Article

The Andersen-Forbes Computational Analysis of Biblical Hebrew Grammar

A. Dean Forbes
https://orcid.org/0000-0002-4633-9048
University of the Free State, South Africa
adforbes@post.harvard.edu

Francis I. Andersen
University of the Free State, South Africa

Abstract

The Andersen-Forbes BH database is based upon the text of L, having omitted cantillations, corrected obvious errors, segmented or ligatured orthographic words by rule, and resolved homographs. Each segment has an associated set of grammatical features and an assistive gloss. Our linguistic preferences favour data-driven over theory-driven analyses, language performance over language competence, and quantitative over qualitative language models. In our research, we rely on successive approximations, planning at least one step ahead at each stage. In representing grammatical structure, we opt for simple descriptive features displayed in a single-level environment that allows representation of non-binary, discontinuous, and ambiguous situations. As our work has progressed, refinements and extensions have been added, among them naïve semantic categories, constituent licensing relations, and semantic roles. With proper care, the grammatical data may be productively probed. As we enhance the database's consistency, we are also extending its linguistic coverage and refining its search methodology.

Keywords: Biblical Hebrew grammar; phrase markers; text tagging; searching; Biblical Hebrew style

The Goal and Organisation of this Essay

Our goal is that readers will emerge from reading this essay with a coherent sense of the network of linguistic principles undergirding the Andersen-Forbes database, will have

Journal for Semitics
https://upjournals.co.za/index.php/JSEM
Volume 27 | Issue 1 | 2018 | #2936 | 22 pages

https://doi.org/10.25159/1013-8471/32936
ISSN 0256-8853 (Print)

a feel for its present capabilities as well as its potential for enhancement, and will be in a position to gather tools for reliably exploiting it.

Regarding essay organisation, we document our approach to the grammar of Biblical Hebrew. We then characterise text preparation and tagging, highlighting our linguistic preferences and strategies. We discuss data-driven adjustments to our initial approach that emerged during text parsing. We describe our overall approach, including its representational repertoire and foci. Next, we take up the implications of our approach for data retrieval, cataloguing presently impossible and difficult kinds of queries and indicating additional query types scheduled for release in the mid-term. We illustrate searching both for readily accessible varieties of information and for more elusive types. We briefly address "style" mark up. Finally, we comment on our plans.

The Role of Linguistic Theory

Linguistic Theories Used to Tag Data

"Theory" in Linguistics

Before getting into the specifics of the Andersen-Forbes (A-F) *approach* to grammar, please consider a few words regarding the role of "theory" in linguistics. Heine and Narrog (2010, 4) put matters well:

> [O]ne and the same author may refer to his or her work, as a 'model' in some contexts, as an 'approach' in other contexts, or as a 'framework' in still other contexts. More generally, authors with a generativist orientation tend to phrase their work in terms of a theory, and, for equally good reasons, other linguists avoid this term; for quite a number of linguists with a functionalist orientation, there is some reluctance to recognize 'theory' of any kind as being of use in doing linguistics. The problem with the terminology is that there is not much agreement across the various schools on how to define these terms.

More directly, Dryer (2006, 27) argues: "What [formal linguists] find lacking in much functional work is a proposed *metalanguage* in which languages are analysed or described, a metalanguage in which representations of structure and rules are stated." Over the years, we have devised a metalanguage that we have used both to represent structures (Andersen and Forbes, 2012)[1] and state rules (Andersen and Forbes 1995, 49–75).[2] Yet, our work is not an exemplar of formal linguistics; our approach achieves none of Chomsky's levels of adequacy for formal grammars.

[1] Note, however, that there are many relations in the text that we have not yet represented. These lacks will be overcome, going forward.

[2] Our rule-set correctly parsed around 80% of the constituents found in the Hebrew Bible (HB).

We prefer to view our work as an *approach* to grammar. It has been constrained to varying extents by the limitations and linguistic preferences with which we worked. Over the decades, the limitations and preferences have evolved, and thus so has our approach. The actual and desired statuses of our approach are accessible in our grammar book, *Biblical Hebrew Grammar Visualized*, henceforth *BHGV*. For the purposes of this essay, however, we have gathered its central characteristics into a series of tables, presented with commentary and references in the following pages.

In the Beginning

To understand the A-F linguistic approach, it is helpful to sketch its history. Readers thereby will be alerted to initial decisions that we made. The A-F collaboration began in 1970. Andersen was then a professor at The Church Divinity School of the Pacific/Graduate Theological Union (GTU) in Berkeley, and Forbes was a full-time student at the GTU and a cardiological research consultant at both Hewlett-Packard Laboratories (HPL) and Stanford University School of Medicine. As a consultant, Forbes had weekend access to an early HP computer. He also benefited from the advice of expert HPL programmers, as needed. Meanwhile, Andersen was finding his path between Pike’s tagmemics and Chomsky’s generativism. He remained drawn to structuralism, notwithstanding the withering onslaught the generativists had launched against it. We began our linguistic work with a mixed set of linguistic preferences and emphases. Andersen (1970) had just published *The Verbless Clause in the Hebrew Pentateuch*. Many of the concepts already in that book appeared in our subsequent joint work. Included were:

- Openness to descriptivism
- Reliance on:
  - Immediate constituent (“IC”) analyses
  - Analyses using nucleus and core ideas
- Use, as warranted, of:
  - Non-binary phrase structure
  - Discontinuity
  - Zeroes (in limited contexts)

Constraints and Choices

Table 1 documents our basic text preparation decisions. Table 2 shows our decisions regarding the tagging of the text. Table 3 presents our major linguistic preferences and strategies. Unfortunately, our representational preferences involve various linguistic fine points. The associated footnotes are included to provide an adequate path to understanding for those wishing to appreciate the distinctions involved.

Table 1: Text basics

Use a single instance of the HB, **L**, rather than a text mixture or group.[3]
Opt for single-character transliteration, thereby omitting cantillations.[4]
Correct “obvious errors” in **L** (“sic **L**” in *BHS*); e.g., הַשִּׂדִּים → הַשִּׂדִּים, Gen 14:10.[5]
Make no emendations. Introduce *lapsus calami* part of speech (e.g., for בגד in Gen 30:11) and *nebulous* constituents (e.g., אֵת in 2 Kgs 9:25).[6]
Systematically, let *kethiv* be *kethiv* and *qere* be *qere*. Never swap them.[7]
Following rules, segment and ligature words; e.g., לִי →לְ′ִי *to* + *me*, while בֵּית־אֵל is ligatured into one segment.[8]
Point the *kethiv* text based on Gordis (1971); e.g., מהם → מְ′הֵם in Ezek 8:6.[9]

Table 2: Text tagging

Force resolution of lexical ambiguity. For example, the noun עִיר is usually *city* but also appears as *excitement* (e.g., Jer 15:8), *donkey* (Gen 49:11), and *Ir*, a specific human (1 Chr 7:12). Hence the lexicon has four entries for עִיר.[10]
Provide “type glosses.” For example, all tokens of ִי′ are (type) glossed *me* and all glosses of נתן-verbs include some form of *give*. Consequently, Gen 9:13 reads *bow-me I-gave in-the-cloud* rather than the more standard *bow-my I-put in-the-cloud*.[11]
Tag segments with their grammatical features; e.g., אֶבֶן *stone* has the feature bundle: *common noun/singular/feminine/normal/natural substance*.[12]

[3] For a brief survey of attitudes toward texts in antiquity, the Masoretic traditions, available text choice options, and the option we took, see *BHGV* (§A1.1).

[4] For a discussion of our historical reasons for omitting cantillations, see *BHGV* (§A1.3.1).

[5] On our corrections (and a comparison with Dotan’s practices in Genesis) see *BHGV* (§A1.2).

[6] For lists of *lapsus calami* and nebulous constituents, see *BHGV* (§3.2.1.1 and §9.3.1.2, n. 8).

[7] For a treatment of possible options, see *BHGV* (§1.1.3.2).

[8] On segmentation, see Forbes (2014b, 215–17) which discusses where to segment, how to cut, context-sensitive rules, and bootstrapping. On ligaturing, see *BHGV* (§2.1.3).

[9] For our reasoning, see *BHGV* (§A1.3.2).

[10] *BHGV* (§A1.3.3) treats homography, both multiple part-of-speech and within part-of-speech.

[11] As *BHGV* (xiii) explains, “one gloss is applied to each form [in the database] (the ‘type’) and is used whenever that form appears in the text (as a ‘token’).” See also *BHGV* (§A7.1).

[12] For the repertoire of relevant features, see *BHGV*, Chapter 3. Parts of Speech.

Table 3: Linguistic preferences and strategies

Procedural Preferences[13]
Data-driven over theory-driven.[14]
Language performance over language competence.
Quantitative models over qualitative models.
Strategic Constraints
Iterative analyses with provisions for subsequent refinement and expansion.[15]
Plan “one step ahead” at each research stage.[16]
Representational Preferences
Simple features over complex features.[17]
Single level over multiple level.[18]
Constituent grammar over dependency grammar.[19]
Surface structure over deep structure.[20]
Non-binary branching over forced binary branching.[21]
Multidominance over null nodes to mark scope, gapping, and distributed apposition.22
Discontinuity over continuity, where necessary.[23]
A unified syntax-discourse transition over a disjoint one.[24]

Texts in Context

We decided first to define a character set, create software capable of displaying and printing the resulting pointed Hebrew, enter the book of Ruth into the computer as a pilot text, divide it into “segments”, mark it up with traditional grammatical information,

[13] For these “Emphases Drawn from Modern Linguistics,” see *BHGV* (§1.2.2.1).

[14] For us, the data, not a favoured theory, are king. See Forbes (2006, 123–125).

[15] We represent, rather than supress, ambiguity, see *BHGV* (§20.3).

[16] The goal of this policy is to avoid dead end research. See Forbes (2014b, 228–229).

[17] We use feature bundles, see Forbes (2009, 152–154).

[18] Note, however, that opting for a single-level representation does not preclude eventual inclusion of multidimensional (projection) representations.

[19] Over time, important phenomena will be represented using dependency edges (Forbes 2017).

[20] Also known as “non-derivational over derivational”.

[21] This is a way of avoiding introduction of a verb phrase (VP) constituent (Crystal 2008, 192).

[22] Multidominance allows a constituent to operate across clauses. See *BHGV* (§9.3.4.1) on scoping, *BHGV* (§20.2.3.3) on gapping, and *BHGV* (§7.2.2) on distributed apposition.

[23] Standard phrase markers are trees (*BHGV*, §4.2). However, several phenomena in BH evoke discontinuous non-tree phrase markers (*BHGV*, §7.2.2; Chapter 20. Non-Tree Phrase Markers).

[24] Following Webber, we seek a smooth representational transition from syntax into discourse. We therefore avoid disjoint. See *BHGV* (§21.1.2.1).

extract from it a primitive dictionary, and use these resources to create a keyword-in-context (KWIC) concordance used for checking our work.[25]

Behind these seemingly simple tasks were other necessary decisions and tasks, often not straightforward and alluded to above: transliteration tables, non-representation of cantillations, part-of-speech taxonomy, text-handling policies, segmentation rules, ambiguity suppression, homograph resolution, and lexeme assembly. Throughout the 70s, we transcribed and checked the text of **L**, produced the associated dictionary, and published four KWIC concordances. During the 80s and early 90s, we focused on studies of Biblical Hebrew orthography and continued to refine and extend our database.

We also began work on parsing the HB. We first divided the text into clauses. Aware that some clause boundaries were ambiguous, we specified a "preferred boundary" in such cases as our first-pass division criterion.[26]

Forbes was impressed by progress being made in generalised phrase structure grammar[27] and was on friendly terms with its devisers who also consulted at HP Labs. Partly as a consequence, Andersen and Forbes devised batteries of context-free grammatical rules, incrementally building up the clause constituents. The eventual computer output was a phrase marker ("PM") for each clause in the full text.[28]

Data-Driven Adjustments

As we parsed the biblical texts, the encountered phenomena required that we adjust our procedures and mark up along the lines indicated in Table 4.

[25] A history of our project may be found in Forbes (2014b).

[26] Andersen and Forbes (1992, 181–202) gives a full exposition of this work.

[27] Gazdar, Klein, Pullum, and Sag (1985) introduces generalised phrase structure grammar.

[28] Parsing results for Deut 8 are derived, step-by-step, in Andersen and Forbes (1995a).

Table 4: Data-driven adjustments

Problem	*Solution(s)*
Subject/object confusion[29]	Introduce naïve semantic categories. [30]
Syntactic ambiguity[31]	Force resolution initially.[32]
Limited licensing relations[33]	Increase the repertoire as needed.[34]
Discontinuous constructions[35]	Permit discontinuity in PMs [Figure 1].
Constructions with multi-dominance	Permit multidominance in PMs [Figure 2].
Nulls in gapped constructions	Represent using multi-dominance [Figure 3].
Clause immediate constituents (CICs) partly tagged with semantic roles[36]	Finish introduction of semantic roles.[37]

Note well: The usual practice is to place grammatical functions and semantic roles in differing strata (levels).38 To retain a single-level representation, we have included both grammatical functions and semantic roles in the clause immediate constituent (CIC) node specifications.

[29] Our rule-based parsers were good but often confused subjects and objects.

[30] For a census of semantic categories used, see *BHGV* (§3.3.1 with Table 3.2).

[31] For an introduction to syntactic ambiguity in BH, see Andersen and Forbes (1995b, 356–367). On the problem of constituent attachment preferences, see Andersen and Forbes (2002, 167– 186).

[32] Our initial policy of forced resolution of ambiguity is briefly sketched in *BHGV* (§1.1.3.4). On provisions made to allow representation of ambiguity in the future, see *BHGV* (§20.3).

[33] A *licensing relation* specifies "the grounds according to which a syntactic construction is identified" (*BHGV*, 364). For phrase markers showing various licensing relations (involving discontinuity, as it happens), see *BHGV* (§20.1.2).

[34] Most licensing relations, such as *modification*, are quite straightforward. A few are not, such as the vexed relation *paradoxical*, on which see *BHGV* (§8.2.3 and §21.4).

[35] Ojeda (1987, 257–282) was influential in our handling of discontinuity.

[36] According to Whaley (1997, 290), a semantic role discloses "the semantic relationship that a nominal bears to the rest of the clause. Common semantic roles include agent, patient, locative, and benefactive." For a full discussion of semantic role *repertoire*, *taxonomy*, and *recognition criteria* in BH, see *BHGV* Chapter 10.

[37] For a discussion of the in-process status of semantic roles in the A-F database, see *BHGV* (§9.2).

[38] In lexical-functional grammar, for example, "[t]he syntactic structure of a sentence consists of two formal objects, neither of which is derived from the other: the c-structure, which is a constituent structure tree of the familiar kind, and the f-structure, which carries non-constituent information such as the grammatical functions" (Trask 1993, 156–157).

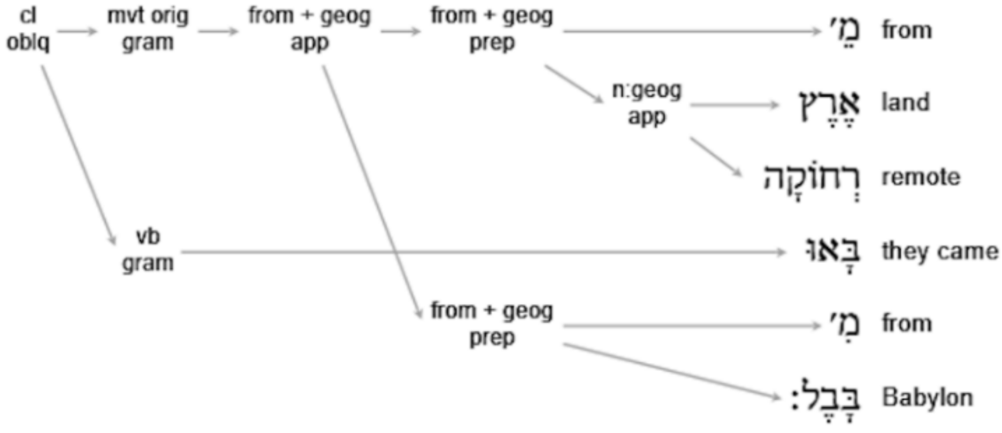


Figure 1: Discontinuity in 2 Kings 20:14c. *The movement origin* CIC is a discontinuous appositional prepositional phrase.

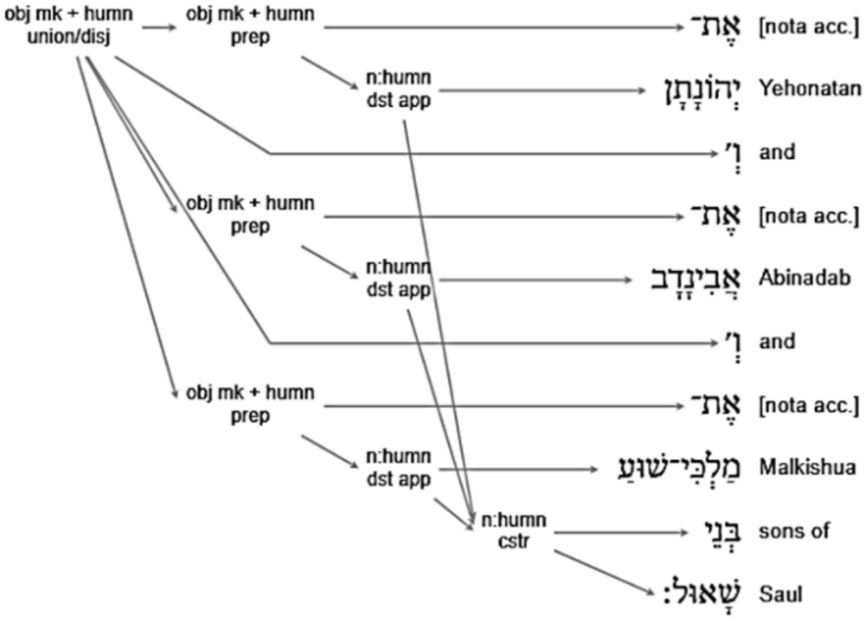


Figure 2: Distributed Apposition in 1 Samuel 31:2b. The three proper nouns jointly are in apposition with the noun phrase בְּנֵי שָׁאוּל.

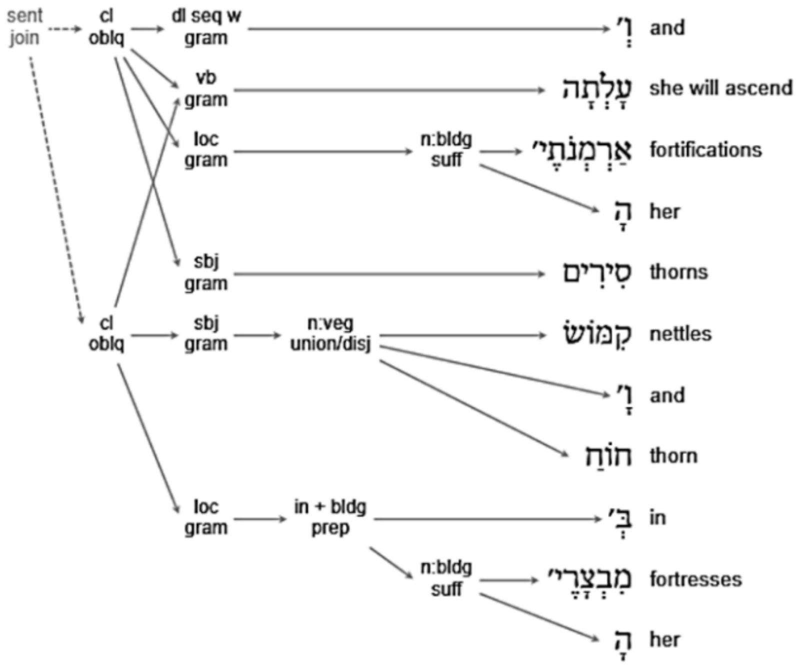


Figure 3: Verb Gapping in Isaiah 34:13. The **vb/gram** CIC has two **cl/oblq** mothers. That is, it is multidominated.

For Additional Insights

BHGV is the best single place to learn about our grammatical approach. It contains extensive material on text preparation and phrase marker creation (Chapters 1–4), as well as our plans (§9.2 and Chapter 21). It presents studies of specific clause constituents (Chapters 5–11) and then examines the overall behaviours of four frequently encountered verbs (Chapters 12–15). It considers the makeup of the various kinds of clause immediate constituents (Chapter 16) and analyses the "distances" between verb corpora (Chapter 17). Finally, it addresses quasiverbals (Chapter 18), verbless clauses (Chapter 19), non-tree phrase markers (Chapter 20), and discourse analysis (Chapter 21).

Implications of Our Chosen Linguistic Approach on Data Retrieval

Some kinds of information are simply *inaccessible* from the A-F database in its present form because we decided to omit them. Our linguistic approach has nothing to do with their inaccessibility. Excluded are: non-**L** texts, cantillations, repaired "sic **L**" instances, emendations, et cetera. Were the inclusion of these sorts of information deemed crucial, they could be added. Indeed, several of them may be introduced eventually.[39]

This leaves the issue at hand: What are the implications of the choice of linguistic approach for answering queries? Naturally, questions must be cast in terms of the categories of the A-F approach. Hence, categories present in *specific grammars* must be mapped onto corresponding categories present in the A-F approach. For many specific grammars, this may not be easy or even possible. In general: The further an approach to grammar departs from ours, the less easy will it be to construct queries for our database derived from that approach.[40]

For specific answers regarding data retrieval, consider Table 5 which tallies some approach-related query issues for A-F, now and in the mid-term.

Table 5: Query capabilities: at present and mid-term

Impossible/Difficult at present	Possible in the mid-term
(1) Complex-feature element	(1) Full representation
(2) Dependency-related	(2) Deixis mark up
(3) Valency	(3) Text type (closed set)
(4) Coindexation	(4) Proper-noun semantic resolution
(5) Verb phrase	
(6) Null node	

Difficult/impossible Query Capabilities now

We consider six impossible-to-difficult types of searches that may be important to certain users.

(1) Complex-feature-element query: Searching for an element in a *complex-feature*[41] structure may simply be out of the reach of the A-F database.

(2) Dependency-related query: Similarly, direct searches for constituents having *dependency relations* typically are not feasible. But one may search for clauses

[39] Decades ago, G. E. Weil advised Forbes to analyse a mixture of texts. Instead of the **L** Torah, he argued for *British Museum OR 4445*, and so on. We will not be following Weil's advice at this late date. Of much greater interest is inclusion of emendations. Then the problem becomes, which ones? Non-**L** MT manuscript readings? Relevant non-MT readings? Conjectural emendations?

[40] There is, however, a broad set of grammars that are theoretically on a par as regards category systems (Gazdar et al. 1988, 1–19; this paper is quite technical). Included are phrase structure grammar, transformational grammar, systemic grammar, et cetera. Queries from the perspectives of these theories should be transformable to conform to our perspective.

[41] Ibid (1988, 152–155).

containing subjects or direct objects plus any related complements, a dependency relation being signalled in their labelling. Similarly, suspended CICs and resumed CICs can be analysed in obvious ways.

(3) Valency query: Since a valency specification characterises a verb across all its tokens, valency attributes are imported from the lexicon onto verb tokens. Until now we have not included valency information. Our reasons may be found elsewhere (Forbes 2016a; also *BHGV*, §11.4).

(4) Coindexation query: Since we do not need to avoid tangling and since we are using other means for performing participant tracking, coindexation is not needed. This, however, makes the study of situations where traditional grammars use coindexation more complicated.[42]

(5) Verb phrase ("VP") query: Both subjects and clause boundaries are identified in the A-F database. Hence, for the definition of the VP that includes both complements and adjuncts, the study of VPs is trivial. For definitions that exclude adjuncts, however, matters become fraught.[43]

(6) Null node query: Dear to the hearts of transformationalists are the concepts of the *trace* and other kinds of empty category (*pro*, PRO, et cetera). Some non-transformationalists have formulated their grammars without resorting to empty categories, as have we.[44] Mapping between the transformationalists' empty nodes and our present analyses is sometimes simple (ellipsis, pro/PRO) but sometimes difficult to impossible (for example, long-range dependencies or non-explicit referents).[45]

Mid-term Query Capabilities

These four capabilities are being implemented or will shortly be implemented. Until they are completed, queries will not yield reliable results:

(1) Full-representation-dependent query (underway): The mixed representation process involved tagging any given CIC preferentially with its argument status, if any: subject, subject complement, direct object, direct object complement, indirect

[42] For a description of ploys others have adopted for avoiding "tangling," see *BHGV* (§7.2.2).

[43] Herbst et al. (2004, xxii) warn: "Given the complexity of the task and the prototypical nature of crucial distinctions … between complements (*Ergänzungen*) und [sic] adjuncts (*Angaben*), it might seem advisable to modify the standard text used in German programmes when the winning lottery numbers are announced, and say: *Alle Angaben und Ergänzungen ohne Gewähr*." [*For all adjuncts and complements, no responsibility taken*.]

[44] Regarding head-driven phrase structure grammar ("HPSG"), (Kathol et al. 2011, 70) reads: "[T]he reliance on … empty syntactic categories is generally considered to go against the spirit of HPSG as a surface-oriented theory."

[45] Forbes (2017) goes into these matters in detail. Ways of formulating such queries await implementation.

object, et cetera. If no argument grammatical function tag was relevant, then the CIC received its appropriate semantic role tag (*BHGV*, §9.2). With full representation in place, each CIC will be tagged with both its grammatical function ("GF") and semantic role ("SR").[46]

(2) Deixis-dependent query (underway): We have always followed the reference grammars and lexicons and classified כה, כן, and ככה (thus) as adverbs of manner. We now prefer to classify them, following Miller (2003, 135) and others, as *discourse deictic pronouns*. Converting to this classification is non-trivial (Forbes 2014a).

(3) Text type (not yet begun): This task has recently been interpolated into the queue. In a recent essay, following clues in the literature and his experience assessing Biblical Hebrew orthography, Forbes (2016c) argued that the effect of closed-set text type on diachronic analyses must be examined. To provide for such analyses, the A-F database will have BH closed-set text types marked up.

(4) Proper-noun-semantic-resolution-dependent query (not yet begun): This is a much-needed refinement. At present, proper nouns are not subdivided into their most refined semantic classes. For example, יִשְׂרָאֵל (Israel) is always marked as a specific human and never as a group.

Grammatical Data Recovery[47]

Readily Accessible Grammatical Data

Examples of answers to the question, "What kinds of grammatical data can be retrieved using A-F?" are abundant in *BHGV*. Some guidelines, however, can be provided. For some time, we have had two ways of accessing our data: via Logos and via Linux. In general, when we know what we are looking for, Logos is the way to go, provided we insure that all possibilities have been considered. When we are seeking previously unnoticed phenomena, then Linux is the way to proceed. In addition to allowing one to make precise tallies, searches often turn up interesting phenomena. For example, consider this pair of curiosities: definite possessives (x7) and definite prepositional phrase (x1). Figures 4a and 4b show PM fragments illustrating these cases (*BHGV*, §6.3.2.1).

[46] Or pseudo-SR for certain families of CICs: *impermanent*, *syntactically-isolated*, *predicator*, and *operator*. See *BHGV* (§9.3.1–§9.3.4).

[47] The topics covered in the section were specified by the commissioners of this essay. The material will be accessible to those already familiar with the Logos rendition of the A-F database. We hope that those unfamiliar with the database and Logos search engine will nonetheless catch glimpses of the sorts of studies these amenities make possible.

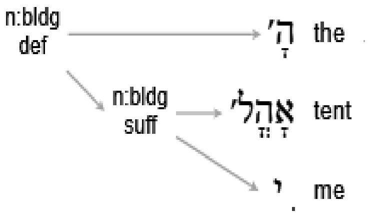


Figure 4a: Definite possessive

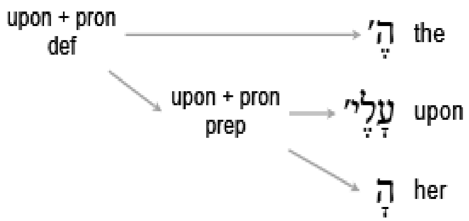


Figure 4b: Definite prepositional phrase

As noted above, the search for oddities is often best carried out using Linux scripts and winnowing the results. By this route, we found this sequence in 2 Sam 23:24–39:

P0P0P0P0P0P0P0P0P0P0P0P0P0P0P0P0P0**S**0P0P0P0P0P0P0P0P0P0P0P0P0P,

where P stands for a phrase, 0 indicates the absence of a connective, and **S** is a simple isolated segment (*BHGV*, §6.5.4, n. 15). Once one knows the form of the PM, one can easily construct the Logos search shown in Figure 5 which finds it alone.[48]

Linux scripts efficiently search for maximal constituents, although one can in some circumstances use Logos iteratively to locate such entities, as will be shown in the next subsection. For example, using a Linux script, we find that the longest complex joined phrase, consisting of 35 joined construct phrases, is in Ezra 2:43-54.

Linux scripts are also an efficient way to gather and consolidate information for publications. Consider Chapter 16 of *BHGV*: “Makeup of Clause Immediate Constituent Subtypes,” a chapter containing thirteen tables holding incidence tallies for nearly eighty semantic roles/grammatical functions (e.g., “movement bearing,” “direct object”). Given a release of the A-F database, one computes and formats all these particulars in seconds by properly executing a simple command. The executed script gathers the needed information and formats it into a single master file. Elusive grammatical data

[48] In the search, phrase$_1$ must be allowed to “appear anywhere” and segment$_1$ may *not* “skip levels”.

Exclusive Grammatical Data

The richness of the A-F database allows well-versed users to assemble data regarding an extensive array of grammatical issues. But the A-F database itself and the Logos search engine can frustrate casual users. In this subsection, we take up three potential traps involving *counting* and two available boons allowing more advanced searches of which too many are unaware.

Constituent Counting

Rule: Be very vigilant when you count constituents! We examine three situations in which error lurks.

(1) Repeated counting of constituents due to embedding: Consider a purposefully devious request. How many times does אֲצַוֶּה (I shall command) occur in L? If we go to page 981, items 371–378, in Even-Shoshan, we obtain his answer: eight times. If we search the A-F text for אֲצַוֶּה, we also find eight, agreeing with Even-Shoshan. Suppose we now go to the A-F phrase markers to make a count. Since segments are always part of clauses, we naively perform the search specified in Figure 6.

The Logos search correctly provides twenty-four hits. Why? Because of embedding. For example, the token of אֲצַוֶּה in Jer 11:4 gets counted repeatedly since it is so deeply embedded in clause within clause within clause within … (BHGV, §6.1). A simple search for the segment specified returns eight hits.

(2) Repeated counting of constituents due to multidominance: Multidominance appears in two contexts: ellipsis and distributed apposition. If we search across 1 Chronicles as specified in Figure 7, the count found is 63. If we delete the clause node and search, the count is eight, the true number of יֵצֵא tokens in 1 Chronicles. The difference is due to verb gapping. The verb appears once in 1 Chr 24:7 and then is used to fill 24 subsequent gaps (in verses 7-18). Figure 8 shows the start of the long, heavily-gapped sentence

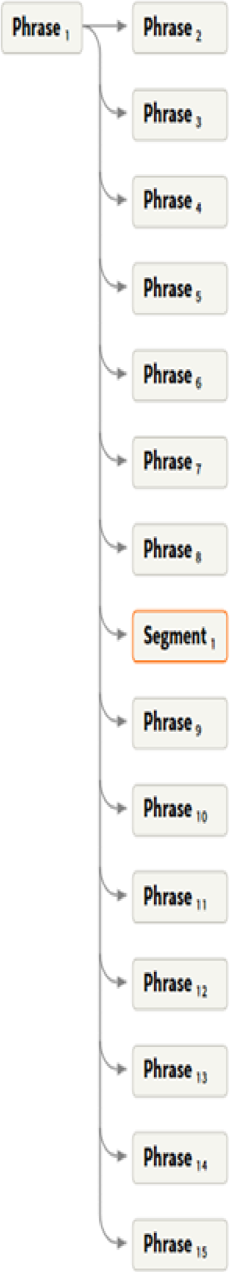


Figure 5. Isolated segment search

(3) Absorbed segments due to ligaturing or lexicalisation: Segment counts are altered when an item is combined with another (“ligaturing”) or is not segmented off because of lexicalisation.

a. Ligaturing: When segments are ligatured to form new lexemes, those tokens no longer contribute to their constituent totals. For example, the total for lexeme בית (house) is reduced by seventy because it is part of בית־אל (Bethel) seventy times.[49]

b. Non-segmenting: Our stated policy is that if a “word” is the result of lexicalising a pair of segments at least one of which is not found elsewhere singly, then those segments are not cut apart. If the pieces are found elsewhere singly, then the two components should be cut apart.[50] When the pieces are left assembled, the counts for those segments will be too low.[51]

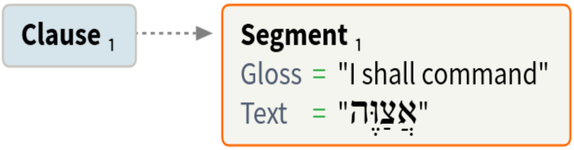


Figure 6. Given Segment in Clause.

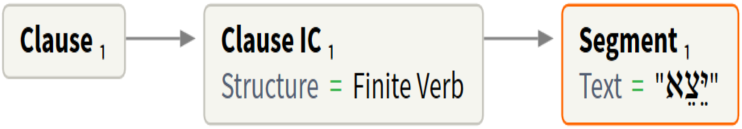


Figure 7. Finite Verb in Clause.

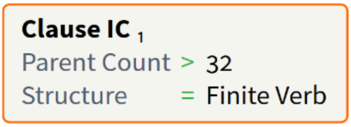


Figure 9. Gap search.

[49] The practice is common: *DCH* (vol. 2, 38) where בית־אל is a lexeme, also Even-Shoshan (pp. 175-76).

[50] Thus, for example, we cut לִפְנֵי into פְּנֵי + לְ. Our policy has been inconsistently applied. For example, the 208 instances of לָכֵן (therefore) are not cut apart. This omission is likely to be corrected as we carry through our work on deixis. That a pair of segments has a good single-word English translation does not warrant leaving its segments assembled.

[51] Hence, for example, the totals for לְ and כֵּן are each presently reduced by 208 counts because לָכֵן has not been divided.

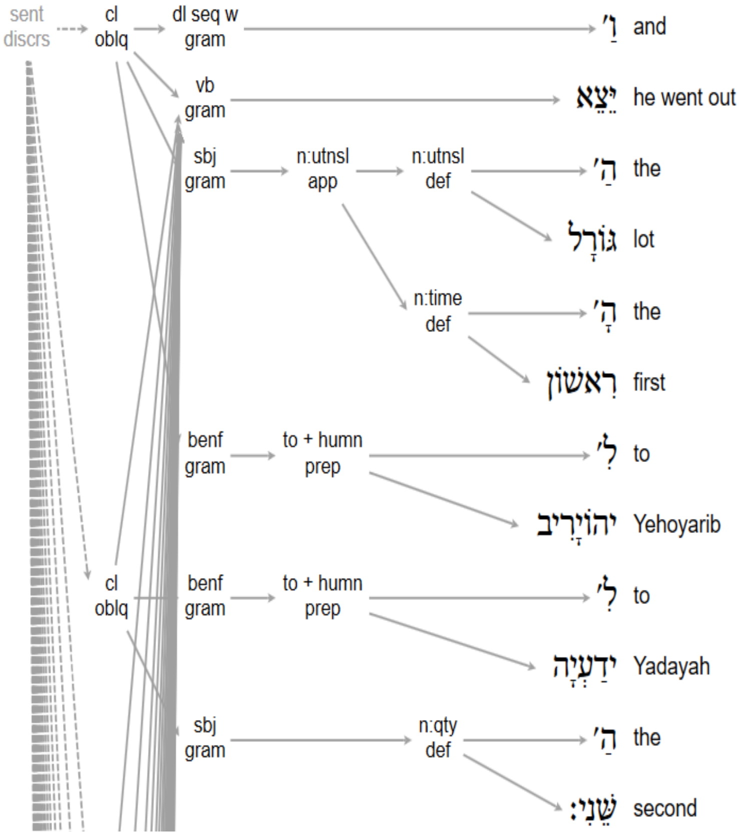


Figure 8. First Two Clauses in Very Long Sentence.

Very Useful Less-Known Search Capabilities

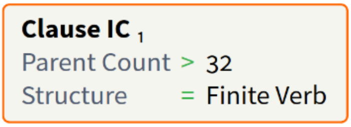
The Logos search includes two underutilised capabilities: *in/out degree* and *agreement*.

In- and out-degree: In-degree ("parent count") and out-degree ("child count") are quite useful.

In-degree: Suppose you want to know how many times we have marked a finite verb as gapped. Under Linux, the answer is easily found in less than two seconds. Under Logos, it takes longer. One proceeds by indirection. If a finite verb CIC is multidominated, it is because ellipsis has been marked. Run the search specified in Figure 9.[52] The result returned will be empty. Change the parent count constraint to "Parent Count > 16" and run the search again. You will get two hits: 1 Chr 24:7 and 1 Chr 25:9. You now have a choice: Go to the passages found and manually count the gaps *or* run the search repeatedly until you find the number of parent clauses involved. Taking the second

[52] The "Parent/Child" search specification is at the bottom of the menu.

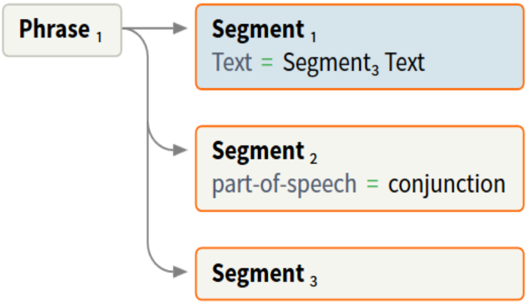
approach, we find that “Parent Count = 24” reproduces our two hits with more than sixteen parents. It follows that the finite verbs are gapped in the text 23+23 = 46 times. (For each hit, one of the edges will correspond to the ungapped clause.) Carry on, iteratively decreasing the number of parents being sought and tallying the results. In all, the searches take only a few minutes.[53]



Clause IC ₁
Parent Count > 32
Structure = Finite Verb

Figure 9. Gap search.

Out-degree: Suppose you want to study long lists of joined noun phrases having human semantics. We search for a phrase having: License = Joined, Phrase = Noun, and Semantics = Human. Performing the search yields 152 hits. If we next add the further constraint that “Child Count > 6”, then we get 24 hits for study.[54]



Phrase ₁

Segment ₁
Text = Segment₃ Text

Segment ₂
part-of-speech = conjunction

Segment ₃

Figure 10. X<conj>X search.

Agreement: Suppose you wish to study contexts in which the text has the pattern [X]<conj>[X]. Using agreement rules, Logos enables this sort of search. Figure 10 shows the search specification required.

The single agreement rule constrains the text of the first conjunct (Segment$_1$) to be identical to the text of the second conjunct (Segment$_3$).

Naturally, we specify that the part of speech of the middle segment is a conjunction. Carrying out the search yields 67 hits for study. If we modify the constraint to require that the match ignore the textual marks (“pointing”), we get 93 hits. Thus, phrases such as דֹּר וָדֹר (Ps 100:5) are now included, but differing *matres lectionis* remain enforced. Hence, phrases such as דֹר־וָדוֹר (Ps 89:5) are absent. If we simply impose the constraint

[53] We currently get a total of 1060 finite-verb gaps.

[54] The two longest hits are at Ezra 2:43–54 with 35 children and Neh 7:46–56 with 32 children.

that the *lexemes* must be identical, we then get 105 hits, including the *defective-plene* phrase in Ps 89:5. The yield is now over-productive, since forms such as שָׁרִים וְשָׁרוֹת (Qoh 2:8) are included. One may impose further agreement rules to filter out such phrases. One may impose constraints involving: text, text (no marks), lexeme, genre,[55] source, semantics, part of speech, number, gender, person, state, stem, voice, and parent.

"Style" Mark Up

As requested, we comment on three aspects of our database involving "style".

Eissfeldt's *Hexateuch-Synopse* document categories[56]

We made use of these sources in studies of Biblical Hebrew attachment preferences in the Primary History and of clause complexity in Genesis (Andersen and Forbes 2002, 176–184; also Andersen and Forbes 1998, 309–314).

Extracted Style Features

Stylistic analysis involves variables thought to be sensitive to differences in style, somehow defined. In a 1997 paper, we examined the variations of three measures of phrase marker ("PM") complexity as functions of source document in Genesis (Andersen and Forbes 1997, 309-14). The measures were: (1) median PM width (number of CICs), (2) median maximal PM depth, and (3) median total node count. By source document, we found: (a) the Lay source ("L") uses fewer CICs than do J ~ E ~ P;[57] (b) depth complexity for L is lowest, and the depth complexities for J, E, and P are essentially identical; (c) total node count complexity increases from L to J ~ E and then increases again to P.

Genres

We tagged the texts with what we shall here term *genre information*. In fact, we include both open-set genres and exchange pairs, as per Table 6a. The genres are given down the first column, and the common exchange pairs are given across the top row, with a handful of uncommon and atypical exchange participants tallied in Table 6b. Empty cells indicate that we have not yet encountered the specified combination of genre and

[55] The current Logos search makes reference to "text type". We formerly called this feature "genre" and will return to calling it that when we assign values to a new text-type feature.

[56] Available on the web via books.google.com by entering *Hexateuch-Synopse*.

[57] We counted median clause length measured in terms of clause immediate constituents ("CICs"). Polak uses "explicit, lexicalised sentence constituents" ("ELCs"). They differ from CICs in that micro-realised arguments are not ELCs, while we count suffixed pronoun arguments. It should be possible to generate data of Polak's sort (2006, 115–162), using Linux.

exchange pair. For example, we know of no instance of judgment pronounced from divinity to divinity.

Table 6a: Genres and exchange pairs

	Exchange pair				
Genre ↓	Author → Reader	Divinity → Divinity	Divinity → Human	Human → Divinity	Human → Human
Title	✓				✓
Genealogy	✓				
Narrative	✓				✓
Quarrel				✓	✓
Accusation			✓		
Judgment			✓		✓
Lamentation				✓	✓
Instruction		✓	✓		✓
Request		✓	✓	✓	✓
Supplication				✓	✓
Blessing				✓	✓
Curse			✓		✓
Prediction/Promise		✓	✓	✓	✓
Woe and Dirge			✓		✓
Prophesy					✓
Greeting					✓
Praise					✓
Wisdom					✓
Situation, “SoA”		✓	✓	✓	✓
Oracle			✓		
Other			✓	✓	✓

Table 6b: Atypical exchanges

	Participants	Citations
✓	Angel ↔ Human	[*passim*]
✓	Clay → Human	Isa 45:9b.
✓	Donkey ↔ Human	Num 22:28-30.
✓	Holy one → Human	Dan 4:11-14, 20; 8:13b-14.
✓	Satan ↔ God	Job 1:7, 9, 10, 11; 2:2, 4, 5.
✓	Snake ↔ Human	Gen 3:1, 4, 5.
✓	Spirit ↔ God	2 Chr 18:20, 21.
✓	Tree ↔ Tree	Judg 9:8-15.

Plans

In Forbes (2014b), our database work was divided into six phases, with Phase VI ("Extend into Discourse Analysis") designated as our current focus. As has repeatedly been the case, launching into a new phase has led to the discovery of loose ends and fresh tasks that must be addressed before proceeding into the new area. Four such tasks were discussed above in the subsection headed **Mid-term query capabilities**. We are actively working on these challenges while concurrently seeking to enlarge "we". To that end, a non-profit has been set up in South Africa ("BH Resources, NPC"). We expect the ownership of the A-F database to be transferred to this entity in January 2018. Future directions and contributors to the database will be coordinated by the new non-profit's board.

References

Andersen, F. I. 1970. The Verbless Clause in the Hebrew Pentateuch. New York: Abingdon.

Andersen, F. I. and Forbes, A. D. 1986. Spelling in the Hebrew Bible. Rome: Pontifical Institute Press.

Andersen, F. I. and Forbes, A. D. 1992. "On Marking Clause Boundaries." In Proceedings of the Third International Colloquium: Bible and the Computer – Methods, Tools, Results, 181–202. Paris-Geneva: Champion-Slatkine.

Andersen, F. I. and Forbes, A. D. 1995a. "Opportune Parsing: Clause Analysis of Deuteronomy 8." In Proceedings of the Fourth International Colloquium: Bible and the Computer – Desk & Discipline, 49–75. Paris: Editions Honore Champion.

Andersen, F. I. and Forbes, A. D. 1995b. "Syntactic Ambiguity in the Hebrew Bible." In Proceedings of the Fourth International Colloquium: Bible and the Computer – Desk & Discipline, 356–367. Paris: Editions Honore Champion.

Andersen, F. I. and Forbes, A. D. 1998. "Approximate Graph-Matching as an Enabler of Example-Based Translation." In Proc. 5th Inter. Colloq.: Bible and the Computer–Translation, 285–314. Paris: Editions Honore Champion.

Andersen, F. I. and Forbes, A. D. 2002. "Attachment Preferences in the Primary History." In Proc. of the Sixth International Colloquium: From Alpha to Byte, edited by J. Cook, 167–186. Leiden: Brill.

Andersen, F. I. and Forbes, A. D. 2012. Biblical Hebrew Grammar Visualized. Winona Lake, IN: Eisenbrauns. (BHGV)

Crystal, D. 2008. A Dictionary of Linguistics and Phonetics. 6th ed. Oxford: Blackwell. https://doi.org/10.1002/9781444302776.

Dryer, M. 2006. "Functionalism and the Metalanguage-Theory Confusion." In Phonology, Morphology, and the Empirical Imperative: Papers in Honour of Bruce L. Derwing, edited by G. E. Wiebe et al., 27–59. Taiwan: Crane Pub.

Forbes, A. D. 2006. "On not Putting Descartes before D. Hume: Balancing Rationalism and Empiricism in Corpus Tagging." In Corpus Linguistics and Textual History, edited by P. S. F. Van Keulen and W. Th. Van Peursen, 123–128. Assen: Van Gorcum.

Forbes, A. D. 2008. "How Syntactic Formalisms Can Advance the Lexicographer's Art." In Foundations for Syriac Lexicography III, edited by J. Dyk and W. Van Peursen, 139–158. Piscataway, NJ: Gorgias.

Forbes, A. D. 2014a. "Discourse Deixis in Biblical Hebrew." SBL San Diego: unpublished.

Forbes, A. D. 2014b. "A Tale of Two Sitters and a Crazy Blue Jay." In Reflections on Lexicography: Explorations in Ancient Syriac, Hebrew, and Greek Sources, edited by R. A. Taylor and C. E. Morrison, 211–232. Piscataway, NJ: Gorgias.

Forbes, A. D. 2016a. "The Proper Role of Valency in Biblical Hebrew Studies." In Contemporary Examinations of Classical Languages (Hebrew, Aramaic, Syriac, and Greek): Valency, Lexicography, Grammar, and Manuscripts, edited by T. M. Lewis, A. G. Salvesen and B. Turner. 95–112. Perspectives on Linguistics and Ancient Languages. Piscataway, NJ: Gorgias Press.

Forbes, A. D. 2016b. "The Diachrony Debate: A Tutorial on Methods." JSem 25 (2), 501–546.

Forbes, A. D. 2016c. “On Dating Biblical Hebrew Texts: Sources of Uncertainty/Analytic Options.” In From Ancient Manuscripts to Modern Dictionaries. Select Studies in Aramaic, Hebrew and Greek, edited by T. Li and K. Dyer, 297–330. Perspectives on Linguistics and Ancient Languages 9. Piscataway, NJ: Gorgias Press.

Forbes, A. D. 2017. “Capitalizing on Dependency Relations in Biblical Hebrew Grammar.” SBL Boston: unpublished.

Gazdar, G. E., Klein, E. Pullum, G.K. and Sag I.A. 1985. Generalized Phrase Structure Grammar. Oxford: Blackwell.

Gazdar, G. E., Klein, E. Pullum, G.K. and Sag I.A. 1988. “Category Structures.” Comp. Ling. 14 (1), 1–19.

Gordis, R. 1971. The Biblical Text in the Making: A Study of the Kethib-Qere. Jersey City, NJ: KTAV.

Heine, B. and Narrog, H. 2010. “Introduction.” In The Oxford Handbook of Linguistic Analysis, edited by B. Heine and H. Narrog, 1–25. Oxford: OUP.

Herbst, T. et al. 2004. A Valency Dictionary of English: A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns, and Adjectives. Topics in English Linguistics 40. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110892581

.

Kathol, A. et al. 2011. “Advanced Topics in HPSG.” In Non-Transformational Syntax: Formal and Explicit Models of Grammar, edited by R. Borsley and K. Börjars, 54–111. Chichester: Blackwell. https://doi.org/10.1002/9781444395037.ch2.

Miller, C. 2003. The Representation of Speech in Biblical Hebrew Narrative. Winona Lake, IN: Eisenbrauns.

Ojeda, A. E. 1987. “Discontinuity, Multidominance, and Unbounded Dependency in Generalized Phrase Structure Grammar: Some Preliminaries.” In Syntax and Semantics Volume 20: Discontinuous Constituency, edited by G. J. Huck and A. E. Ojeda, 257–282. Orlando: Academic Press.

Polak, F. 2006. “Sociolinguistics: A Key to the Typology and the Social Background of Biblical Hebrew.” Hebrew Studies 47, 115–162. https://doi.org/10.1353/hbr.2006.0025.

Trask, R 1993. A Dictionary of Grammatical Terms. London: Routledge.

Whaley, L. 1997. Introduction to Typology: The Unity and Diversity of Language. Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781452233437.