Encoding Biblical Hebrew: Reflections on the Linguistic Theories Underlying the Andersen–Forbes System

Michael LANGLOIS

Review article of Francis I. Andersen and A. Dean Forbes, 2012. *Biblical Hebrew Grammar Visualized*. (Linguistic Studies in Ancient West Semitic 6). Winona Lake, Indiana: Eisenbrauns. xvii + Pp.394. Hardback. ISBN 978-1-57506-229-7.

Biblical Hebrew databases and grammars are not a novelty: numerous medieval treatises deal with grammatical features of the Hebrew Bible, providing statistics as to the number of occurrences of a given phenomenon. This can already be seen in the marginal notes that accompany the biblical text on Masoretic manuscripts.

The development of computer sciences in the twentieth century has paved the way for the creation of extensive computer databases of the Hebrew Bible, starting with the text itself — usually that of the Leningrad Codex rather than an eclectic edition or a text with critical apparatus. Lemmatisation enhances the textual database by identifying the various forms of a given lemma, thus enabling the user to perform lexicological queries. Morphological analysis encodes such features as part of speech, person, gender, number, state, aspect, and so on. The user is then able to search for all occurrences of a given pattern.

The Andersen-Forbes database has all these features, and already differs from other databases in this respect. But more importantly, the Anderson-Forbes database goes beyond the word level so as to encode syntactical relationships. Various types of constructions, phrases, clauses or sentences are identified throughout the biblical text, based on the authors' understanding of Biblical Hebrew syntax. These underlying principles are explained in their latest volume: Francis I. Andersen and A. Dean Forbes, *Biblical Hebrew Grammar Visualized* (Linguistic Studies in Ancient West Semitic 6), Winona Lake, Indiana: Eisenbrauns, 2012.

The first chapter, "Introduction," contains several prolegomena. The authors define "Biblical Hebrew" as the language of the biblical text according to the Leningrad Codex (p. 1). In other words, this codex serves as the textual basis for their database. This comes as no surprise, since L is the earliest complete Hebrew Bible known to us. The *Biblia Hebraica Stuttgartensia* (BHS), and now the *Quinta* (BHQ), chose this codex as their main source; most biblical databases are likewise based on L.

Choosing a biblical manuscript as a main source is one thing; defining "Biblical Hebrew" as the language of a single manuscript is another. The Dead Sea Scrolls have revealed (or rather confirmed) that the Masoretic text belongs to but one of several traditions, some of which are

more ancient and thus closer to the language of the biblical writers. The language of the Samaritan Pentateuch, for instance, is as much *Biblical Hebrew* as that of the Masoretic text — if not more, judging by its vocalisation.

There's more. Assuming that the Masoretic tradition is preferred over others, and that *Biblical Hebrew* is understood as *Masoretic Biblical Hebrew* or *Leningrad Codex Hebrew*, it is quite surprising that the authors opted to keep all "*Kethiv* readings, setting aside the *Qere* variants" (p. 2). *Qere* readings are not only part of the Masoretic tradition; they are what the Masoretes consider to be the proper reading, over — and in spite of — the reading suggested by the consonantal text. By choosing *Kethiv* readings, Andersen and Forbes go against the Masoretic tradition. This inconsistency has immediate repercussions: how should one parse unvocalised *Kethiv* readings? Whereas the Masoretic vocalisation and cantillation system differentiates otherwise homographic forms (*e.g.*, *Qal* or *Piel*), *Kethiv* readings are more often ambiguous. In the Andersen-Forbes database, the latter have "been vocalized in accordance with Gordis" (p. 3). Why is a hypothetical vocalisation preferred over the factual Masoretic vocalisation of the *Qere*? Is it an attempt to recover a "better" or "earlier" text? Why, then, trust the Masoretic vocalisation elsewhere and not suggest a "better" one? Why exclude "earlier" witnesses such as the Dead Sea Scrolls?

The authors are aware of this issue, and emphasise that their database might later be improved so as to "make two representations of each clause that contains a *Qere / Kethiv* pair of words that differ in syntax" (p. 3). Hopefully, other variant readings will be included as well, especially those attested by the Dead Sea Scrolls and the Samaritan tradition.
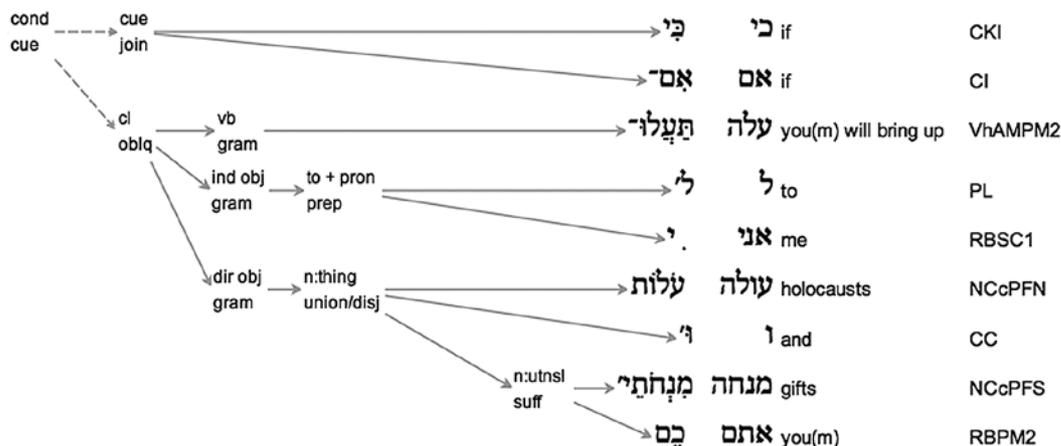
After having defined "Biblical Hebrew," the authors talk about "Grammar" (§ 1.2). Acknowledging the "extensive treatments of morphology" carried out by previous grammarians, Andersen and Forbes see the need for a wider approach: "Our major working units are whole clauses" (p. 5). They are aware of Joüon's or Waltke and O'Connor's recent works on syntax, but "the treatment remains at the level of microsyntax" or "short-range syntactic functions" (p. 8). In order to account for the internal syntax of complete clauses, the authors propose to build enhanced phrase markers linking each constituent to its neighbours according to their syntactical relationships. A graphical representation is then drawn by means of labels and arrows, thus allowing the reader to "visualize" the grammatical structure of the Hebrew text. This is Hebrew grammar "visualized" — hence the volume title.

Chapter 2 deals with "Text Division" (p. 15): on the one hand, words can be composed of several segments (for instance, בַּיּוֹם = preposition + definite article + noun); on the other hand, several words can be ligatured to form a proper noun (for instance, בֵּית־אֵל = Bethel). The authors insist that even such a lexicalised compound as לִפְנֵי "before" is segmented into preposition + noun (p. 16); it seems quite inconsistent, then, that the words כִּי אִם are said to be ligatured on the mere assumption that they function together as a subordinating conjunction meaning "except" (p. 17). But this is not always the case; let us look at Amos 5:22a:

כִּי אִם־תַּעֲלוּ־לִי עֹלוֹת וּמִנְחֹתֵיכֶם לֹא אֶרְצֶה

"‹It is› that if you raise up to me burnt offerings and grain offerings, I will not take pleasure."

Understanding כִּי אִם as a conjunction meaning "except" would imply that Yhwh *does* take pleasure in the people's burnt and grain offerings, which is the opposite of what the author is saying (see vv. 21*ff.*). I checked the Andersen-Forbes database (version 0.97) on this verse:

cond	cue		כִּי if	CKI
cue	join		אִם if	CI
	cl / oblq	vb gram	תַּעֲלוּ you(m) will bring up	VhAMPM2
		ind obj gram → to + pron prep	לְ to	PL
			אֲנִי me	RBSC1
		dir obj gram → n:thing union/disj	עֹלוֹת holocausts	NCcPFN
			וּ and	CC
		n:utnsl suff	מִנְחֹתַי gifts	NCcPFS
			אַתֶּם you(m)	RBPM2

Indeed, כִּי and אִם are not grouped together as a single segment and are both translated "if," which indicates an awareness that כִּי אִם is not understood as meaning "except." But the Andersen-Forbes database allows for multiple translations of a segment; כִּי אִם could thus be taken as a single segment translated "if." Alternatively — and preferably — כִּי and אִם could always be taken as two segments: כ introduces the subordinate clause and אִם emphasises the conditional nature of this clause. I don't see the need to ligature them into one segment, and it seems inconsistent to do so while even לִפְנֵי is divided into two segments.

In Chapter 3, "Parts of Speech" (p. 20), the authors present the reader with their system of segment grammatical categories. Beyond the major traditional categories, such as "conjunctions," "prepositions," "substantives," "verbs", *etc.*, Andersen and Forbes propose to distinguish no less than 37 grammatical categories for Biblical Hebrew. Nouns, for instance, are divided into eight categories, including a dedicated category for a single lexeme, כֹּל *all*; this is due, we are told, to "its odd behavior" (p. 24). Notwithstanding its abundant use, כֹּל is certainly not the only word in the Hebrew Bible exhibiting an "odd" behavior. The very purpose of any taxonomy is to group elements according to their shared features; subcategories are useful as long as they highlight features shared by several but not all elements of a category. Unfortunately, the reader quickly learns that Andersen and Forbes did not stop here: their system ends up (p. 25) with 76 parts of speech! Major prepositions such as בְּ, כְּ, לְ, מִן *etc.*, all seem to deserve their own category. Some segments are even given two categories: אִם, for instance, belongs either to the "*if* אִם" category or to the "[question] אִם" category. In fact, it may even belong to a third category named "*not* אִי / אִם / לֹא." One may wonder, then, why other words are not rewarded with their own part(s) of speech, left as they are with such categories as "other conjunctions" or "other prepositions."

Again, if the purpose is to "facilitate further study," as suggested by the authors about כֹּל (p. 24), a computer database will easily combine multiple search criteria, including the specification of a given lemma such as כֹּל. Likewise, searching for occurrences of אִם tagged as "conjunction" (as opposed to "preposition," for instance) is simple and avoids the creation of an unnecessarily complex and arbitrary taxonomy that includes single-lexeme categories such as "*if* אִם."

Other comments could be made on the Andersen-Forbes 76 parts-of-speech taxonomy. For instance, תּוֹך is considered to be a "preposition" — it even receives its own one-lexeme category ("*inside* תּוֹך" p. 25) — rather than the construct state of תָּוֶך "midst"; this seems inconsistent with the authors' previous statement that לִפְנֵי "before" is to be segmented into preposition + noun (p. 16). Perhaps תּוֹך is to be tagged as a "preposition" only when it functions as such without another preposition; I checked the Andersen-Forbes database: this is unfortunately not the case. Here is an example from Genesis 3:8b, where Adam and Eve hide מִפְּנֵי יְהוָה אֱלֹהִים בְּתוֹךְ עֵץ הַגָּן "from the face of Yhwh Elohim, in the midst of the tree‹s› of the garden." Both פָּנֶה "face" and תָּוֶךְ "midst" are used in the construct state and prefixed with a preposition so as to become prepositional compounds. Yet, the Andersen-Forbes database sees תּוֹך as a preposition, not a noun:

mvt orig gram	from + deity prep		מְ	מן from	PM
		n:deity cstr	פְּנֵי	פן face of	NCcPMC
		n:deity app	יהוה	יהוה Yahweh	NPDSMN
			אֱלֹהִים	אלהים God	NPDSMN
loc gram	in + veg prep	in join	בְּ	ב in	PB
			תוֹך	תוך inside	PT
		n:veg cstr	עֵץ	עץ tree of	NCcSMC
		n:geog def	הַ	ה the	XD
			גָּן	גן garden	NCcSMNH

By failing to identify the parallel syntax, the Andersen-Forbes database will miss a number of occurrences searching for patterns like "preposition + noun in the construct state." This will have repercussions when studying various types of phrases and clauses in Biblical Hebrew syntax, as found later in the volume. For instance, the authors state that "of the 74,058 prepositions in Biblical Hebrew, 24,973 (34%) are part of basic prepositional phrases" (p. 53), 43.8% of which combine with a noun (p. 54). The precision of these numbers and percentages is deceptive, as they are simply inaccurate.
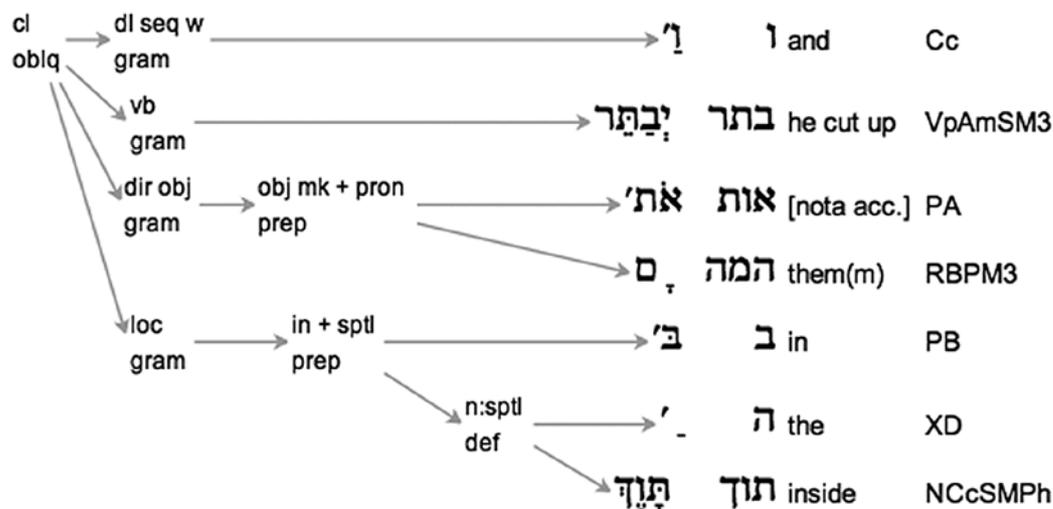
As far as prepositional compounds are concerned, the authors' policy is to leave "compounds unsegmented *if* their nominal components were never attested with nominal functions and their

original literal meaning" (p. 27). This principle itself is questionable: how do we know *a priori* that a nominal component is never attested with a nominal function? Working on a closed corpus certainly helps (especially if one excludes variant readings), but this remains problematic from a methodological and epistemological standpoint. Now, assuming we accept this principle, what about תּוֹךְ? From a morphological standpoint, it is the construct state of תָּוֶךְ, a segholate noun of the *qatl pattern from an ע״ו root (*cp.* מָוֶת → מוֹת "death"; the diphthong *aw contracts to *ô*). Now, תָּוֶךְ *is* attested with a nominal function and its original literal meaning; see for example, Genesis 15:10:

<div dir="rtl">

וַיִּקַּח־לוֹ אֶת־כָּל־אֵלֶּה וַיְבַתֵּר אֹתָם בַּתָּוֶךְ
</div>

"He took for him all these and cut them in the middle."

I checked again the Andersen-Forbes database:

In this verse, תָּוֶךְ is correctly identified as a common noun. In fact, it appears that the Andersen-Forbes database has two distinct lemmas, a preposition and a noun. This is really problematic from both a lexicological and grammatical standpoint, and one can only hope that the Andersen-Forbes database (as well as their underlying grammar) will be corrected as soon as possible.

When it comes to proper nouns, the 76 parts-of-speech taxonomy includes such categories as "land proper nouns," "mountain proper nouns," "city proper noun," or "river proper noun." While I find it very useful to be able to search for specific kinds of toponyms, the purpose of a part-of-speech taxonomy is to group or separate segments according to their syntactical — not semantic — features. The parts of speech mentioned above deal with semantics, not syntax: proper nouns designating cities function the same way as those designating land; they should not be attributed two different parts of speech.

This is not to say that encoding semantics is useless; on the contrary, one of the often underestimated values of the Andersen-Forbes database is its semantics feature presented on pp. 38*ff*.
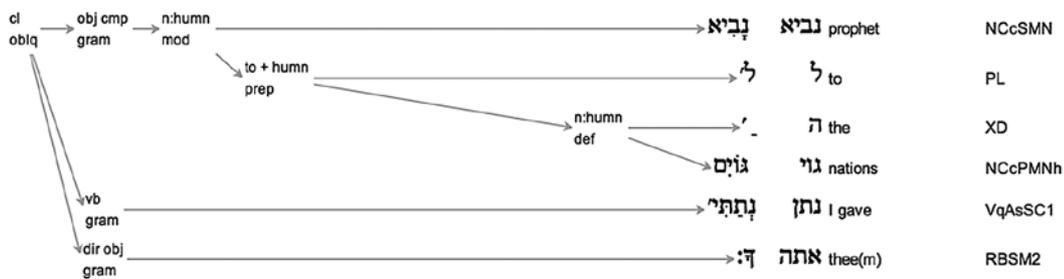
But this feature is *not* the same as the part-of-speech taxonomy discussed above. As a result, semantic data is sometimes encoded twice: עֵדֶן is tagged as "other geog. proper nouns" according to the 76 parts-of-speech taxonomy (p. 25), and as "geographical name or feature" according to the semantic taxonomy (p. 38). This is yet another example showing that the Andersen-Forbes part-of-speech taxonomy is flawed. In their defence, I should mention that the semantics feature was added later, "in the mid-1980s to assist computer parsing" (p. 39). The inclusion of "semantic" subcategories in their part-of-speech system may thus have been an early attempt at encoding semantic features (although this is not clearly stated). Moreover, the authors confess that their new semantics feature is not (yet) based on a principled taxonomy: "When we assigned the semantic codes, principled taxonomies were beyond our ken. The introduction of enriched, even multivalued semantic labels is one of our (too-populated) priorities" (p. 39). I would suggest that, since the semantics taxonomy will be updated, so should the part-of-speech taxonomy.

In the 76 parts-of-speech system, participles are also rewarded with several subcategories: "pure noun participles" is a subcategory of the "substantives" category, while "noun-verb / noun participles," "noun-verb participles," and "pure verb participles" are subcategories of the "verbals" category (p. 25). Later in the chapter (pp. *32ff.*), these various types of participles are explained and illustrated with examples. One of them is 1 Samuel 8:1:

וַיְהִי כַּאֲשֶׁר זָקֵן שְׁמוּאֵל וַיָּשֶׂם אֶת־בָּנָיו שֹׁפְטִים לְיִשְׂרָאֵל׃

"And it was, as Samuel was old, ‹that› he put his sons ‹as› judges for Israel."

According to the authors, שֹׁפְטִים is an example of a noun-verb participle: it "has its own beneficiary" (p. 34), לְיִשְׂרָאֵל. It differs from a "pure noun participle" that exhibits only nominal characteristics. This example is not convincing, as the same sentence could have been constructed with a common noun such as, let's say, נָבִיא "prophet": וַיָּשֶׂם אֶת־בָּנָיו נְבִיאִים לְיִשְׂרָאֵל "he put his sons ‹as› prophets for Israel." Cautious readers might question my example as being theoretical; while such tests are useful to check grammatical theories, let me provide an actual example from the Hebrew Bible. Jeremiah 1:5b reads נָבִיא לַגּוֹיִם נְתַתִּיךָ "I have given you ‹as› a prophet for the nations"; here is the Andersen-Forbes phrase marker:

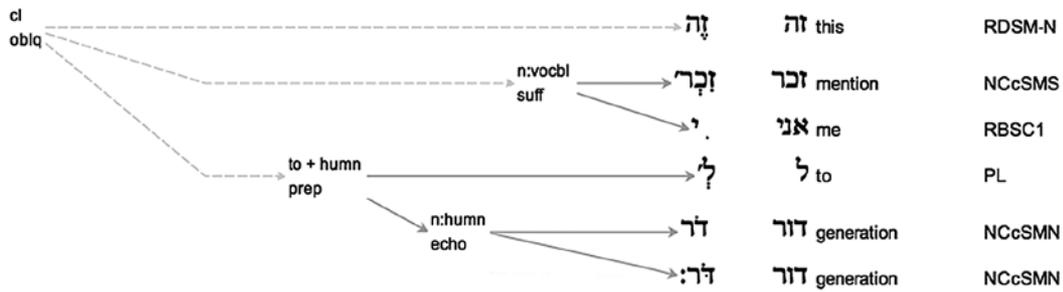|						
cl oblq	obj cmp gram	n:humn mod		נְבִיא	נביא prophet	NCcSMN
			to + humn prep	לְ	ל to	PL
				n:humn def	־ַ	ה the	XD
					גּוֹיִם	גוי nations	NCcPMNh
	vb gram			נְתַתִּי	נתן I gave	VqAsSC1
	dir obj gram			ךָ׃	אתה thee(m)	RBSM2

As in 1 Samuel 8:1, we have an object complement ("obj cmp") consisting of a noun followed by a "to + humn" prepositional phrase. Since such a construction exhibits no verbal characteristic, the occurrence of שֹׁפְטִים in 1 Samuel 8:1 should be a "pure noun participle" and not a "noun-verb participle" according to the Andersen-Forbes part-of-speech taxonomy.

After having discussed "Parts of Speech" in Chapter 3, the authors turn to "Phrase Marker Concepts and Terminology" in Chapter 4 (p. 43). They present in detail their graphical representation system and explain the way in which phrase markers were automatically generated using various rules, taking into account semantic information. Of course, the authors are well aware of the limitations of computer parsing — however sophisticated it may be: "Corrections, extensions, and consistency enforcement are the ongoing work of human over-readers" (p. 49). They are also aware that alternative analyses could be offered, especially when the text is ambiguous: "We plan to restore and represent ambiguity in later releases of our data" (p. 45).
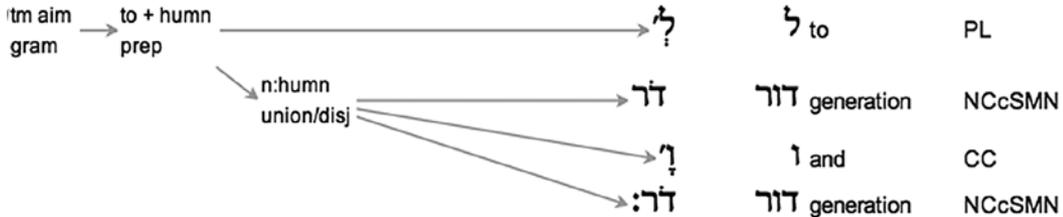
In light of these two limitations, it would be useful for end-users to be able to change a phrase marker if an error is detected or if an alternative parsing is preferred. In fact, alternative parsing would not even need to replace the one provided by the database, since we later learn that the "representational apparatus is designed to allow for multiple parses" (p. 297). Multiple parses would moreover strengthen the reliability of statistical data: I mentioned earlier that giving very precise percentages is deceptive if these figures are inaccurate or subject to variation; taking into account variant parses would lead to a range of percentages and thus to a margin of error. If the margin is too wide, statistics lose their significance; if, on the contrary, the margin is narrow, statistics gain in reliability.

Chapter 5 presents "Basic Phrase Types of Biblical Hebrew" (p. 50). The chapter is well organised, starting with the simplest phrases such as אֵם יַעֲקֹב "mother of Jacob" (p. 50; note the uncorrected change of Hebrew font in the paragraph). Phrases that merely involve the prefixation of a definite article or the suffixation of pronoun are called "tightly joined phrases" (p. 52). Phrases that exhibit no conjoining are called "unconjoined phrases" (p. 53); they include construct phrases, prepositional phrases, *etc*. Conjoined phrases include coordinate phrases (which the authors call "union or disjoint phrases," p. 56, to emphasise that coordinating conjunctions can be disjunctive), juxtaposed phrases (when coordination is implicit; *i.e.*, no conjunction is used), and mixed phrases (when more than two elements are coordinated, some but not all with a conjunction).

The authors then focus on juxtaposed phrases in which the elements share an identical reference; for example, יהוה אֱלֹהִים "Yhwh Elohim" (p. 59). Rather than simply tagging these phrases as "juxtaposed phrases," a new phrase type was created: "apposition." If the elements in apposition also exhibit identical form and function (*e.g.*, מֹשֶׁה מֹשֶׁה "Moses, Moses"), they are rewarded with a new phrase type called "echo." While I find it very useful to tag such phrases, the organisation is problematic: syntactically, "echo" phrases are a subset of "apposition" phrases, which are themselves a subset of "juxtaposed" phrases; but in the Andersen-Forbes taxonomy, they are independent. As a result, searching for "juxtaposed phrases" in their database will *not* return "apposition" or "echo" phrases, even though they *are* juxtaposed phrases. Likewise, looking for "apposition" phrases will wrongly exclude "echo" phrases. But that's not the only problem: if I am interested in the phenomenon of "apposition" (*i.e.*, "a construction wherein two or more constituents have an identical reference," p. 62), why should I exclude union phrases? The same remark goes for the phenomenon of "echo." For instance, the end of Exodus 3:15 reads: וְזֶה זִכְרִי לְדֹר דֹּר "and this ‹is› my memorial for generation-generation." Here is the corresponding Andersen-Forbes phrase marker:

cl oblq		זֶ֫ה זה this	RDSM-N
	n:vocbl suff	זִכְרִ֫ זכר mention	NCcSMS
		יִ֫ אני me	RBSC1
to + humn prep		לְ֫ ל to	PL
	n:humn echo	דֹ֫ר דור generation	NCcSMN
		דֹֽר׃ דור generation	NCcSMN

The last two words, דֹר דֹר, constitute an "echo" phrase, as expected. Now, let's look at the second half of Psalms 33:11: מַחְשְׁבוֹת לִבּוֹ לְדֹר וָדֹר "The plans of his heart ‹are› for generation and generation." Here, דֹר is also written twice, but the two occurrences are joined by a coordinating conjunction. Here is the Andersen-Forbes phrase marker:

tm aim gram →	to + humn prep	לְ֫ ל to	PL
	n:humn union/disj	דֹ֫ר דור generation	NCcSMN
		וָ֫ ו and	CC
		דֹֽר׃ דור generation	NCcSMN

The phrase is tagged as a "union or disjoint phrase" (*i.e.*, a coordinate phrase), because the two nouns are indeed explicitly conjoined. But the fact that the two coordinated terms are identical is not tagged in any way. The Andersen-Forbes system of phrase types is thus inconsistent: the subset of juxtaposed phrases involving identical terms deserves a specific phrase-type ("echo"), but not the subset of coordinate phrases involving identical terms. This inconsistency would not have occurred if "echo" had not been a phrase type *per se* (after all, these phrases could just be parsed as "juxtaposed phrases" or "union or disjoint phrases" without the need for other phrase types) but rather a tag that can be added to either of these phrase types. Earlier, I mentioned problems with the authors' part-of-speech system due to its confusion of morphological and semantic features. Similar problems now appear with their phrase-type system due to its confusion of syntactical and semantic features.

The failure to identify variant expressions such as לְדֹר דֹר and לְדֹר וָדֹר becomes even more obvious when they appear in a parallel context or, worse, in the same verse. For instance, Samaritan manuscripts read לדור ודור at the end of Exodus 3:15, as opposed to Masoretic manuscripts where there is no ו conjunction. Likewise, the end of Proverbs 27:24 is לדור דור according to the consonantal text, but the Masoretes read לְדוֹר וָדוֹר, adding a ו conjunction. It is unfortunate, as I mentioned earlier, that the database does not take into account those variant readings; but even if we limit ourselves to the Leningrad Codex consonantal text, we could at least expect the database to identify similar cases of repetition (or "echo"), whether the elements are separated by a coordinating conjunction or not.

Phrase types remain the focus of Chapter 6, which deals with "Complex Phrases" (p. 64). As opposed to a basic phrase, whose constituents are segments, a complex phrase has among its constituents at least one phrase. Various subtypes of complex phrases are presented, including rare or

interesting cases. The authors point out "layout engine flaws" (p. 71) — hence the arrows that cross other arrows on a phrase marker (p. 72; see also pp. 77, 78, 79, *etc.*), a problem mentioned again on p. 294. This is a minor issue, and does not affect the value of their database; more problematic is the fact that, quite often, "alternate parses are also defensible" (p. 66). The authors are well aware of this issue: "In the building up of complex phrases, multiple equally valid parses are often possible" (p. 76). One may wonder, then, why they provide tables summing up the number of occurrences for all patterns of coordinated phrases having two or three constituents (pp. 75–76). The apparent precision of those numbers (*e.g.*, 134 S+S+S phrases or 141 PoP+S phrases) is misleading and, until a margin of error can be assessed, these statistics can hardly be trusted.

Having discussed segments and phrases, the authors turn to clauses, divided into "main clauses" (Chapter 7, p. 86) and "embedded clauses" (Chapter 8, p. 98). Interacting with previous (including recent) works in the field, Andersen and Forbes discuss such issues as word order, discontinuity, null anaphora, *etc*. They provide statistical data for the ordering of subject (S), verb (V) and direct object (O): the most frequent sequence appears to be VSO (42.7%) followed by SVO (33.5%) and, at a distance, by VOS (14.4%) and the three other possible sequences (p. 89). As I explained above, it would be important to estimate the margin of error, but the authors are correct in concluding that "the oft-repeated assertion that Biblical Hebrew is a VSO language is an insufficient description" (p. 89). Indeed, SVO clauses are less than 10 points away from VSO clauses. Moreover, *wayyiqtol* or *weqatal* verb forms constrain word order and account for a number of VSO or VOS clauses; they are "anchored" predicators (p. 157). Setting those forms aside, the VSO sequence drops to 30.2% while the SVO sequence rises to first place at 44.8% (p. 89). VOS clauses remain in third position at 12.4%; adding those to the VSO clauses, we get 42.6% of clauses in which V precedes S and O, compared to 47.5% of clauses in which S precedes V and O. In other words, when anchored predicators are set aside, Biblical Hebrew seems *not* to favour V-initial sequence (and VSO in particular).

But these numbers are misleading: their margin of error has not been estimated (V-initial and S-initial clauses are just 5 points away from each other), and they account only for clauses that have both a subject and an object; in order to determine the frequency of S-initial and V-initial sequences, one must take into account clauses without an object. And if V-initial clauses are studied without comparison to S-initial clauses, one must take into account clauses without a subject. It is unfortunate that the authors do not provide these numbers; they do, however, indicate in a footnote (p. 89 n. 18) that "when we examine clauses with unanchored predicator plus either a subject or an object, initial-predicator incidence increases substantially. The VO sequence occurs in 80% of cases, and the VS sequence occurs in 62% of cases." If actual numbers (rather than percentages) had been provided, cumulative percentages could have been computed and may have proved wrong the impression that Biblical Hebrew does not favour V-initial sequence.

When discussing syntactically discontinuous expressions, the authors give examples of phrase markers exhibiting tangling (pp. 90–91). While I agree that Biblical Hebrew does exhibit discontinuity, this is not true of some of the cases that are given as examples. For instance, 1 Samuel 31:2 is provided as a case of distributed apposition:

אֶת־יְהוֹנָתָן וְאֶת־אֲבִינָדָב וְאֶת־מַלְכִּי־שׁוּעַ בְּנֵי שָׁאוּל׃

"Jonathan and Abinadab and Malkishua, sons of Saul."

Indeed, בְּנֵי שָׁאוּל refers to יְהוֹנָתָן, אֲבִינָדָב and מַלְכִּי־שׁוּעַ, and since יְהוֹנָתָן is not adjacent, the authors consider it a case of discontinuity. But if אֶת־יְהוֹנָתָן וְאֶת־אֲבִינָדָב וְאֶת־מַלְכִּי־שׁוּעַ are considered a single phrase (more specifically, a complex union phrase), it appears that this phrase is immediately followed by בְּנֵי שָׁאוּל. There is no discontinuity, and no need for tangling.

The same observation can be made for 1 Chronicles 22:13, whose phrase marker fragment exhibits tangling on p. 104. The text reads as follows:

אֶת־הַחֻקִּים וְאֶת־הַמִּשְׁפָּטִים אֲשֶׁר צִוָּה יְהוָה אֶת־מֹשֶׁה עַל־יִשְׂרָאֵל

"the precepts and the judgments that Yhwh commanded Moses upon Israel."

The authors argue that אֲשֶׁר refers not only to מִשְׁפָּטִים but to חֻקִּים as well, hence the extraposition (*cf.* p. 103) and tangling. Yet, by considering the phrase אֶת־הַחֻקִּים וְאֶת־הַמִּשְׁפָּטִים as the head for the nominalised clause introduced by אֲשֶׁר, tangling disappears and, contrary to what the authors believe, the clause is not "inherently extraposed" (p. 104). The irony is that the authors themselves state that numerous cases of extraposition in Holmstedt's list can be resolved by "allowing phrases to be heads for nominalized clauses" (p. 103). It is very unfortunate that on the following page they present 1 Chronicles 22:13 as one of "94 instances of extraposed nominalized clauses that are not in Holmstedt's list"!

In the same chapter, the authors discuss "overtly and covertly headed nominalized clauses" (p. 101), "restrictive and nonrestrictive nominalized clauses" (pp. 101–102), and "resumption" (pp. 102–103). These phenomena are illustrated with well-chosen examples and their accompanying phrase markers. I noticed that some of these phenomena do not appear to be tagged in any way in the Andersen-Forbes system. For instance, it would be very useful to improve the taxonomy so as to distinguish between restrictive and nonrestrictive nominalised clauses. As for resumptive pronouns, I was surprised to find out later (in Chapter 9) that Andersen and Forbes tag "resumption" as a special characteristic of some clause constituents (p. 130). Since the authors are well aware of the "resumption phenomena" (p. 102), it seems inconsistent that their suspension/resumption tagging does not apply here (compare phrase markers 8.6 and 9.27, pp. 106 and 131 respectively). Let's hope that this shortcoming will soon be corrected. Suspension and resumption could even be represented by some kind of tangling in phrase markers: an arrow could connect a resumptive pronoun with the substantive to which it refers.

The issue of tangling also occurs in the discussion of "noun-verb participles" (p. 109); according to the authors, out of the 2,413 occurrences, "29 participles are parts of non-tree phrase markers, 27 having two mothers and two having three mothers" (p. 109). One of these two occurrences is Psalms 106:21–22:

שָׁכְחוּ אֵל מוֹשִׁיעָם עֹשֶׂה גְדֹלוֹת בְּמִצְרָיִם׃ נִפְלָאוֹת בְּאֶרֶץ חָם נוֹרָאוֹת עַל־יַם־סוּף׃

"They forgot El, their savior, doing great ‹deeds› in Egypt, wonderful ‹deeds› in the land of Ham, awesome ‹deeds› upon the Red Sea."

Their phrase marker exhibits tangling, but this is due to the fact that עֹשֶׂה is part of a clause limited to עֹשֶׂה גְדֹלוֹת בְּמִצְרָיִם. But the clause should rather be extended to the end of v. 22: עֹשֶׂה has a threefold complement consisting of three juxtaposed phrases: (1) גְדֹלוֹת בְּמִצְרָיִם; (2) נִפְלָאוֹת בְּאֶרֶץ חָם; and (3) נוֹרָאוֹת עַל־יַם־סוּף. These phrases exhibit a parallel structure, with an indefinite feminine plural adjective followed by a preposition introducing a toponym. I would even embed

them together into a single phrase that complements עֹשֶׂה (*cp.* the superset node p. 274), but if one wants to separate direct objects from locative adjuncts, it is still possible to represent עֹשֶׂה followed by six constituents (dir obj, adjunct, dir obj, adjunct, dir obj, adjunct) without resorting to tangling.

This, of course, depends on the system adopted for classifying clause constituents, which is the subject of Chapter 9, "Classifying Clause Immediate Constituents" (p. 113). Five CIC subtypes are introduced: impermanents, syntactic isolates, predicators, operators, and grammatical functions and semantic roles. These subtypes are said to be "exhaustive and mutually exclusive" (p. 114), but the name given to the fifth subtype, "grammatical functions and semantic roles," betrays its hybrid nature. Indeed, whereas some of the CICs will simply be tagged with "grammatical functions," others will be tagged with "semantic roles." This is unfortunate, especially since CICs that are tagged with "grammatical functions" do in fact have semantic roles. The authors are aware of this issue: they acknowledge that this is a "mixed representation" that is used "on an interim basis" (p. 115). Yet, it is reminiscent of the issues I highlighted earlier concerning the part-of-speech and phrase-type systems because of their confusion of morphological and syntactical features with semantic features. As we now reach the clause level, it is unfortunate that the author's taxonomy exhibits a similar flaw. Indeed, at the end of their discussion of "Semantic Role CICs" in Chapter 10 (p. 135), the authors state in a footnote that "assessing the adequacy of our taxonomies of parts of speech and of semantic roles is an iterative process" (p. 150 n. 48).

Having introduced CICs in their various subtypes and semantic roles, Andersen and Forbes focus on their use (including presence and ordering) in clauses featuring a given verb. Chapter 11 (p. 152) illustrates their methods with חפץ, whose corpus is small, before applying them to very frequent verbs: אמר (Chapter 12, p. 170), היה (Chapter 13, p. 186), עשה (Chapter 14, p. 196), and נתן (Chapter 15, p. 207). These chapters are well illustrated, with few typographical errors; for example, the use of a different Hebrew font in tables on pp. 178–179 or in charts on p. 213 (one of which is too pixelated). Numerous syntactical features are mentioned; for instance, "the אמר corpus exhibits a strong alternation in realising indirect objects: אל (66%) versus ל (34%)" (p. 173), which is very different from נתן, whose "corpus exhibits mild alternation in realizing indirect objects: ל (91.4%) versus אל (4.3%)" (p. 209). Likewise, כֹה is by far the most frequent adverb of manner used in the אמר corpus (p. 174), while the עשה corpus mostly uses כֵן (p. 198). The Andersen-Forbes database also shines in identifying deep speech embedding: four-level embedding occurs more than 20 times, while five-level embedding occurs only in Jeremiah, with three occurrences (p. 177).

The statistical results presented in these chapters depend, however, on the reliability of the data and on the adequacy of the taxonomies. The problematic mixed representation of grammatical functions and semantic roles mentioned above quickly resurfaces, which prompts the authors to conclude that this "makes it all the more important that we implement the full representation as soon as possible" (p. 200).

Chapter 16 examines the "Makeup of Clause Immediate Constituent Subtypes" (p. 218) independently of the verb being used, while Chapter 17 introduces methods for "Computing the Distances Between Verb Corpora" (p. 232). The presentation is clear, except for a few font changes on the charts pp. 233*ff.* More importantly, one should bear in mind that verb clustering derives from, and therefore reflects, the factors used to compute distances between verbs. These factors are in turn related to the part-of-speech, phrase-type and CIC systems. For instance,

VLC ("verbless clauses") and QV ("quasiverbal clauses") form a cluster (p. 249) because they share similar syntactical features; indeed, a quasiverbal is by definition "a segment that does not have verb morphology, but functions as a predicator" (p. 368), so that quasiverbal clauses are verbless clauses — especially since Andersen and Forbes conveniently refrain from identifying as "quasiverbals" occurrences of these terms in verbal clauses (see my remarks below). Failing to bear this in mind leads to circular reasoning, as when the authors state in their conclusion: "We find, for example, that the verbless corpus has much in common with the quasiverbal corpus" (p. 250). What appears to be a conclusion derived from the dendrogram is, in fact, the result of the authors' prior creation of a part-of-speech category defined by such behaviour.

The reason for creating such a part of speech is explained in the following chapter, "The Five Quasiverbals" (p. 251). These five segments are, namely: יֵשׁ, אֵין, עוֹד, הִנֵּה, and אַיֵּה. They do not always function as quasiverbals: "Approximately 720 of these lexemes combine with verbal elements to produce compound predicators … Here we deal with the 1,213 clauses in which these items appear as simple predicators" (p. 251, *sic*; the first sentence should be corrected to something like "720 *occurrences* of these lexemes"). Indeed, these lexemes will be given various parts of speech depending on their function; hence, אֵין is not consistently parsed as the construct state of אַיִן (see p. 251 n. 8 where these forms, and even the pausal אָיִן, are presented as "homographs"). This is another example of the problematic part-of-speech system developed by Andersen and Forbes: I used earlier תּוֹךְ as an example of a lexeme classified as a "preposition" rather than the construct state of תָּוֶךְ "midst"; the same problem occurs here. Moreover, just as תּוֹךְ had its one-lexeme part-of-speech category, אֵין also receives its own one-lexeme part-of-speech category, defeating once again the purpose of creating a category. In fact, each of the five quasiverbals has its own one-lexeme category, but they are not on the same level. For instance, the "אַיֵּה *where?*" subcategory is on the third level — that of the "Adverbial Subclass," as seen on Andersen-Forbes database screenshot below:

Surprisingly, the occurrences of this one-lexeme adverbial subclass do not belong to the "Quasiverbal" adverbial family, but to the "Interrogative" adverbial family. עוֹד, on the other hand, does not appear only as an "עוֹד *still*" adverbial subclass: it also appears as the "עוֹד *again*" adverbial family, and again as the "עוֹד *still*" adverbial family, as the following screenshot indicates:

@A[zsQ]S

Part of Speech	Adverbial Family			Adverbial Subclass	
Adverbial	"Above" (מֵעַל)	"Now" (עַתָּה)	"Together" (יַחְדָּו)	"Behold!" (הִנֵּה)	"Who?" (מִי)
Conjunction	"Afterwards" (אַחַר)	"Only" (אַךְ)	"Underneath" (תַּחַת Or מַטָּה)	"Do Not" (אַל)	"Why?" (מַדּוּעַ)
Noun	"Again" (עוֹד)	"Only" (רַק)	"Upwards" (מַעְלָה)	"Exists" (יֵשׁ)	"Yes Or No?" (הֲ)
Preposition	"Already" (כְּבָר)	"Positively" (אָמְתַם)	"Very" (מְאֹד)	"How?" (אֵיךְ)	Temporal
Pronoun	"Apart" (לְבָד)	[Question] (אִם)	"Yet" (עֲדֶן)	Negation (בְּלִי)	
Verb	"Back" (אָחוֹר)	"Quickly" (חִישׁ)	Emphatic	"Not" (לֹא)	
Miscellany	"Certainly" (לֹא)	"Still" (עוֹד)	Interrogative	"Not Exists" (אֵין)	
	"Eagerly" (אָסְפַּרְנָא)	"Suddenly" (פִּתְאֹם)	Modal	"Still" (עוֹד)	
	"Farther" (הָלְאָה)	"Then" (אֵפוֹא)	Negative	"What?" (מָה)	
	"Here" (הֵנָּה)	"Then" Or (אֲדַיִן / אָז)	Quasiverbal	"When?" (מָתַי)	
	"Inadvertently" (הַצָּדָא)	"There" (שָׁם)		"Whence?" (מֵאַיִן)	
	"Not Yet" (טְרֶם)	"Thither" (שָׁמָּה)		"Where?" (אַיֵּה)	
	"Now" (טֶרֶם)	"Thus" (כֵּן Or כֹּה)		"Where?" (אָנָה)	

As a result, searching for the "עוֹד *still*" adverbial subclass will *not* include the occurrences of עוֹד belonging to the "עוֹד *still*" adverbial family. This behaviour is inconsistent with that of the "אַיֵּה *where?*" adverbial subclass. As for הִנֵּה, not only will the "הִנֵּה *behold!*" subclass *not* include occurrences that do not belong to "Quasiverbal" family, but those other occurrences are not rewarded with a(nother) single-lexeme category. They are not even found within the "Adverbial" top category; one must look for the "Exclamation" subclass within the "Miscellany" category. In fact, they are even said to represent another lexeme altogether. This is acknowledged in a footnote: "We split the forms into three homographs: a spatial adverb (glossed 'here', 268×), an exclamative (glossed 'behold!' 19×), and a quasiverbal (glossed 'behold', 912×)" (p. 252 n. 23). This assertion is problematic on more than one account: the exclamative and the quasiverbal forms seem to be as different from each other as the spatial adverb, whereas in reality there are only two forms, הֵנָּה "here" and הִנֵּה "behold." The so-called third homograph is in fact the result of the authors' desire to distinguish between two uses of הִנֵּה "behold." Creating a new lexeme is both artificial and misleading, as it will affect all searches involving הִנֵּה.

Surprisingly, the authors elsewhere add to the occurrences of יֵשׁ those of אִיתַי, a lexeme found in the Aramaic portions of the Bible. Does this mean that Aramaic is part of Biblical Hebrew? The authors even add: "We observe that the Aramaic form is preceded by לֹא 'not' six times" (p. 255 n. 49). This, of course, is wrong: in these six occurrences, אִיתַי is preceded by לָא and not by לֹא. Aramaic is *not* Hebrew, and the fact that the books of Daniel and Ezra contain Aramaic portions does not justify their use for the study of the Hebrew language, unless one studies the affinities and influences of these two related languages. Earlier, I deplored the fact that the authors

excluded other biblical manuscripts and variant readings; I now have to deplore their inconsistency in including Aramaic data.

Having conveniently excluded occurrences of עוֹד, הִנֵּה or אַיֵּה that do not fit the desired behaviour (either by assigning them another part of speech or by creating another lexeme altogether), the authors present bar charts for each of them, and the resulting dendrogram highlighting their affinities. Andersen and Forbes see it as a proof that these five lexemes should be grouped together, as stated in their conclusion: "As far as we know, our work is the first time that these 5 lexemes have been grouped together as a distinct part of speech" (p. 260). As mentioned above, this is a classical case of circular reasoning. In fact, despite the authors' efforts to exclude unwanted occurrences, the dendrogram nonetheless shows that these five lexemes *do* exhibit different behaviours, except for יֵשׁ and אַיִן, whose affinities as existentials have long been recognised by grammarians. So much for the "quasiverbal" category.

Chapter 19 deals with "Verbless Clauses" (p. 261), which are believed to exhibit a number of properties listed on pp. 262–263. The first feature states that verbless clauses (VLCs) are bipartite; that is, that they contain two CICs. According to the authors, however, only 57% of VLCs have just two CICs, which challenges the "reigning paradigm" of analysing verbless clauses as "subjects and predicates" (p. 263). In reality, the number of CICs depends on how they are parsed. Examples given on pp. 274–275 illustrate this fluctuation well: in Jeremiah 1:1, the expression הַכֹּהֲנִים אֲשֶׁר בַּעֲנָתוֹת בְּאֶרֶץ בִּנְיָמִן "the priests that ‹are› in Anathoth, in the land of Benjamin" features two parallel phrases that should be combined into one CIC; however, these phrases were initially parsed in the Andersen-Forbes database as two independent CICs. Likewise in Song 2:14: יוֹנָתִי בְּחַגְוֵי הַסֶּלַע בְּסֵתֶר הַמַּדְרֵגָה "My dove, in the clefts of the rock, in the covert of the cliff"; the two parallel phrases (= preposition בְּ + hiding place in construct state + location preceded by definite article) should be combined.

Since the number of CICs in a verbless clause depends on their parsing, the authors' initial statement that only 57% of VLCs have two CICs is just inaccurate, and the mention of exact numbers ("5,453" clauses out of "9,500") is deceptive. Yet, the authors organise the rest of their chapter according to the number of CICs: § 19.9 (p. 275) deals with one-CIC verbless structures and clauses; two-CIC VLCs are discussed § 19.10 (p. 283), followed by three-CIC VLCs in § 19.11 (p. 287) and, finally, multi-CIC VLCs (§ 19.12 p. 290). Such an organisation is questionable given the unreliability of the number of CICs per clause. The authors are aware of this problem and admit that "the grouping of verbless clauses by number of CICs is not precise" (p. 275); why, then, did they organise the rest of their chapter according to the number of CICs, giving specific numbers (*e.g.*, "1,466 one-CIC verbless structures"; "5,320" two-CIC VLCs; "1,534" three-CIC VLCs; "349" four-CIC VLCs, *etc.*)?

An example of a four-CIC VLC is given on p. 290: וְכֹל שֶׁרֶץ הָעוֹף טָמֵא הוּא לָכֶם "and all the swarm of the winged-animals, it is unclean to you" (Deuteronomy 14:19). However, this clause may be seen as bipartite, with a subject (כֹל שֶׁרֶץ הָעוֹף) and a predicate (טָמֵא הוּא לָכֶם). It is unfortunate that the Andersen-Forbes phrase marker does not "visualize" this structure, and it is even more unfortunate that the authors fail to realise that a bipartite clause does not necessarily correspond to what they call a two-CIC clause. In their conclusion, they condemn the "binarism of the prevailing definitions of VLC" since "we have seen, however, that Biblical Hebrew contains

VLCs with as many as ten CICs" (p. 291). I could include here a discussion of the alleged ten-CIC VLCs, but it would only confirm that: (1) the number of CICs depends on parsing and can often be reduced; (2) a bipartite clause is *not* necessarily a two-CIC clause given the limitations of the Andersen-Forbes CIC system.

The final chapters focus on more complex structures such as non-tree phrase markers and supra-clausal structures. Chapter 20 (p. 294) gives cases of discontinuity and multiple mother constructions, such as construct participles, distributed apposition or ellipsis. I've already discussed construct participles and distributed apposition, so let's look at an example of backward ellipsis given on p. 308, taken from Psalm 77:2:

<div dir="rtl">

קוֹלִי אֶל־אֱלֹהִים וְאֶצְעָקָה קוֹלִי אֶל־אֱלֹהִים

</div>

The authors note that "the two clauses are identical except for the verb" (p. 309) and thus conclude that the verb in the second clause is a case of backward ellipsis. Here is the corresponding phrase marker:
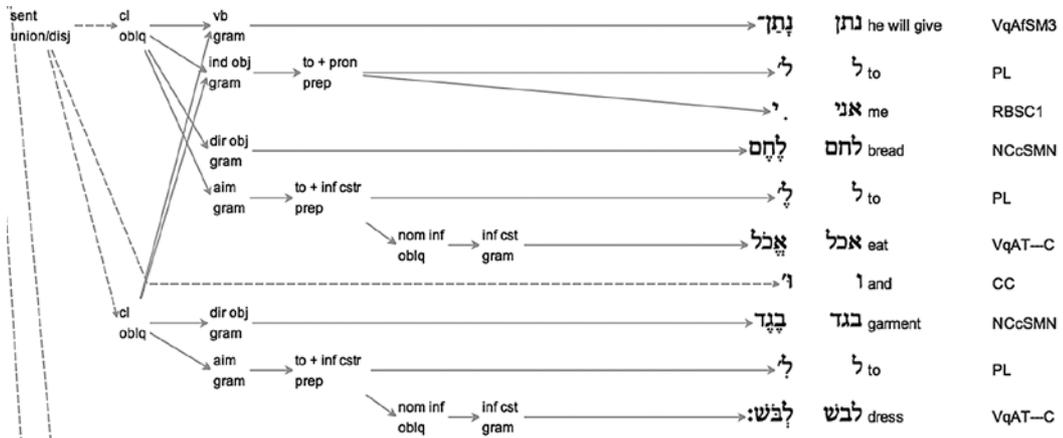

Unfortunately, this parsing is wrong in multiple places. First, the second clause is not אֶצְעָקָה קוֹלִי אֶל־אֱלֹהִים "I cry ‹with› my voice unto Elohim"; taking into account the end of the verse (וְהַאֲזִין אֵלָי), which also contains a verb introduced by וְ, we realise that אֶצְעָקָה must be disconnected from קוֹלִי אֶל־אֱלֹהִים, which in turn becomes an echo of the first clause. Moreover, in the first and (now) third clause, קוֹלִי is not an instrument CIC, but the subject of the verbless clause. The four clauses form two pairs: verbless / verb // verbless / verb. The two pairs are juxtaposed, while the two clauses inside each pair are coordinated by וְ. Moreover, the first clause in each pair is identical. Here is a literal translation showing the structure:

<div dir="rtl">

קוֹלִי אֶל־אֱלֹהִים וְאֶצְעָקָה קוֹלִי אֶל־אֱלֹהִים וְהַאֲזִין אֵלָי:

</div>

"My voice ‹is› unto Elohim, and I want to cry; my voice ‹is› unto Elohim, and he will listen to me."

Note that the *atnāḥ* confirms this structure; it is thus all the more surprising that the authors failed to recognise its pattern. Now, the BHS apparatus notes that some Hebrew manuscripts do not have a ו conjunction before אֶצְעָקָה, which is reflected in ancient versions such as the Old Greek. In that case, a point could be made for parsing קוֹלִי אֶל־אֱלֹהִים אֶצְעָקָה as a single clause: "‹With› my voice unto Elohim I want to cry." This is what the Greek version does: Φωνῇ μου πρὸς κύριον ἐκέκραξα (note the dative Φωνῇ). But since the authors insisted that "Biblical Hebrew" is, in their terminology, limited to the Leningrad Codex (see my comments earlier) and that variant readings are not taken into account, their parsing should indeed reflect the text of the Leningrad Codex. Moreover, their parsing system could be improved so as to reflect the phenomenon of "echo" exhibited by the first clause of each pair (see my comments earlier on the "echo" phenomenon).

The following example on p. 309, borrowed from Genesis 28:20, is given as a case of "multiple ellipsis":

As the phrase marker indicates, the authors believe that "both the verb and the indirect object are ellipted." Indeed, they parse בֶּגֶד לִלְבֹּשׁ as a separate clause introduced by ו, a reading that fails to see the parallel structure between לֶחֶם לֶאֱכֹל and בֶּגֶד לִלְבֹּשׁ (= an absolute noun followed by an infinitive construct introduced by ל). The ו conjunction coordinates those two phrases — not what the authors parse as two clauses. What we have here is, in fact, a single clause made up of a verb, נָתַן, followed by an indirect object, לִי, and a direct object made up of two union phrases, לֶחֶם לֶאֱכֹל וּבֶגֶד לִלְבֹּשׁ. This is *not* a case of multiple ellipsis; in fact, this is not a case of ellipsis *at all*.

I mentioned earlier that the authors did not pay attention to the position of the *atnāḥ* in Psalms 77:2. This question is dealt with in Appendix 1, "Text Choice, Corrections, and Reductions" (p. 326). The authors state that, in the early development stages of their work, they "experimented

with the *atnāḥ* as the most likely to assist in the mapping of clause boundaries, since at least in poetic texts it divided many one-bicolon verses into two clauses" (p. 330). They nonetheless decided to ignore these cantillation marks, except at the segment level (for instance to distinguish בָּאָה "she is coming" from בָּאָה "she came"). They conclude by stating: "Pause is purely elocutionary, and its significance for grammar is minimal" (p. 333). As the example in Psalms 77:2 has made clear, however, this is simply not true. In fact, cantillation marks are an integral part of the Masoretic tradition; it seems quite inconsistent, therefore, to opt for a study of Biblical Hebrew in this single tradition — at the expense of others texts such as the Dead Sea Scrolls or Samaritan manuscripts — while depriving it of some of its major components. If pauses in the Masoretic traditions are not to be trusted, why should vowels be different?

To sum up my reflections, the Andersen-Forbes system is problematic at all levels. The textual basis excludes the most ancient witnesses while setting aside Masoretic *Qere* variants and even cantillation marks, against principles of corpus linguistics — which the authors nonetheless claim to adopt.

At the segment level, the Andersen-Forbes system is inconsistent in ligaturing words to form new segments while maintaining the division of other words. It artificially creates homographs and multiplies the number of part-of-speech categories. Many categories contain a single lexeme *by design*, which defeats the purpose of creating a taxonomy. Moreover, single-lexeme parts of speech are not all on the same level, as seen with the so-called quasiverbals, whereas semantic features are sometimes mixed in the part-of-speech system.

Several phrase types are also created on the basis of semantic features, but without consistency, as seen in the "apposition" and "echo" phrase types. Moreover, alternate parses are often possible, which undermines the reliability of the statistical data provided throughout the volume; the precision of these numbers is deceptive because they are not accompanied with their margin of error, against fundamental principles of statistical analysis.

The CIC-type taxonomy, like the part-of-speech taxonomy and the phrase-type taxonomy, mixes syntactical and semantic features, which affects subsequent syntactical analyses and studies. Various problems arise, including circular reasoning.

Although the Andersen-Forbes system aims at "visualizing" grammar (as emphasised in the volume title), it fails to identify obvious parallel structures in phrases and clauses, while obscuring phrase markers with unnecessary (and sometimes incorrect) tangling.

These drawbacks should not, however, eclipse the potential value of the Andersen-Forbes system. I know of no other Biblical Hebrew database that contains so many grammatical and semantic features. In this article, I have expressed concerns about its underlying linguistic theories and suggested ways of improvement which, I hope, will be addressed in a future version of their database.

Michael LANGLOIS
université de Strasbourg, Institut universitaire de France
E-mail: michael.langlois@unistra.fr