

text-fabric

text + annotations
represented for
slicing and dicing - collaboration - versioning

Dirk Roorda
SBL San Antonio
2016-11-18

Data Archiving and Networked Services



EEP TALSTRA CENTRE
FOR BIBLE AND COMPUTER

בְּרֵאשִׁית בָּרַא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ:

הַ

art

אֶת

prep

אֱלֹהִים

subs

בָּרָא

verb

רֵאשִׁית

subs

בְּ

prep

הַ

art

NA NA NA

אֶת

prep

NA NA NA

אֱלֹהִים

subs

m pl NA

בָּרָא

verb

m sg p3

רֵאשִׁית

subs

f sg NA

בְּ

prep

NA NA NA

בָּרָא

bār'ā

בָּרָא בָרָא]

B.@R@74>

B.@R@> BR>[

create

none verb verb hbo 3

m sg p3 NA perf qal

absent absent absent absent absent

48 746 15 2341

רֵאשִׁית

rēš,îṭ

רֵאשִׁית רֵאשִׁית/

R;>CI73JT

R;>CIJT R>CJT/

beginning

none subs subs hbo 2

f sg NA a NA NA

n/a absent absent n/a n/a

51 707 45 868

בְּ

bə

בְּ ב

B.:—

B.:— B

in

none prep prep hbo 1

NA NA NA NA NA NA

n/a n/a absent absent n/a n/a

15542 3 14194 3

<p>בָּרָא</p> <p>bār'ā</p> <p>בָּרָא בָרָא]</p> <p>B.@R@74></p> <p>B.@R@> BR>[</p> <p>create</p> <p>none verb verb hbo 3</p> <p>m sg p3 NA perf qal</p> <p>absent absent absent absent absent</p> <p>48 746 15 2341</p>	<p>רֵאשִׁית</p> <p>rēš,îṭ</p> <p>רֵאשִׁית רֵאשִׁית/</p> <p>R;>CI73JT</p> <p>R;>CIJT R>CJT/</p> <p>beginning</p> <p>none subs subs hbo 2</p> <p>f sg NA a NA NA</p> <p>n/a absent absent n/a n/a</p> <p>51 707 45 868</p>	<p>בְּ</p> <p>bə</p> <p>בְּ ב</p> <p>B.:–</p> <p>B.:– B</p> <p>in</p> <p>none prep prep hbo 1</p> <p>NA NA NA NA NA NA</p> <p>n/a n/a absent absent n/a n/a</p> <p>15542 3 14194 3</p>
NA Pred VP NA NA 2 2	und Time PP NA NA 1 1	und Time PP NA NA 1 1
? xQtX NA 0 1 0 1 1	? xQtX NA 0 1 0 1 1	? xQtX NA 0 1 0 1 1
1 1	1 1	1 1

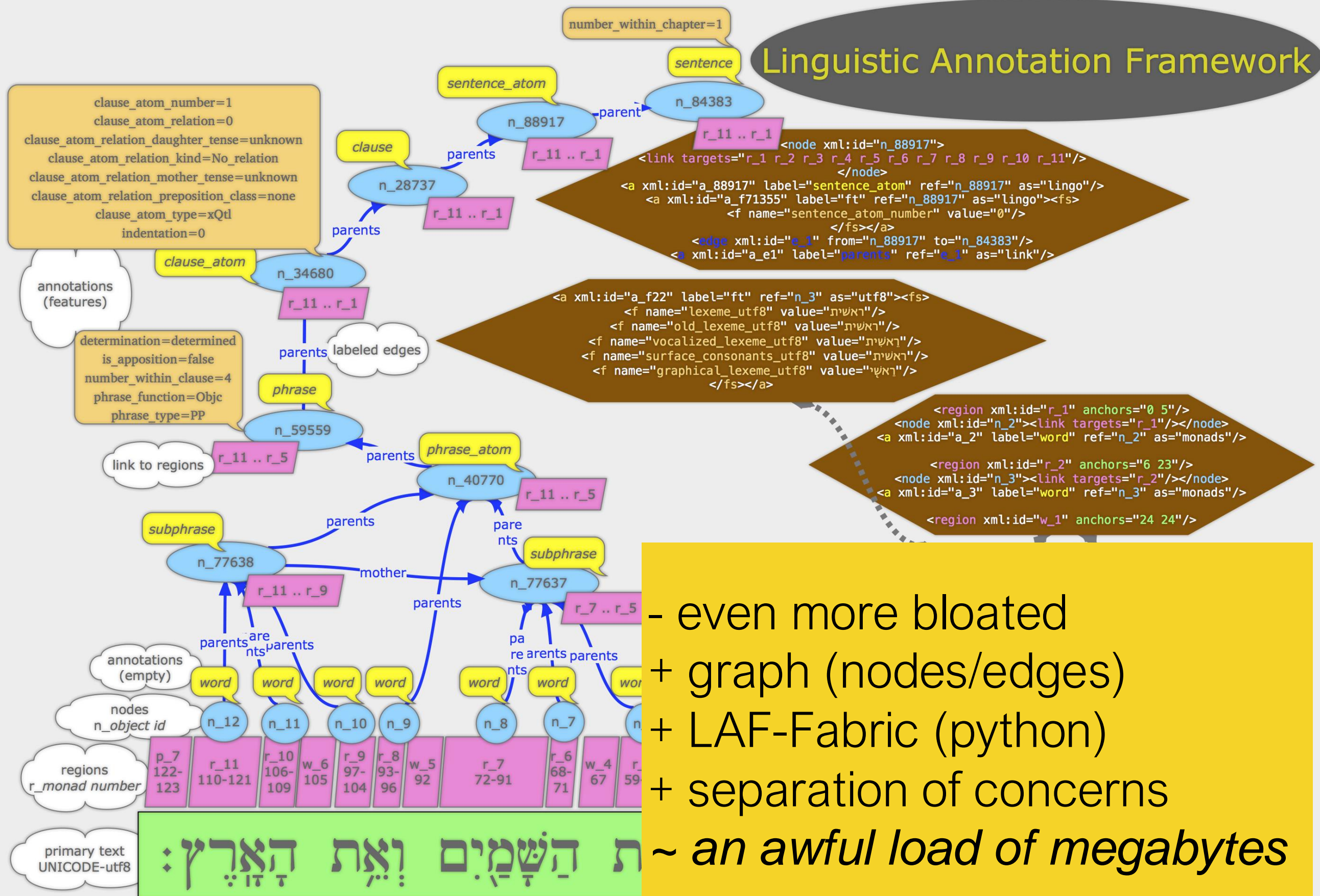
וְ	רָקִיעַ	הָ	אֶת־	אֱלֹהִים	יַעַשׂ	וְ
and	firmament	the	<object marker>	god(s)	make	and
Conj 65 1	Objc 64 4	Objc 64 4	Objc 64 4	Subj 63 3	Pred 62 2	Conj 61 1
NA 20 1	NA 19 1	NA 19 1	NA 19 1	NA 19 1	NA 19 1	NA 19 1
תַּחַת	מִ	אֲשֶׁר	מַיִם	הַ	בֵּין	יִבְדֵּל
under part	from	<relative>	water	the	interval	separate
PreC 69 2	PreC 69 2	Rela 68 1	Cmpl 67 3	Cmpl 67 3	Cmpl 67 3	Pred 66 2
Attr 21 2	Attr 21 2	Attr 21 2	NA 20 1	NA 20 1	NA 20 1	NA 20 1
מַיִם	הַ	בֵּין	וְ	רָקִיעַ		לְ
water	the	interval	and	firmament	the	to
Cmpl 71 5	Cmpl 71 5	Cmpl 71 5	Conj 70 4	PreC 69 2	PreC 69 2	PreC 69 2
NA 22 1	NA 22 1	NA 22 1	NA 22 1	Attr 21 2	Attr 21 2	Attr 21 2
וְ	רָקִיעַ		לְ	עַל	מִ	אֲשֶׁר
and	firmament	the	to	upon	from	<relative>
Conj 74 1	PreC 73 2	PreC 73 2	PreC 73 2	PreC 73 2	PreC 73 2	Rela 72 1
NA 24 1	Attr 23 3	Attr 23 3	Attr 23 3	Attr 23 3	Attr 23 3	Attr 23 3
					כֵּן:	יְהִי־
					thus	be
					Modi 76 3	Pred 75 2

straightforward XML

```
...  
<clause id="3" rela="Objc">  
  <phrase id="27" type="VP">  
    <word id="153" sp="verb" lex="BR&gt;[" phono="bār'ā"  
      lex_heb="אָרַב" trans="B.@R@74&gt;"  
      gloss="create" partofspeech="verb"  
      trailer=" ">אָרַב</word>  
    <word ...>...</word>  
    ...  
  </phrase>  
  <phrase ...>...</phrase>  
  ...  
</clause>  
...
```

- bloated
- hierarchy problems
- inferior tooling (xslt)
- no separation of concerns
- *a horrible mess*

Linguistic Annotation Framework



- even more bloated
- + graph (nodes/edges)
- + LAF-Fabric (python)
- + separation of concerns
- ~ *an awful load of megabytes*

LAF-Fabric binary

Fn0(etcbc4,ft,language)

Fn0(etcbc4,ft,lex_utf8)

Fn0(etcbc4,ft,lex)

Fn0(etcbc4,ft,ls)

Fn0(etcbc4,ft,mother_object_type)

Fn0(etcbc4,ft,lex)

~ reasonably compact
- a bit opaque
+ column storage
+ python (pickle + gzip)
+ separation of concerns
~ *starting to look decent*

Fn0(etcbc4,ft,st)

Fn0(etcbc4,ft,tab)

Fn0(etcbc4,ft,trailer_utf8)

0: "B"

1: "R>CJT/ "

2: "BR>["

3: ">LHJM/ "

4: ">T"

5: "H"

6: "CMJM/ "

7: "W"

8: ">T"

9: "H"

10: ">RY/ "

11: "W"

12: "H"

13: ">RY/ "

14: "HJH["

15: "THW/ "

16: "W"

17: "BHW/ "

18: "W"

19: "XCK/ "

first 20 words

B
R>CJT/
BR>[
>LHJM/
>T
H
CMJM/
W
>T
H
>RY/
W
H
>RY/
HJH[
THW/
W
BHW/
W
XCK/

from 200000

200000 BN/
KL/
H
JWM/
B
JWM/
PC<[
>DWM/
MN
TXT/
JD/
JHWDH/
W
MLK[
<L
MLK/
W
<BR[
JWRM/
Y<JR=/

Text Fabric

- + compact
- transparent
- + column storage
- + pure python
- + separation of concerns

~ with the syntax

*bureaucracy out of the way,
we can start to look into the
real problems:*

- * **slicing and dicing**
- * **collaboration**
- * **versioning**

Text Fabric (inside)

LAF-Fabric, with the LAF replaced by Text

Inside TF, the LAF-Fabric API just works:

```
Genesis 1:1 ḇrēš,îṭ bār'ā ʔēlōh'îm ʔet haššām'ayim w̱əʔet haʔāreš .
Genesis 1:2 w̱əhāʔāreš hāyʔt,ā ṭ'ōhû wāv'ōhû w̱əh,ōšek ʕal-p̱an'ē ṭ'h'ôm w̱ər'ûah ʔēlōh'îm m̱arah,efet ʕal-p̱an,ē hamm'āyim .
Genesis 1:3 wayy,ōmer ʔēlōh'îm y̱əh'î ʔ'ôr w'ay̱əhî-ʔ'ôr .
Genesis 1:4 wayy'ar ʔēlōh'îm ʔet-hāʔ,ôr kî-ṭ'ôv wayyavd'el ʔēlōh'îm b,ên hāʔ,ôr ûv,ên haḥ'ōšek .
Genesis 1:5 wayyiqr,ā ʔēlōh'îm lāʔôr y,ôm walah,ōšek q'ārā l'āyālā w'ay̱əhî-ʕerev w'ay̱əhî-v,ōqer y,ôm ʔeh'ād .
Genesis 1:6 wayy'ōmer ʔēlōh'îm y̱əh'î raq,îaʕ ḇet-ōk hamm'ayim w'ay̱əhî māv'd'î b,ên hamm'ayim lām'ayim .
Genesis 1:7 wayy'aʕaš ʔēlōh'îm ʔet-hārāqî,aʕ wayyavd'el b,ên hamm'ayim ʔaš,er mitt'ahat lārāq îaʕ ûv'ên hamm'ayim ʔaš,er mēš'al lārāq'.
Genesis 1:8 wayyiqr'ā ʔēlōh'îm l'ārāq,îaʕ šām'ayim w'ay̱əhî-ʕerev w'ay̱əhî-v,ōqer y,ôm šēn'î .
Genesis 1:9 wayy'ōmer ʔēlōh'îm yiqqāw,û hamm'ayim mitt'ahat haššām'ayim ʔel-māq'ôm ʔeh'ād w̱əṭērāʔ,eh hayyabbāš'ā w'ay̱əhî-k'ên .
Genesis 1:10 wayyiqr,ā ʔēlōh'îm layyabbāš,ā ʔ'ereš ûḻamiq,w,ē hamm,ayim qār'ā yamm'îm wayy,ar ʔēlōh'îm kî-ṭ'ôv .
Genesis 1:11 wayy'ōmer ʔēlōh'îm ṭ'adl'ē hāʔāreš d'eše 'ʕēsev mazr'îaʕ z'eraʕ ʕ'eš p̱ar'î ʕ'ōseh p̱ar,î ḻamîn'ô ʔaš,er zarʕô-v,ô ʕal-
Genesis 1:12 wattôš,ē hāʔāreš d'eše 'ʕēsev mazr'îaʕ z'eraʕ ḻamîn'ēhû w̱əʕ'eš ʕ,ōseh p̱ar'î ʔaš,er zarʕô-v,ô ḻamîn'ēhû wayy,ar ʔēlōh'
Genesis 1:13 w'ay̱əhî-ʕerev w'ay̱əhî-v,ōqer y,ôm š̱lîš'î .
Genesis 1:14 wayy'ōmer ʔēlōh'îm y̱əh'î m̱əʔôr,ôṭ birāq'îaʕ haššām'ayim ḻəhavd'îl b,ên hayy,ôm ûv'ên hall'āyālā w̱əhāy'û ḻəʔôṭ,ôṭ ûḻam
Genesis 1:15 w̱əhāy'û ḻim'ôr,ôṭ birāq'îaʕ haššām'ayim ḻəhavd'îl ʕal-hāʔāreš w'ay̱əhî-k'ên .
Genesis 1:16 wayy'aʕaš ʔēlōh'îm ʔet-š̱an,ē hamm̱əʔôr,ôṭ hagg̱dōl'îm ʔet-hammāʔ'ôr hagg̱d,ôl ḻəmemš'elet hayy'ôm w̱əʔet-hammāʔ'ôr haqqi
» .
Genesis 1:17 wayyitt,ēn ʔet-ām ʔēlōh'îm ḇirāq'îaʕ haššām'ayim ḻəhāʔ,îr ʕal-hāʔāreš .
Genesis 1:18 w̱əlim̱əš,ôl bayy'ôm ûvall'aylā 'ûḻəhavd'îl b,ên hāʔ,ôr ûv'ên haḥ'ōšek wayy,ar ʔēlōh'îm kî-ṭ'ôv .
Genesis 1:19 w'ay̱əhî-ʕerev w'ay̱əhî-v,ōqer y,ôm ṭ'arʕîf'î .
Genesis 1:20 wayy'ōmer ʔēlōh'îm yišṟəš'û hamm'ayim š̱ereš n'efeš hayy'ā w̱əʕôf y̱əʕôf'ēf ʕal-hāʔāreš ʕal-p̱an,ē raq,îaʕ haššām'ayim
Genesis 1:21 wayyivr'ā ʔēlōh'îm ʔet-hattannîn,îm hagg̱dōl'îm w̱əʔ'et kol-n'efeš h'ahayy'ā h'ārôm'ešet ʔaš,er šāṟəš,û hamm'ayim ḻəm'.
»r ʔēlōh'îm kî-ṭ'ôv .
Genesis 1:22 way̱əv'āreḵ ʔôṭ'am ʔēlōh'îm ḻərm'ôr p̱ar'û ūṟəv'û ūmilʔ'û ʔet-hamm'ayim bayyamm'îm w̱əhāʕ,ôf y,irev bāʔāreš .
Genesis 1:23 w'ay̱əhî-ʕerev w'ay̱əhî-v,ōqer y,ôm ḥ̱amîš'î .
Genesis 1:24 wayy'ōmer ʔēlōh'îm ṭ'adl'ē hāʔāreš n'efeš hayy,ā ḻəmin'āh ḇəhēm,ā wār'emeš w̱əh'ay̱əṭô-ʔ,ereš ḻəmin'āh w'ay̱əhî-k'ên .
Genesis 1:25 wayy'aʕaš ʔēlōh'îm ʔet-hayy,at hāʔāreš ḻəmin'āh w̱əʔet-habḇəhēm,ā ḻəmin'āh w̱əʔ'et kol-r,emeš h'āʔādām,ā ḻəmin'ēhû wayy
Genesis 1:26 wayy'ōmer ʔēlōh'îm n'aʕaš,eh ʔād'ām ḇəšalm,ēnû kiḏəmut'ēnû w̱əyird,û viḏəḡ,at hayy'om ūv̱əʕ'ôf haššām'ayim ūvabḇəhēm,ā
»āʔāreš .
Genesis 1:27 wayyivr,ā ʔēlōh'îm ʔet-h'āʔād,ām ḇəšalm'ô ḇəš,elem ʔēlōh'îm bār'ā ʔôṭ'ô zāk,ār ūṉəqēv,ā bār,ā ʔôṭ'ām .
Genesis 1:28 way̱əv'āreḵ ʔôṭom ʔēlōh'îm wayy'ōmer ḻəh'ēm ʔēlōh'îm p̱ar,û ūṟəv'û ūmilʔ,û ʔet-hāʔ,āreš w̱əḵivš'uhā ūṟəd'û biḏəḡ'at hayy,
»ṭ ʕal-hāʔāreš .
Genesis 1:29 wayy'ōmer ʔēlōh'îm hinn,ē nāt,attî lāk'em ʔet-kol-ʕ'ešev zōr'ēaʕ z'eraʕ ʔaš,er ʕal-p̱an'ē kol-hāʔāreš w̱əʔet-kol-hāʕ'ē:
»hy,eh ḻəʔokl'ā .
Genesis 1:30 'ûḻəkol-hayy'at haʔāreš,ûḻəkol-ʕ'ôf haššām'ayim ûḻək'ôl rôm'ēs ʕal-hāʔāreš ʔašer-b,ô n'efeš hayy'ā ʔet-kol-y,ereq ʕ
Genesis 1:31 wayy'ar ʔēlōh'îm ʔet-kol-ʔaš'er ʕaš'ā w̱əhinnē-ṭ,ôv m̱əʔ'ôḍ w'ay̱əhî-ʕerev w'ay̱əhî-v,ōqer y,ôm haššišš'î .
»))
```

Text Fabric (skeleton)

A TF dataset is a bunch of text files with at least *otype* and *monads*

node feature
node \Rightarrow otype

otype

```
0-426580 word
426581-514580 clause
514581-605142 clause_atom
605143-858316 phrase
858317-1125831 phrase_atom
1125832-1189401 sentence
1189402-1253740 sentence_atom
1253741-1367532 subphrase
1367533-1367571 book
1367572-1368500 chapter
1368501-1413680 half_verse
1413681-1436893 verse
```

This is the
skeleton:

- the positions
- the containment of all text objects

monads

```
1367533 1-28762
1367534 28763-52510
1367572 1-673
1367573 674-1167
1413681 1-11
1413682 12-31
1368501 1-4
1368502 5-11
1125832 1-11
1125833 12-18
1189402 1-11
1189403 12-18
426581 1-11
426582 12-18
514581 1-11
514582 12-18
605143 1-2
605144 3
858317 4
9-11
```

edge feature
node \times node \Rightarrow

Text Fabric (flesh)

information **content**

node features
node \Rightarrow otype

text_full	trailer	p-s-p
בְּ		prep
רֵאשִׁית	—	subs
בְּרָא	—	verb
אֱלֹהִים	—	subs
אֶת	—	prep
הַ		art
שָׁמַיִם	—	subs
!		conj
אֶת	—	prep
הַ		art
אָרֶץ	:_	subs
!		conj
הַ		art
אָרֶץ	—	subs
הִיְתָה	—	verb
תְּהִי	—	subs
!		conj
בְּהוֹ	—	subs
!		conj
חֹשֶׁךְ	—	subs

edge feature
node \times node \Rightarrow

parent

0-1,858317605143
2,858318 605144
3,858319 605145
4-10,858320 605146
426581,1189402 1125832
514581,605143-605146426581

Edges (c't'd)

parent

0-1, 858317605143
2, 858318 605144
3, 858319 605145
4-10, 858320 605146
514581, 605143-605146426581
426581, 1189402 1125832

edge feature
node × node ⇒

1125832

sentence

426581

clause

605143 605144 605145

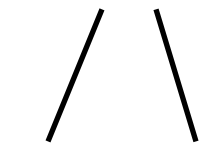
605146

phrase

858317 858318 858319

858320

phrase_atom



0 1 2 3 4 5 6 7 8 9 10

word

bərēš, îṭ bār' ā ʔelōh'îm ʔ, ēṭ haššām, ayim wəʔ, ēṭ hāʔ' āreṣ

skeleton

flesh

slicing 'n dicing

monads

otype

text_full

trailer

p-s-p

typ

vertical:
feature
selection

horizontal:
object
selection

both:
modules

word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word

בְּ
רְאִישִׁית
בְּרָא
אֱלֹהִים
אֶת
הַ
שָׁמַיִם
!
אֶת
הַ
אָרֶץ
!
הַ
אָרֶץ
הַיְּתֵה
תְּהוֹ
!
בְּהוֹ
!

—
—
—
—
—
—
:_
—
—
—
—
—
—
—
—
—
—
—
—
—

prep
subs
verb
subs
prep
art
subs
conj
art
subs
verb
subs
conj
subs
conj
subs

PP
NP
VP
PP
—
Objc
Resu

600000 0-1
600001 2
600002 3
600003 4-10
400001 0-10
400002 11-21
400003 22-34

phrase
phrase
phrase
phrase
clause
clause
clause

חֲשׂוֹן

—

skeleton

flesh

module

collaboration

monads

6000000	0-1
6000001	2
6000002	3
6000003	4-10
4000001	0-10
4000002	11-21
4000003	22-34

otype

word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
phrase
phrase
phrase
phrase
clause
clause
clause

text_full

בְּ
רֵאשִׁית
בְּרֵא
אֱלֹהִים
אֶת
הַ
שָׁמַיִם
!
אֶת
הַ
אֲרֶץ
!
הַ
אֲרֶץ
הִיטָה
תְּהִי
!
בְּהִי
!
חֲשֹׁךְ

p-s-p

prep
subs
verb
subs
prep
art
subs
conj
prep
art
subs
conj
art
subs
verb
subs
conj
subs
conj
subs

strong

8675
7225
1254 a
430
853
8676
8064
8678
853
8676
776
8678
8676
776
1961
8414
8678
922
8678
776

module:
feature *strong*
for *words*

dependency:
on the implicit
monad order!

skeleton
@20170101

flesh
@20170101

versioning

module
@20161118

monads

9

600000**1** 0-1

600000**2** 2

600000**3** 3

600000**4** 4-1**1**

400000**2** 0-1**1**

400000**3** 1**2**-2**2**

400000**4** 2**3**-3**5**

otype

word

word

word

word

word

word

word

word

word

word

word

word

word

word

word

word

word

word

word

word

word

phrase

phrase

phrase

phrase

clause

clause

clause

text_full

בְּ
רֵאשִׁית
בָּרָא
אֱלֹהִים
אֶת
הַ
שָׁמַיִם
!
אֶת
שֶׁ
הַ
אֲרֶץ
!
הַ
אֲרֶץ
הִיְתָה
תְּהוֹ
!
בְּהוֹ
!
חֲשׁוּךְ

A diagram illustrating the structure of the text 'text_full'. A red arrow points to the word 'שֶׁ' (she), which is highlighted in blue. Four blue arrows point to the words 'שָׁמַיִם' (shamayim), 'אֶת' (et), 'אֲרֶץ' (aretz), and 'הִיְתָה' (hiyeta), which are also highlighted in blue. The text is written in Hebrew with vowel points.

@20170101 ← @20161118

	0 0
	1 1
	2 2
	3 3
	4 4
	5 5
	6 6
	7 7
	8 8
	10 9
	11 10
	12 11
	13 12
	14 13
	15 14
	16 15
	17 16
	18 17
	19 18
	20 19
	21 20
600001	600000
600002	600001
600003	600002
600004	600003
400002	400001
400003	400002
400004	400003

strong

8675
7225
1254 a
430
853
8676
8064
8678
853
8676
776
8678
8676
776
1961
8414
8678
922
8678
2822

old module
works **as is** on
new data

**old module
works as is on
new data**

the end of the beginning

etcbc

Stephen Ku

**David van
Acker**

**James
Cuénod**

flesh
@20160101

skeleton

@20160101 <=
@20170101

flesh
@20170101

skeleton

@20170101 <=
@20180101

module **strong**
@20161118

flesh
@20180101

skeleton

@20180101 <=
@20190101

module **accent**
@20170202

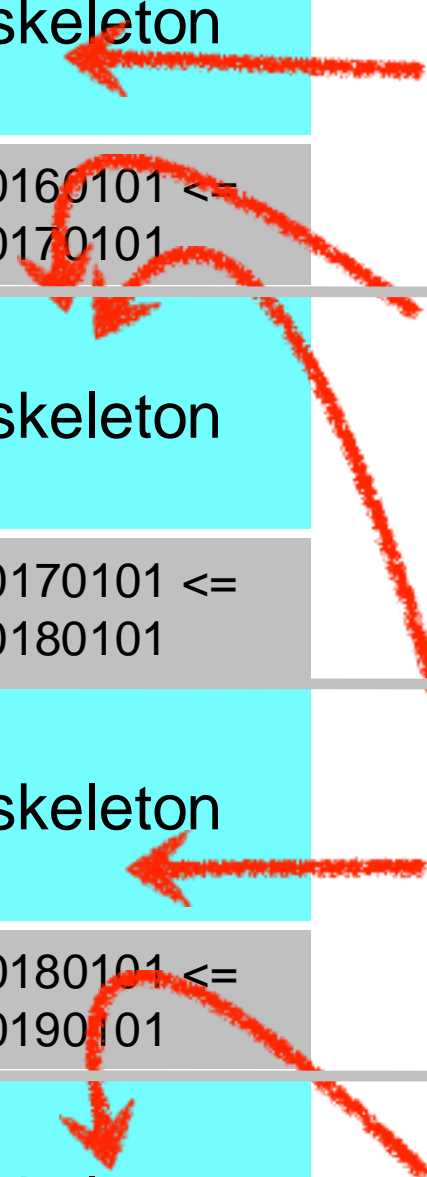
flesh
@20190101

skeleton

@20190101 <=
@20200101

module **strong**
@20171118

module **assoc**
@20180303





text-fabric

thank ye



dirk.roorda@dans.knaw.nl

shebanq@ancient-data.org

Data Archiving and Networked Services

DANS



EEP TALSTRA CENTRE
FOR BIBLE AND COMPUTER