How to Establish a Verbal Paradigm on the Basis of Ancient Syriac Manuscripts

W.Th. (Wido) van Peursen
Leiden Institute for Religious Studies
P.O. Box 9515
NL-2300 RA Leiden
w.t.van.peursen@religion.leidenuniv.nl

Abstract

This paper describes a model that has been developed in the Turgama Project at Leiden University to meet the challenges encountered in the computational analysis of ancient Syriac Biblical manuscripts. The small size of the corpus, the absence of native speakers, and the variation attested in the multitude of textual witnesses require a model of encoding—rather than tagging—that moves from the formal distributional registration of linguistic elements to functional deductions. The model is illuminated by an example from verb inflection. It shows how a corpus-based analysis can improve upon the inflectional paradigms given in traditional grammars and how the various orthographic representations can be accounted for by an encoding system that registers both the paradigmatic forms and their attested realizations.

1 Working with ancient documents

1.1 Challenges

If we wish to make a linguistic analysis of ancient texts, in our case the Hebrew Bible and its Syriac translation, the Peshitta (ca. 2nd century CE), we are confronted with a number of challenges:

- There is no native speaker of the languages involved. We do not know in advance what categories are relevant in the linguistic analysis, what functions a certain construction has, or what functional oppositions there exist in the language system. For this reason we should avoid as much as we can any model that presupposes knowledge about the language.
- We have only written sources. Hence we are challenged by the complex interaction between orthographic conventions and morphological phenomena. There are even some orthographic practices which, it is claimed, have never been supported by a phonological or morphological realization (see section 4.5).
- We are dealing with multiple unique documents. In philology, *the* text of the Hebrew Bible or its Syriac translation is an abstract notion, a scholarly construct. The corpus that we enter into our database consists of the concrete *witnesses* to the abstract text. Textual variants provide useful information about language variation and development (section 4.5).
- We are dealing with a small corpus. The Hebrew Bible contains about 300.000–400.000 words (depending on whether we count graphic words or functional words); the vocabulary consists of about 8.000 lexemes.

Moreover, because of the context in which our research takes place, at the boundary of linguistics and philology, our aim is the construction of a database with a correctly encoded text. Because we want to understand the text, rather than merely collect knowledge about the language system, we have set high standards of accuracy for the encoding of the text.

1.2 Dilemmas

These challenges lead to the following dilemmas for the computational analysis of ancient texts:

- Data-oriented or theory-driven? Since approaches that presuppose linguistic knowledge are problematic, we want to be

Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, pages 1–9,
Athens, Greece, 31 March, 2009.

data-oriented, rather than theory-driven. However, approaches that merely try to extract knowledge from the corpus with a minimum of human input are insufficient because of the size of our corpus and because we want knowledge about the text, not just about the language.

- Priority for the corpus or the language? Due to the lack of native speakers, the sole basis for our knowledge about the language is the corpus, but, at the same time, the corpus can only be accessed through some linguistic knowledge and some basic understanding of the text. We cannot start from scratch, avoiding any preliminary understanding of the text, its language, its features, and its meaning. This understanding is shaped by our scholarly and cultural tradition. It is based on transmitted knowledge. But we have to find ways in which the computational analysis does not only imitate or repeat traditional interpretations.

1.3 Requirements

The challenges and dilemmas mentioned above require a model that is deductive rather than inductive; that goes from form (the concrete textual data) to function (the categories that we do not know a priori); that entails registering the distribution of linguistic elements, rather than merely adding functional labels—in other words, that involves encoding rather than tagging; that registers both the paradigmatic forms and their realizations; that allows grammatical categories and formal descriptions to be redefined on the basis of corpus analysis; and that involves interactive analytical procedures, which are needed for the level of accuracy we aim for.

In the distributional analysis at word level, for example, we mark prefixes and suffixes, rather than tagging a form as "imperfect 2ms" etc. Similarly on clause level we identify patterns such as "subject + participle + complement", as against the usual practice of giving functional clause labels such as "circumstantial clause".

2 Analytical procedure

In our project the analysis of Hebrew and Syriac involves a bottom-up linguistic analysis at the following levels:

2.1 Word level

This level concerns the segmentation of words into morphemes, the functional deductions from the morphological analysis, and the assignment of lexically determined word functions. It will be described in detail in section 3.

2.2 Phrase level

At this level words are combined into phrases (e.g. noun + adjective). This entails the morpho-syntactic analysis and the systematic adaptations of word classes in certain environments (e.g. adjective → noun), and the analysis of phrase-internal relations (e.g. apposition).

2.3 Clause level

This level concerns the combination of phrases into clauses (e.g. conjunction + VP + determinate NP), and the assignment of syntactic functions (e.g. subject, predicate).

2.4 Text level

This level concerns the determination of the relationships between clauses and the assignment of the syntactical functions of the clauses within the text hierarchy (e.g. object clause).

3 Workflow of word-level analysis

In the following discussion we will restrict ourselves to the morphological analysis. At the higher linguistic levels the same principles are applied, although the consequences are somewhat different (see section 5).

3.1 Running text

As an example we take the Syriac translation (Peshitta) of the book of Judges. The starting-point of the analysis is a transliterated running text, called P_Judges, which reflects the Leiden Peshitta edition. Sample 1 contains the first verse of this text. The variant notation between square brackets indicates that the first word, `whw'`, 'and it happened', is missing in a number of manuscripts. Between the angle brackets a comment has been added.

Even this first step involves a number of disambiguating decisions, for example, as to whether a dot above a letter is a vowel sign, a delimitation marker, or just a spot in the manuscript.[1]

[1] One has to take similar decisions if one transcribes the text of a manuscript to Unicode, because the definitions of the Unicode characters include both a for-

```
1 [whw'/ -6h7, 8a1c, 10c1, 11c1, 12a1fam]
<check reading in 6h7> mn btr dmyt y$w` brnwn
`bdh dmry'; 1 $'lw bn:y 'ysryl bmry' w'mr:yn;
mnw nsq ln `l kn`n:y' bry$'; lmtkt$w `mhwn
bqrb';
```

Sample 1: P_Judges (running text)

3.2 Production of graphic text ('pil2wit')

The program pil2wit transforms the running text into the so-called graphic text, a transliterated text according to an established format that enables the subsequent steps in the analysis (sample 2). It has another transliteration system;[2] instructions to select variants have been executed; comments have been omitted; and the markers of book, chapter and verse have been added.

```
  1  %bookname Jd
  2  %language syriac
  3
  4  %verse 1,1
  5    WHW> MN BTR DMJT JCW< BRNWN <BDH DMRJ>
C>LW BN"J >JSRJL BMRJ> W>MR"JN MNW NSQ LN <L
KN<N"J> BRJC> LMTKTCW <MHWN BQRB>
```

Sample 2: P_Judices (graphic text)

3.3 Production of encoded text ('Analyse')

The grapic text is the input file for the program Analyse, which concerns the segmentation of the Syriac words into morphemes (as far as concatenative morphemes are involved[3]). For this segmentation we use a system of encoding, rather than tagging. Thus the imperfect form *neqtol* "he will kill" is encoded as `!N!QV&WL[`, in which the exclamation marks `!…!` indicate the prefix, the ampersand `&` a paradigmatically unexpected letter—the round bracket `(` indicates an expected but absent letter—and the square bracket to the right `[` a verbal ending. Sample 3 provides the interface in the interactive procedure of Analyse.

```
  1,1 WHW>      W-HW(J&>[, W-HW(J&>[/
  1,1 MN        MN, MN=
  1,1 BTR       BTR
  1,1 DMJT      D-M(W&JT[, D-M(W&JT[/:p
  1,1 JCW<      JCW</
  1,1 BRNWN     BR/-NWN=/
  1,1 <BDH      <BD=/-H, <BD[-H, <BD==/-H
  1,1 DMRJ>     D-MRJ>/
```

Sample 3: P_Judices.an (analysed text; automatically generated file)

The first column contains the verse number, the second the graphic words (which may contain more than one functional word; thus the first graphic word contains the conjunction `W` and the verb `HW>`) and the third column contains proposals for the morphological segmentation. These proposals are generated from the 'Analytical Lexicon', a data file containing the results of previous analyses (sample 4).

```
9308  WCKR>     W-CKR/~>
9309  WCLWM     W-CLWM/
9310  WCLX      W-CLX[
9311  WCLX      W-CLX[(W
9312  WCLXW     W-CLX[W
9313  WCLXT     W-CLX[T==
```

Sample 4: Excerpt from the Analytical Lexicon

It appears, for example, that up to the moment that sample 4 was extracted from the lexicon, the form `WCLX` had received two different encodings (lines 9310 and 9311; see below, section 4.3).

The human researcher has to accept or reject the proposals made by Analyse or to add a new analysis. We cannot go through all details, but in the second line of sample 4, for example, a choice has to be made between the preposition *men* (`MN`) and the interrogative pronoun *man* (`MN=`; the disambiguating function of the equals sign is recorded in the lexicon [section 3.6], where both `MN` and `MN=` are defined). Likewise, in the case of `<BDH`, the human researcher has to decide whether this is a verb (hence the verbal ending `[`), the noun 'servant' (`<BD=`), or the noun 'work' (`<BD==`).

For these disambiguating decisions in the interactive procedure the human researcher follows a protocol that describes the relative weight of diacritical dots in the oldest manuscripts, the vowel signs that are added in some manuscripts, the vocalization in printed editions, and grammatical and contextual considerations.

```
  1,1 WHW>      W-HW(J&>[
  1,1 MN        MN
  1,1 BTR       BTR
  1,1 DMJT      D-M(W&JT[
  1,1 JCW<      JCW</
  1,1 BRNWN     BR/-NWN=/
  1,1 <BDH      <BD=/-H
  1,1 DMRJ>     D-MRJ>/
```

Sample 5: P_Judices.an (analysed text; outcome of interactive procedure)

After the interactive procedure the analysed text contains the 'correct' analysis for each word of

---

mal description and a functional analysis. There is not a character for 'a dot above the letter', but rather for 'vowel sign above the letter' etc.

[2] Transliteration alphabet: > B G D H W Z X V J K L M N S < P Y Q R C T.

[3] Non-concatenative morphemes are marked with a colon at the end of a word. We use `:p` for the vowel pattern of the passive; `:d` for the doubled verbal stem and `:c` for the construct state vocalization of nouns.

the graphic text (sample 5). As we shall see below, we do not consider this as the definitive analysis, but rather as a hypothesis about the data that can be tested in the following steps of the analytical procedure.

3.4 Reformatting and selection ('Genat')

The next step concerns the selection of a chapter and the reformatting of the document. This is done automatically by the program Genat. The result is e.g. P_Judices01.at (sample 6).

```
1,1  W-HW(J&>[ MN BTR D-M(W&JT[ JCW</ BR/-
NWN=/ <BD=/-H D-MRJ>/ C>L[W BN/J >JSRJL/ B-
MRJ>/ W->MR[/JN MN=-(HW !N!S(LQ[ L-N <L
KN<NJ/(J~> B-RJC/~> L-!M!@(>T@KTC[/W:d <M-
HWN= B-QRB=/~>
```

Sample 6: P_Judices01.at (analysed text, reformatted)

3.5 Functional deductions ('at2ps')

The next step concerns the functional deductions from the morphological analysis (e.g. person, number, gender) and the assignment of lexically determined word functions (e.g. part of speech). For this purpose the program at2ps uses three language definition files: a description of the alphabet, a lexicon (section 3.6), and a description of the morphology ('Word Grammar'; section 3.7).

3.6 The Lexicon

Each line in the Lexicon contains the lexeme, a unique identification number, lexically relevant characteristics such as a part of speech (`sp`) or a lexical set (`ls`), a gloss (`gl`), which is only intended for the human user and, optionally, a comment added after the hash (#).

```
CLWM   6577:sp=subs:ls=prop:st=abs:gn=m:gl=
       Shallum
CLX    10753:sp=verb:gl=to send, PA to strip,
       to despoil
CLX=   15359:sp=subs:ls=prop:st=abs:gn=m:gl=
       Shilhi
CLX==  32679:sp=subs:de=CLX>:gl=swarm (bees),
       skin (lamb)       # Judges 14,08
```

Sample 7: Extract from the Lexicon

3.7 The 'Word Grammar'

The encoded text is read by the Word Grammar. In this auxiliary file are registered (1) the types of morphemes recognized; (2) the individual morphemes of each morpheme type; (3) a list of grammatical functions; and (4) rules for the functional deductions (see samples 8–11).

```
prefix =
    pfm:  {"!","!"} "preformative"
    pfx:  {"@","@"} "passive stem formation
          prefix"
    vbs:  {"]","]"} "verbal stem"
core =
    lex:  {} "lexeme"
suffix =
    vbe:  {"["} "verbal ending"
    nme:  {"/"} "nominal ending"
    emf:  {"~"} "emphatic marker"
pattern =
    vpm:  {":"} "vowel pattern"
functions
ps: "person" =
    first: "first", second: "second", third:
    "third"
nu: "number" =
    sg: "singular", du: "dual", pl: "plural",
    unknown: "unknown"
gn: "gender" =
    f: "feminine", m: "masculine"
```

Sample 8: Extract from the Word Grammar, section 1: Morpheme types

```
vbe =   "",   "W",   "WN",   "J",   "J=",   "JN",
        "JN=",  "N",  "N=",  "T",  "T=",  "T==",
        "TWN", "TJ", "TJN"
```

Sample 9: Extract from Word Grammar, section 2: Individual morphemes for morpheme types

```
ps: "person" =
    first: "first", second: "second", third:
    "third"
nu: "number" =
    sg: "singular", du: "dual", pl: "plural",
    unknown: "unknown"
gn: "gender" =
    f: "feminine", m: "masculine"
```

Sample 10: Extract from the Word Grammar, section 3: Grammatical functions

```
shared { exist(pfm) && exist(vbe) && not ex-
    ist(nme) :: vt=ipf }
    shared { pfm == "N" :: ps=third }
        vbe == "" :: gn=m, nu=sg
        vbe != {"", "WN", "N="} :: reject
    end
shared { pfm == "T=" :: ps=third }
        vbe == {""} :: gn=f, nu=sg
                vbe != "" :: reject
    end
```

Sample 11: Extract from the Word Grammar, section 4: Rules for functional deductions

Each rule concerns the pairing of a morphological condition and an action. The condition is phrased as a Boolean expression yielding true or false indicating whether the condition is met or not. If the condition is met, the listed actions are undertaken. An action is usually the assignment of a value to a word function, but can also involve accepting or rejecting a form, or jumping to a rule further down. Thus the rule

```
vbe == "W"    :: gn=m, nu=pl
```

can be read as: *if* there is a verbal ending W, *then* assign the values gender = masculine and number = plural.

3.8 Result: the ps2 file

The result is a ps2 file. Each row contains a verse reference, the lexeme, and a list of lexical and morphological features such as the lexical set, part of speech, verbal prefix, verbal stem, verbal ending, nominal ending, verbal tense, person, number, gender, nominal state. Thus the second line of sample 12 shows that the second word of Judges is HWJ, 'to be', which has the lexical set 'verb of existence' (-2); it has the part of speech 'verb' (1); it has no verbal prefix (0); it comes from the simple verbal stem Peal or Qal (0); it has an empty verbal ending (1); it has no nominal ending (0); it is a perfect form (2) 3rd person (3) singular (1), without personal suffix (-1),[4] masculine (2); and the notion of 'state' does not apply to it (-1), because this notion is only used in the case of nominal endings.

```
01,01 W     0    6 -1 -1 -1 -1 -1    -1 -1 -1 -1     -1
01,01 HWJ -2     1  0  0  1  0 -1     2  3  1  2     -1
01,01 MN    0    5 -1 -1 -1 -1 -1    -1 -1 -1 -1     -1
01,01 BTR   0    5 -1 -1 -1 -1 -1    -1 -1 -1 -1     -1
01,01 D    -1    5 -1 -1 -1 -1 -1    -1 -1 -1 -1     -1
01,01 MWT   0    1  0  0  1  0 -1     2  3  1  2     -1
01,01 JCW<  0    3 -1 -1 -1  1 -1    -1 -1  0  2      2
01,01 BR    0    2 -1 -1 -1  1 -1    -1 -1  0  2      0
01,01 NWN=  0    3 -1 -1 -1  1 -1    -1 -1  0  2      2
```

Sample 12: P_Judices.ps2

From this file two files are automatically generated: an encoded surface text (xxx.ct) and a data description in human readable form (xxx.dmp).

```
1   RICHT01,01 W-HW> MN BTR D-MJT JCW< BR-NWN
<BD-H D-MRJ> C>LW BNJ >JSRJL B-MRJ> W->MRJN
MN-W NSQ L-N <L KN<NJ> B-RJC> L-MTKTCW <M-HWN
B-QRB> *
```

Sample 13: P_Judices01.ct

```
1,1 W        W    W     sp=conj
1,1 HW(J&>[ HW>   HWJ   vbe="",sp=verb,vo=act,vs=pe,
                        vt=pf,ps=third,nu=sg,gn=m,
                        ls=vbex
1,1 MN       MN   MN    sp=prep
1,1 BTR      BTR  BTR   sp=prep
```

[4] This column comes from an earlier phase of our project. In our present encoding the value is always 'inapplicable' (-1), because we now treat the suffix pronoun as an independent lexeme. Its lexeme status appears from its own grammatical functions, which are different from those of the word to which it is attached. The traditional lexicographical practice, however, does not treat it as a lexeme (Sikkel, 2008).

```
1,1 D        D    D     ls=pcon,sp=prep
1,1 M(W&JT[ MJT   MWT   vbe="" sp=verb,vo=act,vs=pe,
                        vt=pf,ps=third,nu=sg,gn=m
1,1 JCW</    JCW< JCW<  nme="" sp=subs,+nu,gn=m,st=
                        abs,ls=prop
1,1 BR/      BR   BR    nme="" sp=subs,+nu,gn=m,+st
1,1 NWN=/    NWN  NWN=  nme="" sp=subs,+nu,gn=m,st=
                        abs,ls=prop
```

Sample 14: P_Judices01.dmp

3.9 Summary of the workflow

The workflow can be summarized as follows:

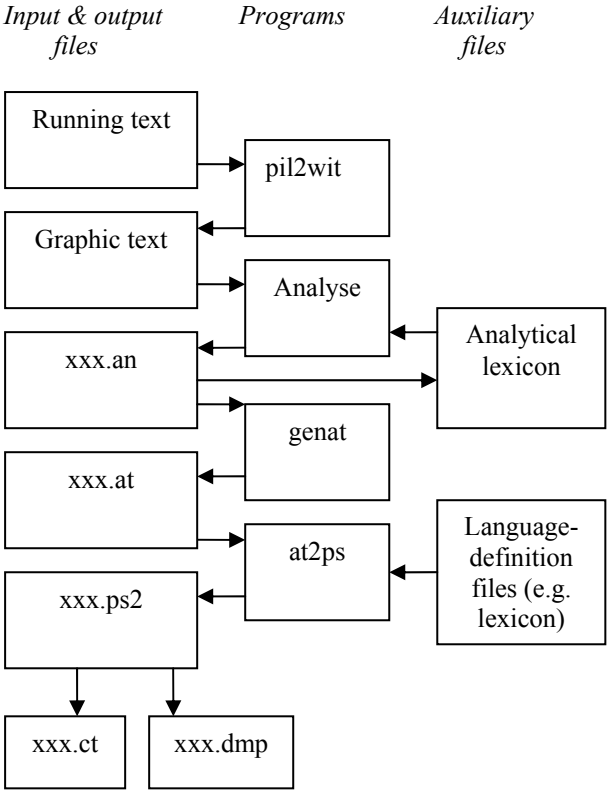

Table 1: workflow of word level analysis

It follows that the following programs and data sets are used:

- Programs that recognize the patterns of formal elements that combine to form words, phrases, clauses and textual units (e.g. at2ps).
- Language-specific auxiliary files (e.g. Lexicon, Word Grammar).
- Data sets, built up gradually, containing all patterns registered in the analysis (e.g. Analytical Lexicon)
- Programs that use the data sets and the auxiliary files to make proposals in the interactive procedures for the linguistic analysis (e.g. Analyse).

3.10 Relation with requirements

Some typical features of the workflow serve to meet the requirements defined in section 1.3. The procedure of encoding rather than tagging guarantees consistency in the analysis of morphemes, because the functional deductions are produced automatically. It has the advantage that not only the interpretation of a word, but also the data which led to a certain interpretation can be retrieved, whereas the motivation behind a tagging is usually not visible. It also has the advantage that both the surface forms and the functional analysis are preserved.

By using the language-specific auxiliary files we take our starting-point in the scholarly tradition of Semitic studies, but the encoding system allows us to test alternative interpretations of the data (see below, section 4.5).

4 The verbal paradigm

4.1 Traditional grammars

We will now illuminate our model by taking a look at the verbal paradigm. For the moment we will restrict ourselves to the paradigm of the suffix conjugation. In the traditional grammars we find the following inflection paradigm:

	singular		plural	
3 m	–		*w* – *wn*	[silent] (*un*)
3 f	*t*	(*at*)	– *y*	 [silent]
2 m	*t*	(*t*)	*twn*	(*ton*)
2 f	*ty*	(*t*)	*tyn*	(*ten*)
1 c	*t*	(*et*)	*n* *nn*	(*n*) (*nan*)

Table 2: Paradigm of the perfect in Classical Syriac according to traditional grammars

4.2 Manuscript evidence

Since we work with real manuscripts, we have to deal with the forms that are actually attested. As appears from the paradigm in table 2, for example, the perfect 3mp sometimes has no verbal ending. What is not recorded in the traditional grammars is that there are also forms 3ms with the ending -*w*. This may be due to the fact that the -*w* in the plural, even if it were represented in writing, was not pronounced.[5] Traditionally the singular forms with -*w* are taken as errors, due to the confusion with the silent -*w* in the plural.

The Leiden Peshitta edition takes such readings as 'orthographic variants'. They do not appear in the critical apparatus to the text, but in a separate *Index Orthographicus*. The general preface to the edition contains a long list of categories of orthographical variation (cf. sample 15).

2.2 varietates inflectionis
2.2.1 afformativa
2.2.1.1 perfectum
e.g. 3 msg + *waw*
3 f.sg *yodh*
2 m.sg + *yodh*
2 f.sg *om yodh*
3 m.pl *om waw*
3 f.pl + *waw*
3 f. pl + *yodh*
1 pl cum des -*nan* etc., etc.

Sample 15: Excerpt from General Preface of Leiden Peshitta Edition: categories of *Index Orthographicus*

These categories are referred to in the *Index Orthographicus* of each volume. Thus we find in the text of Song of Songs in the sample edition:

2.2 varietates flectionis:
2.2.1.1. afformativa *perfecti*
2 f. sg + *yodh*
√ ܪܥܐ (I); √ ܢܚܐ (II)
1_7 II 9l2
8_{10} I 16g6 19 < ?a1

3 f. pl. + *yodh*
√ ܗܠܟ (I); √ ܐܠܡ (II)
4_2 II 10m1.3 11m1.2.4-6 13m1 15a2 17a1.2.4.5.10 18h3
5_5 I 13c1 15a2 16g2.3[1].8.9 17a1-8.10.11 17c1(*vid*) 17g2.6 17h2 18c2[1] 18h3 19g5[1].7

Sample 16: Excerpt from *Index Orthographicus* to Song of Songs in sample volume of Leiden Peshitta edition

Unfortunately, the Peshitta project soon abandoned the inclusion of the *Index Orthograhicus*. It appears only in the sample edition and one other volume (Vol. IV/6, containing Odes, the Prayer of Manasseh, Apocryphal Psalms, Psalms of Solomon, Tobit and 1(3) Esdras).

[5] Admittedly, it can be problematic to make claims about the pronunciation on the basis of written sources, but there are strong arguments for this claim, including the fact that the final letter is ignored in punctuation, that it is frequently omitted in writing (Nöldeke, 2001:§50), and that it does not affect poetic patterns (Brockelmann, 1960:45).

4.3 Encoding the attested forms

In the word-level analysis (cf. section 2.1) the forms listed in table 2 are encoded as follows:

	singular	plural
3m	KTB[	KTB[W KTB[(W KTB[W&N
3f	KTB[T==	KTB[(J= KTB[J=
2m	KTB[T=	KTB[TWN
2f	KTB[TJ	KTB[TJN
1c	KTB[T	KTB[N KTB[N&N

Table 3: Encoded forms of Classical Syriac perfect

As we said above, the square bracket to the right marks the verbal ending and the ampersand a paradigmatically unexpected letter. Thus our encoding in table 3 implies that we take the verbal ending *-wn* as an allomorph of *-w* with an additional *-n*. Alternatively we could decide to introduce a separate morpheme *-wn* besides *-w*. The equals sign is used for the disambiguation of forms that have the same consonantal representation. We use it to distinguish the three verbal endings *-t* and for distinguishing the *-y* of the perfect 3fs from the *-y* of the imperative 3fs.

A round bracket marks a paradigmatically expected but absent letter. Thus we have taken the imperfect form 3fs KTBJ as the paradigmatic form, although KTB occurs as well.

4.4 Paradigmatic forms and their realizations

To deal with this material in an appropriate way it is important to use an encoding system in which both the attested surface forms and the abstract morphemes can be retrieved. Thus *'wqdw* 'they burnt (it)' (Judges 1:8; our transcription: >WQDW) is a form of the verb *yqd* (JQD), with the causative prefix *'-* (>). We mark the causative stem morpheme with two square brackets to the left (cf. sample 8), indicate with the round bracket to the right that the first letter of the lexeme is absent, and mark with the ampersand the *w* that has come instead. The square bracket to the right marks the verbal ending. This results in the following encoding:

```
Encoding:              ]>](J&WQD[W
Paradigmatic forms:     >  JQD   W
Realizations:           >  WQD   W
```

4.5 Language variation and language development

This way of encoding the verb forms attested in multiple textual witnesses provides us with a large database from which language variation data can be retrieved. In some cases language development is involved as well, and the data can be used for diachronic analysis. For this research we can build upon the work done by the Syriac scholar Sebastian Brock. One of the phenomena Brock (2003:99–100) observed was that in West Syriac Biblical manuscripts some orthographic innovations are attested, including the addition of a *-y* to the perfect 3fp, the imperfect 3fs and, on analogy, the perfect 3fs. It is a debated issue whether this ending reflects a morpheme that was once pronounced (thus Boyarin, 1981) or just an orthographic convention (thus Brock, 2003; cf. Van Peursen, 2008:244).

4.6 An experiment

Our approach enables us to deploy a practice that is completely new in Syriac scholarship, namely the possibility of testing assumptions upon the data (cf. Talstra & Dyk, 2006). We can test, for example, what happens if we redefine the distribution of *ktb* and *ktbw* (cf. section 4.2) and take the zero ending and the *-w* as allomorphs for the 3rd person masculine.

In our model such a reinterpretation of the material can be registered formally by changing the relevant sections in the Word Grammar. Since the lemmatization is done automatically on the basis of the morphologically encoded text and a functional description of the morphemes, there is no need to change the lemmatization in all separate instances manually.

We have done this experiment for Judges 1 in nineteen manuscripts. This chapter contains 54 perfect forms 3m (except for third-weak verbs). In the bottom-up analysis (cf. section 2) the effect is that the decision on whether a 3m verb is singular or plural is not taken at word level, but at a later stage of the procedure, in which the verb is matched with a subject or another element that reveals its number.

At first sight the results of our experiment were not exciting. In those 26 cases where the grammatical number of the subject is unambiguous, the 'regular' forms are dominant: Only three times is there an irregular form (singular *ktbw* or plural *ktb*), once in one manuscript, twice in two

manuscripts.[6] Nevertheless, our experiment yielded some interesting observations.

In the first place we discovered that in 28 cases the grammatical number remained ambiguous even in the clause-level analysis because the subject was a collective noun (which in Syriac can take either a singular or a plural).

In these ambiguous cases the traditional analysis of *ktb* as a singular and *ktbw* as a plural implies a striking alternation of singular and plural forms, e.g. 1:10 'and Judah went (*w'zl*, singular) … and [they] *killed* (*wqṭlw*, plural)'. In our experiment, this became mere orthographic variation. Consequently, in the final stage of the bottom-up analytical procedure, the text hierarchical analysis (section 2.4), we arrived at a more elegant text hierarchical structure, because many of the recurrent subject changes caused by the singular/plural alternation had been resolved.

Secondly, the experiment overcame the rather arbitrary division between 'real' and orthographic variants in the Leiden Peshitta edition. In this edition, whenever there may be some doubt as to whether the verb is in the singular or in the plural, variation between *ktb* and *ktbw* forms is taken as 'real' and the variant is included in the critical apparatus; whenever there is no doubt, the variation is considered orthographic and the variant is listed in the *Index Orthographicus* (sample edition and vol. IV/6) or not mentioned at all (other volumes; cf. Dirksen, 1972:vii-ix).

This editorial policy leads to the somewhat arbitrary decision that *nḥt* 'descended' in 1:9 (Ms 16c1, 16g3; other manuscripts: *nḥtw*) is an orthographic variant, because the subject is the plural *bny yhwd'* 'sons of Judah, Judahites', whereas in 1:10, where the subject is just *yhwd'* 'Judah', *'zlw* 'went' (Ms 17a3; other manuscripts: *'zl*) is a real variant. In 1:26, the same form *'zlw* (Ms 19c1; other manuscripts have again *'zl*) is taken as orthographic, because the subject is the singular noun *gbr'* '(the) man'. In our experiment all these variant readings are treated equally.

5 Conclusions

We hope to have shown how the analytical procedure (section 2) and the workflow of the word-level analysis (section 3) meet the challenges of working with ancient documents (section 1), due to their form-to-function approach, their use of encoding rather than tagging, their distinction between paradigmatic forms and their realizations, and because of the exigencies of accuracy in the case of an ancient limited corpus.

In the word-level analysis we lean heavily on existing grammars. For that reason our approach could be regarded as theory-driven, even though we consider it one of our main tasks to revise and refine the paradigm on the basis of the actual corpora. Our encodings should be considered as hypotheses about the data that can be subjected to testing and experiment (section 4.6).

Unlike projects that concern the acceleration of POS tagging (Ringger *et al.,* 2007; Caroll *et al.*, 2007) we start one level below, with the morphology. 'Encoding rather than tagging' is not just a practical, but a crucial methodological characteristic of our model. (For new insights that it produced regarding Syriac morphology see the publications by Bakker, Van Keulen and Van Peursen in the bibliography). We differ from the computer implementation of morphological rules (Kiraz, 2001) in that our work is more deductive and focused on the interaction between orthography and morphology, because we start with the actual forms attested in the manuscripts. Our position in relation to these other projects is mainly determined by the philological demands that direct our research (see section 1).

Whereas at the morphological level the information provided by traditional grammars is relatively stable, at the higher linguistic levels they provide much less solid ground. The gradually built up datasets (analogous to the Analytical Lexicon at word level) of phrase patterns, clause patterns, or verbal valence contain much information that is not properly dealt with in traditional grammars. At these levels the analysis becomes more data-oriented. Thus in the analysis of phrase structure Van Peursen (2007) found many complex patterns that have not been dealt with in traditional grammars.

We have taken our examples from Syriac, but the same analytical procedures have been applied to other forms of Aramaic (Biblical Aramaic and Targum Aramaic) and Biblical Hebrew. Because of the separation of the analytical programs and the language-specific auxiliary files, it should be possible to apply it to other languages as well. This would mainly require writing the appropriate language definition files. Although our model is in principle language-independent, the morphological analysis presented in this paper is especially apt for Semitic languages because of their rich morphology.

[6] 26 forms × 19 manuscripts = 494 forms in all the manuscripts together. Accordingly, the 5 (1+2×2) irregular forms make up 1%.

Acknowledgments
This paper has benefited much from the valuable input of other members of the project 'Turgama: Computer-Assisted Analysis of the Peshitta and the Targum: Text, Language and Interpretation' of the Leiden Institute for Religious Studies; from the documentation of the *Werkgroep Informatica* of the Vrije Universiteit, Amsterdam; and from the collations of variant readings in the Peshitta manuscripts to Judges by Dr P.B. Dirksen in the archive of the Peshitta Institute Leiden.

References

Bakker, D., 2008. Lemma and Lexeme: The Case of Third-Aleph and Third-Yodh Verbs. Pp. 11–25 in FSL3.

Boyarin, Daniel. 1981. An Inquiry into the Formation of the Middle Aramaic Dialects. Pp. 613–649 in *Bono Homini Donum. Essays in Historical Linguistics in Memory of J. Alexander Kerns*, Vol. II. Edited by Y.L. Arbeitman and A.R. Bomhard. Amsterdam Studies in the Theory and History of Linguistic Science; Series IV: Current Issues in Linguistic History 16. Amsterdam: John Benjamins.

Brock, Sebastian P. 2003. Some Diachronic Features of Classical Syriac. Pp. 95–111 in *Hamlet on a Hill: Semitic and Greek Studies Presented to Professor T. Muraoka on the Occasion of his Sixty-Fifth Birthday*. Edited by M.F.J. Baasten and W.Th. van Peursen. Orientalia Lovaniensia Analecta 118. Leuven: Peeters.

Brockelmann, Carl. 1976. *Syrische Grammatik*. 12th edition. Leipzig: Verlag Enzyklopädie.

Carroll, James L., Robbie Haertel, Peter McClanahan, Eric Ringger, and Kevin Seppi. 2007. Modeling the Annotation Process for Ancient Corpus Creation. in Chatressar 2007, *Proceedings of the International Conference of Electronic Corpora of Ancient Languages (ECAL),* Prague, Czech Republic, November 2007.

Dirksen, P.B. 1972. *The Transmission of the Text of the Book of Judges*. Monographs of the Peshitta Institute Leiden 1. Leiden: Brill.

Dirksen, P.B. 1978. 'Judges', in *The Old Testament in Syriac according to the Peshiṭta Version* Vol. II/2 *Judges, Samuel*. Leiden: Brill.

Dyk, Janet W. and Wido van Peursen. 2008. *Foundations for Syriac Lexicography III. Colloquia of the International Syriac Language Project*. Perspectives on Syriac Linguistics 4; Piscataway, NJ. Gorgias. [= FSL3]

Heal, Kristian S. and Alison Salvesen. Forthcoming. *Foundations for Syriac Lexicography IV. Colloquia of the International Syriac Language Project*. Perspectives on Syriac Linguistics 5. Piscataway, NJ: Gorgias. [= FSL4]

Keulen, P.S.F. van, 2008. Feminine Nominal Endings in Hebrew, Aramaic and Syriac: Derivation or Inflection? Pp. 27–39 in FSL3.

Keulen, P.S.F. van and W.Th. van Peursen. 2006. *Corpus Linguistics and Textual History. A Computer-Assisted Interdisciplinary Approach to the Peshitta*. Studia Semitica Neerlandica 48. Assen: Van Gorcum.

Kiraz, George Anton. 2001. *Computational Nonlinear Morphology. With Emphasis on Semitic Languages*. Studies in Natural Language Processing. Cambridge: Cambridge University Press.

Nöldeke, Theodor. 2001. *Compendious Syriac Grammar*. Translated from the second improved German edition by J.A. Crichton. Winona Lake: Eisenbrauns.

Peursen, W.Th. van and Bakker, D. Forthcoming. Lemmatization and Grammatical Categorization: The Case of ܫܡܫ in Classical Syriac. In FSL4

Peursen, W.Th. van. 2008. Inflectional Morpheme or Part of the Lexeme: Some Reflections on the Shaphel in Classical Syriac. Pp. 41–57 in FSL3.

Peursen, W.Th. van. 2007. *Language and Interpretation in the Syriac Text of Ben Sira. A Comparative Linguistic and Literary Study.* Monographs of the Peshitta Institute Leiden 16. Leiden: Brill.

Peursen, W.Th. van. 2008. Language Variation, Language Development and the Textual History of the Peshitta. Pp. 231–256 in *Aramaic in its Historical and Linguistic Setting*. Edited by H. Gzella and M.L. Folmer. Veröffentlichungen der Orientalischen Kommission 50. Wiesbaden: Harrassowitz.

Peursen, W.Th. van. Forthcoming. Numerals and Nominal Inflection in Classical Syriac. In FSL4.

Ringger, Eric, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi and Deryle Lonsdale. 2007. Active Learning for Part-of-Speech Tagging: Accelerating Corpus Annotation. Pp. 101–108 in *Proceedings of the ACL Linguistic Annotation Workshop, Association for Computational Linguistics*. Prague, Czech Republic, June 2007.

Sikkel, Constantijn J. 2008. Lexeme Status of Pronominal Suffixes. Pp. 59–67 in FSL3.

Talstra, Eep, and Dyk, Janet W. 2006. The Computer and Biblical Research: Are there Perspectives beyond the Imitation of Classical Instruments? Pp. 189–203 in *Text, Translation, and Tradition*. Edited by W.Th. van Peursen and R.B. ter Haar Romeny. Monographs of the Peshitta Institute Leiden 14. Leiden: Brill.