Abstract

This paper investigates the visualization of the mapping of semantic and syntactic functions that were marked up in an XML-database containing linguistic data of the Biblical Hebrew text of Genesis 1:1-2:3. It focuses on two-dimensional topic maps as a graphical data-mining utility. The visual information is used to prompt the reconsideration of some existing assumptions and hypotheses about Biblical Hebrew syntax and semantics. Although some of the interesting results may be ascribed to tagging errors, the data-mining process demonstrates the rigor enforced by computer-assisted research. In addition, a number of cases are identified that challenge existing hypotheses and suggest possibilities for further research. This demonstrates the idea that text mining not only helps linguists to test hypotheses, but also prompts new ones.

1. Introduction

The study of the mapping of various linguistic modules has been highlighted by a number of scholars as research that may significantly benefit from a computer-assisted approach. Witt [26] highlights the importance of being able to discover the relations between different tiers of annotation as a part of the computation of a linguistic knowledge representation. "But the interrelations of annotation layers are of interest for many persons concerned with structuring and modelling of information" [27]. Bayerl et al. [1] compared, for example, the structural, thematic and rhetorical levels of a corpus of scientific texts. According to Burnard [3], studying the interplay of analyses of the various language modules is crucial for many grammatical and literary research studies, for example, "the extent to which syntactic structure and narrative structure mesh, or fail to mesh, ... or the extent to which phonological structures reflect morphology". These authors all identified mark-up systems, such as XML, as serviceable technology for the study of linguistic mapping.

This paper investigates the visualization of the mapping of semantic and syntactic functions that were marked up in an XML-based database containing linguistic data of the Biblical Hebrew text of Genesis 1:1-2:3. This is facilitated by the fact that the data in the underlying data cube is highly structured. Every phrase is tagged on various linguistic levels. The original text is thus not marked up only with inline elements [24] – every item is tagged for each level and all the data is stored in a three-dimensional data cube. If a phrase does not have a certain linguistic function, a null value (represented by a dash) is inserted between the opening and closing tags. A detailed discussion of the XML data cube is available at [13].

The mark-up provides the semantics that facilitates not only sharing, exchange and manipulation of data [18], but also its exploration. The core of information visualization is indeed to allow "people at all levels of an organisation to converse with their data and, from these conversations, glean the patterns and trends that will help them become more efficient, productive and successful" [9]. Visualization should also enable computational linguists to converse with their data in such innovative ways.

According to Manning et al. [15] there have been surprisingly few attempts to use visualization techniques to enhance the use of electronic dictionaries. "Despite decades of highly creative and sophisticated innovation, and a plethora of claims for obvious superiority of the visualization approach, we do not see visual maps of verbal information in popular and effective use" [14]. The same situation is probably still true of other ventures in electronic linguistics projects. This paper attempts to make a small contribution to fulfill this need.

Most Biblical information systems, which are currently used by researchers, are limited to morphological and
syntactical data. They may show syntactical tree-diagrams of clauses or even indicate hierarchies of clauses, but do not facilitate aggregate functions across the various linguistic modules. The authors hope to stimulate ideas for exploiting the rich amount of Biblical Hebrew linguistic data that has already been captured over the past forty years.

The paper is organized as follows: after a general discussion on the contribution of visual data mining, various visualization approaches and requirements are highlighted. Finally, some of these ideas are implemented on a linguistic data cube, and the results of this experiment are discussed.

2. Visualization and data mining

Visualization is a graphical display of subsets of a dataset, based on attributes that are linked by means of keys, array indexes or mark-up tags in order to facilitate a preferably interactive exploration of the data. It is an interdisciplinary activity that has links to the information and communication technologies of Information Science, Information Systems and Computer Science [8]. This paper concentrates on the ties between visualization and databases, building on the underlying principle of the use of XML to develop an exploitable database of linguistic data. The underlying data to be visualized should, of course, be stored in some or other databank, such as a relational database [17]. XML file or multi-dimensional array. One has to remember that much theory is already encoded into the structure of the databank and that its use will be restricted to these confines [16]. In this project these assumptions are encoded in the names and definitions of word groups, syntactic and semantic roles. These are based largely on the insights of SC Dik's Functional Grammar [5; 6], especially in the case of semantic functions, and Biblical Hebrew reference grammars.

Using visualization techniques in a project like this is a way of adopting a more holistic approach that is in line with an "externalist" view of good science, which approves of the incorporation of insights from other disciplines, especially in a diverse discipline like Information Systems [4]. (An internalist view, on the other hand, argues "that a core set of knowledge and shared scientific paradigms generated internal [sic] to the discipline are hallmarks of mature science, and thus diversity is to be avoided" [4]).

A graphical visualization tool uses all of these underlying technologies to present the selected data as a picture. This facilitates the exploration of the data, preferably by providing an interactive modus operandi. It therefore comes as no surprise that various authors refer to the data-mining operations made possible by visualization tools. According to Keller et al. [12] information visualization is the interactive, graphical rendering of abstract data to enhance information retrieval, data mining and learning. Many data-mining ventures start with a "hunch", a nagging feeling that there just might be an interesting relation between some of the elements in a dataset. Visualization is a way to make explicit these beliefs and assumptions of a researcher, a way of "organizing information so as to facilitate making the recommended inferences" [22].

The relationship between data mining and visualization is reciprocal. Data mining may be used to facilitate visualization, and visualization may be used to undertake interactive data mining. Interactive data mining requires cooperation between the database management system, the data-mining tool and the visualization tool [21].

Besides its obvious applications for analysis by the intelligence community and for knowledge management in businesses information, visualization may also be used for "exotic applications" by genealogists, lawyers and museums [9]. If humanities computing qualifies for the "exotic-application" tag, linguists may also use visualizations to highlight hierarchies, taxonomies and correlations in their datasets. Text analysis may be regarded as a balancing act between formal and interpretive tasks. An algorithm performing analytic functions on language may be regarded as a tool that takes responsibility for the more formal tasks and frees the hands of the human analyst who can then focus on the more non-deterministic activities [2].

Visualization of linguistic data may be regarded as the third step of computerized text analysis. After an archive or database has been built during the initial meta-linguistic phase to create a marked-up version of a literary text, software is developed in the algorithmic phase to analyze the source materials. These phases are followed by the representational phase, which presents the interpreted data in a way that satisfies the needs of the user [16]. In more advanced approaches visualization may also be used to facilitate data exploration.

Like other text-analysis tools, visualization tools can simply be used as an interface both to find evidence to verify or falsify a theory [19]. Ideally, a visualization tool should allow interactive operations so that the user can try out various scenarios and make adjustments to change or refine questions. Such an iterative process provides an experimental, almost "playful", way to do data mining in texts and this helps the researcher to question and even circumvent stereotyped hypotheses. Although not all results will be useful, this trial and error process could lead to the discovery of new, coherent patterns which would not be suggested by existing theory [19].

3. Various approaches of visualization

A graphical visualization tool uses various related ICT technologies to present the data, for example, as a graph of connected nodes. The relations are based on data attributes
that are linked by means of keys, array indexes or markup tags. The nodes and links form a picture that visually represents the interrelated data attributes. Other types of graphical visualizations are animation, visualization of a DTD (Document Type Definition) as a tree structure, and visualization of an archive as a lattice [22]. These graphical visualizations could still be two-dimensional, but also three-dimensional or multi-dimensional. Although a computer screen is, like paper, essentially a two-dimensional medium, it can be used inventively to simulate three-dimensional models.

Although a multi-dimensional approach could be a better approach, it is not necessarily always the case. One should remember that readers are more used to two-dimensional representations, which are also easier and less expensive to build [8]. Keller et al. [12] found that, although two-dimensional representations and the use of color-coding indeed enhance data mining and learning in comparison to pure text-based renderings, multi-dimensional approaches lead to cognitive overload on the user, which nullifies any additional benefits. However, they leave room for three-dimensional visualization of datasets where integration is important: "...three-dimensional displays are superior to two-dimensional ones only for specific tasks requiring integrating information over three dimensions" [12]. Since the Genesis 1:1-2:3 data cube does integrate various linguistic levels (e.g. morpho-syntax, syntax and semantics), a three-dimensional visualization should be a viable option. However, this paper focuses only on two-dimensional graphs as a data-mining utility.

4. Requirements of a visualization tool

The characteristics of a tool should differ depending on the purpose, target audience and education level of the users. If an interactive interface is built for unsophisticated users, too much detail could lead to confusion and it would be better to use a simple and clean graphical layout [15]. This could be a valid requirement even if the users do have a lot of knowledge regarding the underlying linguistic data, but not about computing, as is often the case in the humanities.

The interface should also be user-friendly, for example by providing meaningful and readable labels. It should allow end-users to visually rearrange the data to create suitable information [8]. The analyst must be able to refine his/her query to focus more sharply on an uncovered pattern in order to better understand the relationship. Such an interface, which is easy to use, could help to involve more people "to take an active role in data mining activities" [9].

Furthermore, a visualization tool should allow the user to adapt queries in an interactive way by dynamically mapping the underlying data and the resulting graphs in real time [9]. This requires the underlying database to be integrated with the GUI.

A visualization tool should also allow scalability. The user should be able to work with anything from small sets of static data to large sets of changing data [8]. The user should be able to adjust the resolution accordingly, because "too much information can cause the screen to resemble a giant hairball". The tool should also be able to visualize the results of both qualitative and quantitative investigations [9]. The visualization of qualitative data is one of the challenges for software creators [8].

The reporting module should include facilities to efficiently and easily communicate findings to other persons concerned [9]. The reports should be customizable so that it can be adjusted for different audiences. A one-dimensional text-based version should be provided as an alternative for non-visually oriented users [8].

Although the application discussed in the next section meets a number of these requirements, not many tools, if any, will have all of these characteristics.

5. Application: a graphical topic map of semantic and syntactic mappings

In this section the mapping of the semantic layer onto the syntactic layer in Genesis 1:1-2:3 will be explored. This information will then be used to test some existing assumptions and hypotheses about Biblical Hebrew syntax and semantics. Bradley [2] discusses topic maps as an example of electronic tools that support the creation of mental models regarding literary analysis. A topic map contains a spatial element and is therefore suitable for graphical visualization. The researcher, for example, identifies various topics in a series of literary texts and draws a picture with the help of a visualization tool linking these topics to the texts where they appear. Associations between the topics are also shown.

In this experiment the concept of a topic map is applied to grammatical categories. Topic maps are used to indicate the associations between selected semantic and syntactic functions. The mapping of semantic functions onto syntactic functions forms a complex network of associations in a text. A traditional interlinear paper-based analysis cannot show this network. A visualization tool could make these associations visible just as it would enable a better understanding of the semantic networks in a dictionary [15].

The idea of a topic map was applied to the linguistic data of Genesis 1:1-2:3. The topic map program was programmed in Java. When one opens the program, the data file that has been used in the previous session is opened. One may click on the “File” menu to browse for the required file. In this case, the XML database, referred to above, is selected and opened (see Figure 1).
Concepts (the semantic and syntactic functions) are represented as nodes in a two-dimensional picture. All the semantic functions appearing in Genesis 1:1-2:3 are shown in the upper block; the syntactic functions are displayed in the middle-block and the phrases in the lower block. All the phrases in the database are shown with links to their semantic and syntactic functions. Based on their collocations, lines are used to indicate the mapping of semantic functions onto syntactic functions, for example, agent, positioner, processed and zero are all first arguments, expressed by subjects in the surface structure of clauses. Patient is a second argument in the logical structure, which may be expressed, *inter alia*, by a/an (direct) object in an active realization, or by a subject in a passive realization. Similarly, other arguments and satellites are linked to the syntactic functions realizing them in the surface structure. The data is still unfiltered and, therefore, looks like a hodgepodge of links. In order to provide a drill-down facility, the user may hover the mouse over any one of the phrases to activate a textbox containing detailed information about the clause.

The "View" menu allows the researcher to view the constituents' data in a textual format (see Figure 2). Another, more important, option in the "View" menu is the filter management function. It allows the researcher to experiment in a trial and error way by adding, removing and moving various filters in order to focus on required aspects. This makes the tool interactive and enables the researcher to look at a dataset from various perspectives. When the researcher clicks on "Manage Filters", a new window opens allowing the definition and fine-tuning of filters (see Figure 3).

The researcher may, for example, isolate phrases with the syntactic function of adjunct by selecting the relevant options on the drop-down lists and entering the name of the required function in a textbox (located towards the bottom of the screen). The filter is inserted in the window by clicking the "Add" button. The "OK" button will use the defined filter(s) to create a topic map. The results, produced by applying the current filter, are shown in Figure 4. It shows that, in Genesis 1:1-2:3, the syntactic function of adjunct is used to realize the semantic functions of time, manner, purpose, location and reason. This confirms the definition of an adjunct as an optional, adverbal element in the predicate [7; 23; 25]. When the user hovers with the mouse over the first phrase, more clause detail of Genesis 1:1a is shown in a pop-up window.

Underlying this visual representation is the slicing off of the phonetic, syntactic and semantic levels in the data cube. To fine-tune the results, the researcher may also include more filters that add or remove parameters on all three these levels. For example, if one would like to add more information on the display regarding the semantic function of location, the following filter may be appended: "ADD phrases with SEMANTIC FUNCTIONS equal to 'location'". The updated graphical display is shown in Figure 5.

The user may also simplify the graph by deleting irrelevant information. For example, if the researcher now wants to focus on data about the semantic function of location, (s)he may now define filters to delete links and fields pertaining to the semantic functions of time, manner, purpose and reason. The result is shown in Figure 6.

The graph now shows that location may be expressed, *inter alia*, either by complements or by copula-predicates in the data set. The researchers suspected that some copula-predicates could have been tagged as complements since it is a specific subtype of complement. Indeed, in Gen. 1:29c one instance was found where the coding was done incorrectly. However, working through all the listed hits revealed that the tagging was done consistently in all other places. With reference to location, copula-predicate has been used as the second argument in a nominal clause, while complement has been used as the third argument in nominal or verbal clauses. Location may also be expressed by adjuncts. This confirms the hypothesis of Functional Grammar that location may be expressed by arguments or satellites [5].

Since the order in which filters are applied, may have an effect on the eventual output, the user is also allowed to move them up or down. An existing filter may be removed and even the whole filter window may be cleared to make a fresh start. If the user wants to save a filter or group of filters for later re-use, these may be saved and reloaded later (see Figure 3).

Using the visualization tool also reveals the following interesting mappings:

- **Patient expressed by indirect object** (see Figure 7)

  Various examples occur in the data set where a preposition phrase expresses the patient, e.g. Gen. 1:5a: *vayikra elohim la'or yom* (God called (to) the light day). Since it is strange to regard a preposition phrase as direct object, these phrases have been tagged as indirect objects. However, this is incompatible with the traditional definition of an indirect object as the third argument or second complement of the main verb [10; 11; 23; 25]. The simplest solution would be to allow preposition phrases like these to be regarded and tagged as direct objects. Alternatively, the definition of indirect object could be changed to allow this syntactic function as a second argument. More in-depth research is needed to explore these hypotheses prompted by the data-mining venture.

---

1 No example of the semantic function of force was found in the data set.

2 In passive clauses, agent and positioner may be expressed as adjuncts on the syntactic level, but no examples were found in the data set.
Figure 1. Topic map of all phrases' syntactic and semantic functions as marked up in Genesis 1:1-2:3, based on an idea for literary analysis by Bradley [2].
Figure 2. A textual representation of the phrases in the database, viewable in the visualization program.
Figure 3. Interface used to define and fine-tune filters in the visualization tool.

Figure 4. A screen shot of a visualization of the network linking the semantic functions that may be expressed by an adjunct, as found in various clauses in the dataset.
Figure 5. Updated graph showing the network linking the semantic functions expressed by adjuncts, as well as other syntactic functions used to express location.

Figure 6. Simplified graph, showing only information about the semantic function of location.
• **Manner expressed by complement** (see Figure 8)
  In a number of identical clauses (vayehi xen — and it was so; see, e.g., Gen. 1:7e) the adverb xen is used as a complement. It suggests that the Functional Grammar theory should be adjusted. Dik [5] defines manner as a satellite that occurs in actions, positions and processes. If the tagging as manner is correct in this experiment, the theory should be adapted to include manner as an argument in states. Alternatively one could reconsider the tagging — maybe xen could be tagged as quality, but even this would prompt an adjustment in Functional Grammar’s description of semantic relations in non-verbal predications — “Property Assignment” is allocated only to adjectival and bare nominal predicate types [5].

• **Purpose expressed by copula-predicate** (see Figure 9)
  In Gen. 1:29e (laxem yiheyeh le’oxla — to you it will be as food) a copula-predicate (le’oxla) is expressing a purpose satellite. Since purpose satellites should be constructions embedded within controlled predications, one should rather consider tagging le’oxla as classification, which, however, in turn prompts further research into the type of predicates that may express “Class Inclusion”. Dik [5] only mentions an “indefinite term”, but it is not clear whether this should include preposition phrases.

• **Quality expressed by attribute** (see Figure 10)
  In various clauses (e.g. Gen. 1:5d) the semantic function of quality is allocated to attributes. For example, in the clause vayehi voker yom exad (and it was morning, day one), yom exad is a noun phrase in apposition to the subject and functions as an adjectival modifier. Also compare Gen. 1:27c (zaxar unkeva bara otam — male and female, he created them). Zaxar unkeva is an adjectival phrase consisting of two adjectives that describe the direct object in the clause. In both examples, however, the attributes are rather loosely coupled to the main clause and cannot simply be regarded as part of the noun clauses that they describe. Although the construction is slightly different from normal “Property Assignment” constituents — they are not predicates — they do seem to fit Dik’s [5] requirement of being adjectival or bare nominal elements. Dik [6] discusses similar extra-clausal constituents on a pragmatic level and calls them “tails”. The function of these “loosely adjoined constituents” is to “add a further specification to a term which is already contained in the clause”. Since pragmatics is excluded from this study, these cases have provisionally been tagged as attributes with the semantic function of quality, but the analysis and semantic tagging of this type of phrases should be researched in more detail.

Although some of these “interesting” mappings may be ascribed to tagging errors, the data-mining process has demonstrated the rigor enforced by visualization as a form of computer-assisted research. In addition, the topic maps visualized a number of cases that challenge existing hypotheses and suggest possibilities for further research. This demonstrates the idea that text mining not only helps linguists to test hypotheses, but that they can also prompt new ones: “The computer can deal with far more information than you can, and even though it can’t (yet) reason, it can show you opportunities for reasoning you would never find without it” [22].

**6. Conclusion**

The paper discussed the use of a graphical topic map as a visualization tool for linguistic data. After discussing the need for visualization in linguistic studies, some basic concepts of visualization have been covered. Some of these requirements and goals have been practically demonstrated by a Java program that creates topic maps linking phrases in the Hebrew text of Gen. 1:1-2:3 to their underlying semantic functions and the syntactic functions expressing these in the surface structure. The application illustrates that graphical visualization may be used as a powerful, experimental way of searching for patterns in a linguistic dataset.

The ideas discussed in this paper and the suggestion of a visualization implementation were submitted to make a small contribution to the search for humanities’ ways of digitally exploring texts, as formulated inimitably by Sinclair [20]: “I navigate through a text with the same blend of fascination, anxiety, and excitement as I explore the streets of an unfamiliar city: I do not hesitate to venture down mysterious pathways and streets, even though they may lead to a dead end. Various things along my journey may prompt me to change directions, and although I often do not know where I am going, I know that I am somehow accumulating a broader representation of the terrain. If I were given a detailed map and path to follow, I would be robbed of the enjoyment of exploration and serendipitous discovery. If I were given a list of the monuments and features of the city, I would still only follow, I would be robbed of the enjoyment of exploration and serendipitous discovery. If I were given a list of the monuments and features of the city, I would still only have limited understanding of it. Similarly, lists of words and other components of text can be very useful and informative, but to truly experience the text I need other means of exploring it.”

**References**


Figure 7. Topic map showing patient semantic functions expressed by indirect objects.

Figure 8. Topic map showing manner semantic functions expressed by complements.
Figure 9. Topic map showing purpose semantic functions expressed by copula-predicates.

Figure 10. Topic map showing quality semantic functions expressed by attributes.


