

Understanding and Using the New Syntax Searching Capabilities in Accordance 9

Robert D. Holmstedt (The University of Toronto)
Dallas, September 25, 2010

0. Preface

Good afternoon! Welcome to the discussion of the new syntactic features in Accordance 9. I am Robert Holmstedt, professor of Ancient Hebrew and Northwest Semitic languages at the University of Toronto. My research and teaching focuses on the use of theoretical linguistics to better understand the ancient Semitic languages, especially Hebrew and Phoenician. I am largely responsible for the syntactic data upon which the syntax add-ons are built. What I will do in this session is explain some of the background to this project and then guide you through the syntactic principles upon which the tagging scheme is built. Finally, we will do some sample searches together.

The new syntax modules in Accordance have their origin in my own reaction to the release of the syntactic database developed by Francis Andersen and Dean Forbes and released within Logos' Bible software in 2006 (although I had the privilege of seeing the pre-release version months before the public release). To simplify somewhat, my reaction was and has remained two-fold: 1) on the one hand, this is the type of information I've longed for in an electronic database -- syntactic (and more) information from noted Hebraists, and 2) on the other hand, the complexity of the data -- of which, as a Hebrew linguist specializing in syntax, I'm all too aware -- should ***not*** be matched by the complexity of the search interface.

Spurred partially by my opinion of the search interface and partially by a

desire to see syntactic databases for non-biblical texts, along with a good dose of the conviction that I could develop a better database (if I lacked this conviction, why would I go the trouble?), I began in 2007 to apply for research funds for my own project. In 2008, by a happy turn of events, I partnered in this quest with long-time Accordance associate and Dead Sea Scrolls expert, Professor Martin Abegg. Before long, Prof. Abegg had arranged for me to meet with Roy Brown of Accordance at the 2008 SBL meeting. The desire for Accordance to add syntactic capabilities and the maturation of our own project vision coincided beautifully, producing the first stage that you now have in Accordance 9.

This project's vision and ultimate goals are ambitious. While we have released just Genesis in the Hebrew Bible and the Gospel of John in the New Testament, by SBL we will be adding well over a dozen books, including the syntactic database for the Hebrew Inscriptions module, and by end of 2012 -- just three full years into the actual tagging process of this project -- we intend to be finished, leaving Accordance users with syntactic information for the Hebrew Inscriptions, Hebrew Bible, Ben Sira (Hebrew), the Dead Sea Scrolls, and the Greek New Testament.

In the rest of this session I will first briefly discuss the underlying linguistic principles in the tagging scheme. Then we will go over the syntactic terms, before moving on to a set of simple searches and then a small set of complex searches.

1. Introduction to Underlying Structural Principles

One of the first challenges to this project concerned data entry, that is, how we

would represent the relationships between words in the actual process of tagging the text. The bracketing approach used widely in linguistics was an easy fit and it was from there that we quickly developed a tagging scheme both flexible enough and adequately explicit to produce a usable database.

1a. Constituents and Phrasal Hierarchy

The key to bracketing in tagging a text is balance. For every opening bracket there must be a closing bracket. Each complete bracket set represents a **constituent**. A constituent is a single syntactic unit that has a place within the hierarchy of a larger syntactic unit. It is important to recognize that morphological words and constituents *may* overlap but are not always identical. That is, a single word may represent more than one syntactic constituent, such as English User's, in which the constituent User has a syntactic role that is distinct from the syntactic role of the possessive 's. This is true in Hebrew, too; moreover, the converse is also true: occasionally multiple words represent a single syntactic constituent. This is the case with many proper nouns, such as בֵּית לְהָם, but also true of complex prepositions, such as מֵעַל פְּנֵי, which is decomposable morphologically as 'from.upon face.of' but syntactically is taken as a single preposition 'from'.

The highest level constituent is a **clause**. A clause is a single constituent consisting of a subject and predicate. Main clauses (or "independent" in Accordance) are self-contained and thus do not function within a larger *syntactic* hierarchy, while subordinate clauses are contained within a phrase, typically a Predicate phrase. A **phrase** may consist of one word or many words and functions within the hierarchy of a clause. A phrase lacks the subject-

predicate nature of the clause and is, at its core, a projection of a single constituent, often referred to as the phrasal "head." For example, a prepositional phrase is the projection of the hierarchy around a preposition, a noun phrase is the projection of a noun, etc.

To illustrate the hierarchical nature of a clause and the constituents within, consider the classic example used by Chomsky and Halle in their book, *The Sound Pattern of English* (1968:372):

(1) *Syntax*: [This] [is [the cat [that [caught [the rat [that [stole [the cheese]]]]]]]]

Prosody: (This is the cat) (that caught the rat) (that stole the cheese)

Chomsky and Halle used this example to illustrate the difference between syntactic hierarchy and intonational structure. Although the contrast is interesting, the purpose here is to show the hierarchical nature of constituent relations and how this is manifested in bracketed tagging. In the syntactic representation in example (1), the subject "This" is bracketed by itself and the entire predicate "is the cat that caught the rat that stole the cheese" is a single constituent bracketed by itself. Within the predicate, the bracketing distinguishes the hierarchy, such that "the cheese" belongs within the verb phrase headed by "stole," which belongs within the relative clause headed by "the rat," which belongs within the verb phrase headed by "caught," which belongs within the relative clause headed by "the cat," which belongs within the verb phrase headed by the main verb "is." Now, just imagine if you threw in a few adverbs, prepositional phrases, etc. Syntactic hierarchy can and quickly

does become complex.

Above all, the point of this is to demonstrate that constituents *are contained within* larger constituents, all the way up to the clause level. For each word, a decision has to be made regarding its location in the syntactic hierarchy -- within what other constituent does it reside? And for that resulting complex constituent, the same question must be answered, until there are no more constituents and one is left with a clause. For example, a subject noun is contained within the subject phrase brackets in (2), the verb phrase within the predicate brackets, and both together form a clausal constituent.

(2) [_{CLAUSE} [_{SUBJ} God] [_{PRED} made the firmament]]

At a basic level the hierarchy that we have followed is binary in nature. That is, a clause consists of a single subject phrase (no matter how complex) and single predicate phrase (no matter how complex).

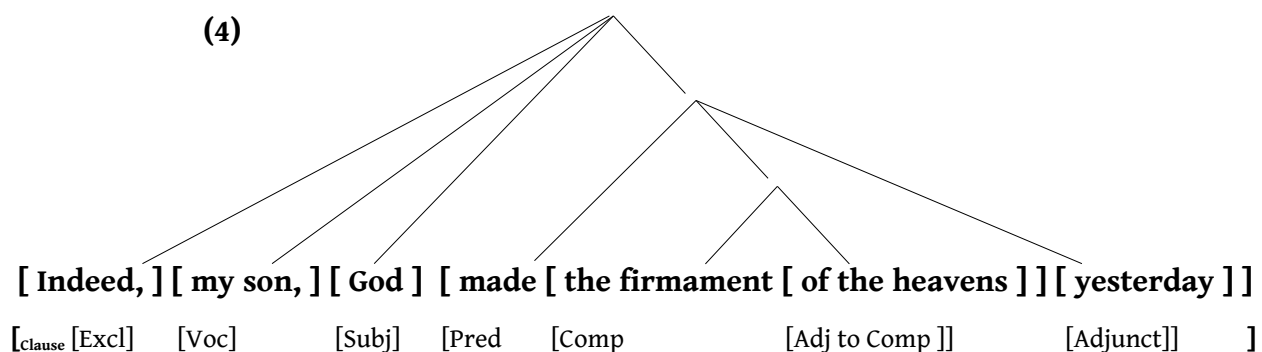
(3)

```
graph TD; A[ ] --- B[ God ]; A --- C[ made the firmament ]
```

Binary-branching phrase structure is 'built in' to the minimalist program of Chomskyan generative linguistics and is a core principle of many other generative frameworks.

But the addition of clause-edge constituents, such as dislocations (casus

pendens), vocatives, and exclamatives results in a tree that is not easy to fit into a binary structure and requires a good deal of theoretical machinery, so to speak. Moreover, within the subject and predicate, the phrasal head often appears to be modified in non-binary ways.



Our phrase structure in the tagging scheme thus departs from the strict binary-branching phrase structure of much generative linguistic syntax. This departure reflects our principled decision to utilize linguistic theory but not allow the tagging scheme to be bound to a specific theory.

1b. Data and Theory

For the databases to be broadly appealing and thus usable, we were committed to tagging principles that allowed the data to modify theory-driven decisions rather than vice versa. Our database is not "theory-neutral," which would be both naive and scientifically impossible, but "data-primary, theory-wise."

Two examples where an awareness of linguistic theory has influenced our tagging principles are with regard to discontinuous constituents and null constituents. Although we eschewed building our bracketing principles on the linguistic notion of 'constituent movement', we were forced to deal with

discontinuous constituents, that is, constituents that are divided into parts separated by un-related constituents. This happens less in English than in Hebrew, although it does occur with some English relative clauses, as in (5).


(5) [A new king] arose over Egypt, [who had not known Joseph]



In (5) the relative clause clearly modifies the NP 'a new king', and yet it is separated from this NP by the VP 'arose over Egypt'.

In Hebrew, discontinuity is extremely common, since many narrative clauses begin with the *wayyiqtol* narrative verb, switch to a subject, and then continue with the rest of the predicate.

(6) וַיִּרְא אֱלֹהִים אֶת־הָאוֹר
and.saw God OBJ - the.light



'and God saw the light' (Gen 1:4)

The challenge of constituent discontinuity is that, based on the hierarchy and projection principles described above, a verb and its modifiers together make up a *single constituent*. In the case of the verb and its complement above, this single constituent is often referred to as a *Verb Phrase*, which we have labeled the *Predicate*.

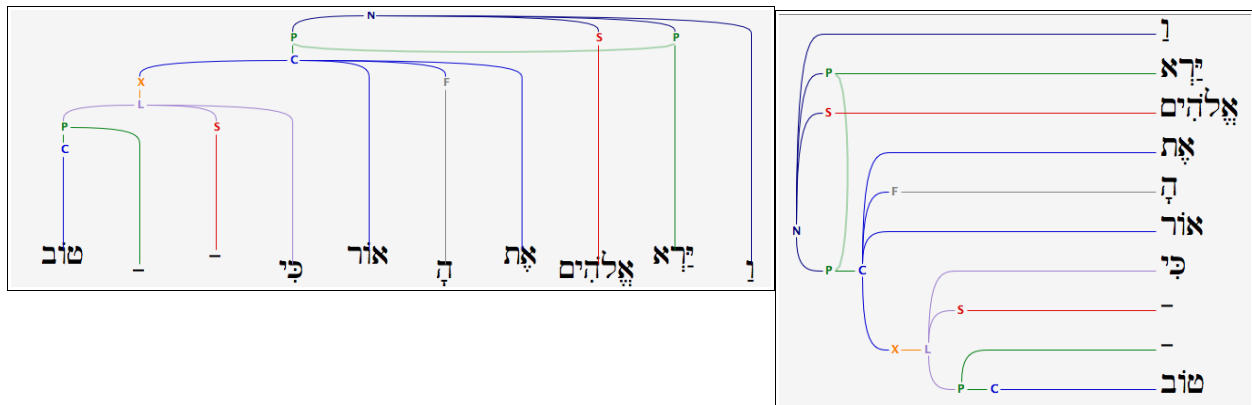
To account for discontinuous constituents we have employed a simple numerical cross-referencing, so that the tagging of the same Gen 1:4 example above looks like what is given in example (7):

(7) וַיִּרְא אֱלֹהִים אֶת־הָאוֹר

[N and [P1 saw] [S God] [P1 [C OBJ - the.light]]]

'and God saw the light' (Gen 1:4)

In (7), the two parts of the Predicate are connected by the alpha-numeric tag P1. The implications of this are that search results for which the hits include discontinuous constituents will highlight all parts of the constituent, with non-highlighted intervening material. In the tree display, the discontinuity is signaled with lighter colored connecting lines, as with the tree for Gen 1:4.



We have used this basic cross-referencing to provide information about three other phenomena: dislocation (casus pendens), resumption in relative clauses, and ellipsis (or 'gapping'). Example (8) illustrates dislocation.


(8) מִקְנֵהֶם וְקַבְיָנָם וְכָל־בְּהֵמָתָם הֲלוֹא לָנוּ הֵם

[[1 their cattle] [Q not ours are [1 they]]

'their cattle and their property and all their beasts -- are they not ours?'
(Gen 34:23)

In (8) the initial compound NP **מְקַנְהֵם וְקַנְיָנָם וְכָל־בְּהֵמָתָם** cannot be a formal syntactic part of the clause, which already has a subject, **הֵם**, and a (verbless/null copula) predicate with **לָנוּ**. Yet, the pronoun **הֵם** refers back to the syntactically ‘hanging’ or ‘dislocated’ NP and, in fact, by this anaphoric co-reference, **הֵם** connects the dislocated constituent to the clause.

Relative clause resumption is also indicated by a similar cross-referencing, as (9) shows.

(9) כָּל־הָעֵץ אֲשֶׁר־בוּ פְּרִי־עֵץ
 [[every 1 **the=tree** [A that [P in 1 it] [S fruit.of tree]]

 'every tree that in it (is) tree-fruit' (Gen 1:29)

In example (9), the numeric cross-referencing connects the head of the relative clause **הָעֵץ** "the tree" with the 3ms pronoun **ִּ** "it" that resumes or picks it back up within the relative.

Ellipsis is dealt with similarly, although it also involves the combination of the cross-referencing numeral with a null tag (0) marking the place of the elided constituent, as in (10)

(10) יָדַע שׁוֹר קִנְיָהּ וְחֹמֹר אֲבוּס בְּעֶלְיוֹ
 [[01 **knows** ox owner=its] [and=ass 01 trough.of master=its]]
 'an ox knows its owner // and an ass (knows) its master's trough' (Isa 1:3)

In (10), the first half of the poetic line-pair (=bicola, distich) has the verb **יָדַע**,

which is then assumed in the second line. In the second line, the null copy of ידע is given both a 0 (for the null finite verb) and a cross-reference 1, to tie it to ידע.

The use of a 0 for a null constituent in (10) brings up another critical -- and theory-informed -- component of the tagging scheme. On the principle that every phrase has a 'head', whether a 'verb' for a Predicate or a noun or similar nominal(ized) constituent for a Subject, we have inserted a null marker (0) in every phrase that lacks an overt head. The use of null constituents is most common in the Subject position (11), since Hebrew allows an overt subject to be omitted, and in the Predicate position (12), since Hebrew utilizes a "verbless clause" copular strategy.

(11) וַיִּשְׁבֹּת ____ בַּיּוֹם הַשְּׁבִיעִי מִכָּל־מְלַאכְתּוֹ
and=rested 0/(he) on.the=day the=seventh from=all.of work=his
'and (he) rested on the seventh day from all his work' (Gen 2:2)

(12) וַחֲשֹׁךְ ____ עַל־פְּנֵי תְהוֹם
and darkness 0/(was) upon face.of deep
'and darkness (was) upon the face of the deep' (Gen 1:2)

In addition to null subjects and predicates, Hebrew also allows null complements and null relative clause heads. All of these null items have been included and tagged appropriately in our databases.

1c. Narrow Syntax

A final defining principle of the Accordance syntax database that I'll mention here is a *narrow focus* on syntax. That is, the tagging scheme provides phrasal, clausal, and inter-clausal information to the exclusion of semantic judgments,

discourse relationships, and implicational pragmatics. For example, when the particle ׀ is a subordinator, our database makes no distinction between its use as a temporal ('when') subordinator or a clausal ('because') subordinator. Those distinctions are left to the user to determine. What our database provides is the distinction between ׀ as an adjunct subordinator (temporal or causal), a complement subordinator ('that'), a conjunction ('but'), and an exclamative ('indeed!').

2. Basic Definitions

Each of the terms below represents a specific label we used in the bracketing of the constituents. That is, the bracketing itself only provides hierarchical information. To complete the necessary syntactic information, such as whether a given constituent is a subject, predicate, vocative, etc., we use abbreviated labels for the syntactic constituent types listed in the Accordance Help.

A caveat is important here: our labels are not meant as linguistic statements. That is, the use of "predicate" instead of "verb phrase" simply reflects an issue of convenience -- it is simpler in the tagging process to use a single character label than a double character label, such as "VP."

2a. Clausal Labels

Clause : a unit of grammatical organization, consisting of a subject and predicate.

Types:

- **N:** Independent, Non-Speech: A Sentence or Independent Clause (the top level): a set of words that is complete in itself, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses.

Ex. 'And God called the light Day.' (Gen 1:5)

- **NA, NB, etc.:** Direct Speech: A clause that is both the complement of a verb of speaking and yet also an independent clause within the speech event.
Ex. 'And God said: "Let light be!" ' (Gen 1:3)
- **LA, LB, etc.:** Direct Speech: similar to NA, NB, but sometimes used in poetic doublets and triplets to preserve the parallelism. *Note:* if a search for all direct speech clauses is performed, to capture these LA, LB, etc., clauses alongside the NA, NB, etc., clauses, it is necessary to choose the "Any" clause option in the Clause search pane.
- **L:** Subordinate, Non-Speech: A clause, often introduced by a subordinating conjunction, that forms part of and is dependent on a main clause.

Sub-Types:

- **C:** Complement -- a clause that "completes" the requirement of the verbal semantics.
Ex. 'And God saw that it was good.' (Gen 1:10)
- **A:** Adjunct -- a clause that adds additional, but not syntactically required, information to the verb in the higher clause or, in the case of relative clauses, to a noun within the higher clause.
Ex. 'And God put them in the firmament of the heavens in order to provide light upon the earth.' (Gen 1:17)
Ex. 'And he divided between the waters that were under the firmament.' (Gen 1:7)

2b. Phrasal Labels

Phrase: a small group of words standing together as a conceptual unit, typically forming a component of a clause, and lacking its own predication.

Types:

- **S:** Subject -- the "doer" (agent) or "experiencer" (patient) of the predicate.
Ex. (agent) 'And God said ... ' (Gen 1:3)
Ex. (patient), 'And the earth was formless and void.' (Gen 1:2)
- **P:** Predicate -- the verb and any modifiers.
Ex. 'Let us make man in our image.' (Gen 1:26)
- **C:** Complement -- the phrase(s) that are required by either a verb or a preposition in order to "complete" the semantics of each.
Ex. 'Let us make man in our image.' (Gen 1:26)

- **A:** Adjunct -- the phrase(s) that are not required but add additional information about a verb or noun.
Ex. 'Let us make man in our image.' (Gen 1:26)

2c. Labels for Constituents that may be Clausal or Phrasal

Parenthesis, T: A clause or phrase that interrupt the flow of an 'argument', whether the argument is at its core chronological (i.e., a narrative) or logical (i.e., an exposition, as in, e.g., many psalms).

Ex. 'And the Nephilim were in the land in those days (and also afterwards), when the Sons of God came to the Daughters of Man ... ' (Gen 6:4)

Ex. 'And the sons of Noah who came out from the Ark were Shem, Ham, and Japheth (Ham was the father of Canaan).' (Gen 9:18)

Appositive, X: A clause or phrase that elaborates on a preceding clause or phrase of the same type.

Ex. 'And Cain said to Abel, his brother ... ' (Gen 4:8)

Ex. 'Two by two they came to Noah, to the ark.' (Gen 7:9)

2d. Individual Syntactical Labels

This is the full list of tags that may be attached to individual words in a clause (some necessarily overlap with those above):

- **S:** Subject: see "Phrase, Subject" above.
- **P:** Predicate: see "Phrase, Predicate" above.
- **C:** Complement: see "Phrase, Complement" above.
- **A:** Adjunct: see "Phrase, Adjunct" above.
- **F:** Specifier: the definite article.
- **X:** Appositive: see "Appositive" above.
- **V:** Vocative: a word or phrase of direct address that stands apart from the subject and predicate of the clause.
Ex. 'And Abraham said: "O Lord Yhwh, what will you give me?' (Gen 15:2)
- **E:** Exclamation or interjection: a word or phrase that interrupts the normal syntax to orient the attention of the addressee (the reader or a character in the narrative).
Ex. 'And Yhwh God said: Look/Behold -- the man has become like one of us.' (Gen 3:22)

- **D:** Casus pendens (dislocation): noun or pronoun placed outside a following clause and resumed within the clause by a retrospective pronoun.
Ex. 'And the fourth river -- it is the Euphrates.' (Gen 2:14)
- **T:** Parenthesis: see "Parenthesis" above.
- **U:** Unknown: used for cases, mainly in Qumran and Inscriptions, where text is missing and the syntactical tagging of the remaining words is uncertain or unknown.

2e. Additional Syntactic information included in the Tagging

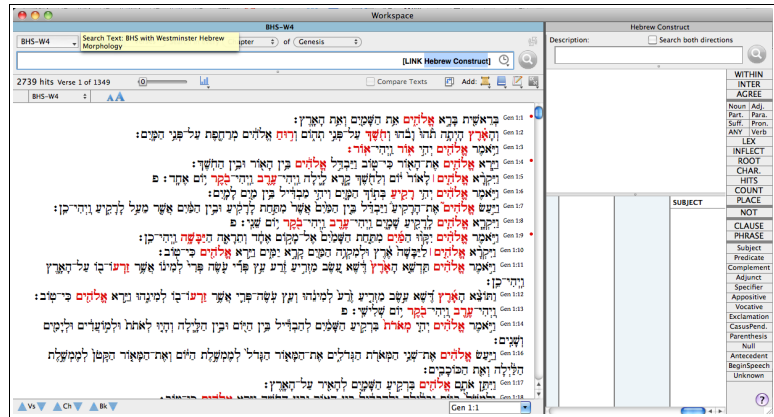
- **Null:** used to mark an implied word such as a subject or verb (indicated by a dash in the syntax display)
Ex. 'And darkness (was/∅) upon the face of the deep.' (Gen 1:2)
Ex. 'And God saw that (it) (was) good.' (Gen 1:10)
- **Antecedent:** a word to which another word (such as a following relative pronoun) refers (indicated by numerals in the syntax display), in addition to the syntactical tag.
Ex. 'And he divided between 01/the waters that (01/they) were under the firmament.' (Gen 1:7)
- **Begin speech:** used to indicate the beginning of direct speech.
- **Compounds:** Each of the individual syntactic categories (i.e., those in above in 2d) can be specified as "any," "single," or "compound." The Compound designation covers those constituents that have more than one head, that is, multiple constituents at the same hierarchy sharing the same syntactic role.
Ex. 'And the heavens and the earth and all their host were completed.' (Gen 2:1)

Now that we have covered the definitions of terms we have used in the database, let us together consider a few simple and complex searches available in Accordance's new syntactic capabilities.

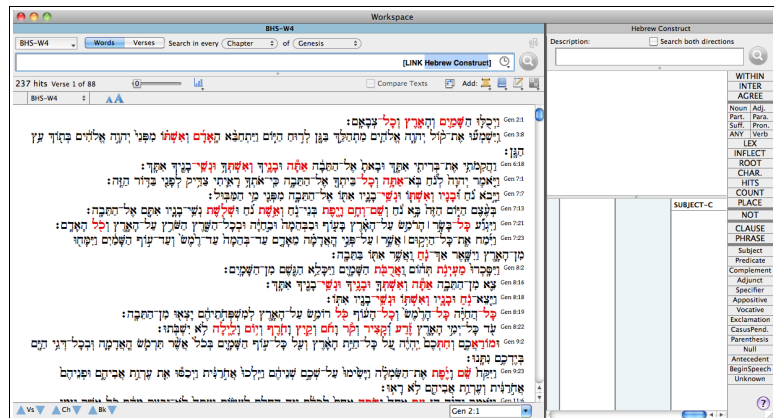
3. Simple Syntax Searches

(13) Subject

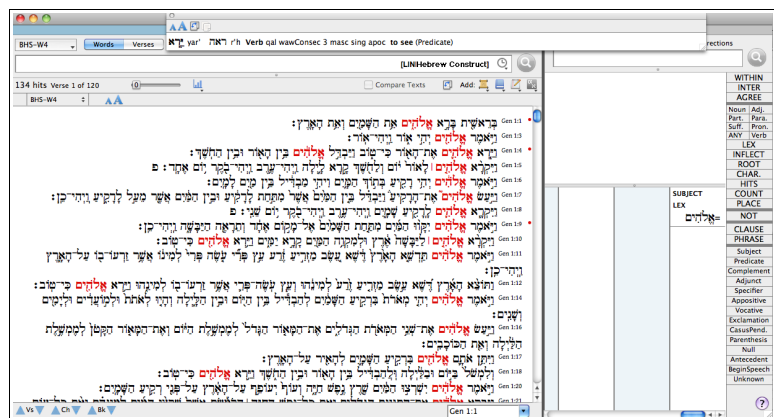
A. Any



B. Compound



C. Subject = אלהים



(14) Vocative

The screenshot shows the BIS-W4 software interface. The main window displays Hebrew text from Genesis 4:23. The right-hand pane, titled 'Hebrew Construct', shows a list of grammatical categories. The 'VOCATIVE' category is highlighted in blue. Other categories visible include WITHIN, INTER, ACREE, Noun, Adj., Part., Suff., Prom., ANY, Verbl., LEX, INFLLECT, ROOT, CHAR., HTS, COUNT, PLACE, NOT, CLAUSE, PHRASE, Subject, Predicate, Complement, Adjunct, Specifier, Appositive, Vocative, Exclamation, CasusPend., Parenthesis, Null, Anecdotal, RegioSzech, and Unknown.

(15) Appositive

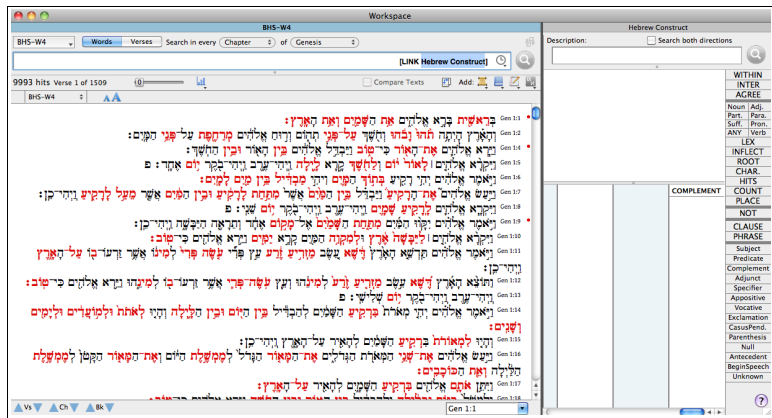
The screenshot shows the BIS-W4 software interface. The main window displays Hebrew text from Genesis 1:11. The right-hand pane, titled 'Hebrew Construct', shows a list of grammatical categories. The 'APPOSITIVE' category is highlighted in blue. Other categories visible include WITHIN, INTER, ACREE, Noun, Adj., Part., Suff., Prom., ANY, Verbl., LEX, INFLLECT, ROOT, CHAR., HTS, COUNT, PLACE, NOT, CLAUSE, PHRASE, Subject, Predicate, Complement, Adjunct, Specifier, Appositive, Vocative, Exclamation, CasusPend., Parenthesis, Null, Anecdotal, RegioSzech, and Unknown.

(16) Predicate - compound

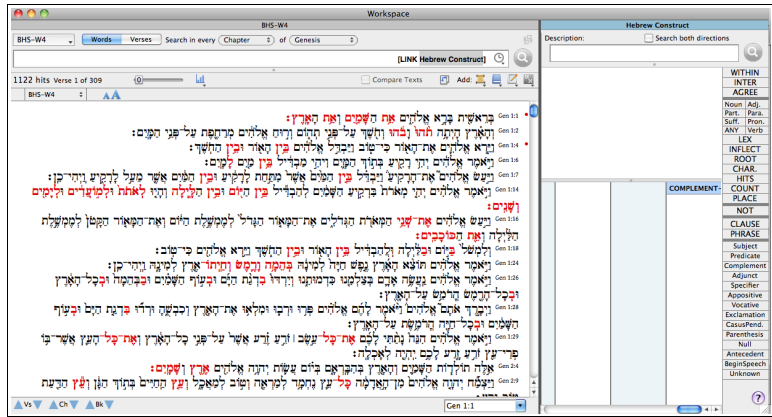
The screenshot shows the BIS-W4 software interface. The main window displays Hebrew text from Genesis 24:4. The right-hand pane, titled 'Hebrew Construct', shows a list of grammatical categories. The 'PREDICATE-C' category is highlighted in blue. Other categories visible include WITHIN, INTER, ACREE, Noun, Adj., Part., Suff., Prom., ANY, Verbl., LEX, INFLLECT, ROOT, CHAR., HTS, COUNT, PLACE, NOT, CLAUSE, PHRASE, Subject, Predicate, Complement, Adjunct, Specifier, Appositive, Vocative, Exclamation, CasusPend., Parenthesis, Null, Anecdotal, RegioSzech, and Unknown.

(17) Complement

A. Any

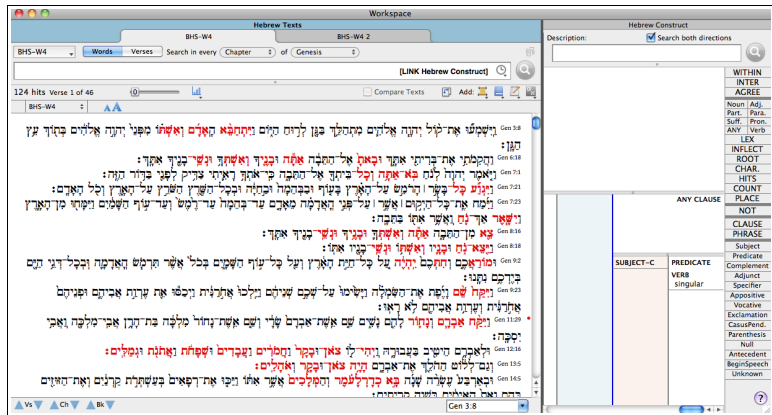


B. Compound

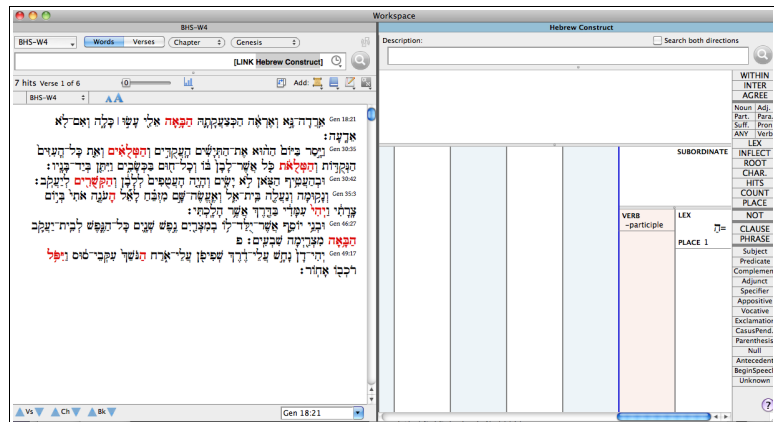


4. Complex Syntax Searches

(18) Compound Subj and Sg Verb:



(19) ךַ relatives versus ךַ as specifier.



(20) ךִּ and אֲשֶׁר used to introduce complement clause rather than relative or causal/temporal clause.

