	
30	120	65	234		101	308	
31	74	66			102	309	
32	73	67			103	305	
		68			104		
iii. <i>Citta</i>		69	(283)		105		
33	136	70	313		106	310	
34	137b	71			107	319, 320	
35		72			108	321	
36	138a	73					

Fig. 2:

Concordance of DP and GDP

- 3.1.2. A thorough content-based segmentation for the Gāndhārī Dharmapada is not easy to establish for the same reasons as with the Udānavarga.
- 3.1.3. As the text can hardly be treated without permanent comparison of its Pāli and Sanskrit equivalents (it is the only larger Gāndhārī Prakrit text preserved at all), cross-referencing is especially important. Reliable segmentation of all three texts is presupposed.
- 3.2. Example 2: Tocharian (A) text relating the tale of the conversion of Upaka (Upage) by Buddha (obviously fragment from the Udānālamkāra, i.e., an Udānavarga commentary): THT 850 sq. / Toch. A (edition Sieg-Siegling), No. 217 sq.
- 3.2.1. "Internal" segmentation as used in the printed edition: lines of manuscript pages (e.g., 217a 5) vs. content-based segmentation suggesting itself from verse structure which is indicated in the manuscript by numbers. Problem: The fragmentary status does not enable us to establish higher units (chapter 21, corresponding to the "Tathāgatavarga" of the Udānavarga as indicated in F. Bernhard's edition [p. 278]?)

No. 217 = T III Š 79. 15	
Der Länge nach ziemlich vollständiges, teilweise aber stark beschädigtes Blatt. Nach Herstellung der Photographie ist noch der sehr zerstörte Rest der linken Seite gefunden worden. Vgl. Tafel 29.	
Vorderseite	
1 (nicht erhalten)	217a
2 kñ[ā]ññā [tā] --- [p·] sne y· · y· wāryāñc· [śś·] --- ptā	
3 ··· e skākā wārpāt, p(tā)ñkāt, ārkīsoṣṣis kranṭ, markampal, āksis(s)i -----	
	[p]ūk, knānmām tmaṣ, bram poñcām wāltsa -
4 --- sn· · · k(ā)ckeyo 3 paklyoṣas wrasañā, pūk kācke parsācā, pūk, knānmām - k· ñkāt,	
	parko parnont, māgat ṣiṃ ypeyaṃ 3 wārpā
5 - ks[i]ssī 3 krañcām markampal, māryu praṣṭam okñāṣ nāktas napenās sam, oñkraci 3 8	
	palskāt, pūk, knānmām ke maltw āksisam, lyāklyām kṣpā

- 3.2.2. Manifold necessity of cross-referencing both with poetic texts (Udānavarga, Dhammapada, Gāndhārī Dharmapada) and prose texts (Sanskrit Catuspariṣatsūtra, Lalitavistara, Mahāvastu; Pāli Vinayapiṭaka and many more); cp. E. Sieg / W. Siegling, Festschrift M. Winternitz, Leipzig 1933, 167 sqq. or, vice versa, F. Bernhard's edition of the Udānavarga (l.c.) or E. Waldschmidt's edition of the Catuspariṣatsūtra (Vol. I, Berlin 1952). E.g., 217b 6 sq. (≈ UA 21, 12) corresponds to Uv. 21,1, DP 353, CPS 10,5, MV III, 326, 5-8, Vin. I, 6,8 and others.

4. Cross-referencing and electronic retrieval

- 4.1. Task: Simple indication of parallels as in many printed editions (cp. Fig. 4 showing Brough's edition of the GDP with parallels from Uv. and DP indicated and Fig. 5 showing cross references indicated in Bernhard's edition of Uv.) should be overcome by immediate automatic access to parallel texts when treated electronically.

THE GĀNDHĀRĪ DHARMAPADA		121
15 na bramaṇasediṇa kiji bhodi yo na nisedhe maṇasa pri'āpi yado yado y-asa mano nivartadi tado tado samudim aha saca. (15) O. 16	390 t'na brāhmaṇass' etad akiñci seyyo yadā nisedho manaso piyehi yato yato hiṃsamano nivattati tato tato sammati-m-eva dukkhaṃ.	
	Uv. xxxiii. 87 ji ltaṃ mi sdug pa las yid ldog pa de hdra bram ze rams la cuñ zad med ji lta ji lta de yi yid ldog pa de lta de lta kun rdzob bden par zad.	
16 brahetva pavaṇi bramaṇaṃ sama'rya śramaṇo di vucadi parvahi'a atvaṇo mala tasa parva'ido di vucadi. (16) O. 17	388 bāhita-pāpo ti brāhmaṇo samacariyā samaṇo ti vuccati pabbājayaṃ attano malaṃ tasmā pabbajito ti vuccati. xi. 15 (T. 16)	
17 na aho bramaṇa bromi yoneka-matra-sabhamu bho-va'i namu so bhodi sayi bhodi sakijaṇo akijaṇa aṇadaṇa tam aho bro mmaṇa. ¹ (17) O. 18	396 na cāhaṃ brāhmaṇaṃ brūmi yonijaṃ matti-saṃbhavaṃ bho-vādi nāma so hoti sace ² hoti sakiñcano akiñcanaṃ anādaṇaṃ tam ahaṃ brūmi brāhmaṇaṃ. xxxiii. 18 (R. 17)	
18 niha'i dāpa bhudeṣu traseṣu thavaṣeṣu ca yo na hadi na ghadhedi tam aho bromi bramaṇa. (18) O. 19	405 nidhāya daṇḍaṃ bhūtesu tasesu thāvaṣeṣu ca yo na hanti na ghāṭeti tam ahaṃ brūmi brāhmaṇaṃ. xxxiii. 47 (R. 45)	
19 yo du drigha ci rasa ji aṇu-thulu suhaṣuhu loki aḍiṇa na adī'adi tam aho bromi bramaṇa. (19) O. 20	409 yo dha dighaṃ va rasaṃ vā aṇu-thūlaṃ subhāsubhaṃ loke adinnaṃ nādiyati tam ahaṃ brūmi brāhmaṇaṃ. xxxiii. 29 (R. 28)	
20 yo du kama prahatvaṇa aṇakare parivaya kama-bhoka-parikṣiṇa tam aho bromi bramaṇa. (20) O. 21	415 yo dha kāme pahatvāna anāgāro paribhaje kāma-bhava-parikkhiṇaṃ tam ahaṃ brūmi brāhmaṇaṃ. xxxiii. 46 (R. 44)	
21 vari puṣkara-patre va arage-r-iva sarṣava yo na lipadi kamehi tam ahu bromi bramaṇa. (21) O. 22	401 vāri pokkhara-patte va ārage-r-iva sāsapo yo na lippati kāmesu tam ahaṃ brūmi brāhmaṇaṃ. xxxiii. 35 (R. 34)	

i.e. bromi bramaṇa.

² Sn. 620, and Dh. A. ed. ev.

Fig. 4: Indication of cross references in GDP edition

XXI Tathāgatavarga	
sarvābhībhūḥ sarvavid eva cāsmi sarvāś ca dharmāḥ satatam na līptaḥ / sarvaṃjahaḥ sarvabhayād vimuktaḥ svayaṃ hy abhijñāya kam uddīṣeyam // 1 kam uddīṣeyam tv asamo hy atulyaḥ svayaṃ pravaktā hy adhigamya bodhim / tathāgato devamanuṣyaśāstā sarvajñatām prāpya balair upetaḥ // 2	
Mss.: 1 a) AA82 - DH37.t - H. 149. 272.t - P. St. 32.t b) AA82.t - AA83.t - P. St. 32.t c) AA83 - DZ41.r - P. St. 32 - P. H. Ms. tr d) AA83.tr - DH37.t - P. St. 32 - P. H. Ms. II Zähl) P. St. 32 - P. H. Ms. - In der Hs. DH37 ausgelassen. 2 a) DH37.t - P. St. 32.t - P. H. Ms. r b) AA83.t - P. St. 32.t - P. H. Ms. t c) AA83 - DZ41.t - DD24.r - P. St. 32 d) AA83.t - DH37.t - DZ41.t - DD24.t - P. St. 32 Zähl) AA83 - DH37 - P. St. 32	
Var. orthogr.: 1 a) DH37, H. 149. 272: sarvābhībhūḥ-sarvavi(d)ḥ. d) P. H. Ms. (CHAKR.): uddīṣeyam.	
Var. lect.: 1 a) AA82: cāsmiṃ. c) P. St. 32: sarvajahaḥ sarva[m]bhayād. - P. H. Ms. (CHAKR.): --[tṛṣṇākeśa](ye vimuk)ktaḥ ¹ . d) P. St. 32: uttiṣeyam. - P. H. Ms. (CHAKR.): svayaṃ-a(bhi)jñā(ya) ² .	
Parallelen: 1 Mv. III p. 326 l. 5-8. - CPS 10. 5. - a-c) cf. Mv. III p. 118a-c. Dh. 353. - MN I p. 171. - Vin. I: I. 6. 8a-d. - Kv. p. 289. - Th. A. p. 220. - a-c) Sn. 211; SN II p. 284 ³ . 2 CPS 10. 6 ⁴ .	
Anm.: ¹ Vgl. Mv. III p. 326 l. 7: sarvajñaḥ 'haṃ tṛṣṇākeṣe vimukto; Dh. 353c: sabbañ-jaho taṇhakkhaye vimutto. ² Vgl. Mv. III p. 326 l. 8: ahaṃ abhijñāya; Dh. 353d: sayam abhijñāya. ³ Vgl. UA XXI. 12c-d; Toch. A 217b 6-7: pāk. ly[utā]r [nāntu] (pāk kṇānmāṃ ⁵ śkaṇ) nasam. pākya sne loplune uramamny[o] (t) (pāk lek tgrkorag. tsg)l(po) p(r)a[s[k]i[n](t)w)āṣ [ma]t(a)k nā yn(e)ś [karsa](rag.) āṇ ggrpñim. ⁴ Vgl. zu a) UA XXI. 13a; Toch. A 217a7: (āṇ) ggr[p](ā)j(m)j . ⁵ Oder: kgrso.	

Fig. 5: Indication of cross references in Uv. edition

4.2. Simple solution: editing of related text passages side by side within one file (cp. Fig. 6 showing GDP 1, 2 with its equivalent, DP 394, arranged interlinearly)

4.2.1. Shortcoming: Loss of readability

4.2.2. Problem not to be covered easily by retrieval software: Separate treatment of several languages within one text (here: Gāndhārī Prakrit vs. Pāli)

Book #1 - Dharmapada (Gandhari-Prakrit)	
DP 394 DhP 1, 2	
2	
ki di jaḍa'i drumed ^a a	
{kim te jaḍa'i dummē ^a a}	
ki di ayaṇa-śaḍi'a	
{kim te ayaṇa-śaḍi'a}	
adara gabaṇa kitva	
{abb ^a antaram te gabaṇaṃ}	
babire parimajasi. (2)	
{bābiraṃ parimajjasi.}	

Fig. 6: GDP 1,2 contrasted with DP 394

4.3. Sophisticated solution: "Synchronizing of texts" (Wordcruncher solution; cf. <http://www.wordcruncher.com>)

4.3.1. Requirement: Common segmentation of the texts to be synchronized

4.3.1.1. This is easy in, e.g., Bible tradition where the same segmentation of texts has been used traditionally (e.g. Mt. 6,9 [cp. Fig. 7 showing synchronous arrangement of the Armenian and Greek New Testament] or 2.Chr. 13,12) with but a few exceptions (esp. in OT: Jer.)

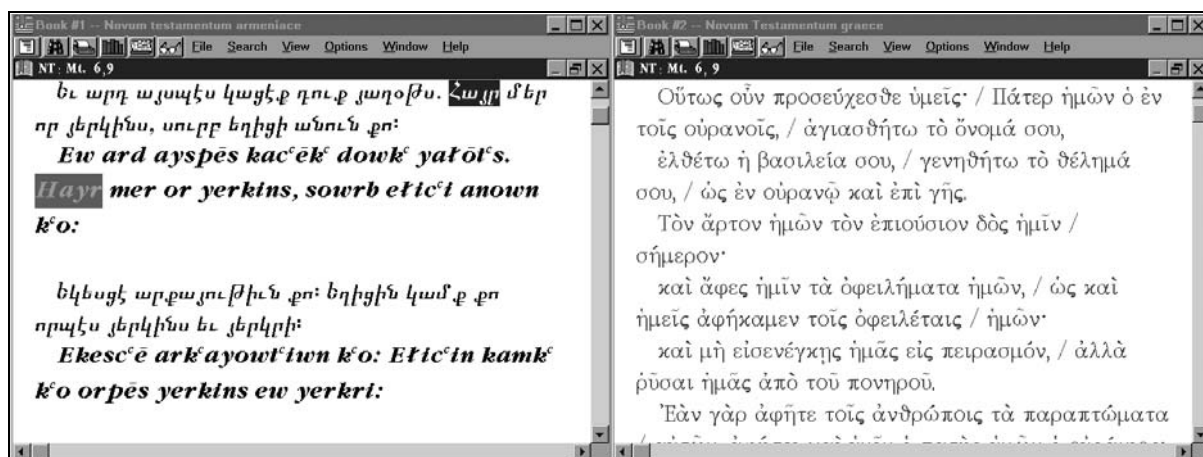


Fig. 7:

Mt. 6, 9 in Armenian and Greek

- 4.3.1.2. Synchronizing is not easy where no such common segmentation exists as in the case of the Dhammapāda equivalents, not to mention the prose texts.
- 4.3.2. Actual solution: Inserting common references as highest order segmentation units in the texts. Cp. Fig. 8 showing GDP 1,2 synchronized with DP 394 and Uv. 33,6; Fig. 9 showing Toch. A 217b 6 [21,12] synchronized [in two arrangements] with DP 353 and Uv. 21,1; Fig. 10 showing GDP 1,6 synchronized with the Pāli prose text Saṃyuttanikāya I, 167; Fig. 11 showing Toch. A 217b 5 [21,12] synchronized [in two arrangements] with CPS 10,2.
- 4.4. Task for the future: Extending the reference system to further texts of the Pāli Canon as well as to other branches of Buddhist tradition (Chinese, Tibetan, etc.).

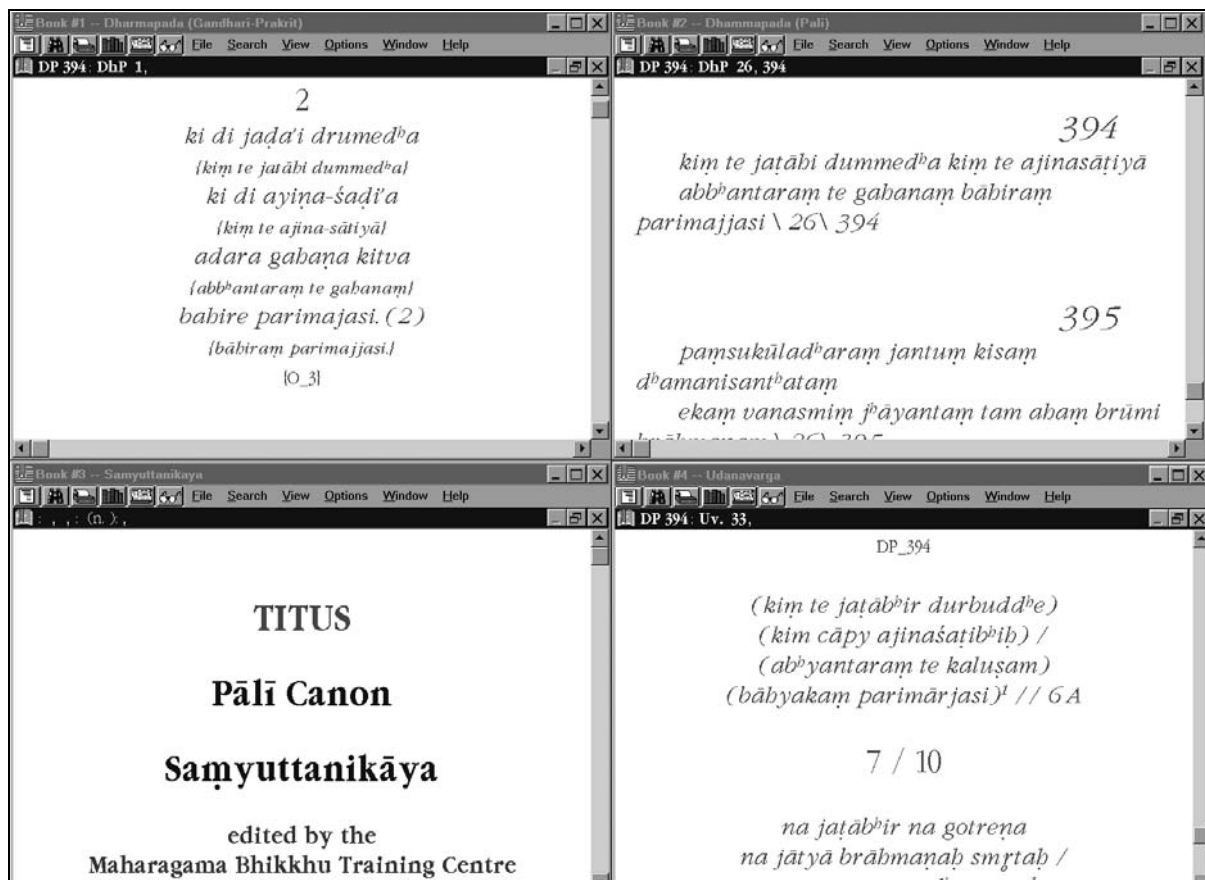


Fig. 8:

GDP 1, 2 synchronized with DP and Uv. 33,6

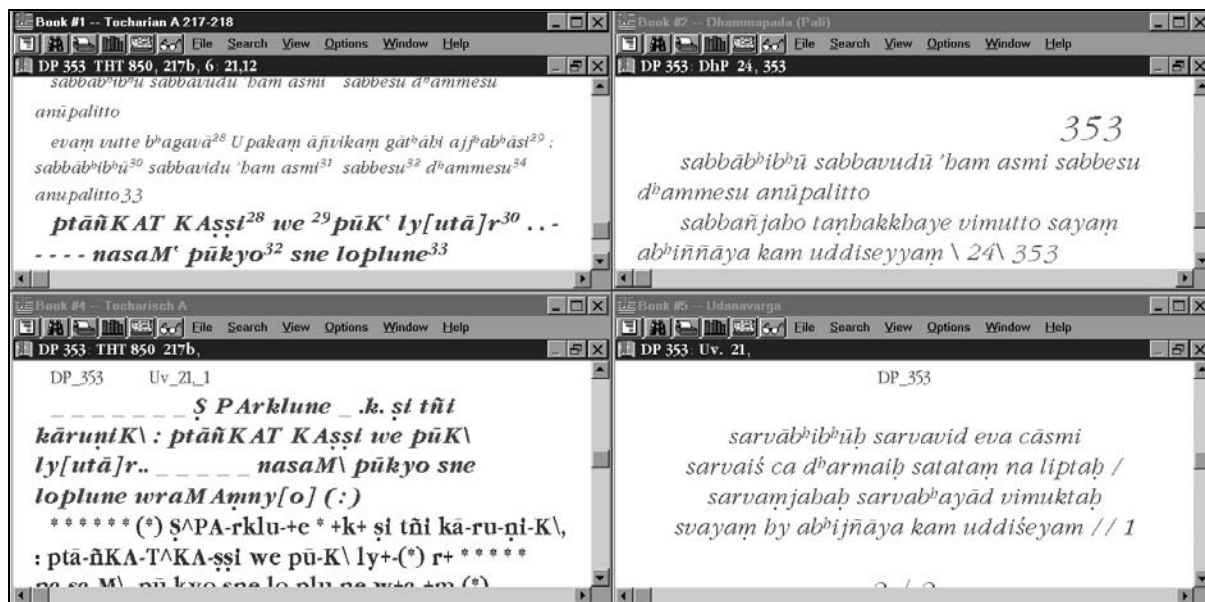


Fig. 9:

Toch. A 217b 6 [21,12] synchronized with DP 353 and Uv. 21,1

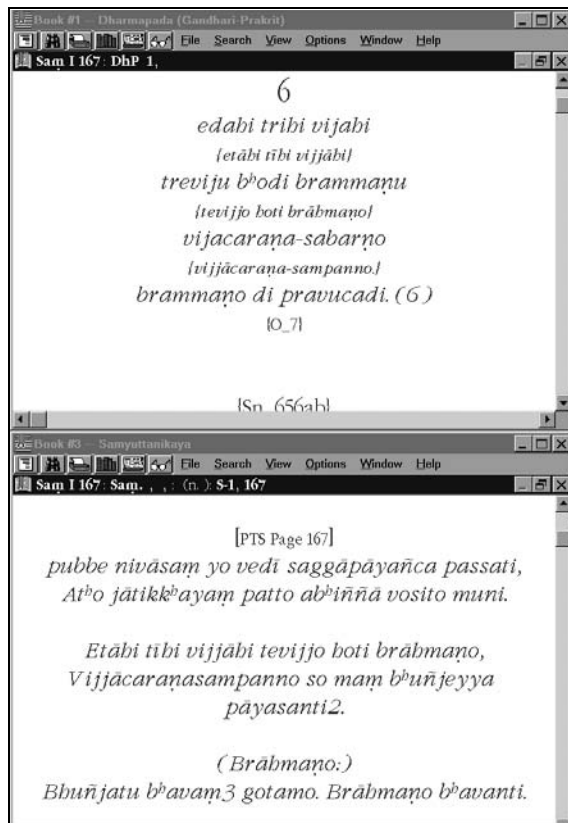


Fig. 10: GDP 1,6 with Sam. I, 167

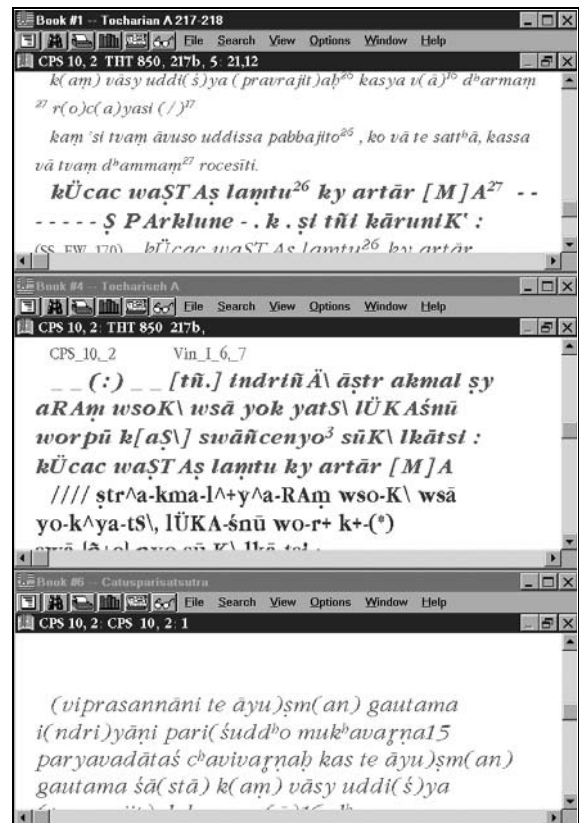


Fig. 11: Toch. A 217b 5 with CPS 10,2